Biometrics, 2002, 58, 2, 387-397

A Generalized Discriminant Rule when Training Population and Test Population Differ on their Descriptive Parameters

Christophe Biernacki,[1,*] Farid Beninel[2,†] and Vincent Bretagnolle[3,‡]

[1]Université de Besançon, UMR CNRS 6623, 25030 Besançon, France
[2]IUT Département STID, 8 rue Archimède, 79000 Niort, France
[3]CEBC-CNRS, 79360 Beauvoir sur Niort, France

March 28, 2003

SUMMARY. Standard discriminant analysis methods make the assumption that both the labeled sample used to estimate the discriminant rule and the non-labeled sample on which this rule is applied arise from the same population. In this work, we consider the case where the two populations are slightly different. In the multinormal context, we establish that both populations are linked through linear mapping. Estimation of the non-labeled sample discriminant rule is then obtained by estimating parameters of this linear relationship. Several models describing this relation are proposed, and associated estimated parameters are given. An experimental illustration is also provided, in which sex of birds which differ morphometrically over their geographical range is to be determined, and a comparison with the standard allocation rule is performed. Extension to a partially-labeled sample is also discussed.

KEY WORDS: model-based discriminant analysis; relationship between populations; model selection; biological variables; sex determination.

1. Introduction

Discriminant analysis usually proceeds in the following manner (McLachlan, 1992, Lachenbruch and Goldstein, 1979, Gnanaddesikan, 1989): a sample of objects is drawn from a population and a partition of this sample is known.

[*] *email:* biernac@math.univ-fcomte.fr
[†] *email:* beninel@univ-poitiers.fr
[‡] *email:* breta@cebc.cnrs.fr

Each object within the population is described by several characters or certain measurements, which together form a feature vector belonging to a suitable feature space. Using the feature vectors and the individual labels of the sample, an allocation rule is established in order to class other non-labeled objects from the previous population.

Fisher (1936) proposed two-class linear discrimination in the Euclidian feature space by using Mahalanobis distance, but since his work, many extensions have been proposed. Rao (1948) extended Fisher's approach to more than two groups (see also Anderson, 1958). In a general case, the multinormal model allows quadratic allocation rules (see for example Tomassone et al., 1988) and some parsimonious models between linear and full quadratic possibilities are conceivable, like models introduced by Banfield and Raftery (1993) and revisited by Celeux and Govaert (1995). Non-parametric methods have also been explored, using kernel procedures (Silverman, 1986), the k-nearest neighbours methods (Fix and Hodges, 1951) or other works on projection pursuit (Friedman and Stuetzle, 1981). The logistic discrimination, first proposed by Anderson (1972), is another direction which may be interpreted as a semi-parametric classification. Adaptation of some previous methods as well as new proposals have also been considered in the case of qualitative data (Celeux and Nakache, 1994). Another important research topic is model selection, e.g. cross-validation (Hand, 1986 for a review) or the AIC criterion (Akaike, 1974).

However, in all previous situations, the labeled and the non-labeled samples come from the same population. Here, we are interested in the multinormal discriminant analysis problem when both populations (learning and testing) may be slightly different, e.g. when they come from different origins or belong to different classes of individuals, a common situation in biology. For instance, morphometrical bird sex discrimination is impaired by the fact that many bird species show considerable variations in size over their geographical range (Zink and Remsen, 1986). Thus, geographical origin may affect mean and variance of normal distributions of species features in each sex group, and the sex discriminant rule may not be applied with efficiency from one sample to another. A very similar problem arises with regard to age differences, since mature birds are often larger, or heavier, than immature individuals (see Bretagnolle et al., 1998 or Genevois and Bretagnolle, 1995 for examples).

Van Franeker and Ter Brack (1993) proposed an empirical solution to this problem (see below). Here we extend their approach by first establishing a distributional linear link between labeled and non-labeled populations in particular situations of interest. Estimation of parameters of this relationship allows then to transform the allocation rule of the labeled population into a new one for the non-labeled population, with generally very few parameters to estimate. It is shown that several constraints (called hereafter models) on the linear relationship can be proposed, including the solution given in Van Franeker and Ter Brack (1993). Extension to a possible partially-labeled "non-labeled" sample is also straightforward in this context.

Layout of this work is the following: Section 2 presents the data in a formal way. Relationship between both populations is constructed in Section 3. Then, models of constraints are presented (Section 4) and parameter estimation in each case is detailed (Section 5). Experiments with birds of geographical difference are shown to provide solid ground for our hypotheses (Section 6). Section 7 is devoted to concluding remarks and possible extensions of this work.

2. The data

Data consist of two samples: a labeled one, S, drawn from a population P and a non-labeled one, S^*, drawn from a population P^*. Note that P and P^* may differ.

The labeled sample S is composed by n pairs $(x_1, z_1), \ldots, (x_n, z_n)$ where x_i is a feature vector of $\Re^d$ for the ith individual, and where z_i is its group number (or label). So $z_i = k$ with $k = 1, \ldots, K$ if the ith individual belongs to the kth class among K possible classes. We consider couples (x_i, z_i) $(i = 1, \ldots, n)$ as independent realizations of the random couple (X, Z) of distribution

$$X_{|Z=k} \sim N_d(\mu_k, \Sigma_k) \quad (k = 1, \ldots, K) \qquad \text{and} \qquad Z \sim B_K(p_1, \ldots, p_K) \quad (1)$$

with $N_d(\mu, \Sigma)$ the Gaussian distribution of mean $\mu \in \Re^d$ and of variance matrix $\Sigma \in \Re^{d\times d}$, and with $B_K(p_1, \ldots, p_K)$ the K dimensional Bernoulli of parameters $p_1, \ldots, p_K$. So, the parameter p_k corresponds to the proportion of the group k in the population P and we have $\sum_{k=1}^K p_k = 1$.

The non-labeled sample S^* consists of n^* individuals, of which only vectors $x_1^*, \ldots, x_{n^*}^*$ are known (variables are the same as in the labeled sample), the corresponding labels $z_1^*, \ldots, z_{n^*}^*$ being unknown or, as we will assume also later, only partially known. We consider couples (x_i^*, z_i^*) $(i = 1, \ldots, n^*)$ as independent realizations of the random couple (X^*, Z^*) of distribution

$$X^*_{|Z^*=k} \sim N_d(\mu_k^*, \Sigma_k^*) \qquad \text{and} \qquad Z^* \sim B_K(p_1^*, \ldots, p_K^*). \quad (2)$$

Our aim is to estimate the n^* unknown labels $z_1^*, \ldots, z_{n^*}^*$ by using information from both samples S and S^*, the challenge being to find a link between P and P^*.

