Journal of Classification 10:25-49 (1993)

The Location Model for Mixtures of Categorical and Continuous Variables

W. J. Krzanowski

University of Exeter

Abstract: Recent research into graphical association models has focussed interest on the conditional Gaussian distribution for analyzing mixtures of categorical and continuous variables. A special case of such models, utilizing the homogeneous conditional Gaussian distribution, has in fact been known since 1961 as the location model, and for the past 30 years has provided a basis for the multivariate analysis of mixed categorical and continuous variables. Extensive development of this model took place throughout the 1970's and 1980's in the context of discrimination and classification, and comprehensive methodology is now available for such analysis of mixed variables. This paper surveys these developments and summarizes current capabilities in the area. Topics include distances between groups, discriminant analysis, error rates and their estimation, model and feature selection, and the handling of missing data.

Keywords: Classification; Discrimination; Distances; Error rates; Feature selection.

1. Introduction

Multivariate data sets containing mixtures of categorical and continuous variables arise frequently in practice. Various simple approaches to the

---

Constructive comments from the anonymous referees are gratefully acknowledged.

Author's Address: Mathematical Statistics and Operational Research Department, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, UK. E-mail: wjk@uk.ac.exeter.msor0 (JANET) or wjk@msor0.exeter.ac.uk (BITNET).

analysis of such data sets are possible: arbitrary categorization of all the continuous variables followed by analysis using standard methods for multivariate categorical data, or arbitrarily scoring all the categorical variables and then using standard methods for multivariate continuous data, or analyzing the categorical variables and the continuous variables separately (each by standard methods) and then attempting to synthesize the two sets of results. None of these options seems satisfactory for comprehensive analysis of the data, however. The first approach loses information in the categorization of continuous variables, the second introduces considerable subjectivity in the numerical scoring adopted, while the third ignores any associations existing between the categorical and the continuous variables.

A much more satisfactory general approach is first to specify a parametric model for mixed variables, then to fit the model to the data at hand and finally to use the parameter estimates for drawing inferences. By parametric model here is meant a suitable joint probability distribution for a set of q categorical variables and c continuous variables. Standard probability theory tells us that a joint distribution of p variables can be expressed as the conditional distribution of any subset of these variables given the values of the remainder, times the marginal distribution of these remaining variables. Thus if we want to specify the joint distribution of q categorical and c continuous variables then there appear to be two routes that we could take: as the conditional distribution of the categorical variables given the values of the continuous variables, times the marginal distribution of the latter; or as the conditional distribution of the continuous variables given the values of the categorical variables, times the marginal distribution of the latter.

The first possibility was briefly raised by Cox (1972), who suggested that the joint distribution of a mixture of binary and continuous variables could be written as a logistic conditional distribution of the binary variables for given values of the continuous variables, times a marginal multivariate normal distribution for the latter. However, this idea appears not to have been pursued any further in the analysis of mixed data sets, almost all work in the area focussing on the second route outlined above. Here it is assumed that the continuous variables have a different multivariate normal distribution for each possible setting of categorical variable values, while the categorical variables have an arbitrary marginal multinomial distribution. This model has been termed the "conditional Gaussian distribution" (CGD), and it forms the central plank of graphical association models for the analysis of mixed categorical and continuous variables. There has been a great deal of interest recently in these models, and full details can be found in the work of Lauritzen and Wermuth (1989), Edwards (1990), Wermuth and Lauritzen (1990) and Whittaker (1990, Chapter 11). We briefly summarize here the relevant technical results for our subsequent purposes.

Suppose that the q categorical variables and c continuous variables are denoted $\mathbf{X} = (X_1, X_2, \ldots, X_q)^T$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_c)^T$. Furthermore, assume that the i-th variable X_i has s_i possible categories so that overall there are $s = \prod_{i=1}^{q} s_i$ possible states, i.e., patterns, of discrete-variable values. The above model thus implies that if $\mathbf{X}$ falls in state j then $\mathbf{Y} \sim N(\mu_j, \Sigma_j)$ while the probability that $\mathbf{X}$ falls in state j is p_j ($j = 1, \ldots, s$; $\sum_{j=1}^{s} p_j = 1$). Hence the joint probability density of observing state j of $\mathbf{X}$ and value $\mathbf{y}$ of $\mathbf{Y}$ is

$$f(j,\mathbf{y}) = p_j (2\pi)^{-c/2} \mid \Sigma_j \mid^{-1/2} \exp \{ -\frac{1}{2} (\mathbf{y} - \mu_j)^T \Sigma_j^{-1} (\mathbf{y} - \mu_j) \} . \qquad (1)$$

By collecting terms and redefining parameters, this density can be rewritten in the form

$$f(j,\mathbf{y}) = \exp \{ \alpha_j + \beta_j^T \mathbf{y} - \frac{1}{2} \mathbf{y}^T \Omega_j \mathbf{y} \} . \qquad (2)$$

The parameters in (1) are called the "moment" parameters of the CGD, the triple (p_j, μ_j, Σ_j) comprising, respectively, the cell probability, the cell mean and the cell dispersion matrix for the j-th state, while the parameters in (2) are the "canonical" parameters of the CGD. Here α_j are scalars (the discrete canonical parameters), the β_j are c-element vectors (the linear canonical parameters) and the Ω_j are $(c \times c)$ positive-definite symmetric matrices (the cell precision matrices). Expanding (2) in terms of vector and matrix elements yields the form

$$f(j,\mathbf{y}) = \exp \{ \alpha_j + \sum_{k=1}^{c} \beta_{jk} y_k - \frac{1}{2} \sum_{k=1}^{c} \sum_{l=1}^{c} \gamma_{jkl} y_k y_l \} . \qquad (3)$$

Since the values of α_j, β_{jk} and γ_{jkl} depend on the state j of the discrete variables, and the latter can be viewed as "factors" in the terminology of design of experiments, then each of α_j, β_{jk} and γ_{jkl} can be expressed as a sum of main effects of the relevant individual discrete variables and interactions of all orders between them. This yields an expansion into terms resembling ANOVA or log-linear models.

A graphical association model is a model with density of the form (3), containing expansions in terms of main effects and interactions, in which all pairs of variables in a specified set are conditionally independent given the remaining variables. (This model is "graphical" because it is a model for multivariate random observations whose independence structure is characterized by a graph, so the word "graphical" should here be interpreted in the

context of mathematical graph theory; for full background details see Whittaker, 1990). Lauritzen and Wermuth (1989) established that two variables are conditionally independent given the rest if and only if all interaction terms involving the two variables are zero. Edwards (1990) defined hierarchical interaction models as the most general densities of form (3) in which the marginality principle is still respected (i.e., if a particular interaction term is set to zero then all interaction terms that "include" it are also set to zero). The goal of graphical modeling is then to determine the most parsimonious such model for a given set of data; the technical aspects concerned with fitting these models (maximum likelihood estimation of parameters with and without constraints, likelihood ratio tests, distributional results) are covered in the references cited earlier.

