JOURNAL OF CHEMOMETRICS, VOL. 10, 189–213 (1996)

DISCRIMINANT ANALYSIS WITH SINGULAR COVARIANCE MATRICES. A METHOD INCORPORATING CROSS-VALIDATION AND EFFICIENT RANDOMIZED PERMUTATION TESTS

PHILIP JONATHAN*, W. V. (MAC) McCARTHY AND ADRIAN M. I. ROBERTS‡
Shell Research Limited, Broad Oak Road, Sittingbourne, Kent ME9 8AG, U.K.

SUMMARY

A computationally efficient approach has been developed to perform two-group linear discriminant analysis using high-dimensional data. The analysis is based on Fisher's method and incorporates two important validation stages: 1, full leave-one-observation-out cross-validation; 2, randomized permutation distribution testing. The resulting algorithm and software are known as CREDIT (cross-validated random-permutation-tested efficient discrimination based on an adjusted generalized inverse for the sample total covariance matrix).

The algorithm has been implemented in the SAS/IML matrix programming language and provides dramatic improvements in computational efficiency compared with existing software for discriminant analysis incorporating validation stages 1 and 2 above. Application of CREDIT to nine multivariate data sets indicates that the predictive performance of the approach, assessed using cross-validation, is comparable with that of other methods for discriminant analysis. Comparisons with two specific methods are included.

Randomized permutation tests show that success rates using the true response classes are almost always better than success rates using random permutations of the classes. This gives confidence that there is a *useful* linear discriminant relationship present in the data being analysed.

For a *randomly* selected training set (used to construct the discriminant rule) the success rates for CREDIT are unbiased predictive success rates for allocating other observations to groups. Predicting group memberships for *future* observations using any discriminant model based on singular estimates of covariance matrices must be performed with great care. A discussion of methods to test the concordance of future observations with the training set is given.

KEY WORDS concordance; discriminant analysis; permutation test; principal components; QSAR

INTRODUCTION

The ability to interpret and utilize high-dimensional data is becoming increasingly important in countless fields of science and technology. In analytical chemistry, for example, spectroscopy provides huge quantities of data to characterize a chemical. Spectroscopy and multivariate statistics have been used in tandem to provide a method to quantify the properties of complex chemicals, such as the octane number of gasolines and the concentration of trace substances in

* Author to whom correspondence should be addressed. Permanent address: Shell Research Rijswijk, PO Box 60, 2280 AB Rijswijk, Netherlands.
‡ Permanent address: Zeneca plc, Jealott's Hill Research Station, Bracknell, Berkshire RG12 9EY, U.K.

Received September 1994
Accepted August 1995

a complex mixture. The major motivation for adopting a statistical solution to these problems is that actual measurement of octane number, for example, is time-consuming and expensive; the combination of spectroscopy and statistics is far more efficient.

We have been particularly concerned with applications of multivariate statistics in molecular design. Here we want to find novel useful performance chemicals. Only in the most elementary of cases is it feasible to generate reasonable theoretical models for the chemical/physical system under examination. In almost every useful application we are again forced to adopt a statistical solution. Molecular modelling can provide high-dimensional theoretical descriptions of chemicals; these descriptors can be used as training data to develop quantitative structure–activity relationships (QSARs) or quantitative structure–property relationships (QSPRs) which characterize a chemical's performance (as a pharmaceutical or gasoline additive, for example) in terms of the theoretical description of that chemical. Various statistical methods for prediction using multivariate data have been used for QSAR studies.[1,2] Discriminant analysis is just one method to establish rules with which to predict the properties of a chemical.

QSAR – a challenging area for chemometrics

The development of useful QSARs is frequently difficult, even when the ratio of observations to variables in the sample is high. In the situation addressed here, the number of variables exceeds the number of observations! In many applications of chemometrics in analytical chemistry, physical theory supports the application of linear statistical techniques; for example, in multivariate calibration using near infrared spectroscopy, we know that the intensity of measured absorbance of a chemical solution should be linearly related to the concentration of solute in the solution provided that the solution is sufficiently weak. In molecular design, however, it is often impossible to justify specific assumptions concerning the nature of the relationship between cause and effect. Consequently, we cannot justify the application of particular statistical methods to establish the QSAR. Instead, the chemometrician is forced to adopt a pragmatic approach: a QSAR (and the statistical methodology used to establish it) is judged by the quality of its predictions. The CREDIT method introduced here is motivated by these considerations.

The generation of data for QSAR modelling is expensive in terms of both the time of biologists and chemists and the expense of performing complex sequences of chemical synthesis and in subsequent evaluation of chemical performance. As a result, chemometricians are often presented with inadequate data. Specifically, the ratio of observations to variables is usually very low and the variability of response data is very high. Even when experimental design methods are used in an effort to ensure efficient use of resources, the eventual data sample is usually incomplete owing to the prohibitive complexity of some syntheses (which often cannot be judged *a priori*). Nevertheless, *decisions will be taken* based on the data gathered; the chemometrician's task is to ensure that the data are used as effectively as possible to influence those decisions. Model validation assumes increased importance in such circumstances. CREDIT incorporates both cross-validation and randomized permutation testing in order to check the usefulness of a particular QSAR prediction rule as rigorously as possible given the limitations outlined here.

Numerous different molecular properties have been used for QSAR, ranging from simple physicochemical measures of molecular size and shape to highly multivariate IR and NIR spectroscopic data. In this paper we consider the application of two-group linear discriminant analysis to IR and NIR data. We also consider applications involving theoretical chemical descriptors derived from molecular modelling, namely CoMFA and EVA.

The CoMFA (comparative molecular field analysis) descriptor[3] characterizes a chemical in terms of the values of electronic interaction between the chemical and a small probe species (such as a carbon cation) placed in turn at points on a grid surrounding the chemical. The chemical is therefore represented as a 3D lattice of numbers. The current method for statistical analysis of such data is to unwrap the lattice into a long 1D *spectrum-like* vector, thereby ignoring the spatial correlation structure of the original descriptor. A number of promising QSAR analyses using CoMFA have been reported.[4,5]

The EVA (eigenvalue) descriptor developed by L. Phillips and co-workers at Shell Research, Sittingbourne is based on a theoretical analysis of molecular vibrational motion. The so-called *normal modes of vibration*, calculated by spectral decomposition of the molecular vibrational Hessian matrix, are used to create a *spectrum-like* vector which can be viewed as a theoretical approximation for an IR spectrum (although EVA was not developed for this purpose).

For each of the applications under consideration here, we are thus faced with analysis of multivariate observations in the form of spectra. In order to establish the QSAR, descriptor and response data are required for a *training set* of chemicals, the members of which are typically chosen *systematically* rather than at random. In general, measuring the response data is time-consuming and expensive (for example, screening a pharmaceutical or engine testing a gasoline additive). As a result, the training set usually contains far fewer observations (<100 typically) than there are variables (>500) in each multivariate observation. Classical techniques of multivariate statistics are therefore not immediately applicable.

The available statistical tools for QSAR using high-dimensional data can be partitioned into two classes: regression methods (for continuous responses) and classification methods (for categorical responses). A number of regression methods have been used for QSAR purposes, in particular techniques such as partial least squares[6] (PLS). Techniques such as projection pursuit regression and MARS[7] (multivariate adaptive regression splines) have received increasing attention. In this report, however, attention is focused on classification.

Probably the most popular methods for classification are discriminant analysis[8,9] and SIMCA[10] (soft independent modelling of class analogy). Quadratic and most forms of linear discriminant analysis assume that the class populations are multivariate normal; these techniques function adequately when the training data provide reasonable estimates of the population means and covariances. SIMCA was specially developed for problems with low numbers of observations compared with variables. Each class is represented by a principal component model; classification of a test observation is made according to its distances from each of the classes. SIMCA has become a popular tool for chemometricians.

Campbell[11] suggested the use of shrunken estimators in discriminant and canonical variate analysis. Friedman[12] and Frank[13] have respectively reported their *regularized discriminant analysis* and DASCO (discriminant analysis with shrunken covariances). These utilize biased estimates of class covariances; this bias helps to overcome the problem of highly variable covariance matrix estimates, typical in applications for which the number of variables greatly exceeds the number of observations. Friedman,[12] Frank[13] and Frank and Friedman[14] report the results of simulation studies to assess the relative predictive performance of the various classification techniques available. Friedman[12] points out that the adoption of the pooled within-group covariance matrix (even if the population class covariances are different) introduces a considerable degree of regularization. The decrease in variance often leads to superior performance of linear discriminant analysis compared with more complex approaches such as quadratic discriminant analysis, especially when the ratio of observations to variables is low; hence the popularity of linear discriminant

analysis. Hastie *et al.*[15] have recently proposed a novel approach to discriminant analysis using optimal scoring.

Background to current developments and the need for rigorous model validation

Work at Shell Research, Sittingbourne has focused on the development of novel techniques for two-group linear discriminant analysis for high-dimensional data. This research is motivated by the need to provide two-way classifications of chemicals in analytical chemistry and molecular modelling. Following initial work based on MacFie's principal component/canonical variate analysis, a method based on ideas proposed by Campbell and Atchley[16] was developed; this method, referred to henceforth as MCA (modified Campbell and Atchley) is the motivation for the new approach, CREDIT, reported here. Subsequently, McCarthy developed a form of generalized ridge discriminant (GRD) analysis. Both the MCA and GRD methods are summarized below and reported in Reference 17 and will be used here to facilitate comparison of predictive performance for CREDIT.

Model validation is an essential step in performing the classification analysis. Because of the sparsity of observations, it is possible to find linear combinations of the explanatory variables which allow the training data to be fitted perfectly during model building; that is, 100% resubstitution success rates can be obtained. The corresponding model possesses impressive descriptive power but often almost no predictive power. Krzanowski *et al.*[17] have demonstrated this for the so-called *zero-variance discriminator*. In view of this problem, leave-one-out cross-validation[18] has become a popular approach to estimating the predictive power of a classifier; almost all the classification techniques discussed above use cross-validatory assessment. Computationally, cross-validation is expensive since it involves the generation of n different classifiers (where n is the number of observations in the training set). For suitably formulated problems, however, Friedman[12] showed that a rank one matrix inverse updating approach can be employed, thereby dramatically reducing the computational burden of cross-validation. Similarly, Dunne and Stone[19] have recently reported the use of rank one matrix pseudoinverse downdating[20]. Unfortunately, matrix inverse up- and downdating are not applicable to CREDIT and MCA since they both involve the selection of a subset of principal components with which to form the discriminant rule. Matrix inverse up- and downdating have not been considered for GRD to date.