3. Relationship between the two populations

3.1 *Linear relationship between measured variables*

We would like to exhibit a distributional relationship ϕ_k ($\Re^d \rightarrow \Re^d$) between random vectors of the same class k but different populations:

$$X^*_{|Z^*=k} \sim \phi_k(X_{|Z=k}) = [\phi^1_k(X_{|Z=k}), \ldots, \phi^d_k(X_{|Z=k})]' \tag{3}$$

with ϕ^j_k a function $\Re^d \rightarrow \Re$ ($1 \leq j \leq d$). In the following, we will make three assumptions on ϕ_k.

First, the jth component $\phi^j_k(X_{|Z=k})$ of $\phi_k(X_{|Z=k})$ only depends on the jth component $X^j_{|Z=k}$ of $X_{|Z=k}$, so we assume now that ϕ^j_k is a function $\Re \rightarrow \Re$ (for simplicity, we preserve the notation ϕ^j_k). In this case, we have

$$\phi_k(X_{|Z=k}) = [\phi^1_k(X^1_{|Z=k}) \ldots \phi^d_k(X^d_{|Z=k})]'. \tag{4}$$

Second, we assume that each ϕ^j_k ($\Re \rightarrow \Re$) is C^1. Consequently, functions ϕ^j_k are necessarily linear (De Meyer et al., 2000, see Theorem 1 in Appendix A) and we have the K relations ($k = 1, \ldots, K$)

$$X^*_{|Z^*=k} \sim D_k X_{|Z=k} + b_k \tag{5}$$

with D_k a diagonal matrix of $\Re^{d \times d}$ and b_k a vector of $\Re^d$. The third assumption is that $b_k = 0$. So, we restrict attention to the specific linear case $X^*_{|Z^*=k} \sim D_k X_{|Z=k}$ for all classes. We will discuss later the plausibility of these assumptions within a biological context.

3.2 *Consequence for the discriminant rule of P^**

If the K diagonal matrices D_1 to D_K are known, it is easy to obtain means and variance matrices of the population P^* from the ones of the population P by the following classical formula: $\mu^*_k = D_k\mu_k$ and $\Sigma^*_k = D_k\Sigma_k D_k$ with $k = 1, \ldots, K$. Then, the discriminant rule of P^* is directly given by these parameters (see, e.g., McLachlan, 1992). In the following section, we discuss issues where the K diagonal matrices are unknown and we propose several scenarios for estimating them.

4. Models of constraints

4.1 *Model definitions*

The main idea is to define some constraints on the matrix transformation D_k to explicitly express some particular links between both populations P

and P^*. We suggest the five following models of constraints on the D_k's, denoted by $M_1, \ldots, M_5$. Definitions of these models are:

(M_1) $D_k = I_d$: both populations are the same (I_d: identity matrix).

(M_2) $D_k = \alpha I_d$: transformation is feature and group independent.

(M_3) $D_k = D$: transformation is only group independent.

(M_4) $D_k = \alpha_k I_d$: transformation is only feature independent.

(M_5) D_k is unconstrained: it is the most general situation.

Note that these models are different from those of Banfield and Raftery (1993), as these authors established links between variance matrices of all classes of a unique population. In our context, their models would correspond to constraints between variance matrices of the K groups inside the population P^*. Our proposal corresponds to constraints on both centers and variance matrices between the two populations P and P^*.

Model M_1 corresponds to the classical discriminant analysis case. Models M_2 and M_3 preserve homoscedasticity and consequently an eventual linearity of the rule: if $\Sigma_1 = \ldots = \Sigma_K$ for P, then $\Sigma_1^* = \ldots = \Sigma_K^*$ for P^*. Lastly, models M_4 and M_5 may transform a linear allocation rule on P into a quadratic rule for P^* with few parameters to estimate (see Section 5 below).

4.2 *Relationship with Van Franeker and Ter Brack's approach*

Our model M_2 is close to the situation modelled in Van Franeker and Ter Brack (1993). These authors considered two groups, males and females, of two seabird populations of the same species that come from different geographical areas (populations P and P^*). They assumed homoscedasticity of the population P ($\Sigma_1 = \Sigma_2 = \Sigma$). In such a situation, the discriminant rule is linear: noting 1 the class number of females and 2 the class number of males, x is classed as male if

$$x'\Sigma^{-1}(\mu_1 - \mu_2) > c \;\text{ with }\; c = 1/2(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) + \ln(p_2/p_1). \quad (6)$$

The fundamental idea of Van Franeker and Ter Brack (1993) is to keep the same discriminant rule for the population P^* by only changing the threshold c into a threshold c^*. The method is attractive and straightforward but their proposed threshold c^* is defined rather empirically: first, they assume that the distribution of the discriminant score (left part of equation (6)) of population P using non-labeled features X^* instead of labeled features X is a mixture of two univariate normal distributions. Second, they estimate parameters of this mixture and, then, they define the new threshold by the point where the two normal densities intersect. Our model M_2 relies on the same basic idea of changing c but, as developped in the following, a theoretical justification for the new threshold c^* is also provided.

As noticed before (Section 4.1), model M_2 preserves the rule linearity, so x^* is male if

$$x^{*'}\Sigma^{*-1}(\mu_1^* - \mu_2^*) > 1/2(\mu_1^* + \mu_2^*)\Sigma^{*-1}(\mu_1^* - \mu_2^*) + \ln(p_2^*/p_1^*). \tag{7}$$

By using the fact that $\mu_k^* = \alpha\mu_k$ $(k = 1, 2)$ and $\Sigma^* = \alpha^2\Sigma$ (Section 3.2), we obtain finally the new threshold: x^* is male if

$$x^{*'}\Sigma^{-1}(\mu_1 - \mu_2) > \underbrace{\alpha\{1/2(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) + \ln(p_2^*/p_1^*)\}}_{c^*}. \tag{8}$$

5. Parameters estimation

We retain a plug-in procedure to estimate matrices $D_1, \ldots, D_K$. So, estimates of $D_1, \ldots, D_K$ will be expressed with true parameters $\mu_1, \ldots, \mu_K$, $\Sigma_1, \ldots, \Sigma_K$ and $p_1, \ldots, p_K$ of population P. Then, when only estimates of parameters for population P are available (the general case), the true parameters of P are to be replaced by their estimates in expressions of $D_1, \ldots, D_K$. This estimation procedure provides consistent estimators with relatively simple calculations.

