WILEY

Discriminant Analysis with Singular Covariance Matrices: Methods and Applications to Spectroscopic Data

Author(s): W. J. Krzanowski, P. Jonathan, W. V. McCarthy and M. R. Thomas

Source: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1995, Vol. 44, No. 1 (1995), pp. 101-115

Published by: Wiley for the Royal Statistical Society

Stable URL: https://www.jstor.org/stable/2986198

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Royal Statistical Society and *Wiley* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the Royal Statistical Society. Series C (Applied Statistics)*

This content downloaded from 148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

Appl. Statist. (1995)
44, *No.* 1, *pp.* 101–115

General Interest Section

Discriminant Analysis with Singular Covariance Matrices: Methods and Applications to Spectroscopic Data

By W. J. KRZANOWSKI†

University of Exeter, UK

P. JONATHAN and W. V. McCARTHY

Shell Research Limited, Sittingbourne, UK

and M. R. THOMAS

CSIRO Division of Tropical Crops and Pastures, Brisbane, Australia

[Received November 1992. Final revision March 1994]

SUMMARY

Currently popular techniques such as experimental spectroscopy and computer-aided molecular modelling lead to data having very many variables observed on each of relatively few individuals. A common objective is discrimination between two or more groups, but the direct application of standard discriminant methodology fails because of singularity of covariance matrices. The problem has been circumvented in the past by prior selection of a few transformed variables, using either principal component analysis or partial least squares. Although such selection ensures non-singularity of matrices, the decision process is arbitrary and valuable information on group structure may be lost. We therefore consider some ways of estimating linear discriminant functions without such prior selection. Several spectroscopic data sets are analysed with each method, and questions of bias of assessment procedures are investigated. All proposed methods seem worthy of consideration in practice.

Keywords: Antedependence modelling; Canonical variates; Cross-validation; Eigenvalues and eigenvectors; Error rate estimation; Partial least squares; Ridge techniques

1. Introduction

Chemical samples are frequently examined by using a test method to assign each sample to one of two or more distinct groups. Sometimes these test methods are expensive or involve a long time delay before the group assignment is determined. However, cheap and readily available alternative information can often be obtained on the samples. With the aid of a prediction rule, such information can be used to predict the group memberships that would have been assigned if the test method had been used.

†*Address for correspondence*: Department of Mathematical Statistics and Operational Research, University of Exeter, Laver Building, North Park Road, Exeter, EX4 4QE, UK.
E-mail: W.J.Krzanowski@uk.ac.exeter

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

A common characteristic of many useful alternative sources of information is that they involve a large number of variables. For example, experimentally measured spectra can be used in this way to identify the sources of complex material such as bitumen. Also, calculated grids of electrostatic potential around molecular structures are similarly used to derive relationships between these structures and their chemical properties. Various groups can be defined corresponding to different multivariate regions of chemical properties.

Discriminant analysis can be performed when the test method assignment is known for each sample in a training set, along with the variables that measure the alternative information. The analysis provides a small number of (linear) combinations of these variables that maximize in some sense the group information in the samples. A classification rule can be developed in terms of the new variables so that the application of this rule to the alternative information on future samples provides a prediction of group assignment. However, application of discriminant analysis by standard packages can run into severe problems when the number of variables exceeds the number of samples. At best, generalized inverses are used in place of ordinary inverses and a warning message is printed, whereas at worst the package simply refuses to provide any analysis at all.

Recent chemometric studies at Shell Research Limited, Sittingbourne, have involved the use of various sets of spectroscopic data. The most common examples are those arising from infra-red (IR) reflectance analysis, in which the reflected energy from the substance is recorded continuously over a range of wavelengths. The continuous trace is preprocessed by discretizing the wavelength scale and taking second differences of log(1/reflectance) at the resulting wavelength values (although differencing is not advocated by some practitioners). Near infra-red (NIR) spectra yield $p = 700$ and IR spectra yield $p = 1738$ correlated variables in this way, different portions of the wavelength range being considered in the two techniques. In most applications at Sittingbourne the number of samples analysed has been less than 50, and the two-group situation has been the most prevalent. Table 1 gives brief details of three such data sets that have been presented recently for analysis, showing for each data set the spectroscopic technique (NIR or IR), the group sizes, the number of variables and the average over these variables of the absolute value of the univariate two-sample t-statistic for testing differences between the group means. This last statistic has values lying between 1 and 2 for all three data sets, reflecting the fact that the two groups are typically distinguishable for at least a subset of wavelengths. For illustration, a plot of data set NIR1 is shown in Fig. 1;

TABLE 1
Summary of data sets examined

Data set	*No. of observations*		*No. of variables*	$\bar{t}$
	Group 1	*Group 2*		
NIR1	18	17	700	1.09
NIR2	22	23	700	1.51
IR1	22	25	1738	1.39