Although we will not be concerned specifically with graphical modeling here, it is pertinent to note that the full CGD model has appeared occasionally in other contexts. One such previous occurrence was in the calculation of distance between two populations (Krzanowski, 1983a). If we suppose that there are g populations, denoted $\pi_i\,(i = 1, \ldots, g)$, and that a different CGD is permitted in each population, then we must introduce an extra subscript into the model parameters to allow for the different populations. Thus p_{ij} now denotes the probability of cell j in population π_i, while μ_{ij} and Σ_{ij} respectively denote the mean vector and dispersion matrix of $\mathbf{Y}$ in cell j of population π_i. The density (1) then generalizes to:

$$f(j,\mathbf{y};\pi_i) = p_{ij}(2\pi)^{-c/2} \mid \Sigma_{ij} \mid^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y}-\mu_{ij})^T \Sigma_{ij}^{-1}(\mathbf{y}-\mu_{ij})\} \quad (4)$$

Krzanowski (1983a) surveyed the various possible general definitions of the distance Δ_{ab} between π_a and π_b, and chose to work with the Matusita (1956) definition, also known as the Hellinger distance. This definition involves calculation of the affinity ρ_{ab} between π_a and π_b, and Krzanowski (1983a) showed that for the case $\mathbf{z} = (j,\mathbf{y})$ and densities in (4),

$$\rho_{ab} = \sum_{j=1}^{s} (p_{aj}p_{bj})^{1/2} 2^{c/2} \mid \Sigma_{aj} \mid^{1/4} \mid \Sigma_{bj} \mid^{-1/4} \mid \mathbf{I} + \Sigma_{aj}\Sigma_{bj}^{-1} \mid^{-1/2} \times$$

$$\exp\{-\frac{1}{4}\sum_{k=1}^{c}[(\nu_{ajk}-\nu_{bjk})^2/(1+\lambda_{kj})]\} \quad (5)$$

where $\lambda_{ij}, \mathbf{l}_{ij}$ are solutions of $(\Sigma_{bj} - \lambda_{ij}\Sigma_{aj})\,\mathbf{l}_{ij} = 0$ and $\nu_{ajk} = \mathbf{l}_{kj}^T \mu_{aj}$.

Since "affinity" is the converse of "distance", possible measures of distance between π_a and π_b are $\Delta_{ab} = \{2(1-\rho_{ab})\}^{\frac{1}{2}}$, $\Delta_{ab} = -\log \rho_{ab}$ or $\tilde{\Delta}_{ab} = \cos^{-1} \rho_{ab}$. The first of these measures was used.

For practical applications, the parameters in (5) may be estimated from data by maximum likelihood, yielding intuitively reasonable estimates. $\hat{p}_{ij}$ is given by the proportion of individuals falling in state j of population π_i, while $\hat{\mu}_{ij}$ and $\hat{\Sigma}_{ij}$ are given by the mean vector $\bar{y}_{ij}$ and covariance matrix S_{ij} of the continuous variable values for these individuals. However, if s is at all large or if sample sizes are small, many of the states will have few observations and some Σ_{ij} will be poorly estimated. In this case it is possible to constrain the model, which will lead to *pooled* estimates. Various levels of pooling are possible:

(i) pool within states for each population (equivalent to assuming that the dispersion matrix is constant over cells in each population separately, i.e., that $\Sigma_{ij} \equiv \Sigma_i$ for $i = 1, \ldots, g$);

(ii) pool within populations for each state (equivalent to assuming that the dispersion matrix is constant over populations in each cell separately, i.e., that $\Sigma_{ij} \equiv \Sigma_j$ for $j = 1, \ldots, s$);

(iii) pool within populations and states (equivalent to assuming that the dispersion matrix is constant over cells and populations, i.e., that $\Sigma_{ij} \equiv \Sigma$ for all i,j).

Krzanowski (1984) provided a Monte Carlo estimation scheme for the null distribution of distance between π_a and π_b in case (iii), a result which enables some inferential procedures to be applied to the analyses of data sets in practice.

Case (iii) above, where the same dispersion matrix Σ is assumed for each combination of categorical variable values (i.e., at each discrete "location") is known as the *homogeneous* CGD case (in which the mixed interaction components of the canonical parameters α_j, β_{jk} and γ_{jkl} are all set to zero). This case was first introduced by Olkin and Tate (1961) under the name "location model" for analysis of mixed binary and continuous variables. These authors looked at canonical correlations between binary and continuous variables for various possibilities involving c and q, established population results connecting these canonical correlations and the continuous variable means μ_{ij}, and investigated the distribution theory for their estimates. Afifi and Elashoff (1969) extended the study of the model to the two-sample case. They investigated the effect of ignoring the binary nature of the x_i in calculating the usual two-sample Hotelling's T^2, and showed that the test was not consistent but that the distribution of T^2 depended on nuisance parameters. They then went on to derive an information-theoretic test of difference between groups and established the null distribution of the test statistic. In this work, they assumed that the parameter estimates $\hat{p}_{ij}$, $\hat{\mu}_{ij}$ and $\hat{\Sigma}_{ij}$ given above would be available for all binary-variable locations; otherwise the test could not be done.

The major practical developments of the location model that have taken place since these two pioneering papers have been almost exclusively in the context of discriminant analysis, and it is with this aspect that the current survey is concerned. In Section 2 we set up the basic location model formulation and summarize the different approaches adopted in practice, while in Section 3 we consider possible extensions of the basic ideas. Section 4 is concerned with model and feature selection aspects and problems, while Section 5 surveys alternative ways of tackling mixed-variable discrimination. In Section 6 we indicate how the graphical modeling ideas considered at the start can point the way to future developments.

2. Discriminant Analysis Methodology

2.1 Bayes Rule

We assume that there are two populations π_1 and π_2, discrimination between which is required. Historically, the location model methodology was developed from the starting point of a mixture of c continuous and q binary variables, and it is convenient to follow this line of development here. In this case we have $s_i = 2$ discrete variable categories for each i, and hence $s = 2^q$ states, or cells, altogether. If we denote the two possible 'values' of each binary variable as 0 and 1, then the s cells can be logically arranged in the order $j = \sum_{i=1}^{q} x_i 2^{i-1}$ where x_i is the value of the i-th binary variable. The location model thus specifies:

$$Pr(\mathbf{X} = j \mid \pi_i) = p_{ij} \text{ and } (\mathbf{Y} \mid \mathbf{X} = j, \pi_i) \sim N(\mu_{ij}, \Sigma)$$
$$\text{for } i = 1,2 \text{ and } j = 1, \ldots, s. \tag{6}$$

By forming the ratio of the joint probability densities in the two populations, it readily follows (see, e.g. Krzanowski 1975) that for equal costs due to the two types of misclassification and equal prior probabilities of group membership the Bayes classification rule is to allocate an individual with $\mathbf{X} = j$ and $\mathbf{Y} = \mathbf{y}$ to π_1 if

$$(\mu_{1j} - \mu_{2j})^T \Sigma^{-1} \{\mathbf{y} - \frac{1}{2}(\mu_{1j} + \mu_{2j})\} > \log(p_{2j}/p_{1j}) \tag{7}$$

and to π_2 otherwise. This allocation rule is, in effect, a different linear discriminant function for each discrete variable location. It is clear that the misallocation probabilities with this rule will therefore be the weighted sums of the misallocation probabilities at each location, these misallocation

probabilities being obtainable from standard linear discriminant theory and the weights being the location probabilities p_{ij}. Denoting by $p(\pi_i \mid \pi_j)$ the probability of allocating to π_i an individual that came from π_j, we have

$$p(\pi_i \mid \pi_j) = \sum_{m=1}^{s} p_{jm} \Phi \{(\log [p_{im}/p_{jm}] - \frac{1}{2} D_m^2)/D_m\} \text{ for } i \neq j . \qquad (8)$$

where $D_m^2 = (\mu_{1m} - \mu_{2m})^T \Sigma^{-1} (\mu_{1m} - \mu_{2m})$ is the squared Mahalanobis distance between π_1 and π_2 in location m.

If there are differential costs c_{12}, c_{21} due to misclassification of an individual, and differential prior probabilities q_1, q_2 of observing an individual from the two populations, the net effect is to add $k = \log (c_{12} q_2 / c_{21} q_1)$ to $\log (p_{im}/p_{jm})$ in both (7) and (8). We will assume $c_{12} = c_{21}$ and $q_1 = q_2$ for simplicity throughout.