Estimation of matrices D_k depends on the model that is being used. For models M_2 and M_3, we can use the least squares estimator since the transformation is group independent in those cases. Moreover, only straightforward computations are needed. We present also a maximum likelihood estimator for these models, which may be particularly useful when labels are partially known in population P^* (see below). For models M_4 and M_5, the maximum likelihood estimator only is used. Our maximum likelihood estimator is based on an adaptation of the maximization step of the EM algorithm (Dempster, Laird and Rubin, 1977). Moreover, for any models, if proportions $p_1^*, \ldots, p_K^*$ are unknown, they are also defined by maximum likelihood and the EM algorithm.

5.1 *Model M_2 ($D = \alpha I_d$)*

5.1.1 Least squares estimator We have the following global relationship between expectation of the two populations:

$$E[X^*] = DE[X]. \tag{9}$$

Estimating $E[X^*]$ by the empirical mean $\bar{x}^*$, the least squares estimator of α is given by (note that $E[X]$ may be simply obtained by $E[X] = \sum_{k=1}^{K} p_k\mu_k$).

$$\hat{\alpha} = \arg\min \|\bar{x}^* - \alpha E[X]\|_2^2, \tag{10}$$

with $\|\cdot\|_2$ being the Euclidian norm. It follows that

$$\hat{\alpha} = \frac{\bar{x}^{*\prime} E[X]}{\|E[X]\|_2^2}. \tag{11}$$

5.1.2 Maximum likelihood estimator When at least some of the labels are known in the population P^*, the least squares estimator method does not take into account this information because it is group-independent. Then, we propose an alternative method that maximizes the likelihood on matrices D_k. In the most general case, it is expressed by

$$\ell(p_1^*, \ldots, p_K^*, D_1, \ldots, D_K) = \prod_{i=1}^{n^*} \sum_{k=1}^{K} p_k^* h(x_i^* | D_k \mu_k, D_k \Sigma_k D_k), \tag{12}$$

where $h(\cdot|\mu, \Sigma)$ is the multinormal density of mean μ and variance matrix Σ. This optimization may be done by using the EM algorithm. Celeux and Govaert (1995) showed that, at the M step of this algorithm, maximizing the likelihood on D_k is the same as minimizing the following functional

$$\begin{aligned} f(D_1, \ldots, D_K) \;\; = \;\; & \sum_{k=1}^{K} \sum_{i=1}^{n^*} t_{ik} \{ \ln |D_k \Sigma_k D_k| \\ & + (x_i^* - D_k \mu_k)' D_k^{-1} \Sigma_k^{-1} D_k^{-1} (x_i^* - D_k \mu_k) \}, \end{aligned} \tag{13}$$

with t_{ik} ($i = 1, \ldots, n^*$, $k = 1, \ldots, K$) given by the previous step E in the following manner. If the label z_i^* of the ith individual is unknown, t_{ik} corresponds to the conditional probability that x_i^* belongs to the class k:

$$t_{ik} = \frac{p_k h(x_i^* | \hat{D}_k^- \mu_k, \hat{D}_k^- \Sigma_k \hat{D}_k^-)}{\sum_{q=1}^{K} p_q h(x_i^* | \hat{D}_q^- \mu_q, \hat{D}_q^- \Sigma_q \hat{D}_q^-)}, \tag{14}$$

with $\hat{D}_k^-$ estimate of D_k obtained at the previous M step of EM. Otherwise, $t_{ik} = 1$ if x_i^* belongs to the group k (i.e. $z_i^* = k$), 0 if not.

Minimizing the function f under the constraint $D_k = \alpha I_d$ leads to solving a second order equation with the only non-negative solution

$$\hat{\alpha} = \frac{1}{2} \left\{ -\frac{\sum_k n_k \bar{x}_k^{*\prime} \Sigma_k^{-1} \mu_k}{nd} + \sqrt{\left[\frac{\sum_k n_k \bar{x}_k^{*\prime} \Sigma_k^{-1} \mu_k}{nd} \right]^2 + 4 \frac{\sum_{k,i} t_{ik} x_i^{*\prime} \Sigma_k^{-1} x_i^*}{nd}} \right\} \tag{15}$$

where $n_k = \sum_{i=1}^{n^*} t_{ik}$ and $\bar{x}_k^* = \sum_{i=1}^{n^*} t_{ik} x_i^* / n_k$.

5.2 *Model M_3 ($D_k = D$)*

5.2.1 Least squares estimator Using equation (9) from model M_2, we solve the system $\bar{x}^* = DE[X]$. It leads to

$$\{D\}_{jj} = \frac{\{\bar{x}^*\}_j}{\{E[X]\}_j} \qquad (j = 1, \ldots, d). \tag{16}$$

5.2.2 Maximum likelihood estimator Similarly to the case of model M_2, we can use a maximum likelihood approach and minimize the function f in order to maximize the likelihood under the constraint $D_k = D$. It is expressed by

$$\begin{aligned} f(D) \quad = \quad & -2n^* \ln |D^{-1}| + \sum_{i=1}^{n^*} x_i^{*\prime} D^{-1} \sum_{k=1}^{K} t_{ik} \Sigma_k^{-1} D^{-1} x_i^* \\ & -2 \sum_{k=1}^{K} n_k \mu_k' \Sigma_k^{-1} D^{-1} \bar{x}_k^* + cst. \end{aligned} \tag{17}$$

Using Theorem 2 of Appendix A, there exists a unique minimum $\hat{D}$ of f. It can be computed by any numerical method by starting from the parameter $D = I_d$ for instance.

5.3 *Model M_4 ($D_k = \alpha_k I_d$)*

Since relationship between variables is no longer group independent, equation (9) cannot be used and thus a least squares method is inappropriate. Therefore, only the maximum likelihood estimator is described.

