PSYCHOMETRIKA—VOL. 52, NO. 3, 371–392
SEPTEMBER 1987
SPECIAL SECTION

IDEAL POINT DISCRIMINANT ANALYSIS

YOSHIO TAKANE

MCGILL UNIVERSITY

HAMPARSUM BOZDOGAN

UNIVERSITY OF VIRGINIA

TADASHI SHIBAYAMA

UNIVERSITY OF TOKYO

A new method of multiple discriminant analysis was developed that allows a mixture of continuous and discrete predictors. The method can be justified under a wide class of distributional assumptions on the predictor variables. The method can also handle three different sampling situations, conditional, joint and separate. In this method both subjects (cases or any other sampling units) and criterion groups are represented as points in a multidimensional euclidean space. The probability of a particular subject belonging to a particular criterion group is stated as a decreasing function of the distance between the corresponding points. A maximum likelihood estimation procedure was developed and implemented in the form of a FORTRAN program. Detailed analyses of two real data sets were reported to demonstrate various advantages of the proposed method. These advantages mostly derive from model evaluation capabilities based on the Akaike Information Criterion (AIC).

1. Introduction

Discriminant analysis (DA) concerns the classification of objects according to some criterion (Hand, 1981; Lachenbruch, 1975). DA is of interest to psychometricians, not only for its practical use in medical diagnosis, psychiatric classification, aptitude diagnosis, and so forth, but also as a model (*albeit* often primitive) of psychological processes underlying such phenomena as categorization, pattern recognition and discrimination learning.

DA consists of two stages. In the first stage a specific formula is developed for classification according to some model. This usually involves estimation of parameters in the model using a so-called training (or learning) sample in which memberships of objects

The work reported in this paper has been supported by Grant A6394 from the Natural Sciences and Engineering Research Council of Canada and by a leave grant from the Social Sciences and Humanities Research Council of Canada to the first author. Portions of this study were conducted while the first author was at the Institute of Statistical Mathematics in Tokyo on leave from McGill University. He would like to express his gratitude to members of the Institute for their hospitality. Thanks are also due to T. Komazawa at the Institute for letting us use his data, to W. J. Krzanowski at the University of Reading for providing us with Armitage, McPherson, and Copas' data, and to Don Ramirez, Jim Ramsay and Stan Sclove for their helpful comments on an earlier draft of this paper.

Requests for reprints should be sent to Yoshio Takane, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, Quebec H3A 1B1, CANADA.

0033-3123/87/0900-SS04$00.75/0

are known a priori. The formula is then applied to initially unclassified objects in a test sample to predict their group membership.

A variety of methods have been developed for DA, ranging from highly parametric methods at one extreme (e.g., Fisher's, 1936, linear discriminant function) to nonparametric methods at the other (e.g., kernel discriminant analysis). These extreme methods have their own advantages and disadvantages (Lachenbruch, Sneeringer, & Revo, 1973; Krzanowski, 1977). While the parametric methods are often the most efficient methods when the parametric assumptions are reasonably accurate, their performance tends to deteriorate as the assumptions become less realistic. The nonparametric methods, on the other hand, are more flexible, but are much less efficient than the parametric methods when the latter are indeed appropriate. In this paper we present a method of DA which lies between these two extremes, and captures the best of both approaches. The proposed method is not as rigid in its assumption as the most parametric methods, and yet not as data-dependent as the most nonparametric methods.

In the proposed method, which we call "ideal point" DA, subjects (cases or any other sampling units) are mapped into a multidimensional euclidean space as a linear function of predictor variables. Criterion groups are assumed to have ideal points in the space that represent the most typical (prototypes) of the groups. It is assumed that the more similar a particular subject's profile is to the prototype of a criterion group, the higher is the probability that the subject belongs to the group. In the space in which both subjects and criterion groups are represented, their similarity is represented by the distance between the corresponding points. The probability is thus stated as a decreasing function of the distance between them.

Ideal point DA is widely applicable. It allows a mixture of continuous and discrete predictors. It can handle three different sampling designs, conditional, joint and separate, with minor modifications from one sampling design to another. In those designs in which the predictor variables are considered as random variables, the method can be justified under the general exponential family of distributions.

In spite of its wide applicability ideal point DA has nearly all the important model evaluation features that only the highly parametric procedures traditionally enjoyed. For example, it allows choice of best dimensionality, optimal subset selection of predictor variables, multiple comparisons of criterion groups, and so forth, which enable us to search for the best specification of the model. The Akaike Information Criterion (AIC; Akaike, 1974; Sakamoto, Ishiguro & Kitagawa, 1986) plays a crucial role in this model identification process.

In the next section (section 2) we present a detailed account of ideal point DA, emphasizing various characteristics of its model, an estimation procedure, and a justification behind the model. We also discuss methodological aspects of model evaluations (section 2.3). In section 3 we demonstrate usefulness of ideal point DA through analyses of two actual data sets. These examples illustrate how the model evaluation procedures described in section 2.3 may be effectively used in practical DA situations. In the final section we place ideal point DA in a broader perspective as a data analysis tool and discuss possible extensions.

2. The Method

2.1 *The Basic Model*

Let us suppose that there are N subjects in a training sample classified into n_G criterion groups. Let X denote an N by n matrix of predictor variables, where n is the number of predictor variables. The predictor variables can be continuous, discrete, or mixed together. All continuous variables are standardized and all discrete variables are coded into dummy variables. Each discrete variable is thus counted as J_i variables, where J_i is the number of observation categories in variable i. Missing data may be coded into zero for continuous variables after the standardization, or a series of zeros in the dummy variables for discrete variables. In this way the missing data will have no effects in discrimination. When interactions among original predictor variables are suspected, appropriate interaction terms are defined, and included in X. (How this is done will be explained in section 2.3).

We assume that the N subjects in the training sample are represented as points in an A dimensional euclidean space. Let Y be an N by A matrix of coordinates of the subject points. We assume that these coordinates are simple linear functions of predictor variables; that is,

$$Y = XB, \tag{1}$$

where B is an n by A matrix of weights analogous to regression coefficients. The weights for discrete variables are sometimes called "quantifications" of observation categories. We impose different restrictions on elements of B, depending on the types of predictor variables. Specific forms of the restrictions will be discussed toward the end of this section. The dimensionality A of the representation space is between 1 and $n_G - 1$ inclusive. However, it cannot exceed the number of nonredundant predictor variables.

Let M denote an n_G by A matrix of coordinates of ideal points of criterion groups represented in the same A dimensional euclidean space. We may either take M as a set of free parameters, or assume that the ideal points of the criterion groups are given by centroids of the groups. Let Z denote an N by n_G matrix of dummy variables indicating group memberships of the subjects. Then in the latter case M is expressed as

$$M = (Z'Z)^{-1}Z'Y = (Z'Z)^{-1}Z'XB. \tag{2}$$

This reduces the number of parameters to be estimated, since M is now a function of B, and only B should be explicitly estimated. All the empirical results reported in this paper were obtained under (2), since this constraint seems to be rather innocuous in most cases. We may also test the empirical validity of the assumption using the model evaluation idea to be presented in section 2.3. For the sake of generality, however, we retain M as containing possible free parameters.

We now define the euclidean distance between the subject points and the ideal points of the criterion groups. For subject k and group g, this is

$$d_{kg} = \left\{ \sum_{a=1}^{A} (y_{ka} - m_{ga})^2 \right\}^{\frac{1}{2}}, \tag{3}$$

where y_{ka} and m_{ga} are elements of Y and M, respectively. The simple euclidean model (3) tacitly assumes that the criterion groups differ only in their most representative points (i.e., m_{ga}). No differential sensitivity of the distance from the ideal points is taken into account. A possible generalization of the distance function to incorporate the differential sensitivity across criterion groups will be suggested in the discussion section.

