Working Paper 92-28
June 1992

Divisi6n de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

A DISCRIMINANT RULE UNDER TRANSFORMATION

Santiago Velilla and Juan A. Barrio*

Abstract

We present a new rule for discriminating among continuous populations which are not multivariate normal. The basic idea is to construct the sample maximum likelihood discriminant rule after transforming the data by a suitable multivariate transformation to normality

Key words:
Cross-validation; Multivariate Box-Cox Transformation.

*Velilla, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid; Barrio, Departamento de Física At6mica, Molecular y Nuclear, Universidad Complutense de Madrid.

1. INTRODUCTION

Discriminant analysis is a widely applied technique in multivariate data analysis. Given g groups, populations or classes, G_1, G_2, ..., G_g, with corresponding probability density functions $f_i(x)$ on $\mathbb{R}^p$, i=1, ...,g, the problem is allocating an individual to one of these groups on the basis of his measurements $\mathbf{x}=(x_1, ..., x_p)'$ which are considered as an observation from a population described by a random vector $\mathbf{X}=(X_1,X_2, ...,X_p)'$ such that $X\sim f_i$ under G_i, i=1, ...,g. The allocation should be "optimal" in the sense of minimizing, on average, the number of incorrect assignments. When the densities $f_i(x)$ are known, a suitable discriminating rule is the *maximum likelihood* (ML) *discriminant rule* (Mardia et al. (1979), p. 301), defined as follows: Assign $\mathbf{x}$ to G_i if

$$f_i(\mathbf{x})= \max_{1\leq j\leq g} f_j(\mathbf{x}). \qquad (1)$$

We will assume, for the rest of this work, that the a priori probabilities of the classes are equal. When this is the case, it can be shown that the rule (1) minimizes the total probability of misclassification (see Seber (1984), p. 331). When the densities $f_i(x)$ depend on unknown parameters θ_i, i.e., if $f_i(x)=f_{\theta_i}(x)$, i=1, ...,g, we use the *sample ML* (SML) *discriminant rule* (Mardia et al. (1979), p. 309), which is obtained by replacing, in rule (1), the parameters θ_i by efficient estimates $\hat{\theta}_i=\hat{\theta}_i(\mathcal{X}_i)$, where $\mathcal{X}_i$ is a n_ixp data matrix from G_i, i=1, ...,g.

An important example of the ideas above appears when $f_i\sim N_p(\mu_i,\Sigma_i)$, and the parameters (μ_i,Σ_i) are unknown, i=1, ...,g. Let $\overline{x}_i$ and S_i be, respectively, the sample mean and the covariance matrix of $\mathcal{X}_i$, i=1, ...,g. The SML rule can adopt two possible different forms: (*i*) If $\Sigma_1=\Sigma_2= ...=\Sigma_g=\Sigma$, say, we assign $\mathbf{x}$ to G_i if

$$\boldsymbol{\ell}_i'\mathbf{x}+c_i = \max_{1\leq j\leq g} [\boldsymbol{\ell}_j'\mathbf{x}+c_j], \qquad (2)$$

where $\boldsymbol{\ell}_i=S_p^{-1}\overline{x}_i$, $c_i=-(1/2)\overline{x}_i'S_p^{-1}\overline{x}_i$, $i=1, \ldots, g$, and S_p is the pooled estimator $S_p=\sum_{i=1}^{g}(n_i-1)S_i/(N-g)$, $N=\sum_{i=1}^{g}n_i$; *(ii)* If the dispersion matrices are unequal, we assign $\mathbf{x}$ to G_i if

$$Q_i(\mathbf{x})= \min_{1\leq j\leq g} Q_j(\mathbf{x}), \qquad (3)$$

where $Q_i(\mathbf{x})=(1/2)\log[|S_i|]+(1/2)[(\mathbf{x}-\overline{x}_i)'S_i^{-1}(\mathbf{x}-\overline{x}_i)]$, $i=1, \ldots, g$.

It is well-known that, when $g=2$, the rule (2) is equivalent to the *Fisher's Linear Discriminant Function* obtained by Fisher (1936) not assuming normality but looking for a "sensible" rule for discriminating between the groups. For $g\geq2$, if we assume that the possible distributions of the random vector $\mathbf{X}=(X_1,X_2, \ldots,X_p)'$ are continuous but not multivariate normal, the rules (2) and (3) above are, in general, no longer appropriate, and alternative procedures should be used. For example (see, e.g., chapter 6 in Seber (1984) or chapter 13 in Krzanowski (1990) for details) we could use a logistic discriminant rule or we could use the ML discriminant rule (1) replacing $f_i(x)$ by a nonparametric density estimator computed from $\mathcal{X}_i$, $i=1, \ldots, g$.