Minimizing the function f under the constraint $D_k = \alpha_k I_d$ leads to solving a second order equation with the only non-negative solution

$$\hat{\alpha}_k = \frac{1}{2} \left\{ -\frac{\bar{x}_k^{*\prime} \Sigma_k^{-1} \mu_k}{d} + \sqrt{\left[\frac{\bar{x}_k^{*\prime} \Sigma_k^{-1} \mu_k}{d} \right]^2 + 4 \frac{\sum_{i=1}^{n^*} t_{ik} x_i^{*\prime} \Sigma_k^{-1} x_i^*}{n_k d}} \right\}. \tag{18}$$

5.4 *Model M_5 (general situation)*

We also minimize the function f in order to maximize the likelihood. This function is expressed by $f(D_1, \ldots, D_K) = \sum_{k=1}^{K} f_k(D_k)$ with

$$f_k(D_k) = -2n^* \ln |D_k^{-1}| + \sum_{i=1}^{n^*} x_i^{*\prime} D_k^{-1} (t_{ik} \Sigma_k^{-1}) D_k^{-1} x_i^* - 2n^* \mu_k \Sigma_k^{-1} D_k \bar{x}_k^* + cst. \tag{19}$$

Using Theorem 2 of Appendix A, there exists a unique minimum $\hat{D}_k$ for each f_k. It can be also computed by a numerical method.

Table 1
Number of estimated parameters for each model.

M_1	M_2	M_3	M_4	M_5	pM_1	pM_2	pM_3	pM_4	pM_5
0	1	d	K	dK	$K-1$	K	$d+K-1$	$2K-1$	$dK+K-1$

5.5 *Estimation of the proportions* $p_1^*, \ldots, p_K^*$

If proportions are not preserved, there is a need to estimate $p_1^*, \ldots, p_K^*$. This can be done by maximizing the likelihood by the EM algorithm again. The M step of EM gives the standard result $\hat{p}_k^* = n_k/n^*$.

5.6 *Choosing among models with the BIC criterion*

There exists five models $M_1, \ldots, M_5$ on D_k matrices and two models on proportions ($p_k^* = p_k$ or p_k^* is unknown, for all k), so, by combination, we obtain 10 models. We will note M_j for model M_j with $p_k^* = p_k$ and pM_j for model M_j with unknown p_k^*'s. One of these 10 models may be choosen by the user himself (the biologist context for example) since models are easily understood. The Bayesian criterion BIC (Schwarz, 1978) may also be used to automatically select one model. This criterion, to maximize, is given by $\mathrm{BIC(model)} = \ln \ell - \nu/2 \ln n^*$, where ℓ is the maximum likelihood value and ν is the number of estimated parameters (see Table 1). For models M_2 and M_3, if the least squares estimators are used, the likelihood associated to these estimators may be retained.

6. A test situation using a real data set

6.1 *Data*

Following the examples given by Van Franeker and Ter Brack (1993), we also chose seabirds from the family Procellaridae (petrels). In our example, the species is the Cory's Shearwater *Calanectris diomedea* (see Thibault and Bretagnolle, 1997 for a review of the biology and the biometrics of this species). Cory's Shearwaters breed in the Mediterranean and North Atlantic, where presumably contrasted oceanographic conditions have led to the existence of marked subspecies differing in size as well as coloration and behavior (Thibault and Bretagnolle, 1997). Subspecies are *borealis*, living in the Atlantic islands (the Azores, Canaries, etc.), *diomedea*, living in the Mediterranean islands (Balearics, Corsica, etc.), and *edwardsii*, from the Cape Verde Islands (Thibault et al., 1997). A sample of *borealis* ($n = 206$, 45% females) was measured using skins in several National Museums. Five morphological

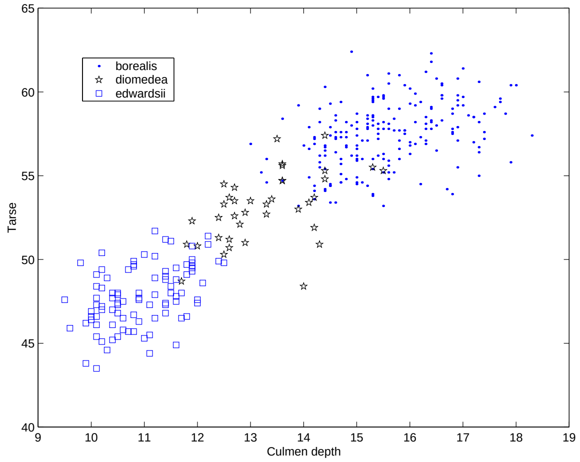


Figure 1. Data *borealis*, *diomedea* and also *edwardsii* for variables culmen depth and tarsus.

variables were measured: culmen (bill length), tarsus, wing and tail lengths, and culmen depth. Similarly, a sample of subspecies *diomedea* ($n = 38$, 58% females) was measured using the same set of variables. In this example, two groups are present, males and females (see Figure 1 for a scatter plot of two of the variables), and all the birds are of known sex (from dissection).

We will first consider the *borealis* sample as being the sexed (labeled) sample S, and the *diomedea* sample as being the non-sexed, or the partially-sexed, sample S^* (in fact, in our data, both samples are sexed but sex of *diomedea* will be only used to measure quality of results provided by the proposed method).

Before using the theoretical approach presented above, we need to verify that the three assumptions we made in Section 3.1 are satisfied. The first requirement is that the distribution of variable j in population P^* is mainly a transformation of the distribution of the same variable j in the population P. In a biological context, this means that the distribution of a morphological variable (e.g., culmen length) depends only on its distribution in the other population, while other factors (such as tarsus length, wing length etc. in this example), have a negligible influence in comparison to the main factor culmen length. The second assumption (each function ϕ_k^j is C^1) seems to be, for reasons of regularity, a desirable property in a real context. The third assumption ($b_k = 0$) can be verified directly in the following way. Since we know the sex of both samples in this example, we estimate, for both populations, parameters $\hat{\mu}_1$, $\hat{\mu}_1^*$ and $\hat{\Sigma}_1$, $\hat{\Sigma}_1^*$ for females and parameters $\hat{\mu}_2$,

Table 2
Cross-validation criterion value.

model	homoscedastic	heteroscedastic
borealis	**9.71**	13.59
diomedea	**15.79**	26.32