We are now in a position to state a model that relates the distance to the probability of group membership of a subject. Let $p_{g|k}$ denote the conditional probability that subject k belongs to criterion group g, given the set of observations on the predictor variables. We assume that this probability is given by

$$p_{g|k} = \frac{w_g \exp(-d_{kg}^2)}{\sum_{h=1}^{n_G} w_h \exp(-d_{kh}^2)}, \tag{4}$$

where w_g is a bias parameter for group g. The bias parameter is like the prior probability of a group, but it is used here to represent a broader concept. It represents whatever makes a certain group more or less likely. Of course, it is also possible to constrain explicitly w_g to be proportional to the group size, in which case w_g is simply written as p_g. We require $\sum_g w_g = 1$ in order to remove scale indeterminancy in w_g.

Model (4) implies $p_{g|k}$ is proportional to $w_g \exp(-d_{kg}^2)$, which in turn implies it is proportional to w_g for a fixed d_{kg} and to $\exp(-d_{kg}^2)$ for a fixed w_g. (The denominator of (4) is just a normalization factor to make $p_{g|k}$ add up to unity over g.) This is in line with the nature of the bias parameter defined above, and also with the basic postulate of the model that the probability of subject k belonging to group g is a decreasing function of the distance between them. A justification of the particular form of the model will be given in section 2.2. An important feature of model (4) is that the overall scale of the subject and ideal point configuration is uniquely determined. It indicates the degree of discriminability among the criterion groups.

Model (4) is a special form of Coombs' (1964) unfolding model combined with Luce's (1959) biased choice model. The speciality lies in that coordinates of subject points are "constrained" as linear functions of the predictor variables. The model can be also viewed as a generalization of Schönemann and Wang's (1972) individual differences preference choice model to multiple-choice situations.

The (conditional) likelihood of the model for the entire set of observations in the training sample is now stated as

$$L = \prod_{k}^{N} \prod_{g=1}^{n_G} (p_{g|k})^{f_{kg}}, \tag{5}$$

where $f_{kg} = 1$ if subject k belongs to group g, and $f_{kg} = 0$ otherwise.

The log of (5) is maximized with respect to model parameters (B, M and w_g) subject to relevant constraints (to be explained below) by an iterative approximation method. We use Fisher's scoring method, which has proven to be extremely efficient in the present context. The scoring method updates parameter estimates by

$$\mathbf{u}^{(q+1)} = \mathbf{u}^{(q)} + \alpha^{(q)} I(\mathbf{u}^{(q)})^{-1} \mathbf{g}(\mathbf{u}^{(q)}) \tag{6}$$

in each iteration, where $\mathbf{u}^{(q+1)}$ and $\mathbf{u}^{(q)}$ are, respectively, new and old parameter estimates (with q being the iteration number), $\mathbf{q}(\mathbf{u}^{(q)})$ and $I(\mathbf{u}^{(q)})$ are the gradient vector and the information matrix, respectively, both evaluated at $\mathbf{u}^{(q)}$, and $\alpha^{(q)}$ is the stepsize parameter. The information matrix is usually singular in the present case. We use the Moore-Penrose inverse of I in (6) (Ramsay, 1978). This ensures uniqueness of the parameter estimates. For initial estimates of B we simply apply the usual canonical DA method.

Once parameters are estimated, they may be used to evaluate $p_{g|k*}$ for a new sample $k*$, and it may be classified according to the $\max_g p_{g|k*}$ rule. (To be more exact, the evaluation of $p_{g|k*}$ requires specification of the sampling design employed, but this will be explained in the next section.)

The predictor variables can be discrete (nominal or ordinal scale level) or continuous (interval or higher). As noted earlier, the different types of predictor variables are distinguished by different constraints imposed on relevant portions of the weight matrix, B. The likelihood is maximized subject to these constraints. For an unordered categorical (nominal) variable i with J_i observation categories we require

$$\sum_{j=1}^{J_i} n_{i(j)} b_{i(j)a} = 0, \tag{7}$$

for $a = 1, \ldots, A$, where $n_{i(j)}$ is the marginal frequency of the j-th category of item i and $b_{i(j)a}$ is the a-th quantification of the category. The above restriction is necessary in order to remove linear dependencies among the categorical variables. The restriction can be incorporated into the above optimization scheme by expressing the quantifications of the last category in terms of the other quantifications. For an ordered categorical variable, we need something similar to (7). We also like category quantifications to satisfy a prescribed order. Consequently we find one set of quantifications, which are multidimensionally weighted to obtain multidimensional quantifications; that is,

$$b_{i(j)a} = r_{i(j)} s_{ia}, \tag{8}$$

where

$$r_{i(1)} \leq r_{i(2)} \leq \cdots \leq r_{i(J_i)}.$$

In order to remove scale indeterminacy between $r_{i(j)}$ and s_{ia}, we further require the quantified categories to have unit variance $(\sum_{s=1}^{J_i} n_{i(j)} r_{i(j)}^2/N = 1)$. When the variable is measured on an interval or higher scale we assume that category quantifications are already given, and obtain only dimensional weights.

2.2 *Further Aspects of the Model*

Model (4) states the conditional probability of a criterion group given a set of observations on the predictor variables. No distributional assumptions are made on the predictor variables. The conditional probability formulation is most natural in the conditional sampling situation (Kalbfleisch, 1984), where we observe frequencies of criterion groups for several fixed values of the predictor variables. Since the predictor variables are fixed, no distributional assumptions are necessary in this case. However, the conditional likelihood, (5), is still valid in other sampling situations, so far as the distribution of the predictor variables leads to the conditional probability stated in (4) (Anderson, 1972, 1982). In the joint sampling design only the total number of subjects is fixed, and the joint frequencies of criterion groups and patterns of observations on the predictor variables are observed. In the separate sampling situation marginal frequencies of criterion groups are fixed, and within each group frequencies of patterns of observations on the predictor variables are observed.

Model (4) can be justified in the joint or separate sampling situation, whenever the conditional distribution of the predictor variables given a criterion group is one of exponential family of distributions (Efron, 1975). This includes, as special cases, multivariate normal distributions with equal covariance matrices across criterion groups, independent binomial and multinomial distributions, the loglinear model with the second and higher order interactions assumed equal across criterion groups, etc. (Anderson, 1972). A mixture of the above distributions is also permissible. The requirement of independence and identical interactions can be easily eliminated by including appropriate interaction terms in the set of predictor variables.

The exponential family of distributions can be generally expressed as

$$f(\mathbf{x}_{k|g}) = c_g h(\mathbf{x}_k)\exp(\mathbf{b}_g' \mathbf{x}_k), \quad g = 1, \ldots, n_G \tag{9}$$

(e.g., Andersen, 1980), where $\mathbf{x}_k$ is the set of values on the predictor variables for subject k, and $\mathbf{b}_g$ and c_g are parameters of the distribution. This leads to the conditional probability of the form,

$$p_{g|k} = \frac{c_g \exp(\mathbf{b}_g' \mathbf{x}_k)}{\sum\limits_h c_h \exp(\mathbf{b}_h' \mathbf{x}_k)}. \tag{10}$$

To see model (4) is a special case of (10), we rewrite (4) into

$$p_{g|k} = \frac{w_g^* \exp(\mathbf{a}_g' \mathbf{x}_k)}{\sum_h w_h^* \exp(\mathbf{a}_h' \mathbf{x}_k)}, \tag{11}$$

where

$$\mathbf{a}_g = 2B\mathbf{m}_g, \tag{12}$$

and

$$w_g^* = w_g \exp(-\mathbf{m}_g' \mathbf{m}_g). \tag{13}$$

In (12) and (13), $\mathbf{m}_g$ is the vector of coordinates of ideal point of group g. Now it is obvious that (11) is a special case of (10). With centroid restriction (2) on $\mathbf{m}_g$ in ideal point DA, the above relationship is no longer strictly true, but an approximate relationship still holds to the extent that the restriction is plausible.