In practice, of course, the population parameters will be unknown but random samples (''training sets'') are generally available from π_1 and π_2. The simplest approach that has been adopted in such cases is to estimate the population parameters from the training sets, and to replace the parameters in (7) and (8) by these estimates. Chang and Afifi (1974) considered the special case of $q = 1$, i.e., one binary variable, and assumed that there was at least one observation in each of the two binary variable locations in each population. Let there be n_{ij} observations in the j-th location of the training set from π_i, and let $\mathbf{y}_{ijk}$ be the k-th continuous variable vector in this location. The situation then corresponds exactly to a 2×2 (location $\times$ population) MANOVA, whence estimators of the population parameters are

$$\hat{\mu}_{ij} = \bar{\mathbf{y}}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} \mathbf{y}_{ijk}$$

$$\hat{\Sigma} = \mathbf{S} = \frac{1}{n-4} \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{n_{ij}} (\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{ij})(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{ij})^T$$

and

$$\hat{p}_{ij} = n_{ij}/n_i$$

where $n_i = \sum_{j=1}^{2} n_{ij}$ and $n = n_1 + n_2$. Chang and Afifi called the resulting allocation rule the ''double discriminant function''; Tu and Han (1982) studied this rule further, in particular discussing an ''inverse sampling'' procedure to ensure non-singularity of matrices.

As Chang and Afifi pointed out, there is no bar in principle to extension of the above approach for the case $q > 1$. However, it is evident that if

sample sizes are small, or if q (and hence s) is at all large, then there are bound to be locations for which no data are present in the training sets. What strategy is then to be followed when this location occurs in an individual to be classified? Also, there will be some locations with only one or two individuals present in the training sets, so the parameters for these locations will be very poorly estimated. It is therefore clear that an alternative to the naive estimation method given above is needed if this model is to have widespread practical utility. A second problem is that the misclassification probabilities (8) are derived under the assumption of conditional normality on the continuous variables. How can the performance of an allocation rule derived from (7) be assessed if this assumption is not satisfied?

Krzanowski (1975) tackled both of these problems, proposing a scheme for obtaining smoothed parameter estimates and outlining steps to make data-based error rate estimation feasible. For the parameter estimation we first note that the binary variables can be treated as if they were factors in a MANOVA context, the 2^q locations being the possible categories of a q-factor experiment where each factor has two possible levels and the multivariate response is $\mathbf{y}$. Then if we denote by ν_i the overall mean of $\mathbf{y}$ in population π_i, by α_{ij} the main effect of X_j, by $\beta_{i,jk}$ the interaction between X_j and X_k, and so on for interactions between the X_j of all orders, then we can express the μ_{ij} as the linear model:

$$\mu_{ij} = \nu_i + \sum_{u=1}^{q} \alpha_{iu} x_u + \sum\sum_{u<v} \beta_{i,uv}\, x_u\, x_v + \ldots + \gamma_{i,1,\ldots,q}\, x_1 \cdots x_q \qquad (9)$$

where x_u is the observed value of X_u in location j.

The above provides a MANOVA structure for the conditional means of the continuous variables. Moving on to the marginal distributions of the binary variables, we now have contingency tables of numbers of occurrences in each of the 2^q locations for the two populations. Thus we again have a 2^q factorial structure defined by the levels of the X_i, but now the responses at each location are incidences rather than realizations of a continuous vector $\mathbf{y}$. A standard approach for the analysis of such data is by formulating an analogous log-linear model for the expected values $\eta_{ij} = n_i p_{ij}$ in each location, so in our case we have the model

$$\log \eta_{ij} = \omega_i + \sum_{u=1}^{q} \delta_{iu}\, x_u + \sum\sum_{u<v} \phi_{i,uv} x_u x_v + \ldots + \psi_{i,1,\ldots,q} x_1 \cdots x_q \qquad (10)$$

where x_u is as before.

Such expansions in terms of the main effects of the individual x_i and the interactions of all orders between them link up with the expansions

discussed in the introduction to graphical modeling above. A current concern of graphical modeling would be to determine which terms of (9) and (10) to retain and which to delete in forming the most parsimonious model that fitted a given set of data. Krzanowski (1975), however, adopted the pragmatic approach of retaining only (and all) main effects and first-order interactions in both (9) and (10); he proposed fitting the resulting second-order models to the continuous variable parameters by multivariate regression and to the discrete variable parameters by iterative proportional fitting. This scheme involves $2q(q + 1) + 4$ parameters altogether. If the data are too sparse to admit such second-order models, then it should be possible to fit first-order models in which just the main effects are retained (giving $4q + 4$ parameters to be estimated); a possible intermediate stage is one in which separate main effects are fitted in the two populations, but the interactions are constrained to be equal across populations (i.e., $\beta_{1,uv} = \beta_{2,uv}$ and $\phi_{1,uv} = \phi_{2,uv}$ for all u,v with $q^2 + 3q + 4$ parameters to be estimated).

This approach ensures that estimates $\hat{\mu}_{ij}$ and $\hat{p}_{ij}$ are available even for those discrete-variable locations that have no observations in the training sets, so that the classification rule (7) can be estimated in all eventualities. What of the estimation of error rates induced by this rule? As mentioned above, using parameter estimates obtained from the second-order models in equation (8) will not give accurate assessment if the continuous variables are not normally distributed at each location, so a data-based method was sought. A suitable such method had earlier been proposed by Lachenbruch and Mickey (1968) in the now familiar leave-one-out method, for which each data point is omitted from the training sets in turn and classified on the basis of the allocation rule computed from the remaining observations; the proportion of individuals misallocated in each of the two training samples gives the two estimated error rates. Naive application of this procedure to large data sets may be feasible with modern computers, but at the time would have been computationally prohibitive with the location model. However, Krzanowski (1975) showed that various matrix identities could be employed advantageously in the multivariate regression, and that the iterative scaling computations could be arranged in sufficiently effective manner for the whole process to be carried out relatively simply and quickly. Various examples, both real and simulated, demonstrated both the efficiency and efficacy of the methodology.

Once a methodology was available for mixtures of binary and continuous variables, extension to general mixtures of categorical and continuous variables was extremely simple and was effected essentially by replacing an m-state categorical variable with $(m - 1)$ dummy binary variables and proceeding as before. Suppose the categorical variable with m states is replaced by the $(m - 1)$ dummy variables $X_1, \ldots, X_{m-1}$. Then state j of the

categorical variable can be indicated by setting $X_j = 1$ and $X_i = 0$ for all $i \neq j$ ($j = 1, \ldots, m-1$), in which case state m would be indicated by setting all X_i to zero. Note, however, that no more than one such dummy binary variable can have value 1 at any location so models (9) and (10) will be over-parameterized. Two extra features therefore had to be incorporated into the estimation scheme: (i) all interaction terms *within* each group of dummy binary variables had to be excluded from the linear model (9) for the μ_{ij} (to avoid break-down of the multivariate regression estimation procedure), and (ii) all multinominal states corresponding to joint incidences $x_u = 1$, $x_v = 1$ within each group of dummy binary variables had to be fixed at zero (to ensure correct iterative scaling estimates in the log-linear model (10)). Full details of this generalization were provided by Krzanowski (1980).

The Bayes allocation procedure (7) derives from the ratio of the two probability densities in the two populations, i.e. the ratio of the likelihoods for the observation to be classified. The problem in practice is to *estimate* this ratio, and the replacing of parameters of (7) by their estimates from the training data is the simplest and most commonly used way of doing so. However, two other general procedures have also been proposed: the hypothesis-testing method and the Bayesian predictive method. These approaches have been discussed in the context of multivariate normal data, and compared with the parameter-replacement approach for such data by Han (1979). We outline their implementation with the location model in the two following sections.

2.2 Hypothesis-testing Rule

Let us suppose that the training sets consist of n_1, n_2 individuals from π_1, π_2, respectively and denote the i-th individual in the training set from π_j by $\mathbf{v}_i^{(j)}$ ($i = 1, \ldots, n_j; j = 1,2$). Then the hypothesis-testing approach says that to allocate an individual $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$, we use the test statistic for the null hypothesis that all the $\mathbf{v}_i^{(1)}$ and $\mathbf{z}$ belong to π_1 while all the $\mathbf{v}_i^{(2)}$ belong to π_2 versus the alternative that all the $\mathbf{v}_i^{(1)}$ belong to π_2 while all the $\mathbf{v}_i^{(2)}$ and $\mathbf{z}$ belong to π_2.