In this paper, we present a new method for discriminating among continuous populations which are not multivariate normal. The method is based on the multivariate generalization of the transformation of Box and Cox (1964) given by Andrews et al. (1971). In section 2, we introduce the general ideas of this new discrimination procedure. In section 3, we explicitly construct the associated discriminating rule and suggest a cross-validation method for assessing its performance. Section 4 contains an example of application and section 5 some final comments.

2. MOTIVATION

Let X be a random variable which takes values denoted by x. If X is not normal, it is often convenient to consider a transformation that might help to normalize the data. A useful family of transformations is the family of Box and Cox (1964):

$$x^{(\lambda)}=\begin{cases}\frac{x^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \log x, & \lambda=0.\end{cases} \tag{4.1}$$

We assume that the following model holds

$$X^{(\lambda)}\sim N(\mu,\sigma^2), \tag{4.2}$$

at least approximately. When we have a random vector $X=(X_1,X_2, \ldots,X_p)'$ whose distribution is not $N_p(\mu_1,\Sigma_1)$, Andrews et al. (1971) have given the following multivariate generalization of the setup (4.1)-(4.2). We have a p-vector $\Lambda=(\lambda_1, \ldots,\lambda_p)'$ of transformation parameters, one for each dimension, such that, approximately, the model

$$X^{(\Lambda)}=(X_1^{(\lambda_1)},X_2^{(\lambda_2)}, \ldots,X_p^{(\lambda_p)})'\sim N_p(\mu,\Sigma) \tag{5}$$

holds. In both (4) and (5), it is assumed that x and the components X_j are positive. If not, x and X_j must be shifted by adding suitable constants.

Write the parameters in model (5) as $\Theta=(\Lambda,\mu,\Sigma)$. Given an observation $x=(x_1, \ldots,x_p)'$ from the model (5), the likelihood of **x** is given by

$$\ell(\mathbf{x};\Theta)=(2\pi)^{-(p/2)}|\Sigma|^{-(1/2)}\exp[-(1/2)(\mathbf{x}^{(\Lambda)}-\mu)'\Sigma^{-1}(\mathbf{x}^{(\Lambda)}-\mu)]J_{\Lambda}(\mathbf{x}), \tag{6}$$

where $J_{\Lambda}(\mathbf{x})$ is the jacobian of the transformation, $J_{\Lambda}(\mathbf{x})=\prod_{j=1}^{p}x_j^{\lambda_j-1}$. If we now face the problem of discriminating among groups $G_1, \ldots,G_g$ such that the respective densities $f_i(x)$ are not multivariate normal but can be, in turn, modelled by (5), i.e. such that $f_i(x)=f_{\Theta_i}(x)=\ell(x;\Theta_i)$, where $\Theta_i=(\Lambda_i,\mu_i,\Sigma_i)$, i=1, ...,g, we can construct the following SML discriminant rule: Assign $\mathbf{x}$ to G_i if

$$\ell(\mathbf{x};\hat{\Theta}_i)=\max_{1\le j\le g}\ell(\mathbf{x};\hat{\Theta}_j), \tag{7}$$

where $\hat{\Theta}_i=\hat{\Theta}_i(\mathcal{X}_i)$ is an efficient estimator of Θ_i, i=1, ...,g. In the next section, we show how to compute explicitly the discriminant rule (7), we relate it to the rules (2) and (3), and suggest a procedure for assessing its performance.

3. CONSTRUCTION AND ASSESSMENT OF THE DISCRIMINANT RULE

3.1 Two groups

We will consider first the case when we have only two groups, G_1 and G_2. For i=1,2, we are given a data matrix $\mathcal{X}_i$ of $n_i$xp from G_i such that for some unknown vector $\Lambda_i=(\lambda_{i1}, \ldots, \lambda_{ip})'$ in $\mathbb{R}^p$ the rows of the $n_i$xp transformed data matrix $\mathcal{X}_i^{(\Lambda_i)}=(x_{ijk}^{(\lambda_{ik})})$, j=1, ...,$n_i$, k=1, ...,p, are i.i.d. $N_p(\mu_i,\Sigma_i)$. We also assume independence between $\mathcal{X}_1$ and $\mathcal{X}_2$.