When both continuous and discrete variables that follow the exponential family of distributions are included in the set of predictor variables, (10) and consequently model (4) are still valid, assuming that the same exponential family of distributions hold for the continuous predictor variables at each subsample defined by the discrete predictor variables. That is, the continuous and the discrete variables should not interact. However, this requirement can be eliminated by including appropriate interaction terms between the continuous and the discrete variables. How the interactions can be defined will be discussed in the next section.

With minor modifications, (11) can be further rewritten into a form that directly suggests its relationship to logistic discrimination (Anderson, 1972; Cox, 1966; Day & Kerridge, 1967; Walker & Duncan, 1967); namely,

$$p_{n_G|k} = \left[1 + \sum_{h \neq n_G} \exp(\boldsymbol{\alpha}_h' \mathbf{x}_h + \beta_h)\right]^{-1}, \tag{14}$$

and

$$p_{g|k} = \exp(\boldsymbol{\alpha}_g' \mathbf{x}_k + \beta_g) \cdot p_{n_G|k}, \tag{15}$$

for $g = 1, \ldots, n_G - 1$, where $\boldsymbol{\alpha}_g = \mathbf{a}_g - \mathbf{a}_{n_G}$ and $\beta_g = \ln(w_g^*/w_{n_G}^*)$ for $g = 1, \ldots, n_G - 1$. Parametrization, (14) and (15), also makes it obvious that the model is a special case of a class of models called "generalized linear models" (McCullagh & Nelder, 1983).

One may well wonder why we use model (4) rather than its alternative parametrization, (14) and (15). There are three important reasons for this. First of all it makes obvious its relationship to multidimensional scaling (MDS). We obtain a spatial representation of individual subjects and criterion groups, thereby visually understand their mutual relationship. Secondly, the parametrization is more natural in ideal point DA. In (14) and (15) the last group is taken as the reference group; all other groups are characterized in reference to the last group. This makes the parametrization "asymmetric," which in turn makes it difficult to understand the relation between two groups, neither of which is the reference group. Thirdly, the form of decomposition of $\mathbf{a}_g$ in (12) tacitly implied by ideal point DA allows possible dimension reduction. Matrix B has A columns, where A is at most $n_G - 1$, but in general we may need much smaller dimensionality than $n_G - 1$. Estimating too many dimensions may even be harmful for obtaining reliable parameter estimates. However, in logistic discrimination no mechanisms are built in for possible dimension reduction. It always takes $A = n_G - 1$. Thus, although the difference looks only subtle, there is an important difference between ideal point DA and logistic discrimination.

Insofar as the distributional assumption (if required) is satisfied, the same estimation procedure can be used without regard to the sampling designs. This is because in model (4), w_g and $\exp(-d_{gk}^2)$ are completely separate (they do not affect each other), and the different sampling designs only affect the bias parameters. In some designs, however, the training sample may not reflect population group sizes. Then some adjustment is necessary on estimated w_g before it is used for prediction purposes. Let p_g be an estimate of the prior probability of group g in the training sample. This is typically taken as N_g/N, where N_g is the observed marginal frequency of group g. Let $\tilde{p}_g$ be some other, presumably more realistic, estimate of p_g. Then the proper adjustment on $\hat{w}_g$ would be $\hat{w}_g(\tilde{p}_g/\hat{p}_g)$. In the joint sampling situation no adjustment is usually required. When no realistic estimate of p_g is available, we may use the minimax type of classification rule. While this rule is hard to incorporate when the number of criterion groups is greater than 2, an approximate solution can be obtained by setting $p_g = 1/n_G$ for all g.

We may also draw boundary hyperplanes in the representation space. The boundary hyperplanes are the traces of points that satisfy $p_{g|x} = p_{h|x}$ $(g \neq h)$. On one side of a hyperplane observations are classified into one group, while on the other side they are classified into the other. The boundary hyperplanes are piecewise linear in the present case, and the hyperplane that devides two groups is perpendicular to the line segments connecting the ideal points of the two groups. (See Figure 1). These properties derive from the fact that we have squared euclidean distances in the exponent.

Finite maximum likelihood estimates can be obtained in most cases. However, there are situations in which no finite maximum likelihood estimates exist (Anderson, 1974). In such cases ideal point DA should not be applied. One obvious case is in which criterion groups are linearly separable. Although this case may sometimes be difficult to detect by merely inspecting the data, it can be easily identified by noting the log likelihood approaching zero, as the iteration proceeds. Similar degeneracy occurs when a category of a discrete variable has no responses (a zero frequency) in one or more groups. This latter case can be screened out prior to the analysis, and the category might be dropped from the analysis.

2.3 *Model Evaluation*

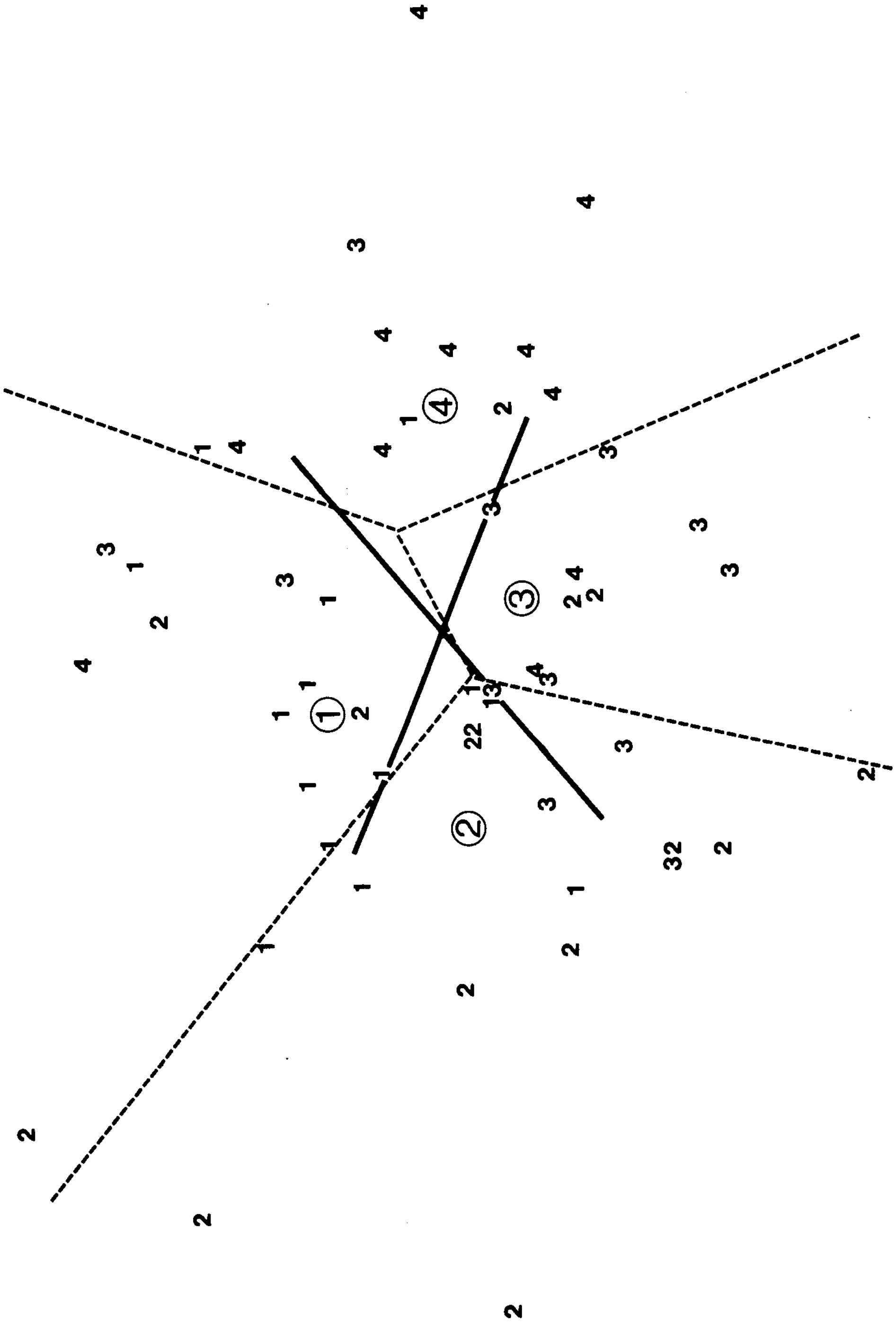
Ideal point DA allows a variety of model evaluations. A general strategy is to fit each candidate model by ideal point DA, calculate the value of the AIC statistic (Akaike, 1974), and choose the model associated with the minimum AIC value. The AIC is defined as