Now the likelihood-ratio test statistic in this case is $T = \dfrac{\sup(L_{1m} \times L)}{\sup(L_{2m} \times L)}$, where L is the joint likelihood for all the $\mathbf{v}_i^{(j)}$ and L_{jm} is the likelihood for $\mathbf{z}$ in π_j given $\mathbf{x} = m$. Using the joint density from model (6), it is easy to show (see Krzanowski 1982) that

$$T = \{ |\hat{\Sigma}^{(1)}| / |\hat{\Sigma}^{(2)}| \}^{\frac{1}{2}(n_1+n_2+1)} \left\{ \prod_{i=1}^{2} \prod_{m=1}^{s} (\hat{p}_{im}^{(2)} / \hat{p}_{im}^{(1)})^{n_{im}} \right\} (\hat{p}_{1m}^{(2)} / \hat{p}_{2m}^{(2)}) \qquad (11)$$

where $\hat{\Sigma}^{(j)}, \hat{p}_{im}^{(j)}$ are the estimates of Σ, p_{im} respectively when $\mathbf{z}$ has been included with the training set from π_j $(j = 1,2)$. For stability, smoothed parameter estimates using second-order linear and log-linear models are again recommended. Krzanowski (1982) showed that simplified estimation of the parameters is obtained if all parameters are estimated for the training set data only, and then some simple algebraic identities are used to update inverses and determinants on including $\mathbf{z}$ successively with the two training sets. The final allocation rule is to classify $\mathbf{z}$ to π_1 if $T \geq 1$ and otherwise to π_2.

Error rates can again be estimated using the leave-one-out procedure, and this requires one initial estimation of all parameters using the training data only together with a re-estimation of all parameters when each individual is removed from its own training set and placed in the other one. Once again, some useful matrix and vector identities are available to enable the latter estimates to be obtained easily from the former ones; full details are given by Krzanowski (1982).

2.3 Bayesian Predictive Rule

The Bayesian approach to the problem is to postulate prior distributions for all the unknown parameters (μ_{ij}, Σ and p_{ij} for all i,j), use the likelihood of the training data under the location model to obtain posterior distributions of these parameters, multiply the joint density of $\mathbf{z}$ in each population by these posterior distributions and then integrate the resulting products with respect to the unknown parameters to obtain predictive densities of $\mathbf{z}$ in π_1 and π_2. The allocation of $\mathbf{z}$ is to the population in which it has the higher predictive density.

Vlachonikolis (1990) adopted the vague prior density $g(\{\mu_{ij}\},\Sigma) \propto |\Sigma|^{-\frac{1}{2}(c+1)}$ for the continuous variable parameters, and prior densities for the p_{ij} of the Dirichlet form $h(\{p_{ij}\} \mid \pi) \propto \prod_{j=1}^{s} p_{ij}^{\alpha_{ij}-1}$ where the α_{ij} are positive constants reflecting prior knowledge about the discrete variable locations. When no such prior information exists, he suggested setting $\alpha_{ij} = \alpha_i$ for all $j = 1, \ldots, s$ and $i = 1,2$. He then obtained expressions for the predictive densities of $\mathbf{z}$ in π_1 and π_2, both when the parameters μ_{ij}, Σ, and p_{ij} are estimated by the "naive" quantities $\bar{y}_{ij}$, S and n_{ij}/n, and also when the second-order models (9) and (10) are employed. As all the resulting expressions are rather complicated they are not given here; for full details the reader is referred to Vlachonikolis (1990).

2.4 Assessment and Comparison of the Rules

Various studies, both empirical and theoretical, have been conducted to establish the features of these three allocation rules and to compare their performances. Here we summarize the main findings.

Average optimal error rates incurred by the Bayes rule (7) (i.e. error rates assuming all population parameters to be known) have been tabulated for the cases $c = 1$ continuous variable and $q = 2,3,4$ binary variables over a range of parameter values in the relatively simple case of independent binaries by Krzanowski (1975) and Knoke (1982). More general situations (correlated binaries and $c > 1$) were considered by Krzanowski (1977). Asymptotic expansions of the parameter-replacement classification rule (using "naive" estimators of parameters) and corresponding expected actual error rates were obtained for the case of one binary variable by Tu and Han (1982) and for the general case of mixed binary and continuous variables by Vlachonikolis (1985), who also provided tabulations for various sample sizes and parameter combinations. These asymptotic expansions depend heavily on the normal-case expansions derived by Okamoto (1963).

For small-sample behavior, only Monte Carlo simulation results are so far available. Krzanowski (1975) conducted a very small and limited study to check on the performance of the location model. Much more extensive investigations were conducted by Vlachonikolis (1986), who obtained estimates of the expected actual error rates for which he had previously derived asymptotic expansions, and by Vlachonikolis (1990) to investigate performance of the Bayesian predictive rule. The parameter ranges and combinations in these two studies were the same as in Vlachonikolis (1985) but this time both "naive" and "smoothed" estimators of parameters were investigated. Finally, empirical assessment of performance of the various allocation rules (by either leave-one-out, resubstitution or test-set estimation of error rates on various real data sets) can be found in Chang and Afifi (1974), Krzanowski (1975, 1980, 1982), Knoke (1982), Tu and Han (1982), Vlachonikolis and Marriott (1982) and Leung (1989). It should be noted that the majority of tabulations, such as those cited above, have various practical drawbacks. They only cater for known population parameters, so can only be used as a general guide on the performance of an allocation rule or to set baselines for the expected level of error rates, and they are very dependent on the situations considered. Later authors very often follow the precedent set by previous ones in terms of situation, parameter settings and combinations, etc., and important cases can be easily missed.

Nonetheless, such tabulations do provide useful information, and the above studies seem to point up the following conclusions. Average expected error rates with the parameter-replacement Bayes allocation rule: (a) increase

as the number of continuous variables increases; (b) decrease in large samples as the number of binary (categorical) variables increases; (c) decrease as the within-location Mahalanobis distances D_m^2 between π_1 and π_2 increase; (d) decrease as the difference in binary incidence probabilities between π_1 and π_2 increases; and (e) increase as the correlation between binary variables increases.

Generally, expected actual error rates are slightly higher than the corresponding optimal error rates (approximately 5% - 30% in magnitude), but the estimated actual error rates in the Monte Carlo studies were nearly always smaller than their asymptotic expansion counterparts. However, the difference between the two was rarely significant and the asymptotic expansion seems to be a good approximation even for sample sizes as small as 50 per group. Virtually no difference was detected between the parameter-replacement Bayes procedure and each of the hypothesis-testing and Bayesian predictive rules respectively.

3. Useful Practical Extensions

In addition to the basic allocation rules and their error rates, described in the previous section, various extra features of the location model have been developed and are now available for use by the practitioner.