3.1.1 Parameter estimation

Standard normal likelihood theory shows (see, e.g. Gnanadesikan (1977), p. 141), that, for i=1,2, the maximum likelihood estimator (MLE) of the parameter $\Theta_i=(\Lambda_i,\mu_i,\Sigma_i)$, denoted by $\hat{\Theta}_i=(\hat{\Lambda}_i,\hat{\mu}_i,\hat{\Sigma}_i)$, is given by $\hat{\mu}_i=\bar{x}_i^{(\hat{\Lambda}_i)}$, $\hat{\Sigma}_i=((n_i-1)/n_i)S_i^{(\hat{\Lambda}_i)}$, where $\bar{x}_i^{(\hat{\Lambda}_i)}$ and $S_i^{(\hat{\Lambda}_i)}$ are, respectively, the sample mean and the covariance matrix for $\mathcal{X}_i^{(\hat{\Lambda}_i)}$. For i=1,2, $\hat{\Lambda}_i$ is obtained by maximizing the corresponding concentrated log-likelihood which is (up to an additive constant)

$$L_{\max,i}(\Lambda_i)=-(n_i/2)\log[|\mathcal{X}_i^{(\Lambda_i)\prime}Q_i\mathcal{X}_i^{(\Lambda_i)}|]+\log[J_{\Lambda_i}(\mathcal{X}_i)], \tag{8}$$

where $Q_i=(I_{n_i}-1_{n_i}1_{n_i}'/n_i)$ and $J_{\Lambda_i}(\mathcal{X}_i)=\prod_{k=1}^{p}J_{\lambda_{ik}}$,

$J_{\lambda_{ik}}=\prod_{j=1}^{n_i}|dx_{ijk}^{(\lambda_{ik})}/dx_{ijk}|=(\prod_{j=1}^{n_i}x_{ijk})^{\lambda_{ik}-1}$, k=1, ...,p, are the jacobian terms. Define, for each column k of $\mathcal{X}_i$, the n_i vector of normalized variables $Z_{ik}^{(\lambda_{ik})}=(z_{ijk}^{(\lambda_{ik})})$, k=1, ...,p, of jth coordinate $z_{ijk}^{(\lambda_{ik})}=x_{ijk}^{(\lambda_{ik})}/J_{\lambda_{ik}}^{1/n_i}$, j=1, ...,$n_i$. We have the following result.

LEMMA 3.1 For i=1,2, let $\mathbf{Z}_i^{(\Lambda_i)}=(Z_{i1}^{(\lambda_{i1})}, \dots, Z_{ip}^{(\lambda_{ip})})$ be the $n_i$xp matrix of normalized variables associated with $\mathcal{X}_i$. We can write

$$\mathbf{L}_{\max,i}(\Lambda_i)=-(n_i/2)\log[|\mathbf{Z}_i^{(\Lambda_i)'}Q_i\mathbf{Z}_i^{(\Lambda_i)}|]. \tag{9}$$

Proof. Define, for i=1,2, the matrix $D_i=\mathrm{diag}(J_{\lambda_{ik}}^{1/n_i}, \dots, J_{\lambda_{ip}}^{1/n_i})$. (9) follows from (8), and by observing that $-(n_i/2)\log[J_{\Lambda_i}^{-2/n_i}(\mathcal{X}_i)]=-(n_i/2)\log[|D_i|^{-2}]$, and that $\mathbf{Z}_i^{(\Lambda_i)}=\mathcal{X}_i^{(\Lambda_i)}D_i^{-1}$. ■

For computational purposes, expression (9) is more convenient than (8) in order to determine the MLE $\hat{\Lambda}_i$. In experience of the authors, the functions $-L_{\max,i}(\Lambda_i)$ are typically convex, a convenient feature for solving the optimization problem by means of a canned numerical routine.

3.1.2 Expression of the rule

Given a new individual $\mathbf{x}=(x_1, \dots, x_p)'$ define the p vector $u(\mathbf{x})=-(\log x_1, \dots, \log x_p)'$. Two possible different situations arise:

(*i*) If $\Sigma_1\neq\Sigma_2$, our proposal is estimating $\theta_i=(\Lambda_i,\mu_i,\Sigma_i)$ with $(\hat{\Lambda}_i,\hat{\mu}_i,\hat{S}_i)$, where $\hat{S}_i=S_i^{(\hat{\Lambda}_i)}$. The rule (7) can be written equivalently in the form: Assign $\mathbf{x}$ to G_i if

$$\hat{Q}_i(\mathbf{x})+\hat{\Lambda}_i'u(\mathbf{x})=\min_{1\leq j\leq 2}[\hat{Q}_j(\mathbf{x})+\hat{\Lambda}_j'u(\mathbf{x})], \tag{10}$$

where $\hat{Q}_i(\mathbf{x})=(1/2)\log[\,|\hat{S}_i|\,]+(1/2)[(\mathbf{x}^{(\hat{\Lambda}_i)}-\hat{\mu}_i)'(\hat{S}_i)^{-1}(\mathbf{x}^{(\hat{\Lambda}_i)}-\hat{\mu}_i)]$, $i=1,2$. We can interpret (10) as the estimated normal theory discriminant rule (3) applied to the transformed data matrices $\mathcal{X}_1^{(\hat{\Lambda}_1)}$ and $\mathcal{X}_2^{(\hat{\Lambda}_2)}$ and shifted by a linear term in $u(\mathbf{x})$ which reflects the effect of the transformation;