$$\mathrm{AIC} = -2 \ln L^* + 2n_p,$$

where $\ln L^*$ is the value of the log likelihood maximized over the parameter space (B, M and w_g) and n_p is the effective number of parameters in the fitted model. The effective number of parameters is sometimes rather complicated to figure out. When the centroid constraint, (2), is imposed on M, it is

$$n_p = n_B - \frac{A(A-1)}{2} + (n_G - 1), \tag{16}$$

where n_B is the sum of $(J_i - 1)A$ if variable i is nominal, $(J_i - 2) + A$ if i is ordinal, and A if i is interval or higher. The $A(A-1)/2$ is subtracted because of the rotational indeterminacy in the euclidean space. The last term, $n_G - 1$, is the effective number of bias parame-


①
②
③
④

ters. When (2) is not imposed, (16) becomes

$$n_p = n_B + n_G A - A(A + 1) + (n_G - 1), \tag{17}$$

where $n_G A$ is the number of parameters in M. This time $A(A + 1)$ is subtracted because the model allows an affine transformation of the A dimensional space. When there are no ordinal variables, n_B is n^* times A where n^* is the effective number of predictor variables (we count only $J_i - 1$ for a nominal variable). When the maximum possible dimensions are taken (i.e., $A = n_G - 1$) assuming $n^* \geq n_G - 1$, (17) becomes

$$n_p = (n^* + 1)(n_G - 1), \tag{18}$$

which is identical to the effective number of parameters in logistic discrimination (Anderson, 1972).

The general model evaluation strategy described above can be applied to a wide range of specific model selection problems, including choice of dimensionality, subset selection of predictor variables, multiple comparisons and assessing the effects of various data transformations, and so forth. The first two of these are relatively straightforward. For the choice of dimensionality we simply obtain solutions with dimensionality systematically varied, and choose the minimum AIC solution. The selection of the optimal set of predictors is similarly done. The last two are a bit more involving, and will be described in some detail.

The multiple comparisons concern whether criterion groups in DA are significantly distinct, and in particular which groups are distinct and which groups are not distinct from each other. Multisample cluster analysis (CA) developed by Bozdogan (1986) provides a general framework for such a procedure. To illustrate let us assume there are three criterion groups. Some of these groups may not be distinct from each other, and consequently better clustered together. There are five possible cases:

(1) $\mathbf{m}_1 = \mathbf{m}_2 = \mathbf{m}_3$,
(2) $\mathbf{m}_1 = \mathbf{m}_2 \neq \mathbf{m}_3$,
(3) $\mathbf{m}_2 = \mathbf{m}_3 \neq \mathbf{m}_1$,
(4) $\mathbf{m}_1 = \mathbf{m}_3 \neq \mathbf{m}_2$, and
(5) $\mathbf{m}_1 \neq \mathbf{m}_2 \neq \mathbf{m}_3 \neq \mathbf{m}_1$.

These possible cases are often called *clustering alternatives*. Ideal point DA is applied under each hypothesis, and the goodness of fit is compared through AIC.

Case (1) is particularly easy to fit. In this case $d_{kg}^2 = d_k^2$ for all g, so that

$$p_{g|k} = \frac{w_g \exp(-d_k)^2}{\sum_h w_h \exp(-d_k)^2} = \frac{w_g}{\sum_h w_h} = w_g$$

The maximum likelihood estimate of $p_{g|k}$ is thus N_g/N in this case, and the maximum log likelihood is given by

$$\ln L = \sum_g N_g \ln N_g - N \ln N.$$

This case is equivalent to no predictor variables case ($d_{kg} = 0$ for all g). It is also equivalent to the independence hypothesis between criterion groups and observations on the predictor variables, namely,

$$p_{kg} = p_g \cdot p_k,$$

where p_{kg} is the joint probability of group g and the set of observations for subject k. In this case $p_{g|k} = p_{kg}/p_k = p_g$.

Fitting Cases (2), (3) and (4) is slightly more complicated. Let $\mathbf{m}_1 = \mathbf{m}_2$. (The other cases are similar.) In this case we combine group 1 and group 2 into one group. Ideal point DA is applied as if there were only two groups. After the maximum log likelihood is obtained, we add to it $N_1 \ln N_1 + N_2 \ln N_2 - (N_1 + N_2) \ln (N_1 + N_2)$. This adjustment of the maximum log likelihood is necessary, because the likelihood for the original three-group case is

$$\prod_k (p_{1|k})^{f_{k1}}(p_{2|k})^{f_{k2}}(p_{3|k})^{f_{k3}},$$

but if the first two groups are combined, the likelihood would be

$$\prod_k (p_{1|k} + p_{2|k})^{f_{k1}+f_{k2}}(p_{3|k})^{f_{k3}},$$

so that to make up for the difference,

$$\prod_k \left(\frac{p_{1|k}}{p_{1|k} + p_{2|k}}\right)^{f_{k1}} \left(\frac{p_{2|k}}{p_{1|k} + p_{2|k}}\right)^{f_{k2}}$$

has to be multiplied to the latter. But

$$\frac{p_{1|k}}{p_{1|k} + p_{2|k}} = \frac{w_1}{w_1 + w_2} \quad \text{and} \quad \frac{p_{2|k}}{p_{1|k} + p_{2|k}} = \frac{w_2}{w_1 + w_2}$$

and the maximum likelihood estimates of $w_1/(w_1 + w_2)$ and $w_2/(w_1 + w_2)$ are, respectively, $N_1/(N_1 + N_2)$ and $N_2/(N_1 + N_2)$. Therefore,

$$\sum_k \left\{f_{1k} \ln \left(\frac{N_1}{N_1 + N_2}\right) + f_{2k} \ln \left(\frac{N_1}{N_1 + N_2}\right)\right\}$$
$$= N_1 \ln N_1 + N_2 \ln N_2 - (N_1 + N_2) \ln (N_1 + N_2)$$

is added to the maximum log likelihood obtained from the two-group analysis (where $N_1 = \sum_k f_{1k}$ and $N_2 = \sum_k f_{2k}$).

Case (5) is what we usually obtain. Once the maximum log likelihoods are obtained, the best fitting model is chosen according to the minimum AIC criterion. The number of possible cases grows quite rapidly as n_G increases. The exact formula for this number is given in Bozdogan (1986). The multisample CA used in combination with ideal point DA generalizes his original proposal based on the multivariate normality to the general exponential family of distributions.