Estimation of prediction success rate alone is insufficient to assess the usefulness of a discriminant rule; its significance should also be estimated. In order to quantify the extent to which a measured cross-validated success rate is 'significant', we use the method of randomized permutation testing. Discriminant models for a large number of random permutations of the responses are developed and the predictive power of each is assessed using cross-validation. A distribution of randomized permutation success rates is then constructed, the actual predictive performance is deemed to be 'significant' if it is in the extreme right-hand tail of the distribution. The application of randomized permutation testing in singular discriminant analysis has been discussed by Krzanowski *et al.*[17].

There are many different measures to assess the predictive performance of the discriminant rule; for example, we could adopt the overall success rate or the classification rate for an individual group or any combination of these, as assessed by cross-validation. These measures can be referred to as utilities. Randomized permutation tests can be used to test whether such utilities are biased.

In practice, the results of the randomized permutation tests are most desirable because they give confidence that there is a *useful* linear discriminant relationship (between explanatory

variables and the response class) in the data being analysed. For example, analyses of two different data sets might yield discriminant rules A and B with unbiased predictive success rates (assessed using cross-validation) of 80% but with tail probability (p-value) estimates of 0·001 and 0·05 respectively. The result for rule A suggests that the actual predictive performance is much better than could be obtained from any other permutation of the class allocations. This gives confidence that a specific link between molecular cause and chemical effect actually exists for these data, which can be identified using discriminant analysis. On the other hand, the result for rule B indicates that the actual predictive performance could be bettered quite easily by choosing a different permutation of class allocations. In turn, this suggests that the link between cause and effect is not particularly special and hints that different molecular descriptors and other methods to establish the predictive rule should be explored.

Randomized permutation tests are computationally very intensive, necessitating an R (>200)-fold increase in the computational complexity of the analysis. The development of algorithms which overcome this heavy computational burden is clearly of some importance.

In this paper we introduce a method known as CREDIT (cross-validated random-permutation-tested efficient discrimination based on an adjusted generalized inverse for the sample total covariance matrix) for two-group discriminant analysis. CREDIT performs efficient randomized permutation testing, facilitated by a reformulation of the discriminant problem. The sections below report, in turn, an overview of the method, application of the method, discussion and conclusions. An algorithm for CREDIT is given in the Appendix. However, first, to put CREDIT in context, we give a brief outline of two-group linear discriminant analysis.

Linear discriminant analysis

We consider a data matrix $\mathbf{X}$ $(n \times p)$ consisting of n observations on a random p-vector taken from two distinct populations π_1 and π_2. After reordering its rows, we can partition $\mathbf{X}$ into two matrices $\mathbf{X}_1$ $(n_1 \times p)$ and $\mathbf{X}_2$ $(n_2 \times p)$ corresponding to observations in groups 1 and 2 respectively:

$$\mathbf{X}^{\mathrm{T}} = [\mathbf{X}_1^{\mathrm{T}} \,|\, \mathbf{X}_2^{\mathrm{T}}] \tag{1}$$

Fisher's linear discriminant function[21] is a linear function of the p variables that maximizes the ratio of the between-group sum of squares to the pooled within-group sum of squares. That is, we seek the linear combination $y = \mathbf{a}^{\mathrm{T}}\mathbf{x}$ of the variables defined by the p-vector which maximizes

$$(\mathbf{a}^{\mathrm{T}}\mathbf{W}\mathbf{a})^{-1}(\mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a}) \tag{2}$$

Here $\mathbf{W}$ is the sample pooled within-group covariance matrix

$$\mathbf{W} = \frac{1}{n-2}(\mathbf{X}_1^{\mathrm{T}}\mathbf{H}_{n_1}\mathbf{X}_1 + \mathbf{X}_2^{\mathrm{T}}\mathbf{H}_{n_2}\mathbf{X}_2) \tag{3}$$

in which $\mathbf{H}_m$ is the $m \times m$ centring matrix

$$\mathbf{H}_m = \mathbf{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^{\mathrm{T}} \tag{4}$$

$\mathbf{T}$ the sample total covariance matrix

$$\mathbf{T} = \frac{1}{n-1}(\mathbf{X}^{\mathrm{T}}\mathbf{H}_n\mathbf{X}) \tag{5}$$

and **B** is the sample between-group covariance matrix, which has the particularly convenient form

$$\mathbf{B} = (n-1)\mathbf{T} - (n-2)\mathbf{W} = \frac{n_1 n_2}{n}\,\mathbf{d}\mathbf{d}^{\mathrm{T}} \tag{6}$$

d is the sample group mean displacement vector

$$\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \tag{7}$$

where $\bar{\mathbf{x}}_i$ $(i = 1, 2)$ refers to the sample mean vector for group i. When **W** has full rank, the discriminant vector **a** can be shown to be the eigenvector of $\mathbf{W}^{-1}\mathbf{B}$, namely

$$\mathbf{a} = \mathbf{W}^{-1}\mathbf{d} \tag{8}$$

In problems for which $n - 2 < p$, **W** is rank deficient and hence the classical solution (8) is not accessible; in numerical analytic terms we have an ill-posed problem.[12,22] In such cases a solution of the form

$$\mathbf{a} = \mathbf{W}^{\cdot}\mathbf{d} \tag{9}$$

is usually sought, where $\mathbf{W}^{\cdot}$ is calculated from the properties of W and performs the function of an inverse. the choice of $\mathbf{W}^{\cdot}$ is critical; for instance, we might use $\mathbf{W}^{\cdot} = \mathbf{W}^{-}$, a generalized inverse[21] for **W**. Once $\mathbf{W}^{\cdot}$ is chosen and calculated, a new observation $\mathbf{x}_t$ (not in the training set) can be allocated to one of the two groups using the following rule:

$$\text{allocate to group 1} \quad \Leftrightarrow \quad |\mathbf{a}^{\mathrm{T}}(\mathbf{x}_t - \bar{\mathbf{x}}_1)| < |\mathbf{a}^{\mathrm{T}}(\mathbf{x}_t - \bar{\mathbf{x}}_2)| \tag{10}$$

CREDIT – AN ALTERNATIVE FORMULATION

In this paper we consider an alternative formulation of the linear discriminant problem. We seek a vector **a** which maximizes the ratio

$$(\mathbf{a}^{\mathrm{T}}\mathbf{T}\mathbf{a})^{-1}(\mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a}) \tag{11}$$

where **T** is now the sample total covariance matrix. The solution to this optimization is the principal eigenvector of $\mathbf{T}^{-1}\mathbf{B}$ (assuming that $\mathbf{T}^{-1}$ exists) given by

$$\mathbf{a} = \mathbf{T}^{-1}d \tag{12}$$

This alternative formulation was reported by Fisher,[23] who observed that linear discriminant analysis can be cast in a regression context by a particular choice of response vector.

The full rank case

It is interesting to compare the formulations represented by equations (2) and (11). When **W** (and hence **T**) is full rank, the respective optimizations (2) and (11) are identical, since by (6), $(n-1)\mathbf{T} = (n-2)\mathbf{W} + \mathbf{B}$. Indeed, it can be shown that

$$\mathbf{T}^{-1}\mathbf{d} = (n-2)^{-1}[(n-1) - n^{-1}n_1 n_2(\mathbf{d}^{\mathrm{T}}\mathbf{T}^{-1}\mathbf{d})]\mathbf{W}^{-1}\mathbf{d} \tag{13}$$

That is, the unit vector solutions of (2) and (11) are the same.

The singular case

When **W** (and possibly **T**) is singular, the two optimizations (2) and (11) are also equivalent, but nevertheless motivate different solutions. The solution to (2) is obtained by solving the

equation $\mathbf{Ba} = \lambda\mathbf{Wa}$, where both $\mathbf{B}$ and $\mathbf{W}$ are now singular. Using (6) to express $\mathbf{W}$ in terms of $\mathbf{T}$ and $\mathbf{B}$, it is trivial to show that $\mathbf{a} = \mathbf{T}^{-1}\mathbf{d}$ provides a valid solution, where $\mathbf{T}^-$ is any one of a large family of generalized inverses of $\mathbf{T}$ that includes the Moore–Penrose generalized inverse. Similarly, $\mathbf{a} = \mathbf{T}^-\mathbf{d}$ is a solution of optimization (11). The basic reason that the 'generalized' solution $\mathbf{T}^-\mathbf{d}$ holds is the result:

$$\mathbf{T}\mathbf{T}^-\mathbf{d} = \mathbf{d} \tag{14}$$

It is apparent that $\mathbf{d}$ lies in the range of $\mathbf{T}$; this result is easily demonstrated by expressing the LHS of (14) in terms of the singular value decomposition of the centred data matrix $\mathbf{HX}$.

Conversely, $\mathbf{d}$ *does not* lie in the range of $\mathbf{W}$ in general and we cannot write $\mathbf{W}\mathbf{W}^-\mathbf{d} = k\mathbf{d}$ where k is some constant. For this reason $\mathbf{W}^-\mathbf{d}$ does not provide an exact solution for (2) (or (11)). The example illustrated in Figure 1 serves to emphasize this difference. Consider a data set consisting of two groups of observations. Suppose that the data are p (>3)-dimensional but that they can be projected into three dimensions without loss of information; hence both $\mathbf{T}$ and $\mathbf{W}$ are singular. Now suppose that the two groups of data occupy the planes $x = -1$ and $x = 1$ respectively and that the means of the two groups lie on the x-axis. It is clear that the eigenvectors of $\mathbf{W}$ (the within-group covariance matrix) will be some pair of orthonormal vectors in the (y, z)-plane. Furthermore, the mean displacement vector $\mathbf{d}$ lies perpendicular to $\mathbf{W}$; that is, $\mathbf{d}$ is in the null space of $\mathbf{W}$. Thus $\mathbf{W}\mathbf{W}^-\mathbf{d} = \mathbf{0}$ and $\mathbf{W}^-\mathbf{d} = \mathbf{0}$ does not provide a useful basis for discrimination! However, $\mathbf{d}$ clearly lies in the range of $\mathbf{T}$, so that the solution $\mathbf{a} = \mathbf{T}^-\mathbf{d}$ holds.