$\hat{\mu}_2^*$ and $\hat{\Sigma}_2$, $\hat{\Sigma}_2^*$ for males. Having $b_k = 0$ is equivalent to have

$$\hat{\mu}_k^* = D_k \hat{\mu}_k \tag{20}$$

and

$$\hat{\Sigma}_k^* = D_k \hat{\Sigma}_k D_k \tag{21}$$

with $k = 1, 2$. First we estimate D_k from (20) by $\{\hat{D}_k\}_j = \{\hat{\mu}_k^*\}_j / \{\hat{\mu}_k\}_j$. Then (21) has to be verified with this $\hat{D}_k$. A first step consists of estimating the correlation matrix $\hat{R}_k$ of $X_{|Z=k}$ which has to be computed: $\hat{R}_k = \hat{S}_k \hat{\Sigma}_k \hat{S}_k$ where $\hat{S}_k$ is a diagonal matrix with inverse of standard deviation of $X_{|Z=k}$ on the diagonal, i.e. $\{\hat{S}_k\}_{jj} = 1/\sqrt{\{\hat{\Sigma}_k\}_{jj}}$. Purpose of this transformation is to normalize variation of each feature. Thus equation (21) is equivalent to $\hat{S}_k \hat{\Sigma}_k^* \hat{S}_k = \hat{D}_k \hat{R}_k \hat{D}_k$. We verify that this equation is true by computing matrix norm $N_k = \|\hat{S}_k \hat{\Sigma}_k^* \hat{S}_k - \hat{D}_k \hat{R}_k \hat{D}_k\|$. We must have norms N_1 and N_2 close to zero to conclude respectively that $b_1 = 0$ and $b_2 = 0$. Choosing the norm taking the maximum eigenvalue of the matrix, we obtain with our data: $N_1 = 2.2749 \times 10^{-16}$ for females and $N_2 = 2.2219 \times 10^{-16}$ for males. Consequently, we cannot reject the hypothesis that $b_1 = b_2 = 0$.

6.2 *Results in the non-sexed case*

We consider in this section that all *diomedea* specimen are non-sexed.

A first step consists in computing the cross-validation criterion value to choose between homoscedastic and heteroscedastic models for both samples by using sex information on each of them (see Table 2 for values). Homoscedastic model is selected in both cases, and therefore, parameters of *borealis* are estimated by the homoscedastic model (with free proportions).

The second step consists now in applying parameters estimated by the *borealis* sample using the 10 models to the non-sexed *diomedea* sample. Results, empirical error rate (deduced from the true partition of *diomedea*) and BIC value, are given for each model for the least squares estimators (first column of Table 3) and for the maximum likelihood estimators (first column

Table 3
Empirical error rate (error) and BIC value (BIC) in the non-sexed case with least squares estimators.

		testing *diomedea*		testing *borealis*		testing *edwardsii*	
model	criterion	learning *borealis*	learning *edwardsii*	learning *diomedea*	learning *edwardsii*	learning *borealis*	learning *diomedea*
M_2	error	28.95	44.74	20.88	45.15	46.74	47.83
	BIC	-502.66	-631.52	-2898.55	-4747.34	-1557.79	-1604.64
M_3	error	21.06	13.16	16.02	14.08	13.05	11.96
	BIC	-451.58	-450.01	-2568.86	-2522.62	-1041.08	-1059.40
pM_2	error	42.11	42.11	34.47	45.15	47.83	47.83
	BIC	-489.90	-608.35	-2855.35	-4599.52	-1488.93	-1559.36
pM_3	error	18.43	13.16	15.54	14.08	10.87	10.87
	BIC	-453.37	-451.82	-2571.28	-2525.28	-1043.19	-1061.55

of Table 4). Moreover, empirical error rate of the cluster analysis situation is reported at the last line of Table 4. The clustering procedure (see for instance Celeux and Govaert, 1995) consists in estimating the Gaussian mixture parameters of the non-sexed sample *diomedea* with EM (after 20 random trials) and the optimal model of *diomedea* (the homoscedastic model with free mixing proportions). It is a situation in which no information is used from the *borealis* sample. Nevertheless, this method provides an optimistic error estimate of the clustering procedure, since the optimal model is used.

High error rates are generally obtained with standard discriminant analysis (models M_1 and pM_1) and with standard cluster analysis, as compared to the other models we propose (Tables 3 and 4). The best model selected by the empirical error rate is pM_3 (for both estimators). This model preserves homoscedasticity, a relevant property since both rules selected by cross-validation criterion were homoscedastic. Moreover it indicates that the proportion of females is not the same in the two samples. Model selected by the BIC criterion is M_3 and the error rate is the second best value. So, transformation from *borealis* to *diomedea* seems to be sex-independent but not variable-independent. It should be noted also that BIC's value for pM_3 is very close to the one for M_3.

Differences exist between error rates obtained using the two different estimators, but the maximum likelihood estimators seem overall to provide the best results. On the other hand, least squares estimators have the advantage of simplicity at least as a first approach.

Figures 2 and 3 display projection using the same variables as Figure 1 (i.e., culmen depth and tarsus) of the discriminant rule of all models (max-

Table 4

Empirical error rate (error) and BIC value (BIC) in the non-sexed case with maximum likelihood estimators.

		testing *diomedea*		testing *borealis*		testing *edwardsii*	
model	criterion	learning *borealis*	learning *edwardsii*	learning *diomedea*	learning *edwardsii*	learning *borealis*	learning *diomedea*
M_1	error	42.11	42.11	42.72	45.15	47.83	47.83
	BIC	-753.49	-1129.46	-4147.72	-15565.30	-4517.47	-2667.80
M_2	error	31.58	44.74	22.34	43.69	46.74	47.83
	BIC	-502.11	-631.17	-2897.08	-4665.91	-1555.87	-1602.38
M_3	error	18.43	13.16	15.54	**13.60**	**9.79**	**10.87**
	BIC	**-451.51**	**-449.99**	-2568.62	**-2522.57**	**-1040.82**	**-1059.29**
M_4	error	28.95	44.74	24.28	43.21	46.74	46.74
	BIC	-503.74	-632.92	-2894.06	-4666.66	-1558.01	-1604.36
M_5	error	21.06	13.16	25.73	17.00	11.96	18.48
	BIC	-457.69	**-455.85**	-2556.58	-2531.71	-1048.54	**-1056.12**
pM_1	error	42.11	42.11	45.15	45.15	47.83	47.83
	BIC	-725.24	-1103.25	-3982.45	-15416.02	-4446.57	-2619.79
pM_2	error	42.11	42.11	37.87	45.15	47.83	47.83
	BIC	-489.43	-608.33	-2842.27	-4522.03	-1486.52	-1557.90
pM_3	error	**15.79**	13.16	**14.57**	14.57	**9.79**	**10.87**
	BIC	**-453.20**	**-451.79**	**-2570.12**	-2525.18	**-1042.31**	**-1061.15**
pM_4	error	42.11	42.11	37.87	45.15	47.83	47.83
	BIC	-491.23	-610.15	-2835.88	-4524.70	-1488.78	-1560.17
pM_5	error	21.06	**10.53**	25.25	19.42	13.05	17.40
	BIC	-459.51	**-457.66**	**-2555.17**	-2533.70	-1049.79	-1058.25
clustering	error	44.74		44.78		11.96	