(*ii*) If $\Sigma_1=\Sigma_2=\Sigma$, it is reasonable to estimate Σ by the pooled estimator $\hat{S}_p= [(n_1-1)\hat{S}_1+(n_2-1)\hat{S}_2]/(N-2)$, where $N=n_1+n_2$. The rule (7) becomes now: Assign $\mathbf{x}$ to G_i if

$$\tilde{Q}_i(\mathbf{x})+\hat{\Lambda}_i'u(\mathbf{x})= \min_{1\leq j\leq 2} [\tilde{Q}_j(\mathbf{x})+\hat{\Lambda}_j'u(\mathbf{x})], \qquad (11)$$

where $\tilde{Q}_i(\mathbf{x})=(1/2)[(\mathbf{x}^{(\hat{\Lambda}_i)}-\hat{\mu}_i)'\hat{S}_p^{-1}(\mathbf{x}^{(\hat{\Lambda}_i)}-\hat{\mu}_i)]$, $i=1,2$.

A criterion for deciding whether (10) or (11) should be used is suggested in the analysis of the example in section 4 below.

3.1.3 Cross-validation assessment

To evaluate the performance of the rules (10) and (11), we need to compute estimates of the error rates $p_{ij}=P[\text{Assign } \mathbf{x} \text{ to } G_i|G_j]$, i.e., the probabilities of allocating an individual to G_i when, in fact, it comes from G_j. A well established method for estimating the p_{ij} is the method of cross-validation proposed by Lachenbruch and Mickey (1968). The general idea in cross-validation is determining the allocation rule using the training data deleting one observation, and then using the rule to classify the observation left out. The estimates are $\hat{p}_{ij}=a_{ij}/n_i$, where a_{ij} is the number of observations from G_i which have been misclassified.

For the rules (10) and (11), we need to compute $\hat{\Lambda}_{i(j)}$, the MLE of Λ_i deleting the jth row of the data matrix $\mathcal{X}_i$, $i=1,2$, $j=1, \ldots,n_i$. This is computationally expensive and, therefore, an approximation is in order.

By (9), the MLE of Λ_i is obtained by minimizing $|M_i(\Lambda_i)|$, where $M_i(\Lambda_i)=Z_i^{(\Lambda_i)'}Q_iZ_i^{(\Lambda_i)}$. $\hat{\Lambda}_{i(j)}$ minimizes $|M_{i(j)}(\Lambda_i)|$. Let $\Lambda_{i,0}$ an initial guess for $\hat{\Lambda}_{i(j)}$ (typically $\hat{\Lambda}_i$) and let $\nabla_{i(j)}(\Lambda_i)=\partial|M_{i(j)}(\Lambda_i)|/\partial\Lambda_i$ and $H_{i(j)}(\Lambda_i)=\partial^2|M_{i(j)}(\Lambda_i)|/\partial\Lambda_i\partial\Lambda_i'$, be, respectively, the gradient vector and the hessian matrix of $|M_{i(j)}(\Lambda_i)|$. In the first-order Taylor expansion $\nabla_{i(j)}(\Lambda_i)\cong\nabla_{i(j)}(\Lambda_{i,0})+H_{i(j)}(\Lambda_{i,0})(\Lambda_i-\Lambda_{i,0})$, we can use the fact that $\nabla_{i(j)}(\hat{\Lambda}_{i(j)})=0$ to obtain a one-step approximation for $\hat{\Lambda}_{i(j)}$,

$$\hat{\Lambda}^1_{i(j)}=\Lambda_{i,0}-(H_{i(j)}(\Lambda_{i,0}))^{-1}\nabla_{i(j)}(\Lambda_{i,0}). \qquad (12)$$

Equation (12) is a multivariate extension of equation (15) in Tsai and Wu (1990).

To compute $\nabla_{i(j)}(\hat{\Lambda}_i)$ and $H_{i(j)}(\hat{\Lambda}_i)$ in (12), we define, for i=1,2; j=1, ...,n_i and k=1, ...,p: (*i*) $a_{ijk}=\alpha_{ijk}^{1/n_i}(\prod_{m\neq j}^{n_i}\alpha_{imk})^{-1/n_i(n_i-1)}$, $q_{ijk}(\lambda_{ik})=(a_{ijk})^{\lambda_{ik}-1}$; (*ii*) $Z^{*(\lambda_{ik})}_{ijk}=q_{ijk}(\lambda_{ik})Z^{(\lambda_{ik})}_{ik}$; and (*iii*) $A_{ij}=Q_iB_{ij}Q_i$, $B_{ij}=I_{n_i}-[1-(1/n_i)]^{-1}e_{ij}e'_{ij}$, where e_{ij} is the jth canonical vector of $\mathbb{R}^{n_i}$. We have the following result.