We can generalize this example and consider mean displacement vectors which do not lie along the x-axis. The difference between $\mathbf{T}^-\mathbf{d}$ and $\mathbf{W}^-\mathbf{d}$ is still apparent, however, since a component of $\mathbf{d}$ will still lie in the null space of $\mathbf{W}$. Writing $\mathbf{d} = \mathbf{d}_1 + \mathbf{d}_2$, where $\mathbf{d}_1$ lies in the space spanned by the eigenvectors of $\mathbf{W}$ and $\mathbf{d}_2$ lies in the null space of $\mathbf{W}$, we have $\mathbf{W}\mathbf{W}^-\mathbf{d} = \mathbf{d}_1 \neq k\mathbf{d}$ for any constant k in general.

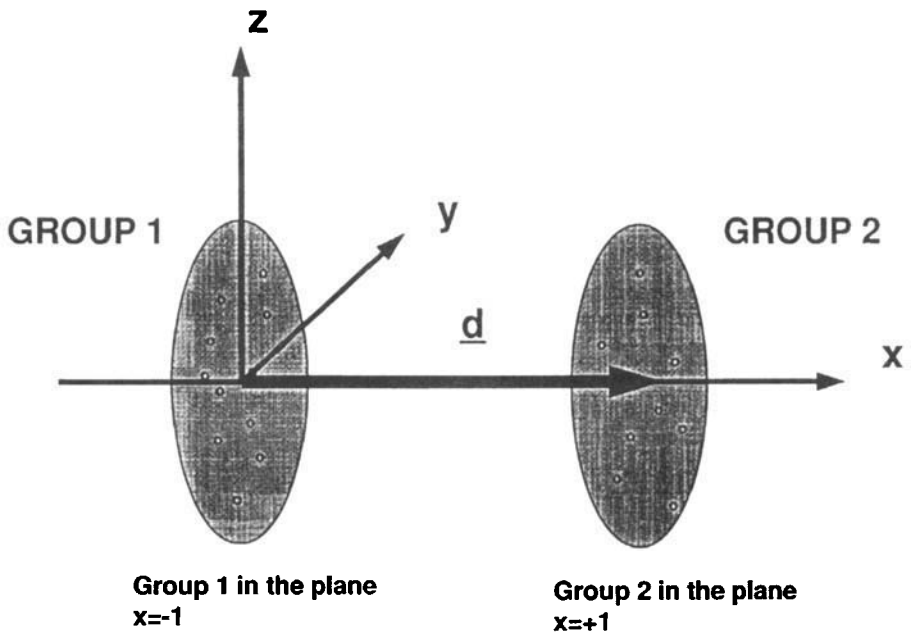


Figure 1. Simple example to illustrate that $\mathbf{T}^-d$ and $\mathbf{W}^-\mathbf{d}$ are different

Sundberg and Brown,[24] in the context of calibration, discuss the dual approaches of, firstly, regression of dummy (population identifier) on spectrum, our approach via equation (11), and, secondly, spectrum on dummy, followed by generalized least squares inversion, akin to our approach via equation (2). In the singular case they show that the two distinct lines of development lead to the same solution space and in particular $\mathbf{a} = \mathbf{T}^{-}\mathbf{d}$ provides a solution to both.

Randomized permutation tests

Another advantage of $\mathbf{a} = \mathbf{T}^{-}\mathbf{d}$ over $\mathbf{W}^{-}\mathbf{d}$ is apparent from consideration of the randomized permutation test. To evaluate the significance of predictive performance, a large number R (>200) of different discriminant rules should ideally be evaluated for random permutations of the original groupings of observations. Application of $\mathbf{W}^{-}\mathbf{d}$ requires recalculation of $\mathbf{W}$ (and hence $\mathbf{W}^{-}$) for each of these permutations. Application of $\mathbf{T}^{-}\mathbf{d}$, however, is much more efficient, since $\mathbf{T}$ is independent of the grouping of observations. Adoption of solution (12) yields a dramatic reduction in the computational complexity of any analysis.

Selection of eigenvectors

As mentioned in the Introduction, equation (9), it is not necessary to adopt a generalized inverse as the basis for the discriminant method. In CREDIT we choose to select a subset of the eigenvectors/principal components of $\mathbf{T}$ which individually provide the best discrimination (as explained below) in order to construct a discriminant rule. In this way we avoid the inclusion of principal components with respect to which the two groups are relatively similar. An evaluation of different approaches to eigenvector selection is given in the Applications section.

Adjustment of eigenvalues

Spectral decomposition of the covariance matrix yields eigenvalues which are biased estimates for the population variances; the smallest variances are underestimated whereas the largest variances are overestimated. The discriminatory importance of eigenvectors/principal components with small eigenvalues can be dramatically inflated because of the negative bias.[12] In order to correct for this effect in CREDIT, we apply a small positive adjustment to all eigenvalues before the assessment of discriminatory importance. Of course, the adjustment also inflates the values of the larger eigenvalues which are already overestimates. However, the relative effect of the adjustment on large eigenvalues is small.

The size of the adjustment is arbitrary and generally application-dependent. However, in our experience a small positive correction of about 1% of the mean eigenvalue is desirable. We present an assessment of the effect of eigenvalue adjustment in the Applications section.

Deriving the CREDIT discriminant rule

An outline of the CREDIT method is given below, highlighting the important steps in the approach.

1. Estimate the eigenstructure of $\mathbf{T}$.
2. Adjust eigenvalues of $\mathbf{T}$, then select eigenvectors with highest discriminatory importance.
3. Establish the discriminant rule.

A more complete algorithm is given in the Appendix.

Sample centred inner product matrix

We start by estimating the eigendecomposition of the sample total covariance matrix **T** given by (5). Direct computation of this $p \times p$ matrix (where $p > 500$) is not necessary, since the eigenstructure can be calculated from knowledge of the eigendecomposition of the sample centred inner product matrix **M**:

$$\mathbf{M} = \frac{1}{n-1}(\mathbf{H}_n\mathbf{X})(\mathbf{H}_n\mathbf{X})^{\mathrm{T}} = \mathbf{E}\mathbf{L}\mathbf{E}^{\mathrm{T}} \tag{15}$$

The second equality follows from the spectral decomposition theorem, where $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_q)$, $\mathbf{L} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_q)$ and $\mathrm{rank}(\mathbf{M}) = q$ for some value of q $(1 \leqslant q \leqslant n-1)$. The eigenvectors $\mathbf{e}_i$ are related to the eigenvectors $\mathbf{f}_i$ of **T** according to the expression:

$$\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_q) = \frac{1}{(n-1)^{1/2}}(\mathbf{H}_n\mathbf{X})^{\mathrm{T}}\mathbf{E}\mathbf{L}^{-1/2} = (\mathbf{H}_n\mathbf{X})^{\mathrm{T}}\mathbf{Z} \tag{16}$$

The matrix **Z** consists of the vectors $\mathbf{z}_i$, $i = 1, 2, \ldots, q$, and is given by

$$\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q) = \frac{1}{(n-1)^{1/2}}\mathbf{E}\mathbf{L}^{-1/2} \tag{17}$$

Note that the eigendecomposition of **T** could have been equally well obtained from the singular value decomposition of the centred data **HX**.

Adjusted eigenvalues

The eigenvalues λ_i, $i = 1, 2, \ldots, q$, of the sample centred inner product matrix **M** provide biased estimates of the population variances as discussed above. In an effort to accommodate this bias, an (arbitrary small) adjustment is made to all the eigenvalues to give a new set of adjusted eigenvalues λ_i^* such that

$$\lambda_i^* = \lambda_i + \frac{1}{100q}\sum_{i=1}^{q}\lambda_i \tag{18}$$

Selection of eigenvectors according to discriminatory importance

Only those eigenvectors/principal components of **T** having highest discriminatory importance are retained for incorporation in the discriminant vector. The discriminatory importance δ_i of the principal component corresponding to eigenvector $\mathbf{f}_i$ is defined as the squared difference in principal component score for the two groups divided by the adjusted variance of the scores derived from the total covariance matrix:

$$\begin{gathered}\delta_i = \lambda_i^{*-1}\psi_i^2 \\ \boldsymbol{\psi}^{\mathrm{T}} = (\psi_1, \psi_2, \ldots, \psi_q) = (n_1^{-1}\mathbf{1}_{n_1}^{\mathrm{T}} \mid -n_2^{-1}\mathbf{1}_{n_2}^{\mathrm{T}})\mathbf{X}\mathbf{F} = \mathbf{d}^{\mathrm{T}}\mathbf{F}\end{gathered} \tag{19}$$

Only those k eigenvectors having the highest discriminatory importance δ are retained to form the allocation rule (analogous to equation (10)). After reordering the eigenvectors and

eigenvalues so that $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_q$, the value of k is selected so that

$$\frac{\sum_{i=1}^{k} \lambda_i^*}{\sum_{i=1}^{q} \lambda_i^*} \approx 0{\cdot}95 \tag{20}$$

ensuring that the retained principal components explain approximately 95% of the adjusted variation in $\mathbf{T}$. Note that (20) is a rather arbitrary method for selecting the number of eigenvectors to retain; other strategies are discussed in the Applications section. The corresponding retained eigenvalue and scaled eigenvector matrices are referred to as $\mathbf{L}_K^*$ and $\mathbf{E}_K$ respectively. $\boldsymbol{\psi}_K$ is used to refer to the group mean displacement vector evaluated in terms of the retained principal components only.

The allocation rule

After some algebra the allocation rule can be expressed in the following form:

$$\text{allocate to group 1} \quad \Leftrightarrow \quad \boldsymbol{\psi}_K^{\mathrm{T}} \mathbf{L}_K^{*-1} \mathbf{Z}_K^{\mathrm{T}} \mathbf{H}_n [\mathbf{X}\mathbf{x}_t - \tfrac{1}{2}\mathbf{X}\mathbf{X}^{\mathrm{T}}(n_1^{-1}\mathbf{1}_{n_1}^{\mathrm{T}} \mid n_2^{-1}\mathbf{1}_{n_2}^{\mathrm{T}})^{\mathrm{T}}] > 0 \tag{21}$$

In (21), $\mathbf{L}_K^*$ represents the variance of principal component scores and consequently is adjusted as in (18) to ameliorate the bias in such variances. $\boldsymbol{\psi}_K$ and $\mathbf{Z}_K$ are calculated using the unadjusted $\mathbf{L}_K$ as in (17).

APPLICATIONS

The CREDIT algorithm derived above has been implemented as an SAS/IML computer program employing the routine EIGEN to calculate spectral decompositions. This program has been used to perform two-group linear discriminant analysis of high-dimensional data for a large number of data sets at Shell Research, Sittingbourne. Here we report the results of analyses of nine different data sets.