Exploring and assessing the effect of data transformations is not technically difficult. It is the variety of transformations that requires some discussion. The transformations of the predictor variables should be broadly construed in this paper, and include such transformations as discretizations (categorizations) of continuous variables, interactions between discrete variables, interactions between continuous and discrete variables, and so on, as well as more standard types of transformations such as power, polynomial and spline transformations.

When a continuous variable is nonlinearly or nonmonotonically contributing to discrimination, we may discretize it into a few observation categories, which are then "requantified" by ideal point DA. A potential danger is that the effect of the continuous variable may indeed be linear. Then we may not only lose some information in the original variable in the process of discretization, but also lose degrees of freedom by estimating extra parameters. However, whether or not a particular discretization scheme on a continuous variable is worth incorporating can be explicitly tested using the general model evaluation strategy given at the beginning of this section. (We have a concrete example of this in the next section.) Our current procedure assumes that we already have specific discretization strategies to try out.

Interactions among the predictor variables are potentially quite important in improving predictability of ideal point DA. The interaction is formally defined as the effect of one variable depending on values of other variables, and is captured in cross product terms among the variables. For two discrete (nominal) variables, i and s ($i \neq s$), they are defined as $x_{ki(j)} x_{ks(t)}$ for all k and for all combinations of j and t (categories of variables i and s, respectively). For a continuous variable i and a discrete variable s, they are defined as $x_{ki} x_{ks(t)}$ for all t. An important thing is that x_{ki} corresponding to $x_{ks(t)} = 1$ should be separately centered for each t. An interaction between two continuous variables is simply defined as $x_{ki} x_{ks}$. Interactions among discrete variables are important in relation to the loglinear model. Interactions between continuous and discrete variables are important, since they allow a different distribution of the continuous variables for each subsample of observations defined by the discrete variables. An implication of this in relation to Krzanowski's (1975) location model will be discussed in the discussion section.

In some cases we might want to compare the performance of ideal point DA with that of other DA methods. What criterion should we use, if an equivalent conditional likelihood is not specified for these methods? The AIC can no longer be used for model comparisons. Rate of misclassification is a useful measure in such situations. Apparent error rate (rate of misclassification in training samples) has been widely used for this purpose. However, it is well known that it tends to underestimate true error rate (Efron, 1986), since the same training sample is used to estimate both model parameters and the error rate. The degree of bias in the apparent error rate depends on particular methods of DA.

A number of methods have been proposed for estimating the true error rate; that is, the rate of misclassification expected to occur when classifying a test sample based on parameter estimates derived from a training sample independent of the test sample. There are two widely used resampling plans used to estimate the true error rate. One is the leaving-one-out method of Lachenbruch (1975), and the other the bootstrap method by Efron (1983). In this paper we use the leaving-one-out method. Krzanowski (1975) used this method to compare the performance of his location model with half a dozen other methods of DA, and we had to use the same method for a direct comparison with his results. In the leaving-one-out method one of the cases (subjects) is eliminated from the training sample in turn to be used as a test sample. Model parameters are estimated from the reduced sample and are used to predict membership of the case eliminated from the training sample, and the frequency of misclassification is counted.

The leaving-one-out method as well as other resampling methods is quite useful in investigating stochastic behavior of a model nonparametrically. It may also be useful in evaluating AIC on the basis of weaker statistical assumptions.

3. Applications

In this section we demonstrate use of ideal point DA in two practical situations, emphasizing various model evaluation features of the method. In each case an extensive search is made for the best specification of the model. Different aspects of a model (dimensionality, predictor variables, clustering alternatives, etc.) all "interact"; the best subset of predictor variables for a specific dimensionality may not be the best for other dimensionalities, etcetra. Consequently a candidate model should be specified for each combination of the different aspects of the model. An exhaustive search for the best fitting

Table 1

Summary statistics from multi-sample cluster analysis and subset selection in the Komazawa data

Clustering Alternatives		K	dim	All Predictors	Logical Minimum	Optimal Set of Predictors	Indices of Predictors
1	1,2,3,4	4	1	133.0(10)	119.0(3)	128.8 (7)	2,4,6,7
			2	135.8(16)	109.8(3)	128.8*(10)	3,4,6,7
			3	140.7(21)	104.7(3)	130.4 (9)	3,5,6
2	(1,2),3,4	3	1	135.2(10)	121.1(3)	130.7 (7)	2,4,6,7
			2	141.6(16)	115.6(3)	133.9 (8)	4,6,7
3	(1,3),2,4	3	1	134.9(10)	120.9(3)	130.5 (8)	2,3,4,6,7
			2	141.9(16)	115.9(3)	133.8 (8)	5,6,7
4	(1,4),2,3	3	1	144.2(10)	130.2(3)$^+$		
			2	149.9(16)	123.9(3)	142.6 (10)	3,4,5,6
5	1,(2,3),4	3	1	142.7(10)	128.7(3)	137.9 (7)	3,5,6,7
			2	150.0(16)	124.0(3)	134.8 (8)	3,5,6
6	1,(2,4),3	3	1	154.1(10)	140.1(3)$^+$		
			2	159.2(16)	133.2(3)$^+$		
7	1,2,(3,4)	3	1	137.7(10)	123.7(3)	133.5 (6)	4,6,7
			2	141.5(16)	125.5(3)	132.4 (6)	4,6
8	1,(2,3,4)	2	1	154.1(10)	140.1(3)$^+$		
9	(1,3,4),2	2	1	140.9(10)	126.9(3)	137.8 (7)	4,5,6,7
10	(1,2,4),3	2	1	156.4(10)	142.4(3)$^+$		
11	(1,2,3),4	2	1	143.4(10)	129.4(3)$^+$		
12	(1,2),(3,4)	2	1	137.3(10)	123.3(3)	134.1 (6)	4,6,7
13	(1,3),(2,4)	2	1	157.2(10)	143.2(3)$^+$		
14	(1,4),(2,3)	2	1	151.2(10)	137.2(3)$^+$		
15	(1,2,3,4)	1	0	149.4 (3)	149.4(3)$^+$		

$^+$The cases in which the logically minimum attainable AIC is larger than the minimum AIC for Case 1 , dim=2.

* Minimum AIC

model is often impossible due to a "combinatorial explosion." In the next two subsections readers are encouraged to pay special attention to what heuristic search strategies we use for cutting down the number of candidate models.

3.1 The Komazawa Data

Komazawa (1982) reported profile data of 52 patients falling into the following four disease categories: 1. cerebral haemorrhage ($N_1 = 14$), 2. cerebral infarction ($N_2 = 15$), 3. myocardial infarction ($N_3 = 12$), and 4. angina (pectoris; $N_4 = 11$). The first two are brain-related and the last two heart-related diseases in the human circulatory system. The profile data were obtained on the following seven measures: 1. opthalmology (normal or abnormal), 2. electrocardiogram (ECG; normal or abnormal), 3. age (49 to 59 years old), 4. systolic (high) blood pressure (98 to 216 mmHg), 5. diastolic (low) blood pressure (62 to 120 mmHg), 6. aortic wave speed (6.3 to 10.2 m/sec), and 7. serum cholestrol (146 to 279 mg/dl). The first two are binary, while the remaining five are continuous.

There are fifteen clustering alternatives in the multisample CA of four criterion groups as defined in Table 1. Case (15) is the null model in which all the four criterion groups are assumed identical. In this case predictions can be made only on the basis of group sizes (which is sometimes called a zero dimensional solution). For each of the remaining 14 nonnull cases, solutions could be obtained from one dimension up to the assumed number of distinct groups minus one. In the table the number of distinct groups hypothesized in each clustering alternative is denoted by K, and the dimensionality of a solution by "dim." Altogether we have 22 possible cases (excluding the null case).