THEOREM 3.1 For i=1,2 and j=1, ...,n_i, define the $n_i$xp matrix $Z^{*(\Lambda_i)}_{ij}=(Z^{*(\lambda_{i1})}_{ij1}, \ldots, Z^{*(\lambda_{ip})}_{ijp})$. We can write:

$$M_{i(j)}(\Lambda_i)=Z^{*(\Lambda_i)'}_{ij}A_{ij}Z^{*(\Lambda_i)}_{ij}. \qquad (13)$$

Proof. We will use the following formula (see Atkinson (1986, p.31)): If a,b are two vectors in $\mathbb{R}^n$ and $P=(p_{uv})$ is an nxn orthogonal projection matrix, we have, for l=1, ..., n,

$$a'(I_n-P)b=a'_{(l)}(I_{n-1}-P_{(l)})b_{(l)}+(1-p_{ll})^{-1}a^*_lb^*_l, \qquad (14)$$

The subscript (l) means that the corresponding quantity has been computed deleting the lth row or coordinate. a^*_l and b^*_l are the lth coordinates of the residual vectors $a^*=(I_n-P)a$ and $b^*=(I_n-P)b$. For

$i=1,2$, $j=1, \ldots, n_i$ and $r,s=1, \ldots, p$, the (r,s) element in the $p \times p$ matrix $M_{i(j)}(\Lambda_i)=(m_{irs(j)}(\lambda_{ir},\lambda_{is}))$ is $m_{irs(j)}=Z_{ir\,(j)}^{(\lambda_{ir})'}Q_{i(j)})Z_{is\,(j)}^{(\lambda_{is})}$. By using an analog of expression (13) in Tsai and Wu (1990) and formula (14) above, we get $m_{irs(j)}=Z_{ijr}^{*(\lambda_{ir})}A_{ij}Z_{ijs}^{*(\lambda_{is})}$. ■

It is important to observe that, by (13), $\hat{\Lambda}_{i(j)}$ minimizes a determinant which is structurally similar as the determinant which minimizes $\hat{\Lambda}_i$. Expressions for the partial derivatives of $|M_{i(j)}(\Lambda_i)|$ can be seen in the appendix.

3.2 More than two groups

When $g>2$, the construction and assessment of the discriminant rule (7) can be done following straightforward extensions of the ideas in sections 3.1.1, 3.1.2 and 3.1.3.

4. A PRACTICAL EXAMPLE

In DELPHI, one of the detectors of LEP, a particle collider at the European Center for Nuclear Research (CERN, Geneva, Switzerland), several variables are observed by measuring the events produced in the particle collisions. The problem is classifying, or "tagging", on the basis of the observed measurements, the event containing a quark into one of three different groups:

G_1: Up and Down (u+d) or Strange (s);

G_2: Charm (c);

G_3: Beauty (b).

Among the set of variables which are observed, we will consider only a special subset of 9 continuous variables $(X_1, \ldots, X_9)$, called Microvertex variables. We are given simulated data matrices $\mathcal{X}_i$ from the groups G_i, such that $\mathcal{X}_i$ is n_ix9, with n_1=182, n_2=52, and n_3=66. We illustrate, with reference to this example, the discriminant techniques presented in section 3. The problem of classification has been also treated by de la Vaissière and Palma-Lopes (1989), two physicists at CERN.

The use of the rule in section 3 arises because of the strong nonnormality of the data. For example, figure 1 below shows the skewed histograms for the variable X_4 for the groups G_1, G_2 and G_3. We decided then, applying the multivariate transformation given in (5). Since, for other variables, negative data appear, we consider, for i=1,2,3, j=1, ...,n_i and k=1, ...,9, the translation $x_{ijk} \longrightarrow x_{ijk}+a_k$, where $a_k=-x_{min,k}+.5$ and $x_{min,k}$ is the global minimum of the kth variable over the three data matrices $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$, k=1, ...,p. We could have considered a shifted version of the transformation (4). Unfortunately, shifting the transformation (4) causes the maximum likelihood estimation of the transformation parameters to become a non-regular problem and not a well established procedure exists yet. However, see a recent work of Atkinson et al. (1991).

Figure 1

By maximizing the respective functions (9) for each class, we found the set of estimated transformation parameters displayed in table 1 below.

Table 1

In order to use (10) or (11) we need now to decide whether or not the dispersion matrices of the transformed data are the same. A reasonable procedure is to test for homoscedasticity treating the transformed data matrices $\mathcal{X}_1^{(\hat{\Lambda}_1)}$, $\mathcal{X}_2^{(\hat{\Lambda}_2)}$ and $\mathcal{X}_3^{(\hat{\Lambda}_3)}$ as data matrices from a multivariate normal distribution. The standard likelihood ratio test, as described for example in Mardia et al. (1979), p. 140, revealed strong evidence again homoscedasticity and, therefore, we decided to use the rule of section 3 in the form (10). It is important to remark that these procedure for choosing between (10) and (11) is not entirely rigorous because we are ignoring the effect of the sampling variability of the $\hat{\Lambda}_1$ in the null distribution of the likelihood ratio test statistic for homoscedasticity, i.e. we are treating the $\hat{\Lambda}_1$ as known constants. This procedure should be therefore taken as merely an indicative guideline for choosing between (10) and (11).