The first eight sets correspond to discretized IR, NIR or EVA spectra for potential commercial compounds. Each chemical has been tested in a physical/chemical screen; screening results permit the chemical to be classified as *active* or *inactive*. The objective of the discriminant analysis is thus to generate a rule for accurate classification of future chemical spectra as active or inactive. The last data set CoMFA1 consists of CoMFA descriptors. Industrial confidentiality prevents further description, of either the data sets or the problems that generated them.

It is apparent from Table 1 that the data are high-dimensional and that the number of variables greatly exceeds the number of observations. It should be noted that the data sets IR1, NIR1 and NIR2 are identical with those analysed by Krzanowski *et al.*,[17] who use the same naming convention. Figures 2(a)–2(d) give typical IR, NIR, EVA and CoMFA 'spectra' respectively for illustration.

In particular, Figure 2 indicates that the IR and EVA spectra are qualitatively similar, taking only positive values. The NIR data, in the form of a derivative of absorbance with respect to frequency, take positive and negative values. The CoMFA spectra typically consist of intervals of consecutive non-zero values separated by intervals where intensity is zero.

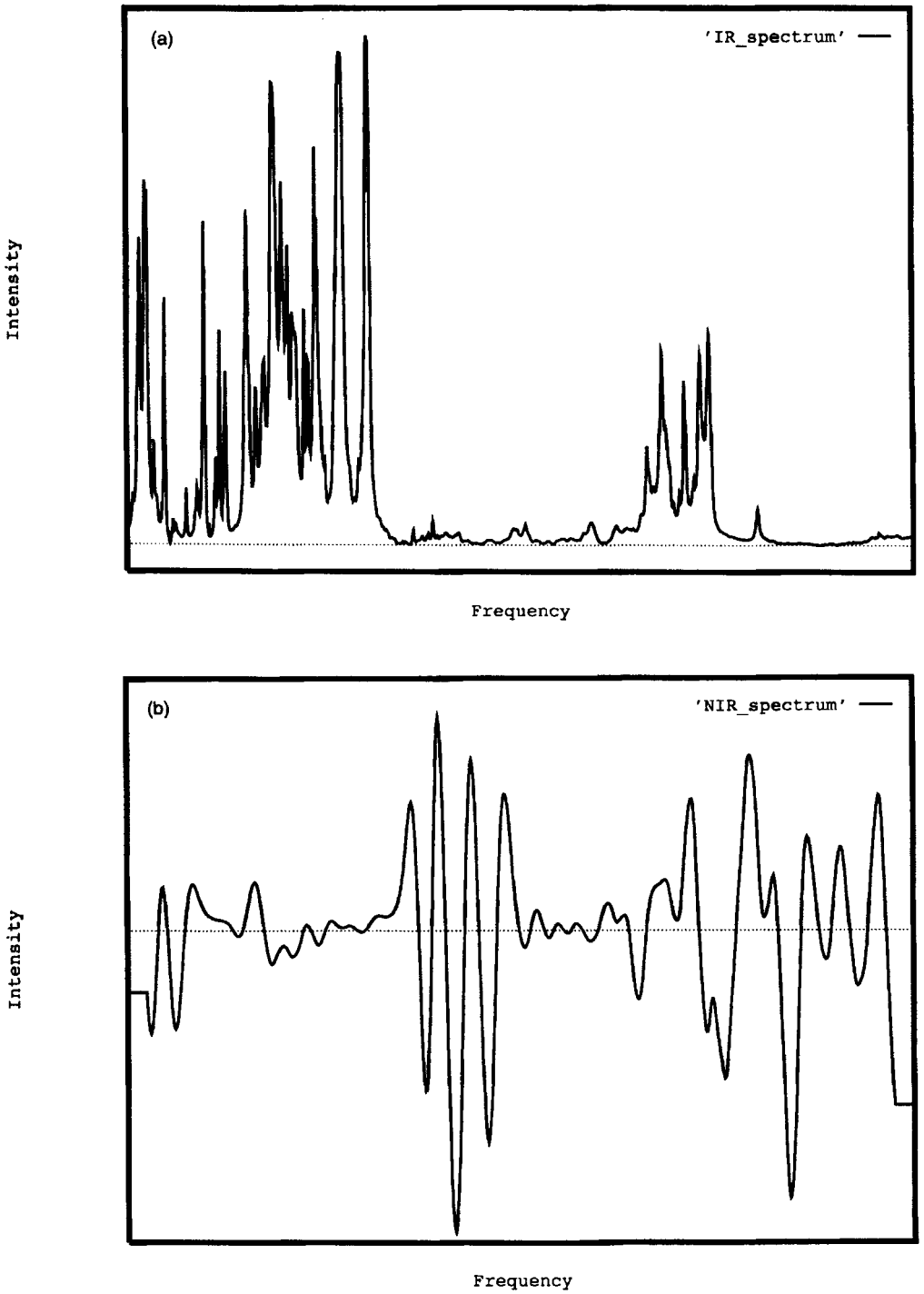


Figure 2. Typical observations: (a) measured infrared (IR) spectrum; (b) moving average second-derivative measured near infrared (NIR) spectrum; (c) calculated EVA descriptor; (d) calculated CoMFA descriptor. The dotted line in each plot indicates zero intensity

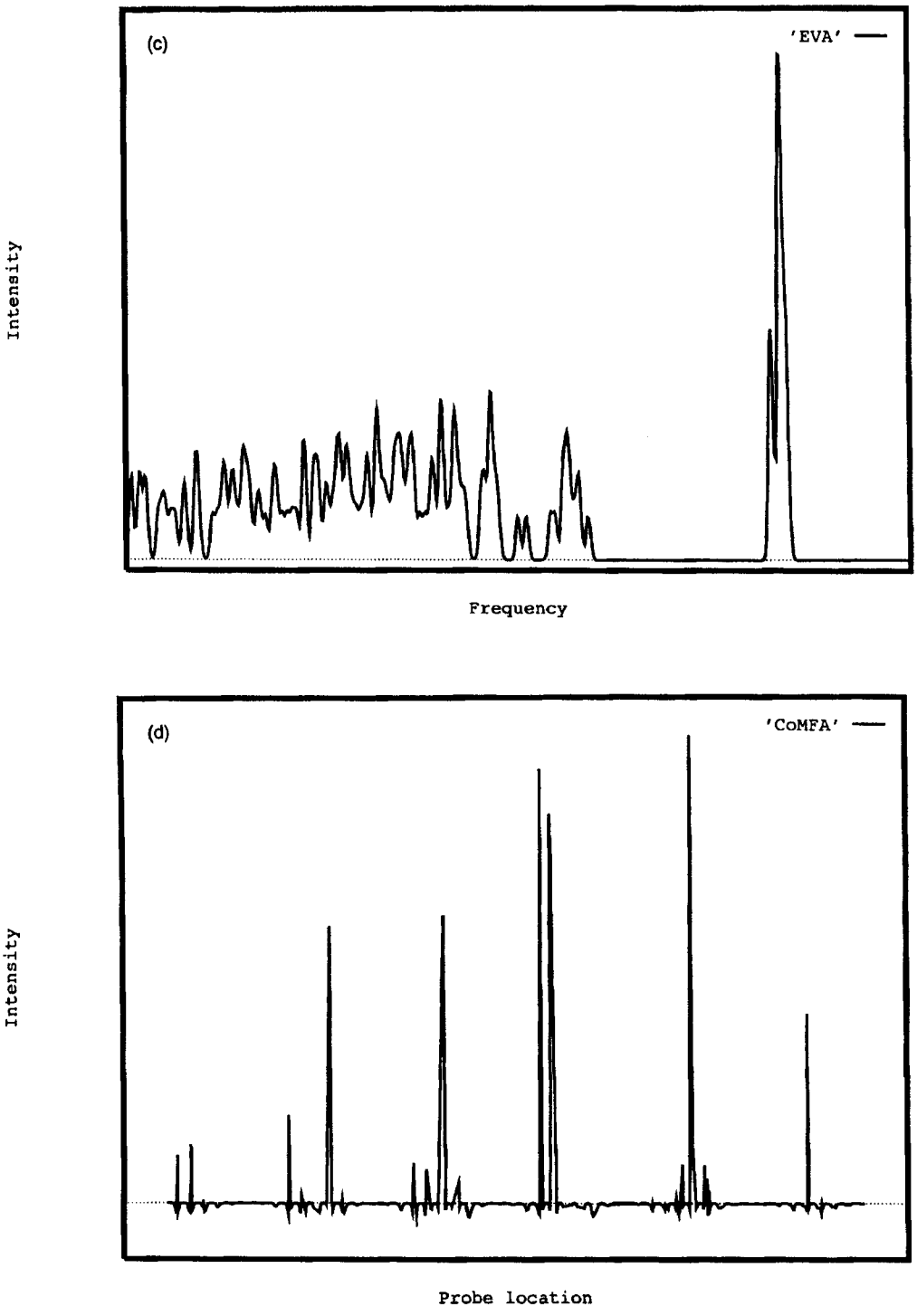


Figure 2. (*continued*)

Table 1. Data sets examined

Reference name for data set	Type of data	Number of variables	Number of observations	Number of active chemicals	Number of inactive chemicals
EVA1	EVA	640	86	46	40
EVA2	EVA	1464	36	23	13
EVA3	EVA	611	148	74	74
EVA4	EVA	514	37	19	18
EVA5	EVA	514	57	19	38
IR1	Infrared	1738	47	25	22
NIR1	Near-infrared	700	35	17	18
NIR2	Near-infrared	700	45	23	22
CoMFA1	CoMFA (steric)	1555	145	73	72

Data pretreatment

It is common practice to attempt some form of data standardization prior to statistical analysis. For example, each explanatory variable might be scaled to have unit variance, on the basis that each variable should be considered as equally informative prior to statistical analysis. In the analyses reported here, no data pretreatment was performed; the data were analysed in the form generated by the relevant molecular modelling or spectroscopic analysis data acquisition systems.

Comparison with other methods

In order to gain a useful appreciation for the predictive performance of CREDIT, we need to compare its predictive performance with that of other methods for two-group discriminant analysis. To achieve this comparison, we have also analysed each data set in Table 1 using two other discriminant procedures. These methods, reported in Reference 17, are a form of linear discriminant analysis motivated by the work of Campbell and Atchley[16] (which we refer to as MCA), and a form of generalized ridge discrimination (called GRD).

The MCA approach is fundamentally very similar to CREDIT, except that it employs the sample within- group covariance matrix rather that the sample total covariance matrix to perform the principal component analysis and to generate the discriminant rule. The relative predictive performance of MCA compared with CREDIT will therefore depend on the amount of discriminatory information in the null space of $\mathbf{W}$. However, randomized permutation tests using MCA are computationally very intensive. The computational efficiency of MCA is therefore expected to be much poorer than that of CREDIT.