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

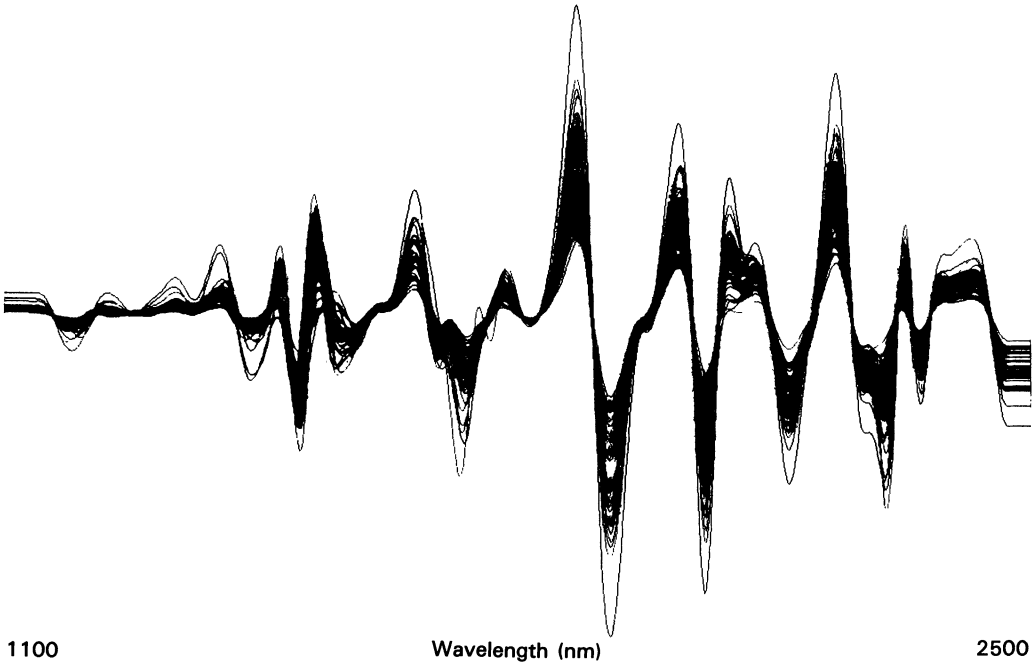

Fig. 1. Plot of the NIR1 data: the horizontal axis is wavelength; the vertical axis is the second-differenced log(1/reflectance) value

areas of low variability of the curves are thought to correspond to wavelengths where the two groups overlap completely, whereas areas of high variability are thought to indicate wavelengths at which the groups are reasonably distinguishable. Industrial confidentiality prevents a further description, either of these data sets or of the problems that generated them; we can only mention that the study of crop enhancement agents, the identification of new varieties and the monitoring of uniformity of production processes are the broad application areas producing data of this type.

In the rest of this paper we identify the problems that arise with standard analyses, outline some alternative approaches, investigate their performance on the above data sets and discuss issues arising in the assessment of these performances.

2. Background Theory

Let n be the number of units in the training set, p the number of alternative information variables and g the number of groups. Denote by $\mathbf{X} = (X_1, \ldots, X_p)^T$ the vector of alternative information variables, by $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ the vector of values on these variables for the jth training set unit in the ith group and suppose that there are n_i units in the ith group. Typical assumptions underlying discriminant analysis are that the training units form random samples from g populations which have different mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g$ but the same dispersion matrix $\boldsymbol{\Sigma}$. Different approaches to the derivation of a classification rule are available, some of which require the additional assumption of normality of data to be made (see,

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

for example, Krzanowski (1988), chapters 12 and 13). If population parameters are known, many of these approaches lead to a classification rule based on the quantities

$$(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \qquad (i = 1, \ldots, g) \tag{1}$$

where $\mathbf{x}$ is the vector of values for the sample to be classified (Mardia *et al.*, 1979). In the special case of two-group discrimination ($g = 2$), the classification rule reduces to a rule based on Fisher's linear discriminant function (Krzanowski (1988), pages 356–358):

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x}. \tag{2}$$

In applications, the population parameters are unknown. Then, although alternative ways of deriving classification rules have been proposed (Krzanowski (1988), p. 338), by far the most common procedure in practice is to replace the unknown parameters in expressions (1) and (2) by their estimates $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}$ from the training data. Thus classification is performed by calculating either

$$(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_i) \qquad (i = 1, \ldots, g) \tag{3}$$

or

$$(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{x} \qquad (\text{if } g = 2) \tag{4}$$

where

$$\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$$

and

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S} = \frac{1}{n - g} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^{\mathrm{T}}.$$

For application of any of these classification rules, both $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ must be non-singular. Although the former is generally so by assumption (but see Mardia *et al.* (1979), p. 304), the latter requirement will fail in most agrochemical applications. This is because we require $n - g \geqslant p$ for non-singularity of $\hat{\boldsymbol{\Sigma}}$, but in such applications n is typically between 30 and 200 whereas p lies between 200 and 4000.

Chemometricians have faced the problem of singular $\hat{\boldsymbol{\Sigma}}$ for many years and have generally circumvented it by a preliminary transformation from $\mathbf{X}$ to $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\mathrm{T}}$ where $m \leqslant n - g$, followed by standard application of expressions (3) or (4) to $\mathbf{Y}$. The most common approaches to selection of $\mathbf{Y}$ have been by choosing the first m components in either a preliminary principal component (PPC) analysis or a partial least squares (PLS) analysis of the data. Some relevant references are Gunst and Mason (1979), Naes and Martens (1985), Manne (1987), Geladi (1988), Hoskuldsson (1988), Yendle and MacFie (1989) and Stone and Brooks (1990). Although the results of such an approach have often been successful in broad terms, the conceptual problems are that an arbitrary decision must be made regarding how many components to retain and how many to discard, and that valuable between-group information may be lost in the later components that are discarded (see Chang (1983), Jolliffe (1986), chapter 9.1, and Krzanowski (1992)). We now consider alternative approaches to a solution of the problem.

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

3. Methodology

3.1. *Augmenting the Within-groups Covariance Matrix*

Since **S** is necessarily singular when $n - g < p$, we have rank(**S**) = $r < p$. One possible approach is to augment **S** in such a way that it retains its major characteristics but it becomes non-singular and so can be used in expressions (3) and (4). In achieving this aim, we clearly want to minimize the perturbation to preserve as much as possible of the original sample information. If principal component analysis is viewed as providing the best r-dimensional approximation to a p-dimensional set of data, then our present objective can be seen as exactly the reverse, namely to provide the 'nearest' p-dimensional non-singular approximation to an r-dimensional singular set of data.

We therefore consider the spectral decomposition of **S**:

$$\mathbf{S} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathrm{T}} \tag{5}$$

where **D** is the diagonal $p \times p$ matrix of ranked eigenvalues $d_1 \geqslant d_2 \geqslant \ldots \geqslant d_r > d_{r+1} = \ldots = d_p = 0$ of **S**, and **L** is the orthonormal $p \times p$ matrix whose columns $\mathbf{l}_1, \ldots, \mathbf{l}_p$ are the corresponding eigenvectors of **S**.