Finally, table 2 shows the estimated error rates by cross-validation. The rule classifies correctly slightly less than 70% of the events containing a beauty quark. This is remarkable because, from a physical point of view, these are the events where correct classification is more important. The relatively poor performance of the rule regarding events containing a quark of type c is not unexpected because G_2 is a transition class between G_1 and G_3 and, therefore, not very well differentiated from G_1 and G_3.

Table 2

5. CONCLUSIONS

This paper presents a new discriminant rule for the case when the groups do not have a multivariate normal distribution but can be modelled approximately as normal after a suitable transformation. From a practical point of view, the rule presented is particularly useful in the presence of long tailed asymmetric marginal distributions which are susceptible of treatment under the framework (4.1)-(4.2). The implementation of the rule should be accompanied with a diagnostic check of normality and heteroscedasticity of the transformed data.

APPENDIX: ELEMENTS OF $\nabla_{i(j)}(\hat{\Lambda}_i)$ AND $H_{i(j)}(\hat{\Lambda}_i)$.

For a general nxp matrix $H=(h_1, \ldots, h_p)$, we will write $H_j(u)$ for the nxp matrix obtained from H by replacing its jth column by the nx1 vector u. We will also write $H_{jl}(u,v)$ for the nxp matrix obtained from H by replacing its jth and lth column by, respectively, u and v.

THEOREM A.1 Write, for i=1,2, $\hat{\Lambda}_i=(\hat{\lambda}_{i1}, \ldots, \hat{\lambda}_{ip})'$ and define, for i=1,2, j=1, ...,p, and k=1, ...,p, the n_ix1 vectors $Z^{*(\lambda_{ik})}_{ijk}=q_{ijk}(\lambda_{ik})Z^{(\lambda_{ik})}_{ik}$ (see section 3.1.3), $W^{*(\lambda_{ik})}_{ijk}=\partial Z^{*(\lambda_{ik})}_{ijk}/\partial\lambda_{ik}$ and $U^{*(\lambda_{ik})}_{ijk}=\partial^2 Z^{*(\lambda_{ik})}_{ijk}/\partial\lambda^2_{ik}$. We will write Z^*_{ijk}, W^*_{ijk} and U^*_{ijk} for the corresponding vectors evaluated at $\hat{\lambda}_{ik}$. If $\mathbf{Z}^*_{ij}=\mathbf{Z}^{*(\hat{\Lambda}_i)}_{ij}$ (see section 3.1.3), we have, for r,s=1, ...,p,

(a) $\partial|M_{i(j)}(\Lambda_i)|/\partial\lambda_{ir}\Big|_{\Lambda_i=\hat{\Lambda}_i}=2|\mathbf{Z}^{*\prime}_{ij,r}(W^*_{ijr})A_{ij}\mathbf{Z}^*_{ij}|$;

(b) $\partial^2|M_{i(j)}(\Lambda_i)|/\partial^2\lambda_{ir}\Big|_{\Lambda_i=\hat{\Lambda}_i}=2[|\mathbf{Z}^{*\prime}_{ij,r}(W^*_{ijr})A_{ij}\mathbf{Z}^*_{ij,r}(W^*_{ijr})|$

$+|\mathbf{Z}^{*\prime}_{ij}A_{ij}\mathbf{Z}^*_{ij,r}(U^*_{ijr})|]$;

(c) $\partial^2|M_{i(j)}(\Lambda_i)|/\partial\lambda_{ir}\partial\lambda_{is}\Big|_{\Lambda_i=\hat{\Lambda}_i}=2[|\mathbf{Z}^{*\prime}_{ij,rs}(W^*_{ijr},W^*_{ijs})A_{ij}\mathbf{Z}^*_{ij}|$

$+|\mathbf{Z}^{*\prime}_{ij,r}(W^*_{ijr})A_{ij}\mathbf{Z}^*_{ij,s}(W^*_{ijs})|]$. (r≠s).

To obtain these expressions we need the following lemma.

LEMMA A.1 Let C, D, and B be three nxp matrices and let E be a symmetric matrix of nxn. If e_i represents the ith canonical vector of $\mathbb{R}^n$, we have, for i=1, ...,n, j=1, ...,p, the following differentiation formulae: a)$\partial|C'ED|/\partial c_{ij}=|C_j'(e_i)ED|$; b)$\partial|C'ED|/\partial d_{ij}=|C'ED_j(e_i)|$; and c)$\partial|B'EB|/\partial b_{ij}=2|B_j'(e_i)EB|$.