For each of the 22 possible cases an optimal subset selection of the predictor variables was conducted. Due to a relatively small sample size no interactions among the original predictor variables were considered. With seven predictor variables, however, the number of possible subsets is quite sizeable. Therefore, it was decided to employ the backward elimination technique. While this heuristic search technique does not ensure global optimality, it is known to work very well in most situations. In the backward elimination technique all predictor variables are initially included in the prediction model. A variable whose elimination decreases the value of AIC most is eliminated at each stage, until no further reduction can be achieved. (Note that the variable whose elimination decreases the AIC value most is the one whose elimination decreases the likelihood least.) In the present case after an "optimal" solution was obtained by the backward elimination all neighboring models were tried to make sure that it was indeed the best solution. In no cases the backward elimination failed to identify the optimal solution. The average number of solutions obtained for each case was approximately 10, a significant reduction from $128 = 2^7$.

A further reduction in the number of possible solutions was achieved by means of a branch and bound process. For each of the 22 cases the full model with all seven predictor variables was first fitted. The resulting AIC values are reported in the column labeled "All Predictors" in Table 1. Then for each case the logical minimum of the AIC value attainable by subset selection was calculated. (See the column labeled "Logical Minimum" in the table.) This value is obtained by subtracting $2n_B$ from the AIC value attained by the full model. Whenever this logically minimum AIC value exceeds the minimum AIC among the full models, the case can be eliminated from further considerations. There is no use to conduct subset selection for this case. Several cases were eliminated this way. We then start the subset selection for the remaining cases. We start with the most promising case, since as soon as we find a solution with the corresponding AIC value smaller than the logically minimum AIC values in the remaining cases, those cases can be eliminated. In the table those cases for which no subset selection was conducted are marked by a plus sign. For all other cases the AIC values and indices of predictor variables corresponding to the optimal subset solution are reported in the last two columns of Table 1.

The minimum AIC solution (marked by an asterisk in the table) was found to be the two dimensional, four-group solution with predictor Variables 3, 4, 6, and 7. The four groups are significantly distinct in the two dimensional space. Apparently, dim = 2 does not imply the number of distinct groups is at most 3, although it implies the number of distinct groups is at least 3. The minimum AIC value of 128.8 is very close to that of $K = 2$, dim = 1 with Variables 2, 4, 6, and 7. The two AIC values are almost identical. In such a case either solution can be chosen on the basis of nonstatistical considerations such as parsimony. We have chosen the solution with a larger number of parameters in the model, since in this particular case the sample size is fairly small, and as we get a larger sample size, the one richer in structure would be more clearly favored by AIC. Also, Variable 3 (age) selected in the two dimensional solution is more readily observed than Variable 2 (ECG). The two solutions are, however, quite similar to each other despite the difference in dimensionality.

The subject and ideal point (group centroid) configuration corresponding to the

Table 2

Effects of discretization in the Komazawa data

	dim 1	dim 2
Original Measures Optimal Predictor Set (Indices of Predictors)	128.8(7) (2,4,6,7)	128.8*(10) (3,4,6,7)
Variable 3 discretized		130.4 (10)
Variable 4 discretized	133.2 (8)	135.1 (12)
Variable 6 discretized	136.4 (8)	139.6 (12)
Variable 7 discretized	132.3 (8)	135.0 (12)

* Minimum AIC

minimum AIC solution is presented in Figure 1. The ideal points of the four criterion groups are indicated by circled numbers, while the subject points are identified by group indices to which they belong. Dotted line segments indicate boundary hyperplanes according to the max $p_{g|k}$ rule.

Perhaps the most illuminating way of looking at the configuration is in terms of two dichotomies: (a) heart-related (Groups 3 & 4) versus (b) brain-related (Groups 1 and 2), and (a) noninfarction (Groups 1 and 4) versus (b) infarction (Groups 2 and 3) diseases. These two dichotomies combined nicely distinguish the four disease categories. They were also neatly borne out in the derived configuration, and indicated by two solid lines crossing with each other at about the center of the configuration. The two heart-related diseases are correlated with a high cholesterol level, while the two brain-related diseases are correlated with high systolic blood pressure and high aortic wave speed. The two noninfarction diseases are correlated positively with age, and the two infarction diseases with high aortic wave speed.

The apparent error rate in this example may look very large (44.2%). However, as argued in the previous section it is not the apparent error rate that is important. In fact the apparent error rate could be made much smaller (36.5%) by selecting a larger model (e.g., the full model with dimensionality 3), but this is not a wise choice.

Komazawa (1982) originally analyzed his data set by the second kind of quantification method (*Q*2) developed by Hayashi (1952) for DA of discrete predictors. This method is similar to the canonical DA applied to dummy coded discrete predictors. A similar method is used as the initialization method in ideal point DA. In order to apply *Q*2, Komazawa had to discretize all the continuous variables. We may test whether the discretization scheme suggested by him improves predictability of ideal point DA. We only consider discretizations of the continuous variables retained in the optimal solution. (For comparison purposes we have done the same for the runner-up solution which was so close to the best solution.) The following discretizations were suggested by Komazawa: (3) age (49 ~ 54, 55 ~ 59), (4) systolic blood pressure (~139, 140 ~ 170, 170~), (6) aortic

Table 3

Summary statistics from multi-sample cluster analysis and subset selection in the Armitage data (3 groups)

Clustering Alternatives	K	dim	All Predictors	Logical Minimum	Optimal Set of Predictors	Indices of Predictors
1 1, 2, 3	3	1	382.0(11)	364.0(2)	373.8 (5)	2,3,5
		2	387.7(19)	353.7(2)	372.2*(7)	5,7,9
2 (1,2),3	2	1	388.8(11)	370.8(2)	380.3 (5)	2,3,5
3 1,(2,3)	2	1	405.9(11)	387.9(2)+		
4 (1,3),2	2	1	386.0(11)	368.0(2)	374.6 (5)	5,7,9
5 (1,2,3)	1	0	398.2 (2)	398.2(2)+		

+The cases in which the logically minimum attainable AIC is larger than the minimum AIC for Case 1 , dim=2.

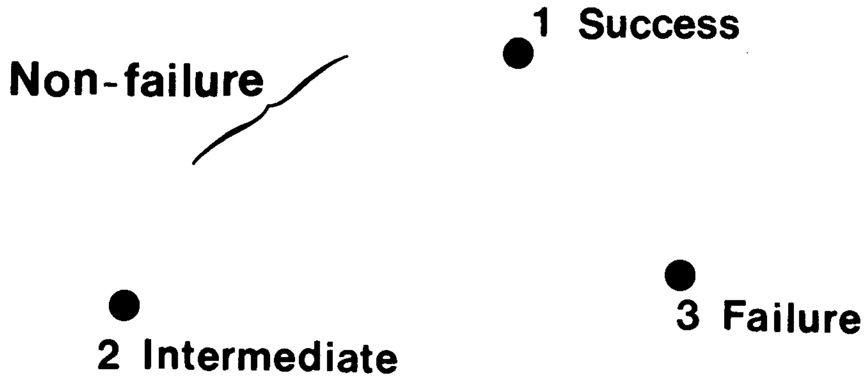
* Minimum AIC

wave speed (~7.4, 7.5 ~ 8.5, 8.5~), and (7) serum cholesterol (~169, 170 ~ 210, 211~). Each of these discretized variables was used as a predictor variable in turn instead of its continuous counterpart. Results are reported in Table 2. None of the AIC values found are as small as that of the original solution. This suggests that we should not discretize the data the way Komazawa did.