Krzanowski (1976) proposed a simple graphical procedure for investigating the worth of the location model discrimination procedure over and above the use of a simple linear discriminant function between two populations. The parameter-replacement version of Bayes rule (7) requires estimates of the continuous variable means μ_{ij} and dispersion matrix Σ. If $\hat{\mu}_{ij}$ and $\hat{\Sigma}$ are the estimates obtained in a particular application (whether by using the naive estimators $\bar{\mathbf{y}}_{ij}$,$\mathbf{S}$ or the smoothed second-order ones), then it is a simple matter to obtain the matrix of Mahalanobis D^2 values between every pair of states in the two populations. This $(2s \times 2s$ symmetric) matrix has entries $(\hat{\mu}_{ij} - \hat{\mu}_{kl})^T \hat{\Sigma}^{-1} (\hat{\mu}_{ij} - \hat{\mu}_{kl})$ where j,l take all values from 1 to s and i,k take values 1 or 2. Use of (metric) scaling on this matrix thus produces a low-dimensional representation of the $2s$ states which (through the ordering of the principal axes) gives an impression of the relative importance of differences between states and between populations. The more compactly clustered are the states within populations, the less difference is there between them in respect of the continuous variable parameters and hence the less benefit will be derived from use of the location model in preference to a simple linear discriminant function. A detailed illustrative example in this paper showed that the major axis of the two-dimensional metric scaling configuration split off all the even-numbered states from the odd-numbered ones, while the minor axis split the populations. Since the even-numbered

states were those for which the first binary variable X_1 took the value zero while the odd-numbered ones were those for which it took the value one, this demonstration showed that the main effect of X_1 was the biggest source of differences in the data. Use of the location model will allow different linear discriminant functions for the two values of X_1, but a simple linear discriminant function will involve averaging over this difference and so will not give as good a final result in this particular example.

This graphical idea was taken one step further and formalized into a hypothesis-testing procedure by Krzanowski (1979). Since the location model methodology will show greatest improvement over a simple linear discriminant function when there is large variability among the cells in respect of the continuous variable means μ_{ij}, a first stage is to look for linear transformations of the continuous variables such that there is as little variation as possible among the cell means in each population for the transformed data. Krzanowski (1979) gave several alternative ways of deriving such linear transformations, and then went on to derive a likelihood-ratio test for equality of the (true) cell means in each population. This procedure is thus a likelihood-ratio test for the adequacy of a simple linear discriminant function in place of the Bayes rule (7) based on the location model (but note that conditional normality of the continuous variables is now a critical assumption). There is also the possibility of using fewer transformed variables than there are original variables in future applications, and this aspect was investigated further by Krusinska (1988b).

One annoying feature of many practical applications of discriminant analysis is the presence of missing values in the data. In a comprehensive and important contribution, Little and Schluchter (1985) provided maximum likelihood estimation schemes for parameters of the location model when some data are missing. Their procedure uses the EM algorithm, embraces both the "naive" and "smoothed" approaches to parameter estimation, and allows constraints to be imposed on some of the parameters if so desired. The authors also discussed general aspects of imputation and discrimination as applications of the technique.

Up to this point all developments had been in terms of two-group discriminant analysis but Krzanowski (1986) extended the location model to multiple-group discrimination. The connection here was made by noting that, in general, the Bayes classification rule with equal costs and equal prior probabilities is identical to the maximum likelihood classification rule while for the continuous-variable-only case where $\mathbf{z} \sim N(\mu_i, \Sigma)$ in π_1, the maximum likelihood rule is identical to the minimum distance rule (i.e., allocating $\mathbf{z}$ to that population π_i for which $(\mathbf{z} - \mu_i)^T \Sigma^{-1} (\mathbf{z} - \mu_i)$ is smallest). For the special case of homogeneous CGD's, and treating $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ as a degenerate "population" in which unit probability is ascribed to the categorical state

defined by $\mathbf{x}$ and zero probability to all other states, and whose continuous component has probability mass unity at the observed value $\mathbf{y}$ and zero elsewhere, Krzanowski (1986) showed that the affinity (5) between $\mathbf{z}$ and π_1 reduced to

$$\rho_i = \{(2\pi)^c \mid \Sigma \mid\}^{-1/4} p_{im}^{1/2} \exp\{-\frac{1}{4}(\mathbf{y}-\mu_{im})^T\Sigma^{-1}(\mathbf{y}-\mu_{im})\} \text{ if } \mathbf{x} = m \quad (12)$$

Since affinity is the converse of distance, a "minimum distance" rule is the same as a "maximum affinity" rule. The multiple-group allocation rule is thus to allocate $\mathbf{z}$ to the population π_j for which ρ_j is greatest; with two groups some simple algebraic manipulation shows that this rule reduces to (7). All the usual features (smoothed parameter estimates, leave-one-out error rates, etc.) are easily implemented. For details, see Krzanowski (1986).

One aspect of the location model that has been tacitly accepted without question in all the developments is the conditional normality of the continuous variables, but what can we do if this assumption is not warranted? A start on answering this question was made by Balakrishnan and Tiku (1988), who developed robust classification procedures for the special cases of one binary and either one or two continuous variables. They used Tiku's modified maximum likelihood estimators (Tiku and Balakrishnan 1984) in which the r smallest and r largest observations are censored and the resulting (normal) likelihood is approximated in a simple fashion, obtained asymptotic error rates for various symmetric non-normal populations and conducted Monte Carlo studies for small n_{ij}. In general, the error rates were shown to be equivalent to the usual ones if normality is appropriate, but they are better and more stable under non-normality of $\mathbf{y}$.

Finally, Leung (1989) has provided an asymptotic expansion of the *studentized* parameter-replacement Bayes allocation rule (7). The asymptotic expansions provided by Vlachonikolis (1985) require knowledge of the true values of p_{ij} and $D_m^2 = (\mu_{1m}-\mu_{2m})^T\Sigma^{-1}(\mu_{1m}-\mu_{2m})$, so that the only use that could be made of them was in the tabulations already described in Section 2. Leung, however, used Anderson's (1973) approach to generalize these expansions by accommodating *estimates* of p_{ij} and D_m^2. Thus, it is now possible to calculate an asymptotic expected actual error rate in any practical application. Leung illustrated the calculation by obtaining this expected actual error rate for Chang and Afifi's (1974) example, and comparing the result with their empirical estimate. Note, however, that this expansion assumes large samples and normality of $\mathbf{y}$.

4. Feature Selection

One major shortcoming of the location model methodology is that the training data becomes very sparsely distributed among the categorical states when the total number of such states s becomes large (either because q is large or because each s_i is large). With sparse data, low-order models have to be fitted in order to obtain smoothed parameter estimates and this might not be satisfactory. It seems better to restrict the number of categorical variables and to fit higher-order models. This point was first made by Krzanowski (1983b), who provided a mechanism for selecting the "most effective" subset of categorical variables for the model. For a *given number* of categorical variables, he argued that the "most effective" choice comprises those categorical variables that yield the largest estimated distance $\hat{\Delta}_{12}$ between π_1 and π_2 (according to the special case (iii) of Equation (5)). Ideally one would conduct an "all subsets" search with $\hat{\Delta}_{12}$ as the objective function but this might not be computationally feasible, so a backward elimination procedure was described instead. Also, since selection is based just on the training data, "naive" estimators can be used instead of "smoothed" ones. Overall the procedure is very fast and easily implemented.

The idea was taken up and extended to more general situations, involving selection of models as well as of features, by a number of authors. The first was Daudin (1986), who extended the conditional distribution of $\mathbf{Y}$ to include "populations" as an extra categorical variable, Z say. Thus if we treat π_1 as the "base-line" population, then all individuals in π_1 are assigned the value $z = 0$ while all individuals in π_2 are assigned the value $z = 1$. The linear and log-linear models (9) and (10) are then extended by including terms such as αz, $\beta x_i z$, $\gamma x_i x_j z$, and so on. Daudin kept to the second-order restriction previously suggested for these models, and hence included only main effects (terms $x_1, x_2, \ldots, x_q, z$) and first-order interactions (terms $x_1 z, x_2 z, \ldots, x_q z, x_1 x_2, x_1 x_3, \ldots, x_{q-1} x_q$). He distinguished two types of model parameters in the linear model (9) for the μ_{ij}, namely those terms that involved the variable Z (α_1) and those that involved only the x_i (α_2), and his aim was to discard in turn the discrete variables, the continuous variables, and the model parameters, that contribute least to discrimination. Selection was to be made on the basis of the Akaike information criterion (AIC: log-likelihood minus the number of independent parameters, Akaike (1973)) thereby making the assumption of normality of $\mathbf{Y}$ an important requirement. He proposed a three-step selection procedure: (i) selection among the continuous variables and α_1 terms, (ii) selection among the α_2 terms, (iii) selection among log-linear terms.