We can therefore write

$$\mathbf{S} = (\mathbf{L}_1 \quad \mathbf{L}_2)\begin{pmatrix}\mathbf{D}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{L}_1^{\mathrm{T}}\\ \mathbf{L}_2^{\mathrm{T}}\end{pmatrix} \tag{6}$$

where $\mathbf{L}_1$ contains the first r columns of **L**, $\mathbf{L}_2$ contains $p - r$ columns that are mutually orthogonal to each other and to those of $\mathbf{L}_1$ but otherwise arbitrary and $\mathbf{D}_1$ is the $r \times r$ diagonal matrix containing the non-zero d_i only.

To achieve the objectives stated earlier, it was decided that the following criteria should be satisfied when forming an augmented covariance matrix $\hat{\mathbf{\Sigma}}$ from **S**:

(a) $\hat{\mathbf{\Sigma}}$ is symmetric,
(b) $\hat{\mathbf{\Sigma}}$ has full rank p,
(c) the first r principal axes of $\hat{\mathbf{\Sigma}}$ are the same as those of **S**, and they are in the same order,
(d) the last $p - r$ principal axes are indeterminate, i.e. the corresponding eigenvalues of $\hat{\mathbf{\Sigma}}$ are identical, and
(e) trace($\hat{\mathbf{\Sigma}}$) = trace(**S**).

These criteria are all upheld by the matrix $\hat{\mathbf{\Sigma}}$ defined by

$$\hat{\mathbf{\Sigma}} = \frac{1}{c}(\mathbf{L}_1 \quad \mathbf{L}_2)\begin{pmatrix}\mathbf{D}_1 + \alpha\mathbf{I} & \mathbf{0}\\ \mathbf{0} & (\alpha + \beta)\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{L}_1^{\mathrm{T}}\\ \mathbf{L}_2^{\mathrm{T}}\end{pmatrix} \tag{7}$$

where α and β are parameters satisfying $\alpha \geqslant 0$, $\beta < d_r$ and $\alpha + \beta > 0$, and c is a normalizing constant given by

$$c = \left\{\alpha p + \beta(p - r) + \sum_{i=1}^{r} d_i\right\} \Big/ \sum_{i=1}^{r} d_i.$$

To conduct discriminant analysis we would need to estimate α and β from the training data and then to insert $\hat{\mathbf{\Sigma}}$ into expression (3) or (4). Given the 'principal axes' representation of $\hat{\mathbf{\Sigma}}$ in equation (7), the inverse follows very readily so that we can telescope these operations. For example, we find for two groups that Fisher's linear discriminant function (4) reduces to

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}}\left(\mathbf{I} - \mathbf{L}_1 \operatorname{diag}\left(\frac{d_1 - \beta}{d_1 + \alpha}, \ldots, \frac{d_r - \beta}{d_r + \alpha}\right)\mathbf{L}_1^{\mathrm{T}}\right)\mathbf{x}. \tag{8}$$

Suppose that $\mathbf{y}$ is the observation vector of an unclassified individual; we allocate this individual to group 1 if the value obtained by inserting $\mathbf{y} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$ in place of $\mathbf{x}$ in expression (8) is greater than 0 and to group 2 otherwise. Analogous simplification results in the g-group case from equation (3). We recommend estimation of α and β by cross-validation on the training data. To do this we found it convenient to set up a mesh of (α, β) pairs over a range of parameter values (in practice a 41×41 mesh with each axis running from 10^{-20} to 10^{20} on the log-scale) and for each point on this mesh to calculate the success rate of the resulting allocation rule using the leave-one-out procedure (Lachenbruch and Mickey, 1968). The estimated values $\hat{\alpha}$, $\hat{\beta}$ are then given by those values that yield the optimum success rate. Fig. 2 shows the pattern of classification success rates across this mesh for one of the data sets discussed in Section 4.

We refer to such parameter estimation by performance optimization as *tuning*. Equation (8) gives rise to some limiting cases of interest.

(a) Taking the limit as both α and $\beta \to 0^+$ gives $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}}(\mathbf{I} - \mathbf{L}_1\mathbf{L}_1^{\mathrm{T}})\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}}\mathbf{L}_2\mathbf{L}_2^{\mathrm{T}}\mathbf{x}$, which can be shown to be the same as the two-group zero-variance discriminator (Section 3.3).

(b) Putting $\beta = 0$ gives

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}}\left(\mathbf{I} - \mathbf{L}_1 \operatorname{diag}\left(\frac{d_1}{d_1 + \alpha}, \ldots, \frac{d_r}{d_r + \alpha}\right)\mathbf{L}_1^{\mathrm{T}}\right)\mathbf{x}.$$

When α is a small positive finite constant then this is the usual expression for ridge discriminant analysis using all p original variables.

(c) Taking the limit as $\alpha \to \infty$ for finite β gives $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}}\mathbf{x}$, the result that would be obtained assuming p independent and identically distributed random variables.

In view of (b), we term the above discrimination procedure as generalized ridge discrimination (GRD). Friedman (1989) also provided discrimination methods based on augmented covariance matrices, but in a somewhat different context to ours. He assumed that groups have different dispersion matrices and formulated the mathematics in such a way that the explicit use of augmented matrices is required. That formulation is impractical in our context, as it would involve the use of huge matrices when applied to spectroscopic or molecular modelling data.

3.2. *Modified Canonical Analysis*

One possible approach to the derivation of classification rules (3) and (4) is by seeking the linear combinations $y = \mathbf{a}^{\mathrm{T}}\mathbf{x}$ which maximize the ratio $V = (\mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a})/(\mathbf{a}^{\mathrm{T}}\mathbf{S}\mathbf{a})$, where $\mathbf{S}$ is as defined earlier and $\mathbf{B}$ is the between-groups covariance matrix

$$\frac{1}{g-1}\sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}}$$

for

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{g} n_i \bar{\mathbf{x}}_i.$$