3.2 *The Armitage Data*

The second example also concerns medical data. The data were originally collected by Armitage, McPherson and Copas (1969). (For simplicity we call them the Armitage data.) The data concern prognosis after ablative surgery of breast cancer, and consist of six continuous predictors, 1. age at mastectomy or when first seen, 2. log time to ablation, 3. 17-hydroxicorticosteroids (mg per 24 hours), 4. androsterone (mg per 24 hours), 5. dehydroepiandrosterone (mg per 24 hours), 6. aetiocholanolone (mg per 24 hours), and three binary variables, 7. presence (1) or absence (2) of mastectomy, 8. type of ablation (andrenalectomy (1) or hypophysectomy (2)) and 9. presence (1) or absence (2) of lesion on breast. There were intially three criterion groups: 1. success (remission of all signs for at least six months after surgery), 2. intermediate (partial or short lived remission) and 3. failure (no remission).

Armitage et al. (1969) applied logistic discrimination to the data. Perhaps due to the limitation in logistic discrimination at the time of their study, they had to reduce the number of criterion groups into two by combining the success and the intermediate groups. (Logistic discrimination was extended to multiple groups by Anderson in 1972.) Using the multiple comparison feature of ideal point DA we may now question its adequacy. A search for the best fitting model was done in a manner similar to that for the Komazawa data, and results are reported in Table 3. The two dimensional solution in the three-cluster case was found to be the best model with Variables 5, 7, and 9 in the optimal predictor set. This suggests that the success and the intermediate groups are sufficiently distinct. Surprisingly the success and the failure groups (1 and 3) are the ones that could



be least harmfully combined. This can be seen by the least increase in the AIC value when these two groups are combined. Figure 2 displays the ideal point configuration corresponding to the best solution. It is the success group, not the intermediate group, that lies somewhat intermediate between the other two groups. The success and the failure groups are closest in distance, confirming our previous observation. It is likely that the major difference that Armitage et al. found between the nonfailure and the failure groups is primarily due to the difference between the intermediate and the failure groups, and not to the difference between the success and the failure groups.

Krzanowski (1975) applied his location model to the Armitage data after the data were reduced to two groups. (In all subsequent analyses we use the two group data for the sake of direct comparisons.) The location model was specifically designed for DA with a mixture of continuous and discrete predictors. It assumes a separate multivariate normal distribution on the continuous variables for each subsample of observations defined by the discrete predictor variables. Krzanowski estimated true error rate by the leaving-one-out method (Lachenbruch, 1975) described earlier. Frequencies of misclassification he obtained are given in the first column of Table 4.

We applied ideal point DA to the same set of data, and estimated the true error rate by the leaving-one-out method. The result is reported in the second column of Table 4. It seems that the location model does considerably better than ideal point DA. However, the latter result was obtained without subset selection of predictor variables; all nine variables were used. When the subset selection was conducted, the AIC improved from 249.6 to 241.2 (see Table 6) with Variables 2, 7, and 9 in the optimal predictor set. The estimated true error rate is reported in the third column of Table 4. Now the advantage of the location model disappears. These results suggest importance of subset selection in general. Inclusion of noninformative predictor variables in DA is indeed harmful for future predictions.

The location model implies interactions between continuous and discrete predictors (as well as those among the discrete variables). Thus, the performance of ideal point DA may be further improved by including those interactions in the predictor set. In the present case there are 6 continuous variables, and 8 levels altogether formed from three binary variables. Consequently there are 48 interaction terms defined between the continuous and discrete variables, and the number of possible models is just enormous. Including meaningful equality restrictions on parameters (this corresponds with a noninteraction hypothesis) this number is something like $21{,}147^6$. An exhaustive search for the optimal model is simply out of the question.

Table 4

Estimates of true error rate by the leaving-one-out method for the Armitage data (2 groups)

Criterion Groups	Location Model	Ideal Point DA		
		All Predictors (Main Effects Only)	Optimal Set of Predictors (Main Effects Only)	Optimal Interactions
1	34	39^+	30	30
2	27	34^+	27	25

+ Obtained by Krzanowski, 1975

Note: Figures in the table indicate frequencies of misclassification (out of N_1=99 for group 1, and N_2=87 for group 2)

The number of possible models to be considered can be cut down to manageable size by the following rather elaborate search strategy. First, the entire sample was divided into eight subsamples defined by the three binary variables. This was done because interactions between continuous and discrete variables imply the effects of the continuous variables varying across the eight subsamples. Ideal point DA was applied to each subsample with the six continuous variables, and a pattern of contributions of the six variables is identified: significantly positive (+), significantly negative (−), and not significantly different from zero. Variables in the third category can be easily identified by subset selection. These patterns are reported in Table 5. Note that in the present case the positive side is associated with the failure group. Note also that coefficients not significantly different from zero are left blank in the table. Considerable variations in the contribution of the continuous variables across the eight subsamples can be observed in Table 5. Interactions were then defined by grouping subsamples for which the effect of a particular continuous variable is in a same direction (either "+" or "−"). In this way eight interaction terms were defined: 1. Variable 1 for Subsample 2 (−), 2. Variable 2 for Subsample 8 (−), 3. Variable 3 for Subsamples 3, 4, 6, and 8 (+), 4. Variable 3 for Subsample 7 (−), 5. Variable 4 for Subsamples 5 and 7 (+), 6. Variable 4 for Subsample 6 (−), 7. Variable 5 for Subsamples 2, 6, and 8 (−), and 8. Variable 6 for Subsample 3 (−). In addition a separate loglinear DA (Andersen, 1980) was performed on the entire sample with the three binary variables. The last two variables (type of ablation and legion on breast) were found to interact significantly. Consequently interactions between the last two binary variables were included in the set of predictor variables.

Ideal point DA was applied to the Armitage data with the nine interaction terms defined above. The AIC value was remarkably improved. (See Table 6). In order to make sure the eight interaction terms between the continuous and the discrete variables were indeed significant, the model was fitted with each of them deleted from the model in turn. The AIC values obtained from the reduced models are also reported in Table 6. None of the AIC values were as small as that of the original model, indicating a significance of every interaction term considered. The true error rate was estimated with the interaction

Table 5

Patterns of contribution of the continuous variables at different levels of the discrete variables for the Armitage data (2 groups)

Sub-sample	Patterns on the Three Binary Variables	Continuous variables					
		1	2	3	4	5	6
1.	1 1 1						
2.	2 1 1	-				-	
3.	1 2 1			+			-
4.	2 2 2			+			
5.	1 1 2				+		
6.	2 1 2			+	-	-	
7.	1 2 2			-	+		
8.	2 2 2		-	+		-	

model using the leaving-one-out method. The results are reported in the last column of Table 4. The rate of misclassification is even lower than that of the best main-effect-only model of ideal point DA. Note that the AIC result agrees very well with that of the leaving-one-out method.

One may argue that the above comparison between the location model and ideal point DA is unfair to the former, since no search was made for the best location model. This argument is well taken. At this point we are not sure whether the superiority of ideal point DA for the Armitage data is due to an extensive search for the optimal model, or to the flexibility in its assumption. Note that the location model assumes a separate multivariate normal distribution for each subsample, while ideal point DA only one of exponential family of distributions. We are not quite sure if the location model allows as extensive model evaluations as are possible with ideal point DA. We then argue that an advantage of ideal point DA lies precisely in its ability to allow an extensive search for the best specification of the model. (But see a recent article of Daudin, 1986, on this topic.)

4. Discussion and Further Prospects

In this paper we proposed a method of DA which enjoys wide applicability. It allows a mixture of continuous and discrete predictor variables in three different sampling designs, conditional, joint and separate. In the latter two cases the method can be justified under the general exponential family of distributions. As has been demonstrated, the method allows various model evaluations such as choice of dimensionality, optimal subset selection of predictor variables, multiple comparisons (significance tests of the difference among criterion groups), and so forth. These are all essential ingredients of DA.