Maximization of AIC was the objective, but there is a problem in step (i) because deleting continuous variables implies non-compatibility of $|\Sigma|$'s

and hence of corresponding likelihoods. For this step, therefore, Daudin proposed the maximization of a modified AIC which, in effect, is the increase in AIC for a given number of continuous variables due to the presence of the population factor Z. Backward elimination or forward selection was advocated in place of a global search, and an illustrative example was considered in some detail.

Further selection strategies were advocated in a series of papers by Krusinska (1988a, 1989a, 1989b). The first of these papers focussed on the two-group case and discussed the selection of those features (i.e. those variables from the complete set of categorical and continuous) that minimize an estimate of $p(\pi_1 \mid \pi_2) + p(\pi_2 \mid \pi_1)$. Various different estimates of this quantity were considered: replacement of parameters in expressions (8) using either naive estimates, smoothed estimates or the U-method (Lachenbruch and Mickey 1968); or empirical estimates via either resubstitution or leave-one-out (again encompassing either naive or smoothed estimation). The second paper allowed multiple-group situations and considered selection of the (minimum number) of features that give significant discriminatory measure $T^2 = \text{Trace}\,(\mathbf{HG}^{-1})$ where $\mathbf{H}$ is the between-states-and-populations sum of squares and products (SSP) matrix while $\mathbf{G}$ is the within-states-and-populations SSP matrix. Some distributional results were provided to check on significance of T^2 and thereby to provide a stopping rule. In both papers, backward elimination was advocated, and both approaches require at least one continuous variable to be present at each stage of the process. Note that both of these approaches involve strictly "discriminatory" criteria, by contrast with Daudin's "adequacy of model" criterion, but normality still plays an important role in definitions (8) and in the T^2 distribution results. (However, the latter may be slightly questionable as the appropriate distribution should be that of the *maximum* T^2 among $g > 1$ values at each step.) The third paper of the set provided a two-step (sub-optimal) branch-and-bound algorithm in place of the backward-elimination process using T^2.

Although each of the papers cited above provided at least one illustrative practical example of the relevant technique, no comparisons have yet been made among the competing proposals, so it is not possible to make recommendations. This is clearly an area that needs further research, but see also the remarks in section 5.3 below.

5. Other Possible Approaches with Mixed Variables

5.1 Linear Discriminant Analysis

The simplest possible practical approach is to ignore the categorical nature of some of the variables by replacing all $m(> 2)$-state categorical

variables by $(m-1)$ dummy binary variables, scoring all binary variables zero and one and using the ordinary linear discriminant function (LDF) as if all the variables were continuous. This procedure was investigated by Krzanowski (1977), who showed that often it will give satisfactory results but clearly will become poorer the more diversity there is among the separate location LDF's (7). Worst results will occur when individual LDF's become 'reversed' between locations. In addition to the techniques already mentioned in section 3 above, changes in binary/continuous correlations between populations provides a useful diagnostic of potentially poor performance with a simple LDF. Knoke (1982) and Vlachonikolis and Marriott (1982) independently showed that considerable improvement could be achieved by including squares of variables and cross-products between them (particularly those involving mixtures $x_i y_j$) in the LDF. With the widespread availability of the LDF and associated variable selection procedures in standard statistical software, these "modified linear discriminant functions" obviously carry considerable practical appeal.

5.2 Distance-based Discrimination

For this section it will be convenient to change notation from that used hitherto. Let us suppose that $\mathbf{v}$ is the individual to be allocated, and that in the two-group case the training sets consist of a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from π_1 and a sample $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m$ from π_2. Write $D(\pi_1,\pi_2)$, $D(\mathbf{v},\pi_i)$, $D(\mathbf{x}_i,\mathbf{x}_j)$ for the distances (however defined) between groups, between an element and a group, and between elements respectively.

One of the oldest distance-based allocation rules can be formally attributed to Matusita (1956), but has been used both formally and informally by many others. This is the intuitively reasonable rule that allocates $\mathbf{v}$ to the "nearer" of the two populations:

$$\text{allocate } \mathbf{v} \text{ to } \pi_j \text{ if } D(\mathbf{v},\pi_j) = \min\left[D(\mathbf{v},\pi_1), D(\mathbf{v},\pi_2)\right]. \tag{13}$$

If populations are multivariate normal with common dispersion matrices and Mahalanobis distances are used, then (13) reduces to the usual simple linear discriminant function. In the mixed-variable case using the location model, Krzanowski (1986) showed that $D(\mathbf{v},\pi_i) = \{2(1-\rho_i)\}^{1/2}$ with ρ_i given by (12) and that (13) then reduced to (7) with this distance function. (Note that in this case, $D(\pi_1,\pi_2)$ is given by Δ_{12} from case (iii) after Equation (5), with all population parameters replaced by their estimates from the training sets.)

A problem arises with use of (13) on multinomial data, since in this case the maximum likelihood rule is again recovered but this rule does not work well with sparse data. In an attempt to overcome the problem, Dillon

and Goldstein (1978) introduced a new distance-based discrimination procedure. If we let $D_{(i)}(\pi_1,\pi_2)$ denote the distance between π_1 and π_2 when $\mathbf{v}$ is included with the sample from π_i, then this new procedure is to allocate $\mathbf{v}$ to that group which yields the greatest separation between π_1 and π_2:

$$\text{allocate } \mathbf{v} \text{ to } \pi_j \text{ if } D_{(j)}(\pi_1,\pi_2) = \max\,[D_{(1)}(\pi_1,\pi_2),D_{(2)}(\pi_1,\pi_2)]\,. \quad (14)$$

Krzanowski (1987) studied this procedure theoretically with the help of influence functions, and showed that for mixed variables with the location model and distance Δ_{12} (14) produced an equivalent rule to (7). Thus neither of these two distance-based approaches seems to offer anything more than the Bayes classification rule for mixed variables with the location model, at least when (5) is used as the basis for distance calculation.

Takane, Bozdogan and Shibayama (1987) adopted a different approach, which they called "ideal point discriminant analysis." They allowed $g > 2$ groups, and assumed that the complete set of training data was contained in the $(n \times p)$ matrix $\mathbf{X}$ (where n is the total number of individuals and p is the total number of variables measured on each individual). The starting point is to suppose that the n individuals can be represented as n points in k-dimensional space, and that the $(n \times k)$ matrix $\mathbf{Y}$ of coordinates in this space is connected to $\mathbf{X}$ by the linear relationship $\mathbf{Y} = \mathbf{X}\mathbf{B}$ for parameters $\mathbf{B}$. Let $\mathbf{M}$ be a $(g \times k)$ matrix of "group ideal points" (typically the group centroids obtained from $\mathbf{Y}$). Then Takane et al. defined the distance from subject s to group t by $d_{st} = \{\sum_{j=1}^{k} (y_{sj} - m_{tj})^2\}^{1/2}$ and postulated the model

$$pr\{s \in \pi_t \,|\, \mathbf{x}\} = \frac{w_t \exp(-d_{st}^2)}{\sum_{h=1}^{g} w_h \exp(-d_{sh}^2)} \quad (15)$$

where $w_1, w_2, \ldots, w_g$ are weights satisfying $\sum_{i=1}^{g} w_i = 1$.

Given the known group membership of individuals in the training data, the (conditional) likelihood of the training data is multinomial with probabilities (15) and observed group frequencies, so iterative approximation methods (e.g. Fisher's scoring) can be used to provide maximum likelihood estimates of all the unknown parameters ($\mathbf{B}$ and the w_i) and hence an individual can be classified to the group for which it has highest estimated probability. Takane et al. advocated model evaluation via AIC, and showed how such additional features as subset selection could be incorporated easily. Note the resemblance of the methodology to logistic discrimination (Anderson 1982), and indeed many of the computational and sampling concerns are the same with

both approaches. However, the authors highlighted what they considered to be the main distinguishing features of ideal point discrimination, namely the multidimensional scaling connection, the more natural parameterization, and the possibility of dimension reduction.

