Received: 27 November 2013, Revised: 28 January 2014, Accepted: 04 February 2014, Published online in Wiley Online Library: 18 March 2014

(wileyonlinelibrary.com) DOI: 10.1002/cem.2609

Partial least squares discriminant analysis: taking the magic away

Richard G. Brereton[a]* and Gavin R. Lloyd[b]

Partial least squares discriminant analysis (PLS-DA) has been available for nearly 20 years yet is poorly understood by most users. By simple examples, it is shown graphically and algebraically that for two equal class sizes, PLS-DA using one partial least squares (PLS) component provides equivalent classification results to Euclidean distance to centroids, and by using all nonzero components to linear discriminant analysis. Extensions where there are unequal class sizes and more than two classes are discussed including common pitfalls and dilemmas. Finally, the problems of overfitting and PLS scores plots are discussed. It is concluded that for classification purposes, PLS-DA has no significant advantages over traditional procedures and is an algorithm full of dangers. It should not be viewed as a single integrated method but as step in a full classification procedure. However, despite these limitations, PLS-DA can provide good insight into the causes of discrimination via weights and loadings, which gives it a unique role in exploratory data analysis, for example in metabolomics via visualisation of significant variables such as metabolites or spectroscopic peaks. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: Partial Least Squares; Discrimination; Classification; Two Class Classifiers

1. INTRODUCTION

Partial least squares discriminant analysis (PLS-DA) was reported formally a decade ago [1] although its first use is purported to be around 20 years ago [2], so the method has been around a long time. The method is now routinely incorporated into most packages used by chemometricians and the results of PLS-DA cited in numerous papers, especially in metabolomics.

Despite a strong theoretical basis [1], the advantages and disadvantages of the method are rarely understood by users. PLS-DA is possibly one of the most misunderstood and misused methods for discrimination in chemometrics. Very few authors of papers understand the importance of the parameters used to obtain and assess the discriminant model and how critical these are for model performance. For example, under certain circumstances, PLS-DA provides the same results as the classical approach of Euclidean distance to centroids (EDC) and under other circumstances, the same as that of linear discriminant analysis (LDA) [3], yet PLS-DA is usually described as a single method, and sometimes, its performance is compared with other approaches such as LDA: in fact, describing PLS-DA as a method in its own right is statistically ambivalent, and it should instead be regarded as an algorithm that is one in a series of steps (such as preprocessing, variable selection, selecting samples for validation, and column centring) in a classification procedure. Barker and Rayens have commented on the relationship between PLS-DA and other statistical approaches for discrimination [1], and this connection is often discussed in the statistically oriented literature but not usually presented in an explicit way for the general user. It is usually not necessary when using common chemometrics software to have to make explicit decisions as to what parameters are required for decision making, and as such, by using default parameters that may well be inappropriate for the problem in hand, it is easy to come to erroneous conclusions without even realising this: chemometric methods are at their most useful when problems are difficult to solve, and it is precisely in such situations that an understanding and appropriate choice of parameters is critical. Only a very small minority of presentations and papers fully describe how the data have been treated prior to modelling, and as such, the description of the classification technique is often of little value to the reader or listener; the author may as well have stated he or she used additions or multiplications in their calculations. In contrast, more simple statistical approaches such as LDA and EDC and quadratic discriminant analysis (QDA) [3,4] have well-known properties and involve well-established assumptions about the data; because fewer decisions have to be made about the parameters (as will be discussed later), they are harder to misuse, and the appropriateness of the model can be more directly related to the statistical properties of the data.

In this paper, we will primarily discuss the role of PLS-DA as a classification technique, that is, a technique to determine what group a sample is most likely to belong to from a set of analytical measurements. The other use of PLS-DA as an exploratory technique will be discussed in the conclusions.

2. PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS ALGORITHM

There are numerous algorithms for partial least squares (PLS) and its enhancements for discriminant analysis. In this paper, for

* Correspondence to: Richard G. Brereton, School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, UK.
E-mail: r.g.brereton@bris.ac.uk

a R. G. Brereton
School of Chemistry, University of Bristol, Cantocks Close, Bristol, BS8 1TS, UK

b G. R. Lloyd
Biophotonics Research Unit, Gloucestershire Hospitals NHS Foundation Trust, Great Western Road, Gloucester, GL1 3NN, UK

simplicity, we describe just one approach with occasional comments about other methods. We will focus on an algorithm PLS1 [5,6] where there are two groups of samples and the aim is to decide which of the two groups a sample belongs to. The extension of PLS1 to more than two groups will be briefly described as well as PLS2, which can also be used when there are more than two groups. However, the main discussion will be constrained to situations where there are two groups, and the aim of the discriminant function is to decide which group a sample belongs to using information such as its spectrum or chemical profile.

Partial least squares discriminant analysis can be regarded as a linear two-class classifier. That is, the method (or a method that includes PLS-DA as one of its steps) aims to find a straight line that divides the space into two regions. Figure 1 illustrates a possible discriminant function for two groups; samples to the left belong to the group represented by blue circles and samples to the right to the group represented by red triangles. The aim of many different algorithms (EDC, LDA, PLS-DA, linear support vector methods, etc.) is to find this discriminator, or separator, or decision function. Of course, when there are more than two variables, it will be represented by a hyperplane in multidimensional space. To simplify, in this paper, we present mainly an example characterised by two variables. Sometimes, the samples are projected onto lines at right angles to this discriminator, often called canonical variates—in which case their distance along this separator is considered a discriminant score, analogous to a principal component (PC) score which involves projecting onto the line of maximum variance.

The difference between various approaches for two-class linear discrimination is the position and slope of the separator, which in turn relates to the criterion used to determine the separator and therefore the assumptions in the model. PLS-DA is no different to any other linear decision function but, because of the historic development of chemometrics, often is described algorithmically rather than statistically.

For brevity, we describe only one PLS algorithm in this paper. However, usually by judicious scaling or centring, other approaches can give exactly equivalent results. It is not the primary purpose of this paper to compare PLS algorithms.

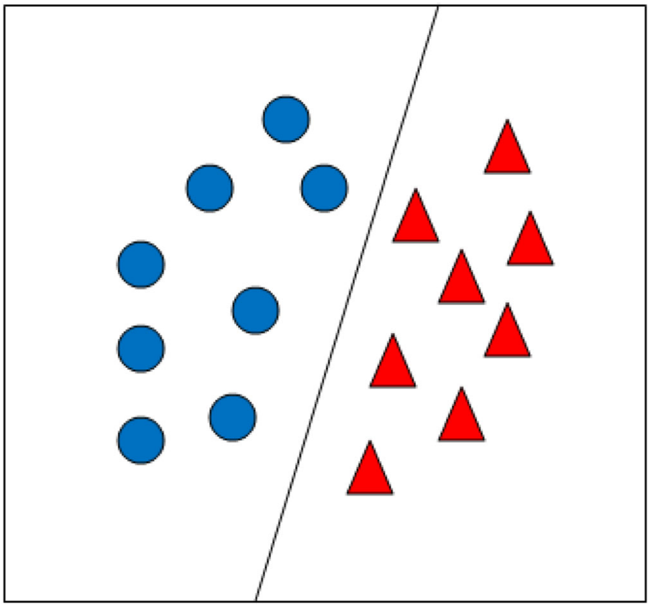


Figure 1. A two-class linear discriminator for two groups characterised by two variables.

2.1. Terminology

For a set of I samples, the $\boldsymbol{X}$ data matrix represents a set of J analytical measurements such as spectra of samples that form two groups. The vector $\boldsymbol{c}$ of length I represents a numerical label for each sample according to its group membership. In the implementation discussed in this paper, we use a label of +1 for I_A samples that are a members of group A and −1 for I_B samples that are a members of group B, where the total number of samples is $I_A + I_B = I$.