Although the proposed method is versatile, we do not claim it is the best method in

Table 6

Optimal interaction set for the Armitage data (2 groups)

	AIC (parameters)
All Predictors (Main Effect Only)	249.6 (10)
Optimal Set of Predictors (Main Effect Only) 2,3,5	241.2 (4)
Optimal Interactions	209.4* (12)
Interaction 1 deleted	210.7 (11)
Interaction 2 deleted	222.3 (11)
Interaction 3 deleted	216.8 (11)
Interaction 4 deleted	217.7 (11)
Interaction 5 deleted	213.1 (11)
Interaction 6 deleted	211.7 (11)
Interaction 7 deleted	226.6 (11)
Interaction 8 deleted	222.6 (11)

* Minimum AIC

all conceivable situations. Needless to say fully parametric methods are more efficient, when the assumptions underlying the methods are satisfied (Efron, 1975). Also, there are situations the exponential family of distributions cannot cover, and consequently more flexible nonparametric methods are called for (Hand, 1982). Rather, we claim that there is a wide range of situations for which ideal point DA is most appropriate, and that because of its intermediate nature, the method will not lose much even when it is not the best method available in particular situations.

Table 7 shows some representative methods of DA in terms of their "parametricity" and data types they cover. Parametric methods are based on the full likelihood. Their rigidty is indicated by the fact that separate procedures should be provided for different data types. Semiparametric methods (Anderson, 1982), on the other hand, are based on the conditional likelihood. These methods do not explicitly specify the marginal distribution of the predictor variables. The relationship between ideal point DA and logistic discrimination has already been discussed. They provide semiparametric alternatives to all the three parametric procedures listed in Table 7. From the form of Model (4) it is obvious that ideal point DA provides the conditional maximum likelihood estimation for Fisher's linear discriminant function method. From the analysis made for the Armitage data it is obvious that ideal point DA, by including the interaction terms between con-

Table 7

Classification of some representative methods of discriminant analysis

	Data Types		
	Continuous	Mixture	Discrete
Parametric	linear discriminant function (LDF)	location model	log-linear model
Semi-parametric	ideal point discriminant analysis logistic discrimination		
Nonparametric	histogram methods kernel discriminant analysis nearest neighbor discriminant analysis		

tinuous and discrete variables, can handle situations covered by the location model. In fact it can be rigorously proven that the conditional likelihood for the location model (if it were derived) reduces to the likelihood for ideal point DA. Similarly, DA by the loglinear model (Andersen, 1980; Kalbfleisch, 1984) reduces to ideal point DA, if appropriate interaction terms among discrete predictor variables are included. In loglinear DA only those terms related to criterion groups are effective in discrimination; all other terms (related to marginal probabilities of the predictor variables) fall out, being common to all criterion groups. In the loglinear model the conditional estimation is effected by including all the interaction terms among the predictor variables, thereby perfectly fitting the marginal frequencies of the predictor variables. When this is done, the loglinear model is completely equivalent to ideal point DA. (Again this can be rigorously proven.)

Ideal point DA is flexible enough to accommodate various modifications that will make the method even more general and versatile. It is relatively straightforward to incorporate ordered criterion groups (Cox, 1966). The distance function in Model (4) may be modified in various ways. For example, the squared euclidean distance in the exponent may be replaced by the straight (unsquared) euclidean distance, or the negative exponential function may be replaced by a negative power function of the distance. These may potentially make the model more robust against outlying observations. Perhaps the most interesting generalization of the distance function is to incorporate groupwise metrics, which allow differential sensitivity to the distance from the ideal points across criterion groups. The distance function in this case is written as

$$d_{kg}^2 = (\mathbf{y}_k - \mathbf{m}_g)' V_g (\mathbf{y}_k - \mathbf{m}_g), \tag{19}$$

where $V_g (g = 1, \ldots, n_G)$ is a symmetric positive definite matrix, called a metric matrix. (The $\mathbf{y}_k$ is the vector of coordinates of subject k). The situation that calls for (19) is analogous to that for the quadratic discriminant function due to nonhomogeneous covariance matrices in the multivariate normal distribution. The model is also similar to individual differences models in multidimensional scaling (Carroll & Chang, 1970; Schönemann, 1972).

There are other desirable features to be incorporated in ideal point DA. The model evaluation process should be automated. At the moment specifications needed for a search for the best fitting model are done manually. Ideally some sort of branch and bound algorithm should be implemented in order to overcome "combinatorial explosions" in the search process. More flexible data transformation methods should be incorporated, for example, spline transformations (Villalobos, 1983) and more elaborate discretizations of continuous variables. Both of these require optimization in the data transformation process. Various diagnostic features (e.g., Pregibon, 1981) such as outlier detection, sensitivity analysis, and residual analysis are also quite important in practical use of the method.

References

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

Andersen, E. B. (1980). *Discrete statistical models with social science applications.* Amsterdam: North-Holland.

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

Anderson, J. A. (1974). Diagnosis by logistic discriminant function. *Applied Statistics, 23,* 397–404.

Anderson, J. A. (1982). Logistic discrimination. In P. R. Krishnaiah and L. N. Kanal (Eds.), *Handbook of statistics 2.* Amsterdam: North Holland.

Armitage, P., McPherson, C. K., & Copas, J. C. (1969). Statistical studies of prognosis in advanced breast cancer. *Journal of Chronic Disease, 22,* 343–360.

Bozdogan, H. (1986). Multi-sample cluster analysis as an alternative to multiple comparison procedures. *Bulletin of Informatics and Cybernetics, 22,* 95–130.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35,* 283–319.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Cox, D. R. (1966). Some procedure connected with the logistic qualitative response curve. In F. N. David (Ed.), *Research papers in statistics: Festschrift for J. Neyman* (pp. 55–71). New York: Wiley.

Daudin, J. J. (1986). Selection of variables in mixed-variable discriminant analysis. *Biometrics, 42,* 473–481.

Day, N. E., & Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics, 23,* 313–323.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association, 70,* 892–898.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association, 78,* 316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association, 81,* 461–470.

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

Hand, D. J. (1981). *Discrimination and classification.* Chichester: Wiley.

Hand, D. J. (1982). *Kernel discriminant analysis.* Chichester: Research Studies Press.

Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics, 2,* 69–98.

Kalbfleisch, J. G. (1984). Aspects of categorical data analysis. In Y. P. Chaubey & T. D. Dwivedi (Eds.), *Topics in applied statistics* (pp. 139–150). Montreal: Concordia University Press.

Komazawa, T. (1982). *Suuryoka riron to deta shori* [Quantification theory and data processing]. Tokyo: Asakurashoten. (In Japanese)

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. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics, 19,* 191–200.

Lachenbruch, P. A. (1975). *Discriminant analysis.* New York: Hafner.

Lachenbruch, P. A., Sneeringer, C., & Revo, L. T. (1973). Robustness of the linear and quadratic discrimination function to certain types of nonnormality. *Communications in Statistics, 1,* 39–56.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models.* London: Chapman and Hall.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics, 9,* 705–724.

Ramsay, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika, 43,* 145–160.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics.* Dordrecht: Reidel.

Schönemann, P. H. (1972). An algebraic solution for the class of subjective metrics models. *Psychometrika, 39*, 441–451.

Schönemann, P. H. & Wang, M. M. (1972). An individual difference model for the multidimensional analysis of preference data. *Psychometrika, 37*, 275–309.

Villalobos, M. A. (1983). Estimation of posterior probabilities using multivariate smoothing splines and generalized cross validation (Technical Report No. 725). Madison: University of Wisconsin, Department of Statistics.

Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika, 54*, 167–179.