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

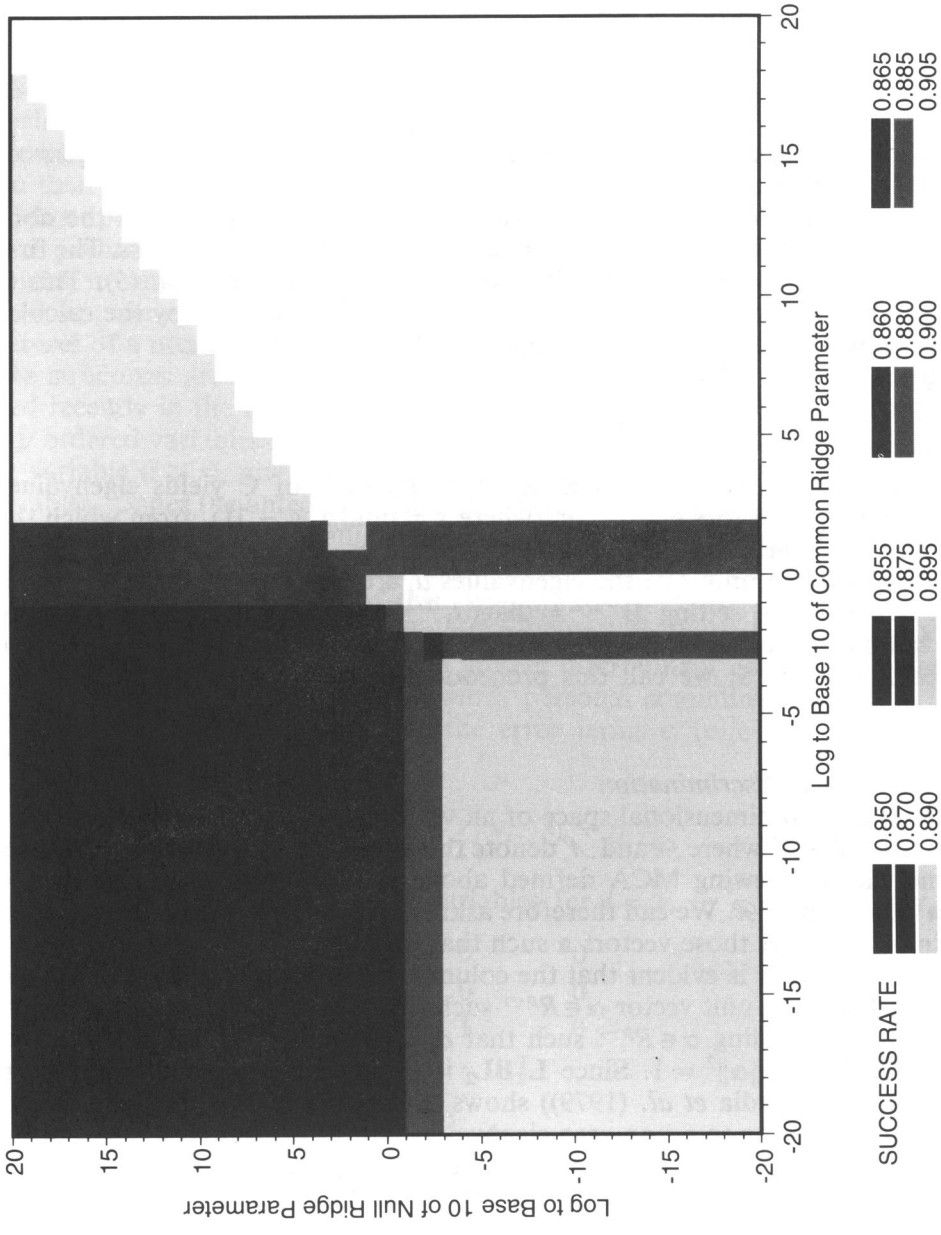

Fig. 2. Overall success rates for the generalized ridge discrimination classification procedure evaluated by the leave-one-out procedure on the IR1 data, over the (α, β) mesh (where α is on the horizontal and β is on the vertical axis)

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

The appropriate coefficients are given by the eigenvectors $\mathbf{a}_i$ corresponding to the non-zero eigenvalues λ_i of the generalized eigenproblem $(\mathbf{B} - \lambda_i\mathbf{S})\mathbf{a}_i = \mathbf{0}$. The resulting variables $y_i = \mathbf{a}_i^T\mathbf{x}$ are the *canonical variates*, and the squares of the Euclidean distances between a point to be classified and the group mean points in the canonical variate space are equivalent to the quantities (3) (see, for example, Krzanowski (1988), chapter 11). When $g = 2$ then

$$\mathbf{B} = \frac{n_1 n_2}{n}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T$$

is of rank 1 so there is only one non-zero eigenvalue λ and the corresponding eigenvector reduces to expression (4).

Campbell and Atchley (1981) demonstrated that canonical analysis in the non-singular case can be achieved by a two-stage spectral decomposition process. The first stage is the spectral decomposition $\mathbf{LDL}^T$ of $\mathbf{S}$, as given by equation (5). This is followed by the transformation from $\mathbf{x}$ to $\mathbf{w} = \mathbf{D}^{-1/2}\mathbf{L}^T\mathbf{x}$, followed by the calculation of the between-groups covariance matrix with respect to $\mathbf{w}$, i.e.

$$\mathbf{C} = \frac{1}{g-1}\sum_{i=1}^{g} n_i(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})^T$$

in obvious notation. Finally the spectral decomposition of $\mathbf{C}$ yields eigenvalues $\lambda_1, \ldots, \lambda_s$ and eigenvectors $\mathbf{c}_1, \ldots, \mathbf{c}_s$ (where $s = \min[p, g-1]$), from which the required canonical variate coefficients are recovered as $\mathbf{a}_i = \mathbf{L}\mathbf{D}^{-1/2}\mathbf{c}_i$.

If $\mathbf{S}$ is singular then only r of the eigenvalues d_i are non-zero so that $\mathbf{D}^{-1/2}$ does not exist. We propose setting $\mathbf{D}^{-1/2} = \text{diag}(d_1^{-1/2}, \ldots, d_r^{-1/2}, 0, \ldots, 0)$ in the standard canonical procedure, which is equivalent to the use of the Moore–Penrose generalized inverse of $\mathbf{S}$; we call this procedure the modified canonical analysis (MCA) method.