The GRD method uses an estimate $\hat{\mathbf{\Sigma}}$ for the true within-group covariance matrix based on an augmented form of the sample within-group covariance matrix $\mathbf{W}$:

$$\hat{\mathbf{\Sigma}} = \frac{1}{\gamma}[\mathbf{G}_1|\mathbf{G}_2]\begin{bmatrix}\mathbf{\Lambda} + a\mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & (\alpha+\beta)\mathbf{I}_{p-q}\end{bmatrix}\begin{bmatrix}\mathbf{G}_1^{\mathrm{T}} \\ \mathbf{G}_2^{\mathrm{T}}\end{bmatrix} \tag{22}$$

where

$$\mathbf{W} = [\mathbf{G}_1|\mathbf{G}_2]\begin{bmatrix}\mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{G}_1^{\mathrm{T}} \\ \mathbf{G}_2^{\mathrm{T}}\end{bmatrix} \tag{23}$$

is the spectral decomposition of $\mathbf{W}$ (see also (3)) and γ is a normalizing constant which is expressed in terms of α, β and the eigenvalues given in the diagonal matrix $\mathbf{\Lambda}$. The values of α and β are chosen to optimize predictive performance. For convenience this was performed via a 41×41 log-scaled grid-searching procedure for $10^{-20} \leq \alpha \leq 10^{20}$ and $10^{-20} \leq \beta \leq 10^{20}$. The GRD method is known to give biased predictions[17] owing to the grid-searching procedure not being cross-validated. We expect to be able to detect this bias using randomized permutation tests.

Predictive performance – first eight data sets

The predictive performance of CREDIT, MCA and GRD, estimated using full leave-one-out cross-validation, for each of the first eight data sets in Table 1 is reported in Table 2. Results for the analysis of the final data set are given later. Table 2 gives classification success rates for each group of observations as well as the overall success rate. The figures in parentheses are p-value estimates for the probability that this value of success rate is obtained by chance; p-values are estimated using 1000 randomized permutations of the response labels.

From the table it can be seen that the predictive performance of CREDIT and MCA is equal for two data sets (EVA5 and NIR1). In three cases (EVA3, EVA4 and IR1) CREDIT performs better than MCA. In the remaining three cases MCA performs better than CREDIT. Overall, therefore, we conclude that there is little to choose between CREDIT and MCA in terms of predictive performance.

The p-values associated with the success rates are generally small. However, there is cause for concern in some instances. For example, the inactive success rate of 0·692 for data set EVA2 is identical for each of the three discriminant methods. However, the p-value associated with GRD is much larger than those associated with CREDIT and MCA. For GRD, 104 of the 1000 randomized permutations yielded success rates greater than or equal to those observed for

Table 2. Results of analyses of first eight data sets

	CREDIT			MCA			GRD		
Data set	Active success rate	Inactive success rate	Overall success rate	Active success rate	Inactive success rate	Overall success rate	Active success rate	Inactive success rate	Overall success rate
EVA1	0·717	0·725	0·721	0·826	0·650	0·744	0·870	0·850	0·860
	(0·015)	(0·002)	(0·001)	(0·00)[a]	(0·03)[a]	(0·00)[a]	(0·00)[b]	(0·00)[b]	(0·00)[b]
EVA2	0·826	0·692	0·778	0·913	0·692	0·833	0·913	0·692	0·833
	(0·056)	(0·012)	(0·010)	(0·001)	(0·025)	(0·000)	(0·015)	(0·104)	(0·007)
EVA3	0·757	0·635	0·696	0·635	0·581	0·608	0·784	0·784	0·784
	(0·000)	(0·012)	(0·000)	(0·01)[c]	(0·09)[c]	(0·01)[c]	(0·0)[d]	(0·0)[d]	(0·0)[d]
EVA4	0·737	0·667	0·703	0·684	0·667	0·676	0·842	0·778	0·811
	(0·024)	(0·081)	(0·018)	(0·083)	(0·089)	(0·050)	(0·016)	(0·065)	(0·011)
EVA5	0·737	0·737	0·737	0·737	0·737	0·737	0·737	0·789	0·772
	(0·005)	(0·092)	(0·008)	(0·004)	(0·077)	(0·006)	(0·045)	(0·075)	(0·019)
IR1	0·880	0·864	0·872	0·880	0·818	0·851	0·960	0·864	0·915
	(0·000)	(0·000)	(0·000)	(0·000)	(0·001)	(0·000)	(0·000)	(0·001)	(0·000)
NIR1	0·824	1·000	0·914	0·824	1·000	0·914	0·882	1·000	0·943
	(0·002)	(0·000)	(0·000)	(0·008)	(0·000)	(0·000)	(0·008)	(0·000)	(0·000)
NIR2	0·739	0·773	0·756	0·913	0·773	0·844	0·826	0·864	0·844
	(0·025)	(0·015)	(0·005)	(0·001)	(0·011)	(0·001)	(0·022)	(0·009)	(0·000)

[a–d] Estimated p-values using [a]500, [b]250, [c]100 and [d]50 randomized permutations.

the true response data. This could be viewed as further evidence of bias in the GRD success rates. For all three methods the p-values for the EVA4 and EVA5 inactive success rates are high.

Bias assessment

For each application the overall predictive performance of GRD is at least as good as the better of CREDIT and MCA. However, this is not surprising in view of the fact that the predictive performance of GRD estimated using cross-validation is known to yield biased results owing to the tuning performed. Evidence for this bias for data set IR1 is given in Figure 3. As the figure shows, the mean active, inactive and overall success rates for randomized responses is approximately 0·60 compared with approximately 0·50 for CREDIT and MCA; the bias is therefore about 0·10 for randomized responses. For set EVA1 the corresponding distributions are given in Figure 4. The bias of GRD success rates is again apparent. Note also that the EVA1 distributions are noticeably narrower than the corresponding distributions for IR1.

We have estimated the bias incurred in cross-validatory assessment of predictive performance using each of CREDIT, MCA and GRD for a number of data sets. Results are given here in Table 3, in terms of mean values from randomized permutation tests, for each of the active, inactive and overall success rates.

If a randomly selected training set from a population is used, then CREDIT and MCA provide unbiased predictive success rates for allocating other observations to groups. GRD provides biased predictive success rates owing to the tuning performed. Krzanowski *et al.*[17] have expressed this bias in terms of the difference between internal ('leave-one-out') success rates and external success rates (obtained by applying the discriminant rule to other observations). They suggest that for data sets having approximately 40 observations, GRD success rates should be reduced by between 0·05 and 0·10 to account for bias. This bias could be removed by using a two-deep cross-validation approach to select α and β in GRD, but this would be computationally very demanding. After allowing for this bias, there is little to choose between CREDIT, MCA and GRD.

Table 3 shows, when group sizes are similar, that CREDIT and MCA have mean permutation active and inactive success rates of approximately 0·5. However, when group sizes are markedly different (for example, EVA2 and EVA5 – see Table 1), then the mean success rate for the larger group is substantially greater than that for the smaller group. This result is initially surprising, perhaps, but is due to the fact that the sample larger group mean provides a better estimate for the true population mean than the sample smaller group mean.

In order to quantify the expected difference in success rates for individual groups, we have performed a simulation study using multivariate normal data with mean zero and identity covariance matrix. We randomly select two groups of observations of sizes n_1 and n_2 and then estimate the expected success rate for the larger group for prediction of a further test observation from the same distribution. Results are given in Figure 5, for the case $n_1 = 2n_2$, as a function of the total number of observations and the dimensionality of the data. The results in Figure 5 are thus approximately comparable with the results in Table 3 for data sets EVA2 and EVA5. For these data sets the effective dimensionality of the data is around 30–60, so that the simulation suggests an expected active success rate of about 0·62 to be compared with the values of 0·637 and 0·592 (for CREDIT and MCA analyses respectively of EVA2) and 0·612 and 0·605 (for CREDIT and MCA analyses respectively of EVA5) from Table 3. The values of Overall success rates from randomized permutation testing for EVA2 and EVA5 using CREDIT and MCA are higher than 0·5 as a result of this difference in group sizes; the value of

Table 3. Mean success rates from randomized permutation tests

	CREDIT			MCA			GRD		
Data set	Active success rate	Inactive success rate	Overall success rate	Active success rate	Inactive success rate	Overall success rate	Active success rate	Inactive success rate	Overall success rate
EVA1	0·521	0·471	0·498	0·517	0·477	0·499	0·585	0·554	0·571
EVA2	0·637	0·339	0·529	0·592	0·399	0·522	0·694	0·496	0·623
EVA3	0·498	0·498	0·498	0·495	0·497	0·496	0·564	0·563	0·564
EVA4	0·500	0·478	0·489	0·500	0·480	0·490	0·593	0·595	0·594
EVA5	0·379	0·612	0·534	0·379	0·605	0·530	0·520	0·670	0·620
IR1	0·511	0·460	0·487	0·513	0·468	0·492	0·591	0·569	0·581
NIR1	0·484	0·505	0·495	0·486	0·497	0·492	0·613	0·609	0·611
NIR2	0·501	0·487	0·494	0·505	0·491	0·498	0·591	0·603	0·597