Proof. See lemma 3.1 in Velilla (1992). ■

Proof of theorem A.1. By the chain rule and part a) of lemma A.1, $\partial|M_{i(j)}(\Lambda_i)|/\partial\lambda_{ir}\Big|_{\Lambda_i=\hat{\Lambda}_i}=2\sum_{l=1}^{n_i}\sum_{m=1}^{p}|Z^{*\prime}_{ij,m}(e_{il})A_{ij}Z^{*}_{ij}|(\partial z^{*(\lambda_{im})}_{ijlm}/\partial\lambda_{ir}\Big|_{\Lambda_i=\hat{\Lambda}_i})=$
$2\sum_{l=1}^{n_i}|Z^{*}_{ij,r}(w^{*}_{ijlr}e_{il})A_{ij}Z^{*}_{ij}|=2|Z^{*\prime}_{ij,r}(W^{*}_{ijr})A_{ij}Z^{*}_{ij}|$. Expressions b) and c) are obtained in a similar way by using, respectively, parts a) and b) of lemma A.1. ■

Finally, if from $Z^{(\lambda_{ik})}_{ik}$ (see section 3.1.1), we define $W^{(\lambda_{ik})}_{ik}=\partial Z^{(\lambda_{ik})}_{ik}/\partial\lambda_{ik}$, $U^{(\lambda_{ik})}_{ik}=\partial^2 Z^{(\lambda_{ik})}_{ik}/\partial\lambda^2_{ik}$ and Z_{ik}, W_{ik}, U_{ik} stand for the corresponding functions evaluated at $\hat{\lambda}_{ik}$, it is easily seen that $W^{*}_{ijk}=q_{ijk}(\hat{\lambda}_{ik})[\log(a_{ijk})Z_{ik}+W_{ik}]$ and $U^{*}_{ijk}=q_{ijk}(\hat{\lambda}_{ik})[(\log(a_{ijk}))^2Z_{ik}+2\log(a_{ijk})W_{ik}+U_{ik}]$. As a conclusion, for the cross-validation assessment of the rules (10) and (11), we need the two MLE's $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$, the array of constants (a_{ijk}) and the n_ix1 vectors Z_{ik}, W_{ik} and U_{ik}. An explicit expression for the functions $Z^{(\lambda_{ik})}_{ik}$, $W^{(\lambda_{ik})}_{ik}$ and $U^{(\lambda_{ik})}_{ik}$, can be found in Atkinson and Lawrance (1989).

ACKNOWLEDGEMENTS

The authors are grateful to the DELPHI Collaboration for the kind permission for using the simulated data matrices $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{X}_3$. The authors also want to devote this work to the memory of R. Llosá.

REFERENCES

ANDREWS, D.F, GNANADESIKAN, R. & WARNER, J.L. (1971). Transformations of multivariate data. *Biometrics*, **27**, 825-840.

ATKINSON, A.C. & LAWRANCE, A.J. (1989). A comparison of asymptotically equivalent test statistics for regression transformation. *Biometrika*, **76**, 223-229.

ATKINSON, A.C., PERICCHI, L.R. & SMITH, A.F.M. (1991). Grouped likelihood for the shifted power transformation. *Journal of the Royal Statistical Society* , **B**, **53**, 473-482.

BOX. G.E.P. & COX, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society* , **B**, **26**, 211-252.

DE LA VAISSIÊRE, CH. & PALMA-LOPES, S. (1989). Multidimensional analysis: a tool for b-tagging?. Unpublished manuscript.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179-188.

GNANADESIKAN, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. (J. Wiley: New York).

KRZANOWSKI, W. J. (1990). *Principles of Multivariate Analysis: A User's Perspective*, Paperback Edition, (Clarendon Press: Oxford, U.K).

LACHENBRUCH, P.A., & MICKEY, M.R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1-11.

MARDIA, K. V., KENT, J.T & BIBBY, J.M. (1979). *Multivariate Analysis*. (Academic Press: London)

SEBER, G.A.F. (1984). *Multivariate Observations*. (J. wiley: New York).

TSAI, C. L. & WU, X. (1990). Diagnostics in transformation and weighted regression. *Technometrics*, **32**, 315-322.

VELILLA, S. (1992). A note on the multivariate Box-Cox transformation to normality. Universidad Carlos III de Madrid, Working Paper 92-08. (Submitted)

CAPTIONS FOR TABLES AND FIGURES

Table 1. Estimated transformation parameters for groups G_1, G_2 and G_3.

Table 2. Estimated error rates by cross-validation.