3.3. *Zero-variance Discrimination*

If $\mathscr{A}$ denotes the p-dimensional space of all vectors $\mathbf{a}$ then when $\mathbf{S}$ has rank r we can write $\mathscr{A} = \mathscr{R} + \mathscr{N}$ where $\mathscr{R}$ and $\mathscr{N}$ denote the range and null spaces of $\mathbf{S}$ respectively. One way of viewing MCA defined above is as the maximization of $V = (\mathbf{a}^T\mathbf{B}\mathbf{a})/(\mathbf{a}^T\mathbf{S}\mathbf{a})$ for $\mathbf{a}$ in $\mathscr{R}$. We can therefore ask: is there a corresponding maximum of $\mathbf{a}^T\mathbf{B}\mathbf{a}$ in $\mathscr{N}$, i.e. over those vectors $\mathbf{a}$ such that $\mathbf{S}\mathbf{a} = \mathbf{0}$?

From equation (6) it is evident that the columns of $\mathbf{L}_2$ form a basis for $\mathscr{N}$. Thus if $\mathbf{a} \in \mathscr{N}$ then there is a unit vector $\boldsymbol{\alpha} \in R^{p-r}$ such that $\mathbf{a} = \mathbf{L}_2\boldsymbol{\alpha}$. Hence our problem reduces to that of finding $\boldsymbol{\alpha} \in R^{p-r}$ such that $\boldsymbol{\alpha}^T\mathbf{L}_2^T\mathbf{B}\mathbf{L}_2\boldsymbol{\alpha}$ is maximized subject to $\|\mathbf{L}_2\boldsymbol{\alpha}\|^2 = \boldsymbol{\alpha}^T\mathbf{L}_2^T\mathbf{L}_2\boldsymbol{\alpha} = \|\boldsymbol{\alpha}\|^2 = 1$. Since $\mathbf{L}_2^T\mathbf{B}\mathbf{L}_2$ is symmetric, standard theory (e.g. theorem A.9.2 of Mardia *et al.* (1979)) shows that $\boldsymbol{\alpha}$ is a normalized eigenvector of $\mathbf{L}_2^T\mathbf{B}\mathbf{L}_2$ corresponding to a non-zero eigenvalue. Thus the eigenvectors $\boldsymbol{\alpha}_i$ corresponding to non-zero eigenvalues of $\mathbf{L}_2^T\mathbf{B}\mathbf{L}_2$ provide a system of canonical variates $y_i = \mathbf{a}_i^T\mathbf{x}$ in the null space of $\mathbf{S}$ via $\mathbf{a}_i = \mathbf{L}_2\boldsymbol{\alpha}_i$. We term this method the zero-variance discrimination (ZVD) procedure. Considerable algebraic simplification can be achieved in the two-group case to expedite computation of this procedure, but details are omitted here.

If $\mathbf{P}_S$ denotes the symmetric projection matrix $\mathbf{L}_1(\mathbf{L}_1^T\mathbf{L}_1)^{-1}\mathbf{L}_1^T = \mathbf{L}_1\mathbf{L}_1^T$ onto the range of $\mathbf{S}$ (theorem A.10.1 of Mardia *et al.* (1979)) and if $\mathbf{P}_S^* = \mathbf{I} - \mathbf{P}_S$ then $\hat{\boldsymbol{\Sigma}}^{-1} =$

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

$\mathbf{P}_S^* = \mathbf{L}_2\mathbf{L}_2^T$ corresponds to the ZVD procedure. In the extreme where there is just one observation per group (i.e. $n_i = 1$ for all i) this rule reduces to $\hat{\boldsymbol{\Sigma}} = \mathbf{I}$, i.e. classification based on Euclidean distance in the sample space of the spectroscopic data.

3.4. *Antedependence Modelling*

All the preceding approaches for obtaining $\hat{\boldsymbol{\Sigma}}$ are either empirical or purely data based. An alternative general approach is to postulate and fit to the data a suitable stochastic model which has fewer parameters, taking care that the dispersion matrix $\hat{\boldsymbol{\Sigma}}$ implied by the model is non-singular. Having estimated the model parameters we can then obtain $\hat{\boldsymbol{\Sigma}}^{-1}$ readily for use in expression (3) or (4).

Most spectroscopic data are obtained by a 'moving window' process which aggregates points within a prespecified width while scanning across the range of wavelengths involved. Since a typical point will appear in three or four successive windows, the resulting data will be serially correlated and will exhibit the general features of a non-stationary time series. A nested series of models suitable for such data structures are the antedependence models introduced by Gabriel (1962) and used recently in the analysis of repeated measurements by Kenward (1987). A set of p ordered variables is said to have an antedependence structure of order r if the ith variable ($i > r$), given the preceding r, is independent of all further preceding variables. Under the antedependence structure of order r, the inverse of $\boldsymbol{\Sigma}$ has non-zero elements only on the leading diagonal and on the r diagonals immediately above and immediately below it. Complete independence ($r = 0$) and general dependence ($r = p - 1$) are special cases of this structure.

Full theory behind these models can be found in Gabriel (1962) and Kenward (1987). For fitting the models in the present discriminant context we merely need the following summary (M. G. Kenward, personal communication).

Suppose that $X_i = \mu_i + e_i$ where the error terms e_i follow an antedependence structure of order r, i.e.

$$e_i = \sum_{j=1}^{r} e_{i-j}\lambda_{ji} + f_i$$

and the f_i are independent $(0, \sigma_i^2)$ variables for $i = 1, \ldots, n$. Let

$$\mathbf{F}_r = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ -\lambda_{11} & 1 & 0 & 0 & \ldots & 0 \\ -\lambda_{22} & -\lambda_{12} & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ -\lambda_{rr} & -\lambda_{r-1,r} & -\lambda_{r-2,r} & & \ldots & 0 \\ 0 & -\lambda_{r,r+1} & -\lambda_{r-1,r+1} & & \ldots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda_{2,p-2} \; -\lambda_{1,p-1} & 1 \end{pmatrix},$$

i.e. a lower triangular matrix with 1s on the leading diagonal and r non-zero subdiagonals. Estimates of λ_{ij} are obtained by setting the values in the lower r

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

bands (ignoring the diagonal) of $\mathbf{F}_r\mathbf{S}$ to 0 and solving the resulting series of linear equations recursively. This gives $\hat{\mathbf{F}}_r$, from which we derive

$$\hat{\boldsymbol{\Sigma}}^{-1} = \hat{\mathbf{F}}_r^{\mathrm{T}}\mathbf{G}_r\hat{\mathbf{F}}_r \tag{9}$$

where $\mathbf{G}_r = \mathrm{diag}(\hat{\mathbf{F}}_r\mathbf{S})$.