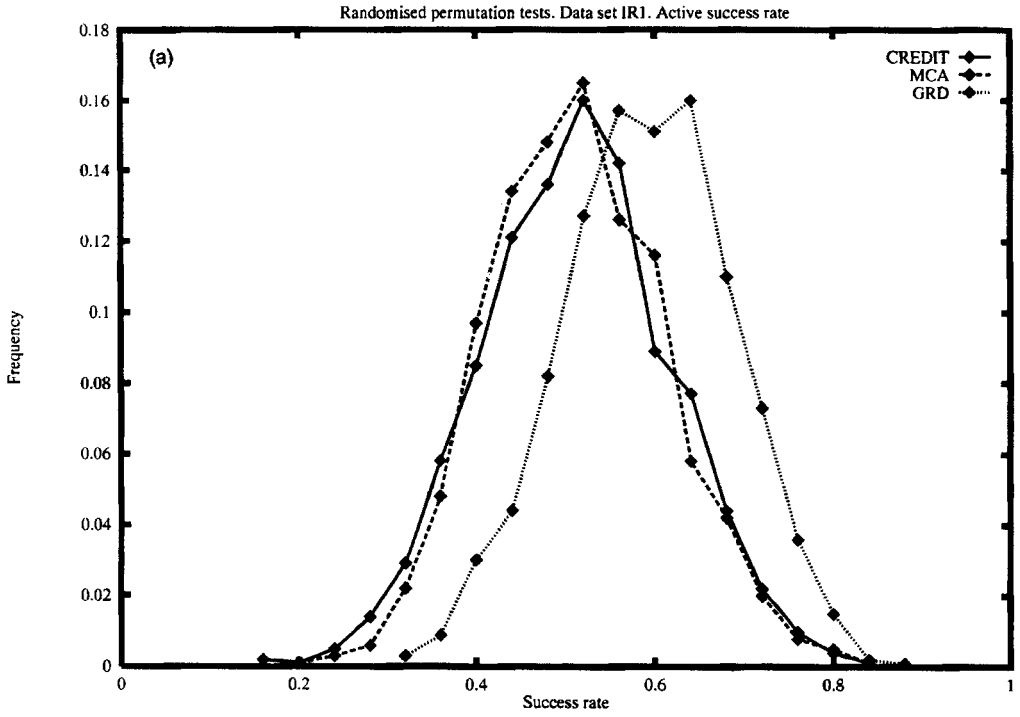
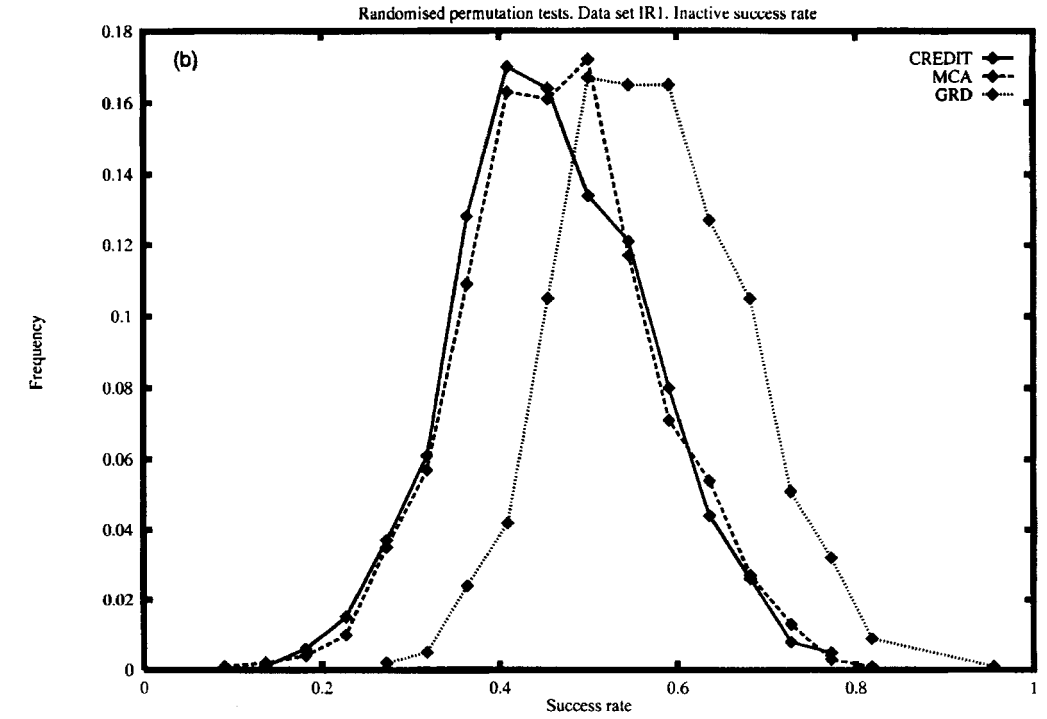


Figure 3. Randomized permutation distributions for data set IR1 for each of CREDIT, MCA and GRD methods: (a) active success rate; (b) inactive success rate; (c) overall success rate. Distributions are discrete and symbols indicate the values of each distribution. Lines are drawn between points as an aid to the eye



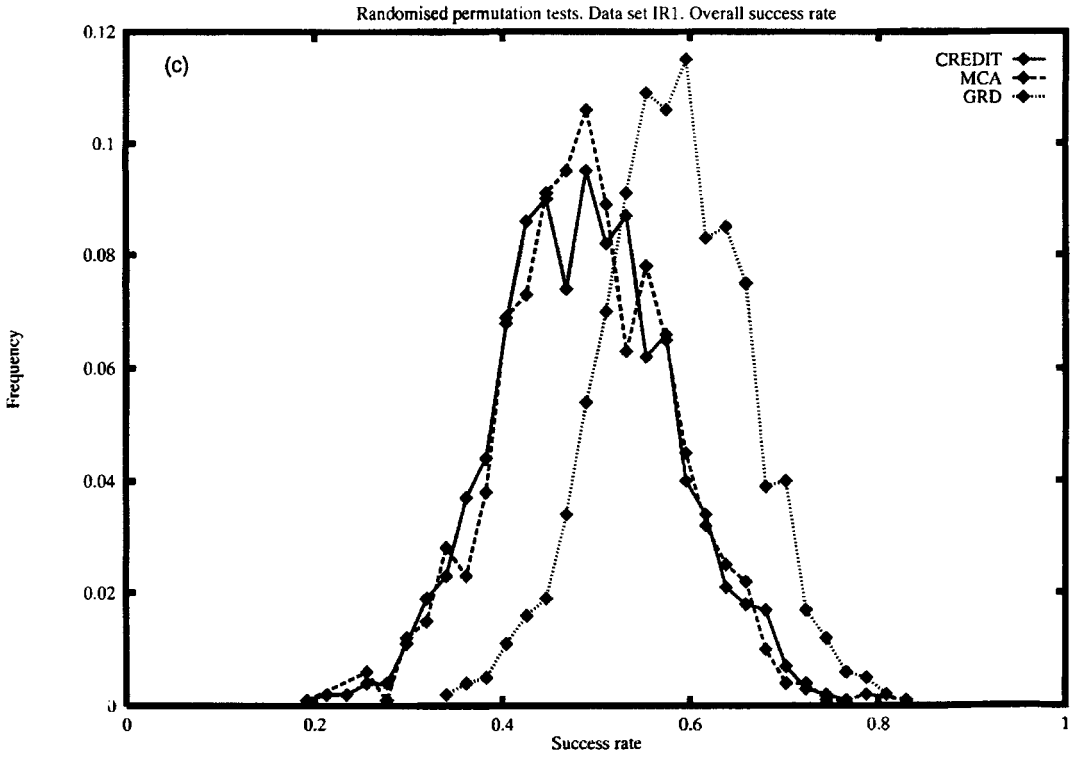


Figure 3. (*continued*)

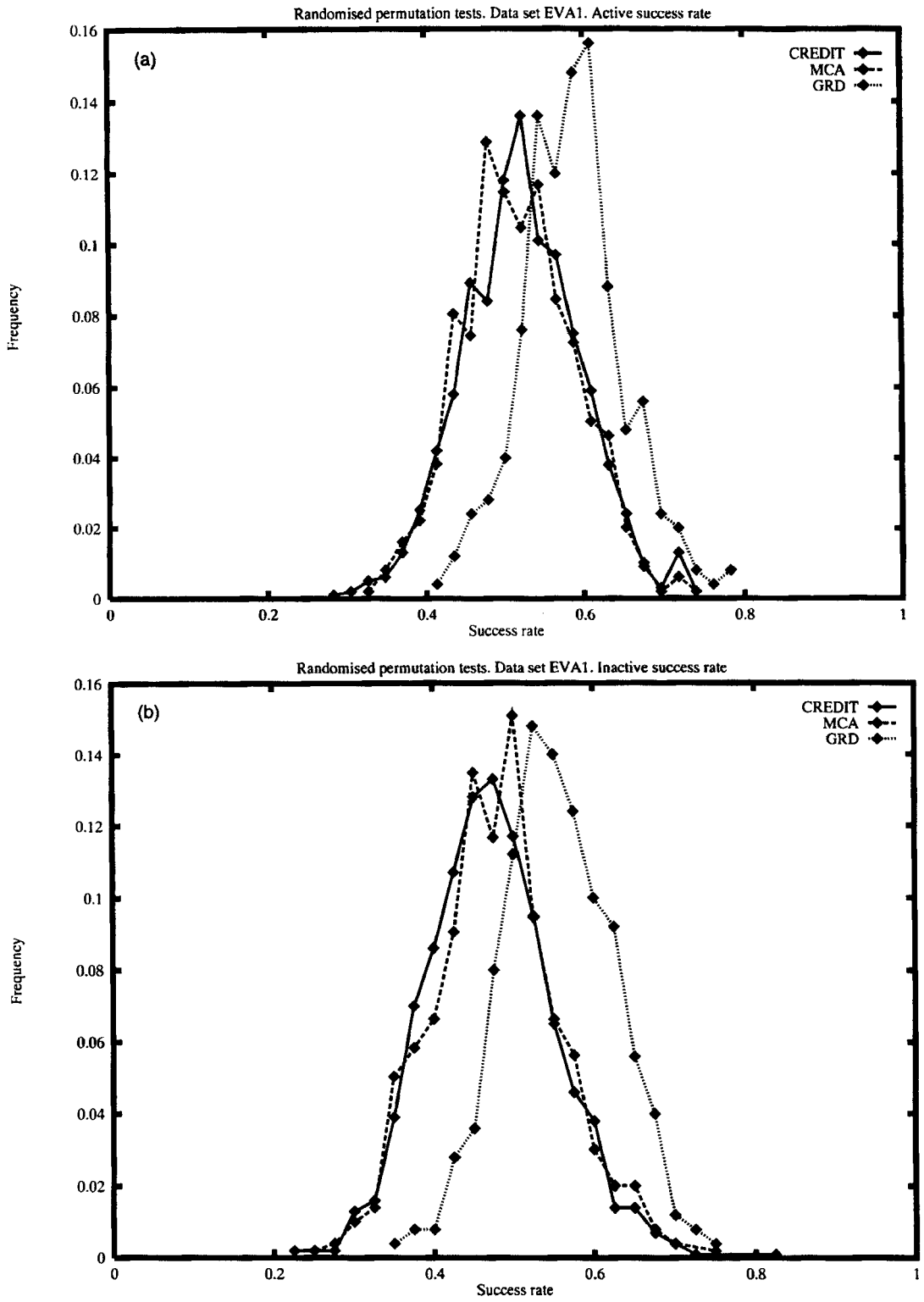


Figure 4. Randomized permutation distributions for data set EVA1 for each of CREDIT, MCA and GRD methods: (a) active success rate; (b) inactive success rate; (c) overall success rate