The most recent distance-based discrimination approach is that due to Cuadras (1989, 1991), who builds on Rao's (1982) diversity indices. Cuadras defines (for the two-group case) the two discriminant functions

$$F_1 = \frac{1}{n}\sum_i D^2(\mathbf{v},\mathbf{x}_i) - \frac{1}{2n^2}\sum_i\sum_j D^2(\mathbf{x}_i,\mathbf{x}_j)\,;$$

$$F_2 = \frac{1}{m}\sum_i D^2(\mathbf{v},\mathbf{y}_i) - \frac{1}{2m^2}\sum_i\sum_j D^2(\mathbf{y}_i,\mathbf{y}_j)$$

and allocates $\mathbf{v}$ to π_j if $R_j = \min(F_1,F_2)$.

The benefit of this approach is that it operates exclusively with distances between *elements* rather than groups, so in the mixed-variable case we can use any of the standard distance measures from cluster analysis that will cope not only with mixtures of variables but also with obstacles such as missing values. A good choice of distance would be the one derived from Gower's (1971) general coefficient of similarity (see also Lerman 1987).

5.3 Empirical Comparison of Results

A limited number of empirical comparisons of different approaches to mixed-variable discrimination has been reported in the literature, and these are first briefly summarized before conclusions are drawn.

Chang and Afifi (1974) reported a study of 43 suicide attempts with $q = 1$ and $c = 2$. They quoted parameter-replacement error rates from (8) for the location model (using naive estimators with both separate and pooled covariance matrices in cells) and corresponding parameter-replacement error rates for the simple LDF. Leung (1989) re-estimated the location model error rates for this data set by means of the asymptotic studentized expansion.

Krzanowski (1975) gave five data sets, all with a medical background, ranging over various values of c and q. He quoted leave-one-out error rates for the location model Bayes rule (7) (with smoothed parameter estimates), the simple LDF, logistic discrimination, and a classification rule based on dichotomized variables.

Knoke (1982) reported a data set comprising 137 patients who had previously recovered from myocardial infarction, with $c = 2$ and $q = 3$. He gave resubstitution, leave-one-out and test set (105 extra patients) error rates for the usual location model rule, the simple LDF, the augmented LDF, and the quadratic discriminant function. Vlachonikolis and Marriott (1982) re-

analyzed Krzanowski's Data Set 4 and also considered a data set comprising 386 medical consultations with $c = 9$ and $q = 5$. Those authors first selected a subset of variables using standard stepwise selection on the simple LDF, the augmented LDF, and the logistic discriminant function, and then they obtained leave-one-out error rates for the chosen subsets.

All the above were two-group problems. Daudin (1986) provided a three-group problem in discriminating between the categories "bad", "acceptable" and "good" for 632 melons with $c = 6$ and $q = 5$. He quoted both resubstitution and leave-one-out error rates for the location model, the simple LDF and the augmented LDF, both with and without prior selection of variables.

Krusinska (1988a, 1989a) used a data set consisting of 164 bronchial asthma sufferers with $c = 6$, $q = 8$ and she reported leave-one-out error rates and T^2 values for various selected subsets and selection strategies based on the location model. Finally, Takane et al. (1987) re-analyzed Krzanowski's Data Set 4 by ideal point discriminant analysis, with and without prior selection of variables.

Nearly all the above comparisons were ones contrasting the location model with some variant of the LDF. In the majority of cases, the location model (without prior variable selection) did as well as or better than the simple LDF (also without prior variable selection). Where the simple LDF did badly compared to the location model, the augmented LDF (including squares and cross-products of variables) had a performance much closer to that of the location model. Prior selection of variables generally improved performances. Daudin's results are the only ones where *all* methods underwent prior selection of variables, and here the location model still performed much better than the other methods (41.9% misclassification as against 44.6% with the augmented LDF and 49.9% with the simple LDF; 7.1% of the "bad" melons allocated to the "good" group as against 8.5% with the augmented LDF and 10.5% with the simple LDF). Thus the above results suggest that the location model is, in general, the best method for mixed variables followed by the augmented LDF and then the simple LDF. Where such comparisons were made, logistic discrimination seemed to be comparable to the simple LDF on mixed data and no particular benefit was derived from a quadratic discriminant function as against the simple LDF.

However, some contradictory results were obtained in those comparisons where some methods had prior selection of variables while other methods did not. For example, the performance of ideal point discriminant analysis was no better than that of the location model in the data set on which they were compared if the full set of variables was used in both methods, but it *did* do better if prior selection was made before ideal point analysis and not before location model analysis. It is the present author's view that all results

involving prior variable selection must be treated with caution for several reasons. Although the leave-one-out procedure guards against bias in the error-rate estimation process, additional bias is being introduced by the variable selection since by definition it is those variables that are "best" for the training data which are being selected. Thus comparison of a method that has not had selection with one that has had prior selection is unfair. Also different methods may react differently to the selection process (for example in Daudin's data, prior selection reduced the error rate for the simple LDF only from 50.3% to 49.9% but for the augmented LDF from 50.0% to 44.6%). Thus unfair comparisons may result even when all methods have undergone prior selection. The whole area of assessing performances of allocation rules with and without variable selection has received very little attention to date, and considerably more needs to be done. A start has been made in the simple LDF context (see Ganeshanandam and Krzanowski 1989), and work on the mixed-variable case is currently in progress.

6. Future Prospects

In addition to the variable selection problem outlined above, where else should effort be concentrated in the mixed-variable discrimination area? It is evident that there is considerable scope for investigating the effect of relaxing assumptions inherent in the location model, and developing suitable modifications of the model in such circumstances. For instance, much remains to be done on the robustness of allocation rule (7) to departures from normality and constant within-cell dispersions. If departures do cause poor performance, then development of robust discriminant functions for the mixed-variable case would be essential. Similarly, is there a call for location quadratic discriminant functions to cater for various types of dispersion inhomogeneity?

A second possible direction of progress brings us back to our starting point, the use of graphical modeling. The whole development of location model methodology to date has assumed a fairly rigid second-order structure for obtaining smoothed parameter estimates via (9) and (10), but now a much wider horizon has opened up with the advent of graphical modeling techniques. The possibility of tailoring best models to each data set is clearly the next step, with development of appropriate software also a top priority. Krusinska (1990) seems to be pointing the way in this direction, but there is clearly still much to be achieved.

Finally we consider the question of availability of software for carrying out the techniques discussed in this paper. Unfortunately, despite the time that has now elapsed since the methods were first proposed, none of the procedures based on the location model has yet found its way into any of the

widely available general statistical software packages. Attempts are being made to interest at least one of the producers in adding suitable routines to a future release, but until such efforts bear fruit potential users will have to be content with acquiring private software. The author has a number of Fortran routines for carrying out many of the location-model-based techniques described above. Although these are not in the most tidy or efficient form (and some are still in a developmental state), he will be happy to send them by e-mail to anyone on request.

References

AFIFI, A. A., and ELASHOFF, R. M. (1969), "Multivariate Two-sample Tests with Dichotomous and Continuous Variables. 1. The Location Model," *Annals of Mathematical Statistics, 40*, 290-298.

AKAIKE, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, Eds., B.N. Petrov and F. Csaki, Budapest: Akademia Kiado, 267-281.

ANDERSON, J. A. (1982), "Logistic Discrimination," In *Handbook of Statistics 2, Classification, Pattern Recognition and Reduction of Dimensionality*, Eds., P.R. Krishnaiah and L.N. Kanal, Amsterdam: North Holland, 169-191.

ANDERSON, T. W. (1973), "An Asymptotic Expansion of the Distribution of the Studentized Classification Statistic W," *Annals of Statistics, 1*, 964-972.