To date we have implemented the fitting of antedependence models of order $r = 1$, 2 and 3 to spectroscopic data. The resulting classification rules (on using equation (9) in expression (3) or (4)) will be termed the AD1, AD2 and AD3 methods respectively.

4. Applications and Comparisons

The data sets described in Section 1 were subjected to all the discrimination procedures above. To compare the results of these new methods with those that would be obtained by chemometricians using existing methodology, PPC and PLS discriminant functions were also calculated for each data set. The comparison was made by calculating the success rate for each classification method on each data set by using the leave-one-out method of Lachenbruch and Mickey (1968). Following an analogous approach to that of Campbell and Rayment (1978), very small eigenvalues were slightly increased when performing PPC analysis to overcome instability associated with the smallest eigenvalues.

To make the preliminary selection of components in PPC analysis, the components were ranked according to the discriminatory measure given by Jolliffe (1986), chapter 9.1, rather than by the more usual eigenvalue criterion. This optimizes the choice of components for discrimination purposes. The smallest number of these ranked components that accounted for at least 95% of the trace of the overall covariance matrix was used in deriving the allocation rule. Fig. 3 shows the rate of decay of eigenvalues of the overall covariance matrix obtained from the complete samples in each data set. Data set NIR1 shows a relatively rapid decay, and 21 of

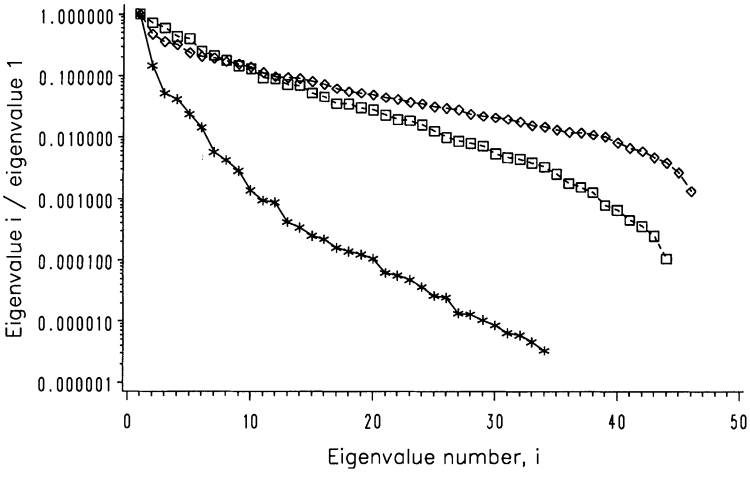

Fig. 3. Eigenvalue decay of the overall covariance matrix for each of the three data sets: *, NIR1; □; NIR2; ◇, IR1

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

the 34 non-zero eigenvalues accounted for the required percentage of the trace. The other two sets show much slower rates of decay, and no reduction of components was possible. This ranking and selection requires only eigenvalue calculations so is fast and could be conducted afresh for each omitted unit in the leave-one-out process, thereby ensuring unbiasedness of assessment of performance.

With the PLS method, however, the number of factors for use in classification was taken to be the smallest number for which the success rate of the resulting allocation rule (as assessed by cross-validation) was a maximum. This is a much slower and more computer-intensive process than eigenvalue calculation; incorporating it into the leave-one-out assessment of performance of the final allocation rule produces two nested leave-one-out procedures and hence leads to a prohibitive increase of computer time. Thus only one selection of factors was made at the outset, using the whole data set; eight factors were selected for NIR1, two for NIR2 and one for IR1. However, since these selections involve performance optimization (i.e. tuning), the final assessment of the method will be necessarily biased.

A similar computational problem was encountered with GRD, where tuning is also embedded into the process, so the parameters were estimated only once for each complete data set. Since this parameter estimation again involves performance optimization, the results with this method will also show optimistic bias. We discuss this problem further in Section 5; all other methods are free of such extraneous sources of bias.

Table 2 presents the success rates for each classification method on each data

TABLE 2
Success rates of the various classification rules

Data set	*Method*	*Group 1 rate*	*Group 2 rate*	*Overall rate*
NIR1	PPC	1.000	0.824	0.914
	PLS†	0.889	0.882	0.886
	GRD†	1.000	0.882	0.943
	MCA	1.000	0.824	0.914
	ZVD	0.667	0.647	0.657
	AD1	0.944	0.706	0.828
	AD2	0.889	0.824	0.857
	AD3	0.722	0.764	0.743
NIR2	PPC	0.773	0.739	0.756
	PLS†	0.864	0.783	0.822
	GRD†	0.864	0.826	0.844
	MCA	0.773	0.913	0.844
	ZVD	0.682	0.478	0.578
	AD1	0.773	0.870	0.823
	AD2	0.773	0.957	0.867
	AD3	0.773	0.957	0.867
IR1	PPC	0.864	0.880	0.872
	PLS†	0.864	0.960	0.915
	GRD†	0.864	0.960	0.915
	MCA	0.818	0.880	0.851
	ZVD	0.864	0.920	0.894
	AD1	0.682	0.960	0.830
	AD2	0.591	0.880	0.765
	AD3	0.455	0.840	0.660

†Biased estimates.