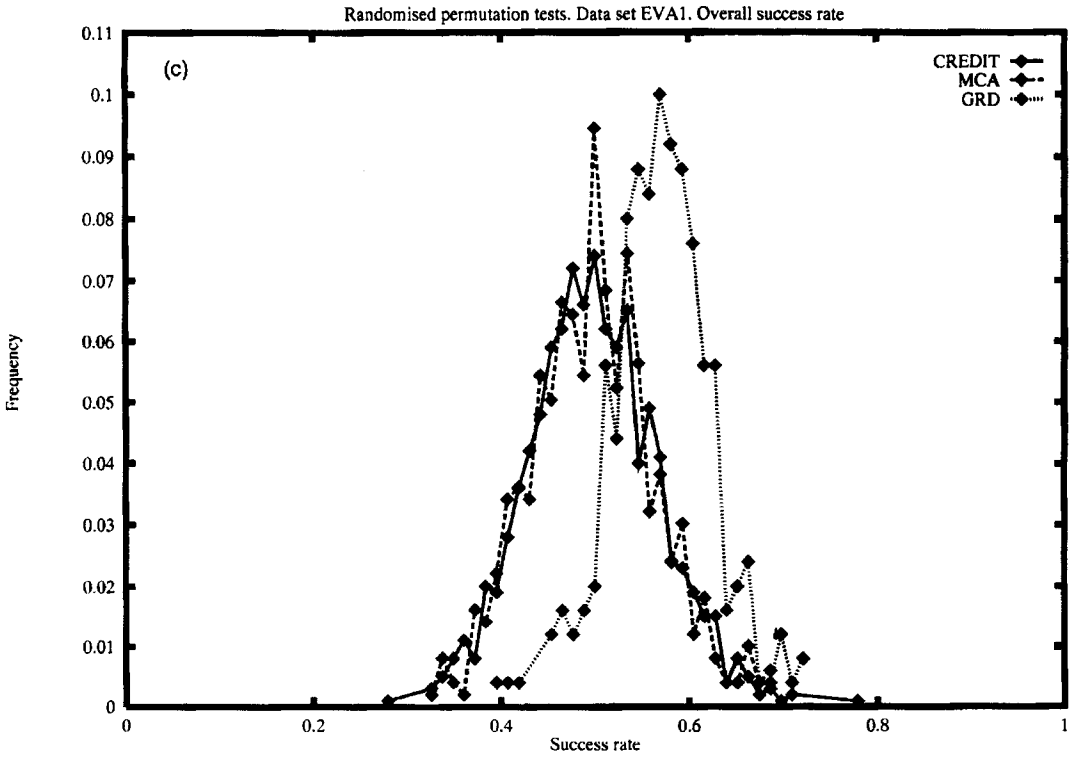


Figure 4. (*continued*)

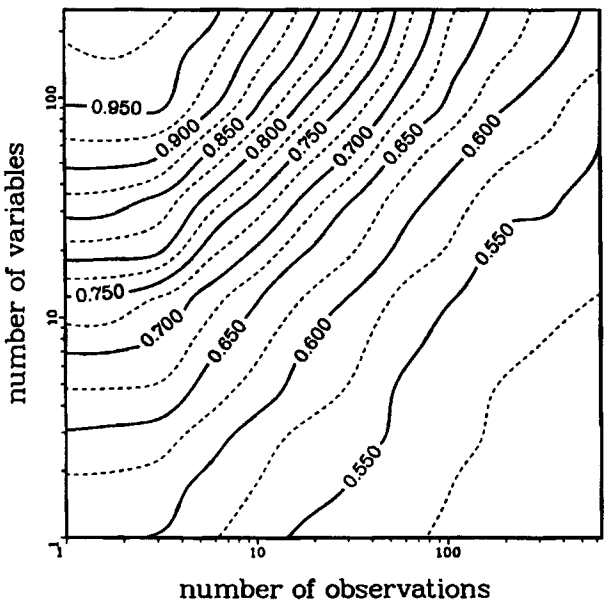


Figure 5. Expected success rates for larger group from randomized permutation testing of multivariate normal $N(\mathbf{0}, \mathbf{I})$ data, assuming one sample group to contain twice as many observations as the other

approximately 0·53 obtained is entirely consistent with our simulation study. However, the overall success rate of approximately 0·62 for GRD analysis of EVA2 and EVA5 is biased owing to tuning.

The effect of eigenvalue adjustment

We have explored the effect of changing the size of the eigenvalue adjustment (18) on the predictive performance of CREDIT discriminant rules, as assessed using cross-validation. Expressing the value of the adjustment as a percentage of the mean eigenvalue, overall success rates were calculated for data sets IR1 and NIR1 using various values of the adjustment. The default method for selection of eigenvectors/principal components to retain based on discriminatory importance (19), (20) was used. Results of our analysis are reported in Table 4. A value of zero for the adjustment is equivalent to adopting no adjustment at all, whereas a very large value has the effect of giving each eigenvector equal adjusted variance (18).

For data set IR1, eigenvalue adjustment has very little effect. For set NIR1, however, predictive performance is rather more sensitive to the precise choice of eigenvalue adjustment. The default value of 1% employed gives reasonable performance in both cases, but the optimal choice of adjustment is clearly application dependent, and cannot be specified *a priori*.

The effect of eigenvalue selection

An essential step in CREDIT is the selection of eigenvectors/principal components of **T** to use in forming the discriminant rule. The default method proposed is based on discriminatory importance (19); this approach is somewhat arbitrary and certainly not optimal, but it does yield useful discriminant models. Nevertheless, more refined methods might produce even better results. Here we explore the effect of various eigenvector selection methods.

One interesting option is to retain *all* the principal components of **T**. It is clear that a substantial improvement in computational efficiency can now be achieved since cross-validation can be performed using matrix up- and downdating operations. We have developed such a discriminant algorithm. The computational efficiency of the resulting SAS/IML code is impressive (for example, for the analysis of set NIR1 using 1000 randomizations, a 60-fold decrease in run time was achieved compared with the corresponding CREDIT analysis);

Table 4. Effect of eigenvalue adjustments

	IR1			NIR1		
Adjustment (%)	Active success rate	Inactive success rate	Overall success rate	Active success rate	Inactive success rate	Overall success rate
0·0	0·920	0·864	0·894	0·588	0·722	0·657
0·1	0·920	0·864	0·894	0·882	0·944	0·914
0·5	0·920	0·864	0·894	0·824	1·000	0·914
1·0	0·880	0·864	0·872	0·824	1·000	0·914
5·0	0·920	0·864	0·894	0·824	0·944	0·886
10	0·920	0·864	0·894	0·824	0·889	0·857
50	0·960	0·864	0·915	0·706	0·889	0·800
100	0·920	0·864	0·894	0·647	0·889	0·771

however, its predictive performance is poorer (overall success rate 0·657 compared with 0·914 for CREDIT). In some applications, however (for example, the analysis of set IR1), the *all-component* method performed as well as CREDIT. Results for analyses of another two data sets are given in Table 5. Since the expected predictive performance of the method is poorer than that of CREDIT, it cannot be recommended for future application.

We have also explored the effect of selecting eigenvectors based solely on variance explained. That is, assuming that eigenvectors are ordered in terms of decreasing variance, we retain a sufficient number to explain 95% of the variation present in the data. Results of the analysis are also given in Table 5.

The table indicates that the selection of a subset of eigenvectors is generally beneficial compared with retaining all components. However, the better method to select eigenvectors is application-dependent. In our experience, selection of eigenvectors based on discriminatory importance gives predictive performance over a range of data sets which is at least as good as selection based on variance explained.

Computational efficiency

The computational efficiency of CREDIT, MCA and GRD is compared in Table 6 for applications in which both CREDIT and MCA analyses were performed using 1000 randomizations. It is clear from the table that CREDIT is more efficient than MCA – by design! GRD is computationally more intensive than both CREDIT and MCA.

Analysis of CoMFA data

Analysis of data set CoMFA1 is particularly illustrative since it provides an opportunity to use CREDIT for a relatively large data set (in terms of both observations and variables). Results of the analysis are given in Table 7.

Again comparisons with MCA and GRD are provided. The predictive performance of CREDIT once more compares well with both MCA and GRD. Note that p-values are only quoted for CREDIT; the reason for this is evident from Table 8, which gives the computational efficiency of the CREDIT and MCA methods.

It is obvious from the table that the computational effort required to perform randomized permutations using MCA (let alone GRD!) is prohibitive.

The CPU timings given in Tables 6 and 8 are of course machine-dependent. In particular, results in Table 8 appear to suggest that MCA is two orders of magnitude less efficient than CREDIT (compared with about one order of magnitude from the results in Table 6). However, a different computer hardware configuration causing considerable processing power to be expended for memory management (paging/swapping) was used for the analyses in Table 8. For this reason the relative timings in Table 8 are not directly comparable with those in Table 6.

Table 5. Overall success rates for various eigenvector selection procedures

Data set	Discriminatory importance	Variance explained	All components
IR1	0·872	0·787	0·894
NIR1	0·914	0·629	0·657
NIR2	0·756	0·800	0·577
EVA4	0·703	0·838	0·703

Table 6. Computational efficiency of analyses

	CREDIT		MCA		GRD	
Data set	Number of randomizations	Run time (CPU Minutes)	Number of randomizations	Run time (CPU Minutes)	Number of randomizations	Run time (CPU Minutes)
EVA2	1000	39	1000	295	100	216
EVA4	1000	40	1000	256	100	223
EVA5	1000	135	1000	1187	25	123
IR1	1000	80	1000	719	50	170
NIR1	1000	28	1000	216	100	206
NIR2	1000	73	1000	524	50	155

Table 7. Analysis of data set CoMFA1

CREDIT			MCA			GRD		
Active success rate	Inactive success rate	Overall success rate	Active success rate	Inactive success rate	Overall success rate	Active success rate	Inactive success rate	Overall success rate
0·808 (0·000)	0·833 (0·000)	0·821 (0·000)	0·767	0·778	0·772	0·822	0·889	0·855

Table 8. Computational efficiency for analysis of CoMFA1

CREDIT		MCA	
Number of randomizations	Run time (CPU minutes)	Number of randomizations	Run time (CPU minutes)
1000	1888	10	5817

The combination of IBM mainframe running VM/CMS operating system and SAS/IML is not ideal for analysis of large multivariate data. For better processing performance a combination such as an RISC workstation running UNIX with FORTRAN or C could be used. Here we are primarily concerned with the development of the approach. If analysis of large data sets such as CoMFA1 was intended on a regular basis, other computational techniques such as generalized cross-validation[25] should be considered.