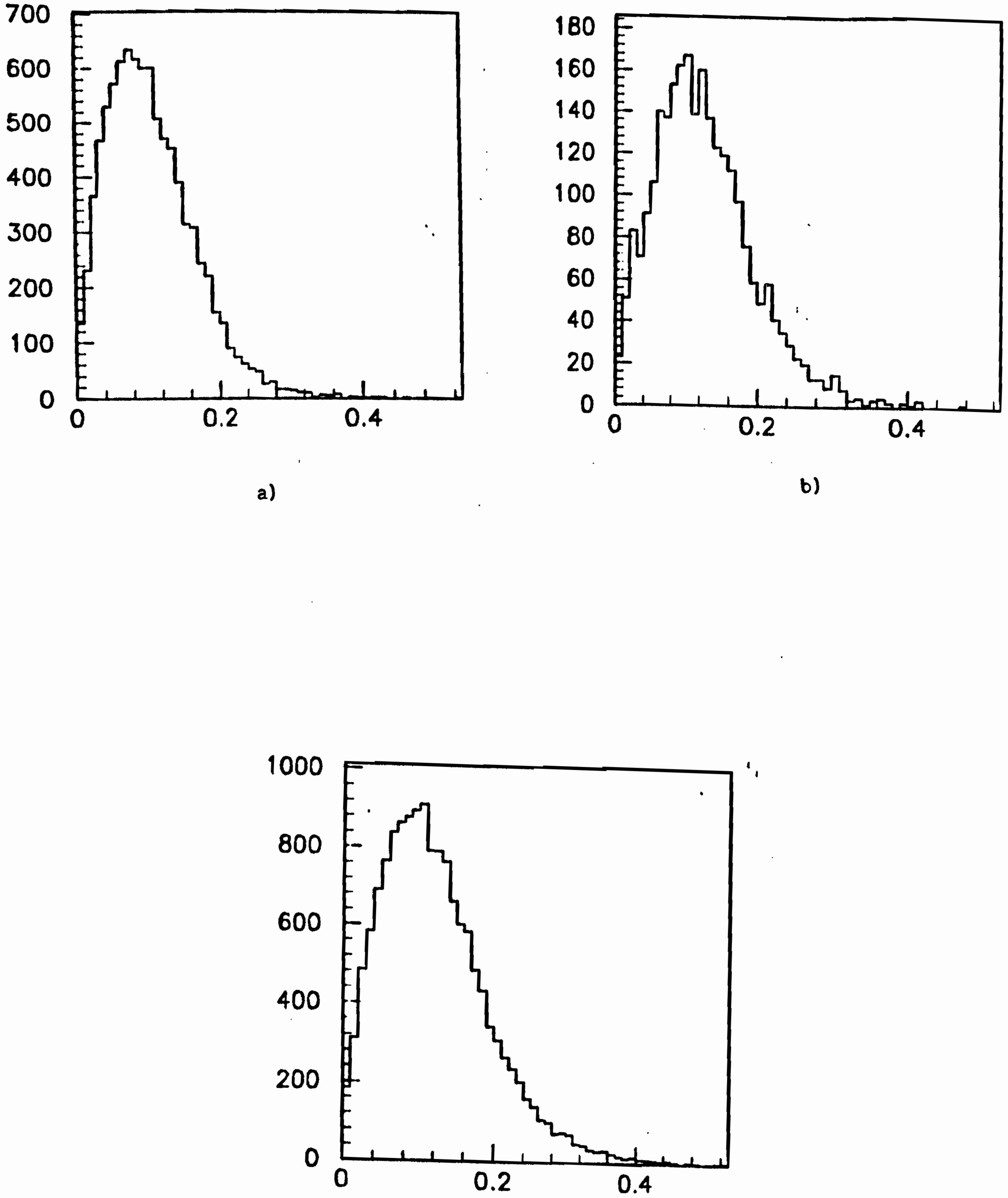
Figure 1. Histograms for X_4 for groups G_1, G_2 and G_3. a) G_1; b) G_2; c) G_3.

	1	2	3	4	5	6	7	8	9
$\hat{\lambda}'_1$	0.985	1.437	0.108	0.485	-2.534	-1.347	1.185	1.025	1.245
$\hat{\lambda}'_2$	0.295	-0.313	0.550	0.558	-1.736	-1.043	1.062	0.853	0.826
$\hat{\lambda}'_3$	-0.622	0.405	0.930	0.582	-0.922	-0.377	0.303	0.695	0.976

Table 1

	G_1	G_2	G_3
G_1	0.621	0.236	0.143
G_2	0.481	0.231	0.288
G_3	0.182	0.136	0.682

Table 2


700
600
500
400
300
200
100
0
0
0.2
0.4
180
160
140
120
100
80
60
40
20
0
0
0.2
0.4
1000
800
600
400
200
0
0
0.2
0.4

a)

b)

c).

Figure 1