BALAKRISHNAN, N., and TIKU, M. L. (1988), "Robust Classification Procedures Based on Dichotomous and Continuous Variables," *Journal of Classification, 5*, 53-80.

CHANG, P. C., and AFIFI, A. A. (1974), "Classification Based on Dichotomous and Continuous Variables," *Journal of the American Statistical Association, 69*, 336-339.

COX, D. R. (1972), "The Analysis of Multivariate Binary Data," *Applied Statistics, 21*, 113-120.

CUADRAS, C. M. (1989), "Distance Analysis in Discrimination and Classification Using Both Continuous and Categorical Variables," in *Statistical Data Analysis and Inference*, Ed., Y. Dodge, Amsterdam: North Holland, 459-473.

CUADRAS, C. M. (1991), "A Distance-based Approach to Discriminant Analysis and Its Properties," Mathematics preprint series no. 90, Barcelona University.

DAUDIN, J. J. (1986), "Selection of Variables in Mixed-variable Discriminant Analysis," *Biometrics, 42*, 473-481.

DILLON, W. R., and GOLDSTEIN, M. (1978), "On the Performance of Some Multinomial Classification Rules," *Journal of the American Statistical Association, 73*, 305-313.

EDWARDS, D. (1990), "Hierarchical Interaction Models," *Journal of the Royal Statistical Society, Series B, 52*, 3-20.

GANESHANANDAM, S., and KRZANOWSKI, W. J. (1989), "On Selecting Variables and Assessing Their Performance in Linear Discriminant Analysis," *Australian Journal of Statistics, 31*, 433-447.

GOWER, J. C. (1971), "A General Coefficient of Similarity and Some of Its Properties," *Biometrics, 27*, 857-871.

HAN, C.-P. (1979), "Alternative Methods of Estimating the Likelihood Ratio in Classification of Multivariate Normal Observations," *American Statistician, 33*, 204-206.

KNOKE, J. D. (1982), "Discriminant Analysis with Discrete and Continuous Variables," *Biometrics, 38*, 191-200.

KRUSINSKA, E. (1988a), "Variable Selection in Location Model for Mixed Variable Discrimination: A Procedure Based on Total Probability of Misclassification," *EDV in Medizin und Biologie, 19*, 14-18.

KRUSINSKA, E. (1988b), "Linear Transformations in Location Model and Their Influence on Classification Results in Mixed Variable Discrimination," *EDV in Medizin und Biologie, 19*, 110-114.

KRUSINSKA, E. (1989a), "New Procedure for Selection of Variables in Location Model for Mixed Variable Discrimination," *Biometrical Journal, 31*, 511-523.

KRUSINSKA, E. (1989b), "Two Step Semi-optimal Branch and Bound Algorithm for Feature Selection in Mixed Variable Discrimination," *Pattern Recognition, 22*, 455-459.

KRUSINSKA, E. (1990), "Suitable Location Model Selection in the Terminology of Graphical Models," *Biometrical Journal, 32*, 817-826.

KRZANOWSKI, W. J. (1975), "Discrimination and Classification Using Both Binary and Continuous Variables," *Journal of the American Statistical Association, 70*, 782-790.

KRZANOWSKI, W. J. (1976), "Canonical Representation of the Location Model for Discrimination or Classification," *Journal of the American Statistical Association, 71*, 845-848.

KRZANOWSKI, W. J. (1977), "The Performance of Fisher's Linear Discriminant Function Under Non-optimal Conditions," *Technometrics, 19*, 191-200.

KRZANOWSKI, W. J. (1979), "Some Linear Transformations for Mixtures of Binary and Continuous Variables, With Particular Reference to Linear Discriminant Analysis," *Biometrika, 66*, 33-39.

KRZANOWSKI, W. J. (1980), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis," *Biometrics, 36*, 493-499.

KRZANOWSKI, W. J. (1982), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis: A Hypothesis-testing Approach," *Biometrics, 38*, 991-1002.

KRZANOWSKI, W. J. (1983a), "Distance Between Populations Using Mixed Continuous and Categorical Variables," *Biometrika, 70*, 235-243.

KRZANOWSKI, W. J. (1983b), "Stepwise Location Model Choice in Mixed-variable Discrimination," *Applied Statistics, 32*, 260-266.

KRZANOWSKI, W. J. (1984), "On the Null Distribution of Distance Between Two Groups, Using Mixed Continuous and Categorical Variables," *Journal of Classification, 1*, 243-253.

KRZANOWSKI, W. J. (1986), "Multiple Discriminant Analysis in the Presence of Mixed Continuous and Categorical Data," *Computers and Mathematics with Applications, 12A(2)*, 179-185.

KRZANOWSKI, W. J. (1987), "A Comparison Between Two Distance-based Discriminant Principles," *Journal of Classification, 4*, 73-84.

LACHENBRUCH, P. A., and MICKEY, M. R. (1968), "Estimation of Error Rates in Discriminant Analysis," *Technometrics, 10*, 1-11.

LAURITZEN, S. L., and WERMUTH, N. (1989), "Graphical Models for Association Between Variables, Some of Which Are Qualitative and Some Quantitative," *Annals of Statistics, 17*, 31-54.

LERMAN, I. C. (1987), "Construction d'un indice de Similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification (1)," *Revue de Statistique Appliquée, 35*, 39-60.

LEUNG, C. Y. (1989), "The Studentized Location Linear Discriminant Function," *Communications in Statistics, Theory and Methods, 18*, 3977-3990.

LITTLE, R. J. A., and SCHLUCHTER, M. D. (1985), "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values," *Biometrika, 72*, 497-512.

MATUSITA, K. (1956), "Decision Rule, Based on the Distance, for the Classification Problem," *Annals of Mathematical Statistics, 8*, 67-77.

OKAMOTO, M. (1963), "An Asymptotic Expansion for the Distribution of the Linear Discriminant Function," *Annals of Mathematical Statistics, 34*, 1286-1301 (with correction in *39*, 1358-1359).

OLKIN, I., and TATE, R. F. (1961), "Multivariate Correlation Models with Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics, 32*, 448-465 (with correction in *36*, 343-344).

RAO, C. R. (1982), "Diversity and Dissimilarity Coefficients: A Unified Approach," *Theoretical Population Biology, 21*, 24-43.

TAKANE, Y., BOZDOGAN, H., and SHIBAYAMA, T. (1987), "Ideal Point Discriminant Analysis," *Psychometrika, 52*, 371-392.

TIKU, M. L., and BALAKRISHNAN, N. (1984), "Robust Multivariate Classification Procedures Based on the MML Estimators," *Communications in Statistics - Theory and Methods, 13*, 967-986.

TU, C. T., and HAN, C. P. (1982), "Discriminant Analysis Based on Binary and Continuous Variables," *Journal of the American Statistical Association, 77*, 447-454.

VLACHONIKOLIS, I. G. (1985), "On the Asymptotic Distribution of the Location Linear Discriminant Function," *Journal of the Royal Statistical Society, Series B, 47*, 498-509.

VLACHONIKOLIS, I. G. (1986), "On the Estimation of the Expected Probability of Misclassification in Discriminant Analysis with Mixed Binary and Continuous Variables," *Computers and Mathematics with Applications, 12A(2)*, 187-195.

VLACHONIKOLIS, I. G. (1990), "Predictive Discrimination and Classification with Mixed Binary and Continuous Variables," *Biometrika, 77*, 657-662.

VLACHONIKOLIS, I. G., and MARRIOTT, F. H. C. (1982), "Discrimination with Mixed Binary and Continuous Data." *Applied Statistics, 31*, 23-31.

WERMUTH, N., and LAURITZEN, S. L. (1990), "On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models," *Journal of the Royal Statistical Society, Series B, 52*, 21-50.

WHITTAKER, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Chichester: Wiley.