Table 5
Mean on the 30 samples of the empirical error rate (error) and the BIC value (BIC) in the partially-sexed case.

model	M_1	M_2	M_3	M_4	M_5
error	42.41	31.94	18.70	29.91	18.98
BIC	-753.49	-502.13	**-451.56**	-503.92	-457.95

model	pM_1	pM_2	pM_3	pM_4	pM_5	clustering
error	42.41	42.69	**15.37**	42.69	20.93	21.13
BIC	-725.99	-489.95	-453.32	-491.77	-460.74	–

imum likelihood estimators only) as well as the clustering procedure. They show that models M_1, M_2, M_4, pM_1, pM_2 and pM_4 (i.e., models that are variable independent) suggest discriminant rules that are actually away from the optimal ones, a result already obtained using error rates.

6.3 *Results in the partially-sexed case*

We consider in this section that two labels (i.e., therefore 5.26% of the data set) are known in the *diomedea* sample. Empirical error rate is obtained for the 36 *a priori* non-sexed birds. The two labels are choosen at random 30 times and, so, it leads to 30 partially-sexed samples.

The 10 models and cluster analysis (using also this new sex information) are applied successively to the 30 partially-sexed *diomedea* samples. Mean of the error rate and the BIC criterion are displayed in Table 5.

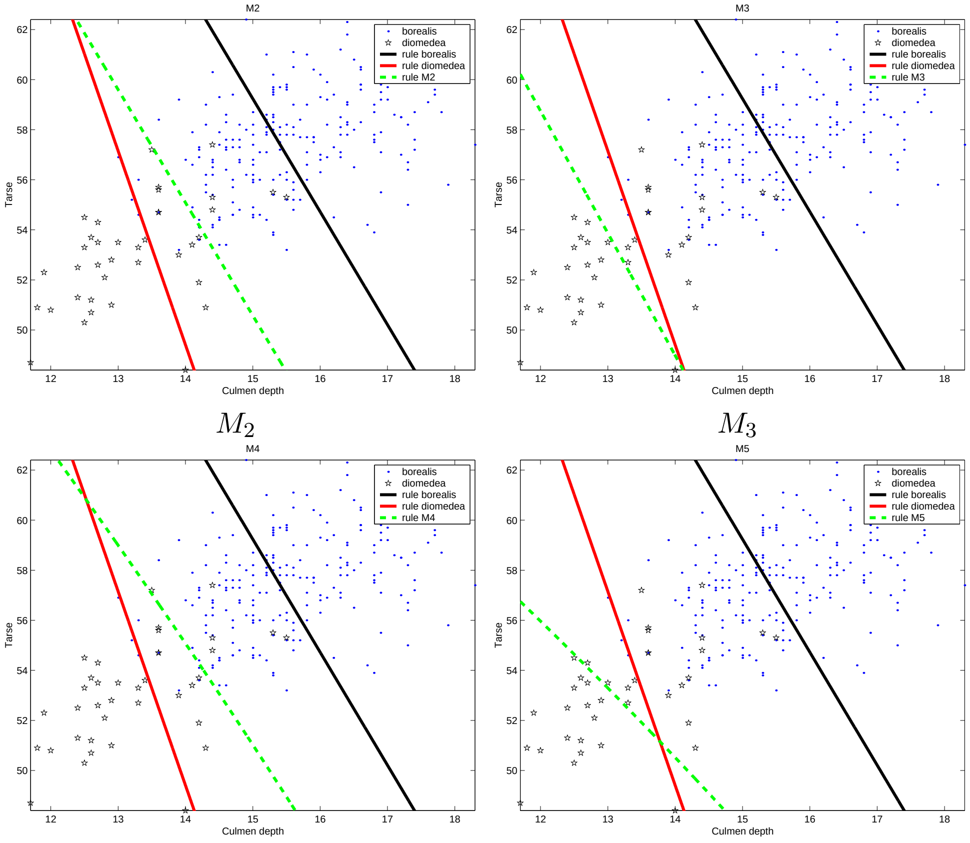
Partial information on sex provides lower error rates in models pM_3, pM_5, M_5 and the clustering method, with the model pM_3 still being the best (Table 5). The BIC criterion still selects the model M_3 (with a low error rate) and then pM_3.

We note that, except model M_5, only adapted models improve thanks to this new label knowledge. Moreover, the more complex the model is, the more the error of classification strongly decreases. This is the case for clustering: It has a good improvement in this example, coming from the last rank to a level close to pM_5.

6.4 *Further tests using the three populations*

We finally extend our approach as follows.