DISCUSSION

Analysis of multivariate data, for which the number of variables greatly exceeds the number of observations, is a precarious task. Great care must be taken while analysing such data. Realistic assessment of the predictive performance of such a model is difficult to achieve. In this paper we use cross-validation in an attempt to compare the predictive performance of various discriminant methods.

The CREDIT results reported above demonstrate that CREDIT provides useful discriminant models. In combination with good predictive performance, the ease with which randomized permutation distributions can be generated provides worthwhile supporting information at little computational cost. For the current analyses the randomized permutation test helps us to demonstrate that the linear relationship between the explanatory variables and the true responses is usually much stronger than that between the explanatory variables and a random permutation of the responses. We find that randomized permutation testing aids our understanding of applications of singular discriminant analysis; the approach has much to offer and its statistical properties deserve more thorough investigation in future.

Once the discriminant model has been established for a training set of observations, the model may then be used to make predictions of response for observations/chemicals of a similar type to those in the training data. A guiding principle in this prediction step is to avoid making predictions for a new observation that lies 'outside' the boundaries of the training set. Classification using the discriminant model should be restricted to chemicals which are *concordant* with the training set. In practice, since the number of observations is much smaller than the number of variables, it is especially difficult to define concordance. The observations occupy a subspace of low dimensionality relative to the number of variables. However, it is essential that some effort be made to test the concordance of future observations for classification.

A crucial step in this process is the specification of the subspace occupied by the data. We typically define the subspace as that spanned by the first r eigenvectors of the sample total covariance matrix. Of course, the choice of r is difficult also; we usually ensure that the subspace explains at least 95% of the total variation present. Once the subspace is specified, we typically use an approach related to Hotelling's T^2 (see Reference 21) and SIMCA[10] in which we estimate

(a) the Mahalanobis distance of the test observation from the centre of the training data in the subspace defined by the training data
(b) the Euclidean distance to the test observation, measured perpendicularly to the subspace defined by the training data.

Cross-validation is used to generate distributions of distances (a) and (b) for the training set. A test observation which yields values of (a) and (b) lying outside the relevant distributions is considered to be discordant with the training set. It is also advisable at this stage to investigate the effect of varying the 'data subspace' dimensionality r. Concordance can also be measured using a non-parametric approach such as nearest-neighbour analysis. In our experience we find that the methods described here to estimate concordance are generally satisfactory but require further development.

ACKNOWLEDGEMENTS

We would like to thank Professor Mervyn Stone (University College London), Professor Wojtek Krzanowski (University of Exeter) and Professor Phil Brown (University of Kent), for valuable discussions. We take pleasure in acknowledging the encouragement and support of numerous former colleagues at Shell Research in Sittingbourne, Dr Laurie Phillips and Dr Mervyn Thomas in particular. We further acknowledge useful comments from three referees!

APPENDIX: THE CREDIT ALGORITHM

An outline of the algorithm is given below, based on the derivation given in the text above.

1. Take the training data $\mathbf{X}$ and the *class allocation vector* $\boldsymbol{a}$, where $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^{\mathrm{T}}$, $a_i = 1$ if observation i belongs to population π_1 and $a_i = 2$ if observation i belongs to population π_2.
2. Generate $R - 1$ random permutations of $\boldsymbol{a}$. Call these $\boldsymbol{a}_2, \boldsymbol{a}_3, \ldots, \boldsymbol{a}_R$. Call $\boldsymbol{a}_1 = \boldsymbol{a}$. Form the matrix $\mathbf{A} = (\boldsymbol{a}_1\ \boldsymbol{a}_2, \ldots, \boldsymbol{a}_R)$.
3. *Start cross-validation loop over* α, $\alpha = 1, \ldots, n$.
4. Reorder $\mathbf{X}$ and $\mathbf{A}$ so that row α becomes the first row.
5. Form

$$\mathbf{X}\mathbf{X}^{\mathrm{T}} = \begin{bmatrix} w & \mathbf{u}^{\mathrm{T}} \\ \mathbf{u} & \mathbf{V} \end{bmatrix}$$

6. Omit first row from $\mathbf{X}$ and $\mathbf{A}$.
7. Form $\mathbf{M} = (n-2)^{-1}\mathbf{H}_{n-1}\mathbf{V}\mathbf{H}_{n-1}$.
8. Perform spectral decomposition of $\mathbf{M}$, retaining only the first q components, $\mathbf{M} = \mathbf{E}\mathbf{L}\mathbf{E}^{\mathrm{T}}$ corresponding to non-null eigenvalues.
9. Form the scaled eigenvectors $\mathbf{Z} = (n-2)^{-1/2}\mathbf{E}\mathbf{L}^{-1/2}$.
10. Adjust the eigenvalues λ_i, $i = 1, 2, \ldots, q$, using

$$\lambda_i^* = \lambda_i + (100q)^{-1}\sum_{i=1}^{q}\lambda_i$$

to give $\mathbf{L}^*$.

11. Extract $\boldsymbol{u}$ from $\mathbf{X}\mathbf{X}^{\mathrm{T}}$.
12. *Start randomized permutation loop over* β, $\beta = 1, \ldots, R$.
13. Calculate $\boldsymbol{\psi}_\beta^{\mathrm{T}} = \mathbf{h}_\beta^{\mathrm{T}}\mathbf{V}\mathbf{H}_{n-1}\mathbf{Z}$, where $\boldsymbol{h}_\beta$ is an $(n-1)\times 1$ column vector with elements $h_{\beta i}$, $h_{\beta i} = n_{1[-\alpha]}^{-1}$ if $a_{\beta i} = 1$ and $h_{\beta i} = -n_{2[-\alpha]}^{-1}$ if $a_{\beta i} = 2$. The subscript $[-\alpha]$ indicates that the values of n_1 and n_2 should be adjusted (when necessary) for the omitted observation α.
14. Calculate the discriminatory importance $\delta_{\beta i} = \lambda_i^{*-1}\psi_{\beta i}^2$, $i = 1, 2, \ldots, q$.
15. Reorder the principal components (for this loop) such that $\delta_{\beta 1} \geq \delta_{\beta 2} \geq \ldots \geq \delta_{\beta q}$.
16. Retain λ_i^* and z_i^*, $i = 1, 2, \ldots, k$, such that

$$\left(\sum_{i=1}^{q}\lambda_i^*\right)^{-1}\sum_{i=1}^{k}\lambda_i^* \approx 0{\cdot}95$$

Alternative criteria might also be used at this stage.

17. Extract $\boldsymbol{\psi}_K$, $\mathbf{Z}_K$ and $\mathbf{L}_K^*$ from $\boldsymbol{\psi}_\beta$, $\mathbf{Z}$ and $\mathbf{L}^*$ corresponding to the retained components.
18. Allocate the omitted observation α to class 1 if

$$\boldsymbol{\Psi}_K^{\mathrm{T}}\mathbf{L}_K^{*-1}\mathbf{Z}_K^{\mathrm{T}}\mathbf{H}_{n-1}[\mathbf{u} - \mathbf{2}^{-1}\mathbf{V}(n_{1[-\alpha]}^{-1}\mathbf{1}_{n_1[-\alpha]}^{\mathrm{T}} \mid n_{2[-\alpha]}^{-1}\mathbf{1}_{n_2[-\alpha]}^{\mathrm{T}})^{\mathrm{T}}] > 0$$

Otherwise allocate it to class 2.

19. Store the success count $s(\alpha, \beta)$; $s(\alpha, \beta) = 1$ if correct allocation with respect to $\boldsymbol{a}_\beta$, otherwise $s(\alpha, \beta) = 0$.
20. *End of randomized permutation loop over* β.
21. Replace first row of $\mathbf{X}$ and $\mathbf{A}$.

22. *End of cross-validation loop over* α.
23. Produce summary statistics.

The algorithm presented here can easily be extended to incorporate prediction of a test observation which is not a member of the training set.

REFERENCES

1. M. Stone and P. Jonathan, *J. Chemometrics*, **7**, 455 (1993).
2. M. Stone and P. Jonathan, *J. Chemometrics*, **8**, 1 (1994).
3. R. D. Cramer, D. E. Patterson and J. D.Bunce, *J. Am. Chem. Soc.* **110**, 5959 (1988).
4. G. Klebe and U. Abraham, *J. Med. Chem.* **36**, 37 (1993).
5. L. B. Bruce and R. B. Nachbar, *J. Comput.-Aided Mol. Design*, **7**, 587 (1993).
6. I. E. Frank and J. H. Friedman, *Technometrics*, **35**, 109 (1993).
7. J. H. Friedman, *Ann. Stat.* **19**, 1 (1991).
8. P. A. Lachenbruch, *Discriminant Analysis*, Hafner, New York (1975).
9. P. A. Lachenbruch and M. Goldstein, *Biometrics*, **35**, 69 (1979).
10. S. Wold, *Pattern Recogn.* **8**, 127 (1976).
11. N. A. Campbell, *Appl. Stat.* **29**, 5 (1980).
12. J. H. Friedman, *J. Am. Stat. Assoc.* **84**, 165 (1989).
13. I. E. Frank, *Chemometrics Intell. Lab. Syst.* **4**, 215 (1988).
14. I. E. Frank and J. H. Friedman, *J. Chemometrics*, **3**, 463 (1989).
15. T. Hastie, R. Tibshirani and A. Buja, *J. Am. Stat. Assoc.* **89**, 1255 (1994).
16. N. A. Campbell and W. R. Atchley, *Syst. Zool.* **30**, 268 (1981).
17. W. J. Krzanowski, P. Jonathan, W. V. McCarthy and M. R. Thomas, *Appl. Stat.* **44**, 101 (1995).
18. M. Stone, *Math. Oper. Stat. Ser. Stat.* **9**, 127 (1978).
19. T. T. Dunne and M. Stone, *J. R. Stat. Soc. B*, **55**, 369 (1993).
20. A. Albert, *Regression and the Moore–Penrose Generalized Inverse*, Academic, New York (1972).
21. K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic, New York (1988).
22. G. H. Golub and C. F. van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD (1989).
23. R. A. Fisher, *Ann. Eugen.* **7**, 179 (1936).
24. R. Sundberg and P. J. Brown, *Technometrics*, **31**, 365 (1989).
25. R. A. Thisted, *Elements of Statistical Computing*, Chapman and Hall, New York (1988).