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

set, as obtained by the leave-one-out method. We show the separate rates for each group as well as the overall rate for the whole data set in each case. Taking the overall rate as the criterion, there is considerable variation in relative performance of the methods between the data sets. GRD comes either top or near the top in each data set, but the strictures regarding bias made earlier must be borne in mind and will be discussed in the next section. The PLS and antedependence methods come out best in one data set each and otherwise perform creditably, although we have to take bias into account for PLS. Moreover, the three orders of antedependence show some variability in performance; whereas there seems to be some genuine competition between AD1 and AD2, AD3 is dominated by AD2 for these data sets and performs very poorly on IR1 so need not be considered further. MCA, while never actually top, is consistently nearly so: second best in one data set, third best in another and fifth best in the remaining data set. The most dramatic variation, however, is shown by ZVD which comes last by a long way in both NIR sets but is beaten only narrowly by PLS and GRD in the IR data set.

Various other features of comparison between the methods were also investigated. One feature was the set of discriminant functions produced by the various methods; since each function has at least 700 coefficients for each data set, they were summarized by computing pairwise inner products. Again, there was much variability across the data sets and the only consistent aspect was the orthogonality (or near orthogonality) of the ZVD functions and the others in most situations. Some of the methods lend themselves to efficient computer organization, and in such a comparison MCA came out strikingly well. However, undoubtedly with extra ingenuity the other methods could also be optimized in this sense so such a comparison is not conclusive. In short, it seems difficult to establish a definitive ranking of the methods in terms of their effectiveness. The question of bias must also be taken up.

5. Investigation of Bias

We conjectured earlier that PLS and GRD will incur bias in the cross-validated estimation of prediction success rate outlined in Section 4, owing to the presence of tuning. One way of checking this conjecture is by finding permutational distributions of the success rates on each data set. To do this, 250 random permutations of the rows of each data matrix were generated, a selection of methods was applied to each such permutation and the success rate of each method was obtained for each permutation by using the leave-one-out procedure. The resulting success rates represent the performances of the methods when differences between the groups have been removed, so that the distribution of success rates should be peaked at 0.5 for an unbiased method. We chose, arbitrarily, MCA and ZVD as two 'unbiased' methods to compare with PLS and GRD, and fitted kernel density estimates to each distribution of success rates (Silverman, 1986). As an example, the densities for the IR1 data set are shown in Fig. 4; those for the other data sets exhibited almost identical patterns. The presence of bias for PLS and GRD is indicated by the shift of the peak in these cases, whereas MCA and ZVD indeed appear to be unbiased. However, since different practitioners of PLS use various methods for selecting the number of factors, the bias associated with some implementations may differ substantially from that reported here.

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

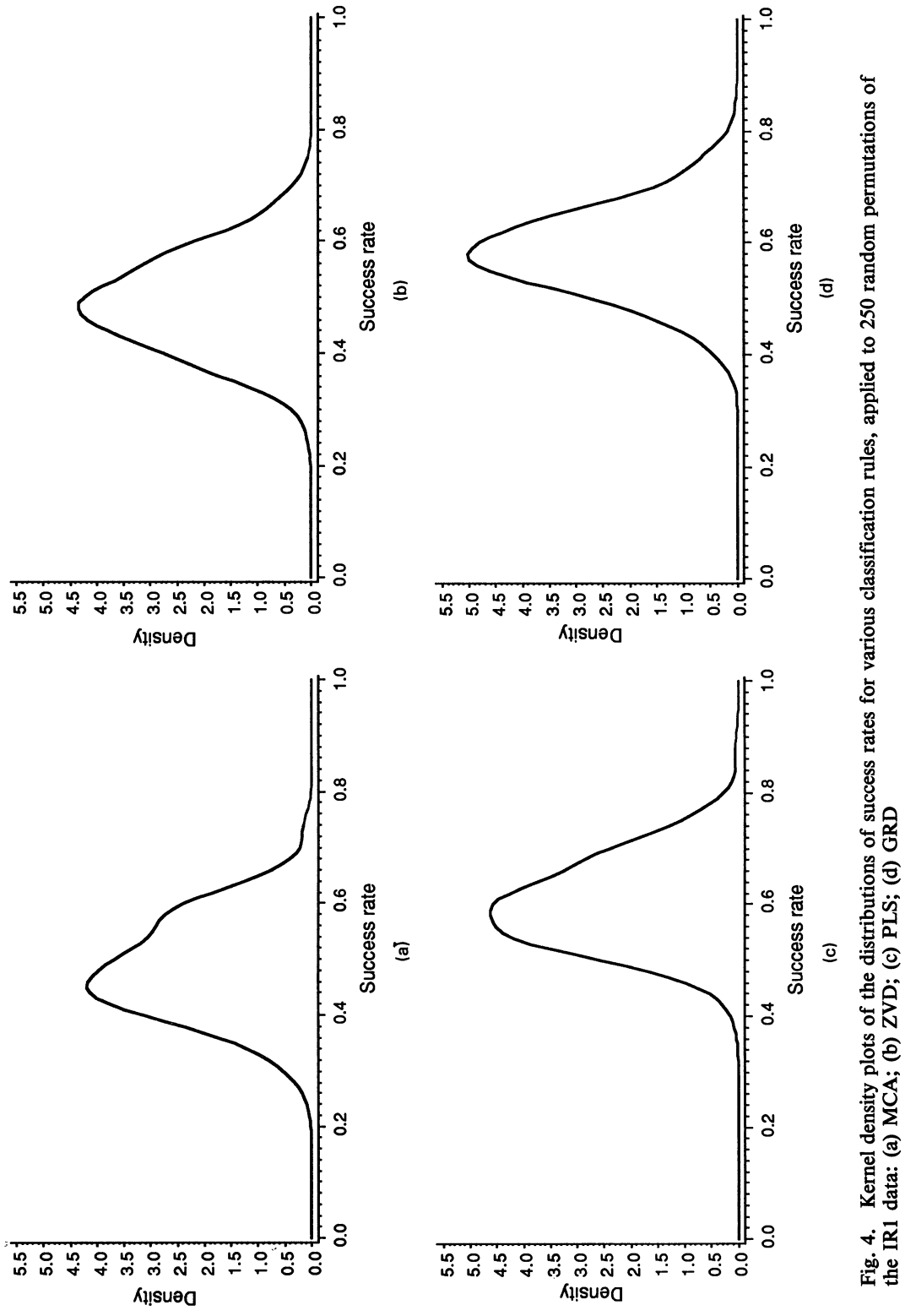

Fig. 4. Kernel density plots of the distributions of success rates for various classification rules, applied to 250 random permutations of the IR1 data: (a) MCA; (b) ZVD; (c) PLS; (d) GRD