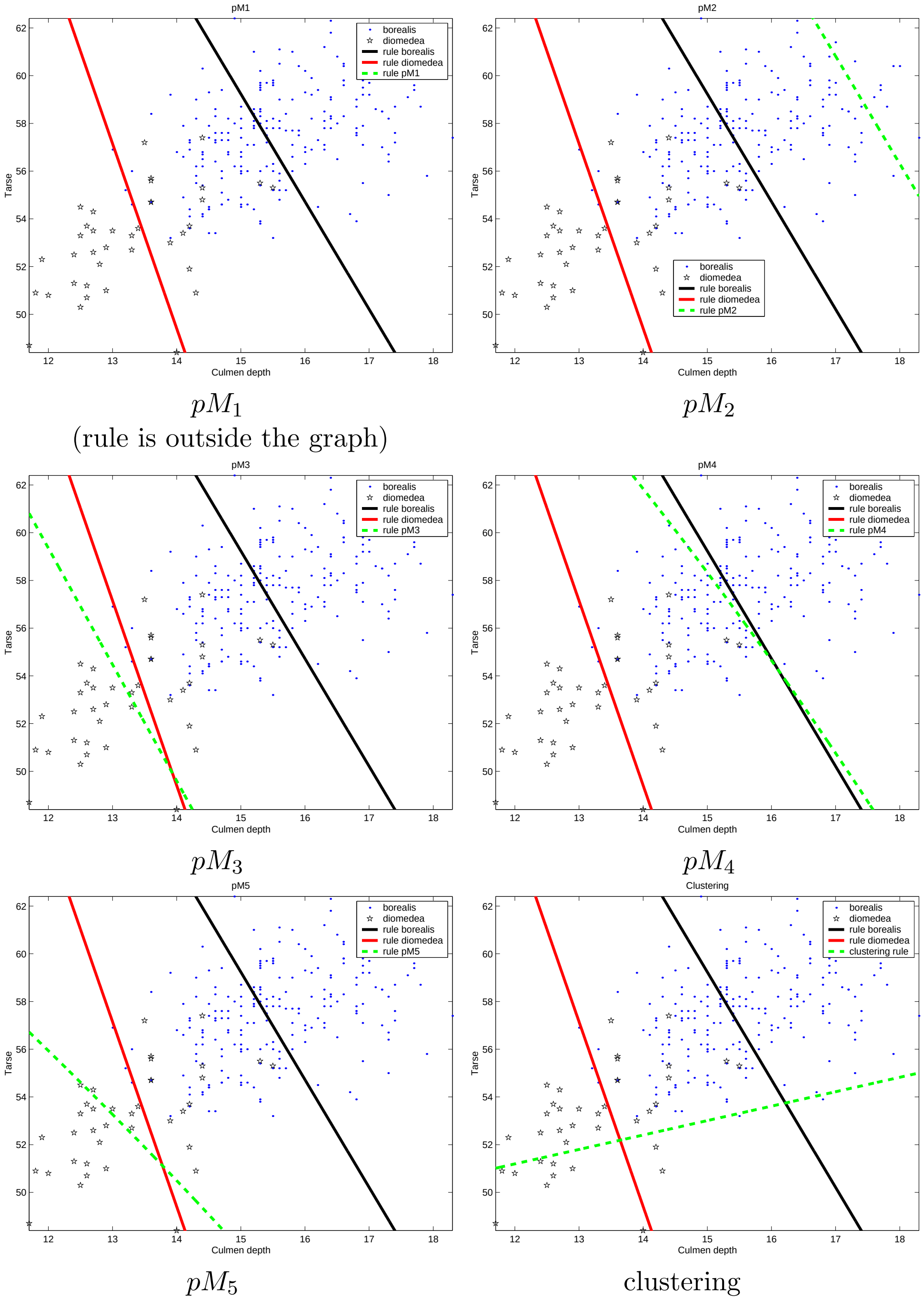
First, we consider the other Cory's shearwater subspecies, i.e. *edwardsii*, living in the Cape Verde Islands. The available sample is composed by 92



M_2 M_3

M_4 M_5

Figure 2. Projection on the two variables culmen depth and tarsus of the discriminant rules obtained from the five morphological variables in the non-sexed case. The right solid line, the left solid line and the dashed line are respectively for the "true" *borealis* rule, the "true" *diomedea* rule and the estimated *diomedea* rule (to be continued).



pM_1
(rule is outside the graph)

pM_2

pM_3

pM_4

pM_5

clustering

Figure 3. Projection on the two variables culmen depth and tarsus of the discriminant rules obtained from the five morphological variables in the non-sexed case. The right solid line, the left solid line and the dashed line are respectively for the "true" *borealis* rule, the "true" *diomedea* rule and the estimated *diomedea* rule (the end).

individuals measured on the same five morphological features as before (52% females are present in this sample, see also Figure 1).

Second, we use all possible pairs from the now three available samples, thus we obtain six different combinations of learning and testing data sets. For example, we have detailed in the previous experiments the pair with *borealis* as the learning sample and with *diomedea* as the testing one. Results (error and BIC values) are displayed for the five other pairs in Table 4 for the maximum likelihood estimators and in Table 3 for the least squares estimators.

Similarly to previous experiments, standard discriminant analysis (model M_1 and its extension, pM_1) shows very high error rates. The four variable-independent models M_2, M_4, pM_2 and pM_4 lead to poor results too. Models M_3 (and pM_3) still give usually better results than models M_5 (pM_5). Models pM_3 (pM_5) always have lower error rates than M_3 (M_5). Nevertheless, this is not the case when *edwardsii* is used as the learning or as the testing sample because proportion of females in *edwardsii* is not too far from female proportions of both *diomedea* and *borealis*.

As noted before, few differences exist between error rates of the two kinds of estimation, maximum likelihood and least squares. Moreover, as pointed out already, the BIC criterion has to be used carefully (as for any information criterion): It is better to retain several models with relatively similar order value of BIC than only the one with the best value. Finally, we note that it is always possible, in these experiments, to obtain a model that is better than the clustering result.

7. Concluding remarks

We present an extension of the standard discriminant analysis in the context of multinormal distributions. The main contribution of the present work is to consider the situation where the learning sample and the testing sample do not necessarily arise from the same population. By establishing, conditionally to the labels, a linear transformation between distributions of both populations, we obtain an allocation rule with few parameters to estimate.

Efficiency of this approach is illustrated by experiments in a biological context: In all tested cases, our method exhibits better performances than classical classification or clustering. In these experiments, best models are generally feature dependent but label independent, and thus models M_3 and pM_3 are retained. The most complex models (M_5 and pM_5) are also satisfactory models since number of estimated parameters is not too high. More experiments would be necessary to confirm these results.

In our likelihood approach, we used a plug-in procedure since the training parameters are estimated solely from the training data and, then, these estimates are plugged into the likelihood function to be used for estimation of additional parameters in the test population. In the place of this approximate likelihood procedure, parameters of both the training population and of the linear transformation may be estimated at the same time by the likelihood formed from these two kinds of parameters. In spite of some foreseeable difficulties of optimization implied by this global approach, comparison of performance with the plug-in procedure is a prospect of interest.

It would be also interesting to extend other classical discriminant methods to the case where the learning population and the testing population are different. Beyond the normal hypothesis treated here, one could consider qualitative data, non-parametric discrimination, logistic discrimination, etc. In each new situation, the main challenge will be to exhibit a realistic relationship between both populations in order to estimate the discriminant rule of the non-labeled population.

Acknowledgements – Authors thank Prof. Michel Weil and the two reviewers for contributing to improve the paper.