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

Once the presence of bias has been confirmed, it is important to estimate its likely magnitude. As GRD involves tuning of two parameters and will probably therefore exhibit the greatest bias of all the methods, we investigated it more intensively. A large data set, having two groups each with 74 observations and 800 molecular descriptors, was analysed by GRD by using the same 41 × 41 grid mentioned earlier. The optimal selection of ridge parameters gave $\alpha = 10^{-20}$ and $\beta = 10^{-7}$, at which values the cross-validated success rates in each group were 0.7838. 35 separate data sets were then randomly selected from this large data set, each set having two groups of 18 and 19 individuals respectively. For each such random selection, an 'internal' success rate was obtained by leave-one-out cross-validation of the 37 training individuals and an 'external' success rate was obtained by calculating the discriminant function from the 37 training individuals and using it to classify the remaining 111 individuals. Table 3 gives the means and standard errors of the 35 internal and external success rates and the differences between them. For comparison ZVD, AD1 and AD2 were assessed in a similar fashion on the same 35 random selections, and the results are also given in Table 3. It is clear that the internal success rates are optimistically biased for GRD but not for the other methods.

We can conclude from these comparisons that the leave-one-out assessment of success rates is a valid procedure for all methods that do not involve tuning, even when dimensionality is very high and sample sizes are small. When tuning is involved, the only safeguard is to conduct the tuning afresh for each omitted unit (as advocated in other contexts by Ganeshanandam and Krzanowski (1989)). Such a procedure may, of course, become prohibitively expensive in terms of computer time (as happened in the present study).

6. Conclusion

Each of the methods described in this paper has something to offer when discriminating between groups using very high dimensional data. After allowance has

TABLE 3
Means of internal and external success rates over 35 random selections from a large data set†

Discriminant method	*Type of success rate*	*Group 1 mean*	*Group 2 mean*
GRD	Internal	0.6415 (0.0187)	0.7859 (0.0154)
	External	0.5992 (0.0156)	0.7228 (0.0148)
	Difference	0.0424 (0.0255)	0.0631 (0.0162)
ZVD	Internal	0.5770 (0.0193)	0.6878 (0.0198)
	External	0.5992 (0.0159)	0.6831 (0.0136)
	Difference	−0.0222 (0.0240)	0.0047 (0.0225)
AD1	Internal	0.5369 (0.0167)	0.5962 (0.0149)
	External	0.5388 (0.0141)	0.6566 (0.0132)
	Difference	−0.0019 (0.0224)	−0.0604 (0.0220)
AD2	Internal	0.5696 (0.0137)	0.6787 (0.0183)
	External	0.5682 (0.0151)	0.7017 (0.0138)
	Difference	0.0014 (0.0229)	−0.0230 (0.0231)

†Standard errors are given in parentheses.

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms

been made in Table 2 for the bias in PLS and GRD, no single method dominates all others in terms of success rates for all data sets and each method performs well on at least one data set. Thus we advocate the inclusion of a selection of methods in the basic toolkit for such problems and building up a database, in the hope that for each different type of data, e.g. NIR, a specific method can ultimately be recommended in view of its historical performance on data of that type.

Acknowledgements

We are grateful to Adrian Roberts and Simon Pack for providing the PPC and PLS analyses respectively, and to the referees for helpful comments on a previous version of this paper.

References

Campbell, N. A. and Atchley, W. R. (1981) The geometry of canonical variate analysis. *Syst. Zool.*, **30**, 268–280.

Campbell, N. A. and Reyment, R. A. (1978) Discriminant analysis of a cretaceous foraminifer using shrunken estimators. *Math. Geol.*, **10**, 347–359.

Chang, W.-C. (1983) On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Statist.*, **32**, 267–275.

Friedman, J. H. (1989) Regularized discriminant analysis. *J. Am. Statist. Ass.*, **84**, 165–175.

Gabriel, K. R. (1962) Antedependence analysis of a set of ordered variables. *Ann. Math. Statist.*, **33**, 201–212.

Ganeshanandam, S. and Krzanowski, W. J. (1989) On selecting variables and assessing their performance in linear discriminant analysis. *Aust. J. Statist.*, **31**, 433–447.

Geladi, P. (1988) Notes on the history and nature of partial least squares (PLS) modelling. *J. Chemometr.*, **2**, 231–246.

Gunst, R. F. and Mason, R. L. (1979) Some considerations in the evaluation of alternate prediction equations. *Technometrics*, **21**, 55–63.

Hoskuldsson, A. (1988) PLS regression methods. *J. Chemometr.*, **2**, 211–228.

Jolliffe, I. T. (1986) *Principal Component Analysis.* New York: Springer.

Kenward, M. G. (1987) A method for comparing profiles of repeated measurements. *Appl. Statist.*, **36**, 296–308.

Krzanowski, W. J. (1988) *Principles of Multivariate Analysis: a User's Perspective.* Oxford: Clarendon.

——— (1992) Ranking principal components to reflect group structure. *J. Chemometr.*, **6**, 97–102.

Lachenbruch, P. A. and Mickey, M. R. (1968) Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–11.

Manne, R. (1987) Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometr. Intell. Lab. Syst.*, **2**, 187–197.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis.* London: Academic Press.

Naes, T. and Martens, H. (1985) Comparison of prediction methods for multicollinear data. *Communs Statist. Simuln Computn*, **14**, 545–576.

Silverman, B. W. (1986) *Density Estimation.* London: Chapman and Hall.

Stone, M. and Brooks, R. J. (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. R. Statist. Soc.* B, **52**, 237–269.

Yendle, P. W. and MacFie, H. J. H. (1989) Discriminant principal components analysis. *J. Chemometr.*, **3**, 589–600.

This content downloaded from
148.88.67.84 on Fri, 16 Jun 2023 16:30:16 +00:00
All use subject to https://about.jstor.org/terms