Résumé

L'analyse discriminante classique fait l'hypothèse intrinsèque que l'échantillon étiqueté provient de la même population que celui dont les labels sont à déterminer. Dans ce travail, nous considérons que les deux populations peuvent être différentes. Dans le cas gaussien, nous établissons une relation en loi, liant linéairement chaque groupe des deux populations. L'estimation des paramètres de cette relation permet alors de déduire une estimation de la règle de classement de la population à étiqueter. Différents modèles concernant cette relation sont proposés et des estimateurs des paramètres sont fournis. Une illustration est fournie par l'estimation du sexe d'oiseaux qui diffèrent de par leur provenance géographique et une comparaison à la discrimination classique est menée. L'extension à un échantillon à classer déjà partiellement étiqueté est aussi discutée et permet de confirmer les premiers résultats encourageants obtenus dans le cas complètement non étiqueté.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**, 716–723.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Bretagnolle, V., Genevois, F. and Mougeot, F. (1998). Intra- and intersexual function in the call of a non-passerine bird. *Behaviour* **135**, 1161–1184.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious models. *Pattern Recognition* **28**, 781–793.

Celeux, G. and Nakache, J. P. (1994). *Analyse discriminante sur variables qualitatives.* Polytechnica, Paris.

De Meyer, B., Roynette, B., Vallois, P. and Yor, M. (2000). On independent times and positions for Brownian motion. Technical Report 1, Les prépublications de l'Institut Élie Cartan, Institut Elie Cartan, Vandœuvre lès Nancy, France.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188. Pt. II.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis - nonparametric discrimination: Consistency properties. Technical report, Report of the U.S.A.F. School of Aviation Medicine, Agrawala (1977).

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76**, 817–823.

Genevois, F. and Bretagnolle, V. (1995). Sexual dimorphism of voice and morphology in thin-billed prions, pachyptila belcheri. *Notornis* **42**, 1–10.

Gnanaddesikan, R. (1989). Discriminant analysis and clustering, panel of experts. *Statistical Science* **4**, 34–69.

Hand, D. J. (1986). Recent advances in error-rate estimation. *Pattern Recognition letters* **4**, 335–346.

Lachenbruch, P. A. and Goldstein, M. (1979). Discriminant analysis. *Biometrics* **35**, 68–85.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification (with discussion). *Journal of the Royal Statistical Society, Series B* **10**, 159–203.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Thibault, J.-. C. and Bretagnolle, V. (1997). A mediterranean breeding population of cory's shearwater which shows behavioral and biometrical characters of the atlantic subspecies. *Ibis* **140**, 523–528.

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

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

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

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

Appendix A
Propositions and proofs

Theorem 1. (De Meyer et al., 2000) *If $Y \sim N(0,1)$ (the standard normal distribution) and $Y \sim \phi(Y)$ with ϕ a C^1 function $\Re \to \Re$, we necessarily have $\phi(y) = \pm y$. We can easily extend this result to $Z \sim \phi(X)$ with $X \sim N(\mu_X, \sigma_X^2)$ ($\sigma_X^2 > 0$) and $Z \sim N(\mu_Z, \sigma_Z^2)$ ($\sigma_Z^2 > 0$). In such case, we obtain the linear relation $\phi(x) = \alpha x + \beta$, where $\alpha, \beta \in \Re$.*

Proof. First, ϕ is strictly monotone. In the contrary, there would exist a point a in $\Re$ with $\phi'(a) = 0$ and so the random variable $\phi(Y)$ would have an infinite density at $b = \phi(a)$. Indeed, noting F the standard cumulative distribution, $F'(\phi(a)) = \phi'(a) f(\phi(a)) = 0$ if the density $f(\phi(a))$ is finite. Second, let us suppose that now ϕ is increasing. We have

$$F(a) = P(Y \leq a) = P(\phi(Y) \leq \phi(a)) = F(\phi(a)) \tag{A.1}$$

and so $\phi(a) = a$. We conclude by assuming that now ϕ is decreasing.

Theorem 2. *The function*

$$f(A) = -\ln|A| + \sum_{i=1}^{n} x_i' A \Gamma A x_i - u' A v \tag{A.2}$$

with A a diagonal non-negative definite matrix $\Re^{d\times d}$, Γ a non-negative definite matrix $\Re^{d\times d}$, u, v, x_i vectors of $\Re^d$, has a unique minimum $\hat{A}$.

Proof. We note

- a the $\Re^d$ vector composed by the diagonal elements $a^1, \dots, a^d$ of A,
- X_i the diagonal $\Re^{d\times d}$ matrices with the elements $x_i^1, \dots, x_i^d$ on the diagonal,
- V the diagonal $\Re^{d\times d}$ matrix with the elements $v^1, \dots, v^d$ on the diagonal.

The function f is now expressed by

$$f(A) = \tilde{f}(a) = -\sum_{j=1}^{d} \ln |a^j| + a' \left[\sum_{i=1}^{n} X_i \Gamma X_i \right] a - u'Va. \tag{A.3}$$

To express the Hessian matrix of $\tilde{f}$, let us proceed to the first and second order derivations:

$$\frac{\partial \tilde{f}(a)}{\partial a^\ell} = -\frac{1}{a^\ell} + 2\sum_{j=1}^{d} \left\{ \sum_{i=1}^{n} X_i \Gamma X_i \right\}_{\ell,j} a^j - \{u'V\}_\ell. \tag{A.4}$$

$$\frac{\partial^2 \tilde{f}(a)}{\partial a^\ell \partial a^m} = \begin{cases} \ell = m : & (1/a^\ell)^2 + 2\{\sum_{i=1}^n X_i \Gamma X_i\}_{\ell,\ell} \\ \ell \neq m : & 2\{\sum_{i=1}^n X_i \Gamma X_i\}_{\ell,m}. \end{cases} \tag{A.5}$$

So the Hessian matrix H is given by

$$H = 2\sum_{i=1}^{n} X_i \Gamma X_i + A^{-2}. \tag{A.6}$$

All matrices $X_i \Gamma X_i$ are non-negative definite since

$$\forall a \in \Re^d, a' X_i \Gamma X_i a = w_i' \Gamma^{-1} w_i > 0 \tag{A.7}$$

with $w_i = X_i a \in \Re^d$ (recall Γ is non-negative definite). Moreover, sum of non-negative definite matrices being non-negative definite, H is non-negative and the function $\tilde{f}$ is strictly convex on each subspace delimited by $a = 0$, so especially on the subspace of interest where each a component is non-negative. Moreover, it is easy to show that $\lim \tilde{f}(a) \to \infty$ when $\|a\|_2 \to \infty$ and also when $\|a\|_2 \to 0$. In conclusion, there exists a unique minimum $\hat{u}$ of $\tilde{f}$ in the subspace of interest, so a unique minimum $\hat{A}$ of f.