Partial least squares discriminant analysis is derived from PLS regression (PLS-R) [6] and involves forming a regression model between the $\boldsymbol{X}$ and $\boldsymbol{c}$ as illustrated in Figure 2. In PLS-R, $\boldsymbol{c}$ (also sometimes denoted $\boldsymbol{y}$) is a set of continuous numbers, for example, the concentration of an analyte. In PLS-DA it contains discrete numbers usually at two levels, one level for what is sometimes called an in-group (A) and the other for the remainder of the data, called the out-group or, in a two-class model, group B. We choose +1 and −1 for the labels in this paper, although 0 and +1 are sometimes employed; however, centring the values of c makes the algebraic derivations very much simpler, hence our choice of labels.

The fundamental PLS-DA equations are as follows

$$\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P} + \boldsymbol{E}$$
$$\boldsymbol{c} = \boldsymbol{T}\,\boldsymbol{q} + \boldsymbol{f}$$

Note the common score matrix $\boldsymbol{T}$ for this implementation. $\boldsymbol{E}$ and $\boldsymbol{f}$ can be considered residuals. In the following algorithm, the successive columns of the score matrix $\boldsymbol{T}$ (PLS components) are orthogonal, but the rows of the $\boldsymbol{X}$ loadings matrix $\boldsymbol{P}$ are not. However, because the scores are orthogonal, the models with successive PLS components are additive. Note also that $\boldsymbol{TP}$ is not the best-fit least squares model for $\boldsymbol{X}$. It can be somewhat confusing in the literature to use a similar notation for PC analysis (PCA) as for PLS when the matrices have quite different properties, for example, X loadings ($\boldsymbol{P}$) in PLS are not orthogonal, unlike in PCA, and the sums of squares of the scores (often called eigenvalues) do not necessarily reduce monotonically with successive components.

2.2. Model building

It is assumed that data have been preprocessed (e.g. standardised or row scaled) previously as appropriate. In this paper, we do not

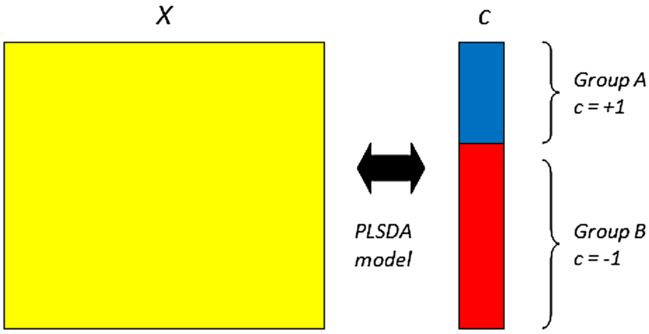


Figure 2. Partial least squares discriminant analysis (PLS-DA) model for two classes.

centre $\boldsymbol{c}$, but there are variants on this theme. If there are equal numbers of samples in each group, $\boldsymbol{c}$ by default will be centred. Extensions to unequal class sizes will be discussed later.

However, $\boldsymbol{X}$ is mean centred down the columns for the standard implementation of PLS as in the following; variants on this will be introduced in Section 4. The PLS1 algorithm we use is as follows.

(1) Calculate the PLS weight vector $\boldsymbol{w}$

$$\boldsymbol{w} = \boldsymbol{X}'\boldsymbol{c}$$

(2) Calculate the scores, which are given by

$$\boldsymbol{t} = \frac{\boldsymbol{Xw}}{\sqrt{\Sigma w^2}}$$

(3) Calculate the x loadings by

$$\boldsymbol{p} = \frac{\boldsymbol{t}'\boldsymbol{X}}{\Sigma t^2}$$

(4) Calculate the c loading (a scalar) by

$$q = \frac{\boldsymbol{c}'\boldsymbol{t}}{\Sigma t^2}$$

(5) Subtract the effect of the new PLS component from the data matrix to obtain a residual data matrix

$$^{resid}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{tp}$$

(6) Calculate the residual value of $\boldsymbol{c}$

$$^{resid}\boldsymbol{c} = \boldsymbol{c} - \boldsymbol{t}q$$

(7) If further components are required, replace both $\boldsymbol{X}$ and $\boldsymbol{c}$ by their residuals and return to step 1. We will not discuss the criteria for deciding how many components to retain in this paper.

A weights matrix $\boldsymbol{W}$ can be obtained, each successive column corresponding to a successive PLS component.

2.3. Prediction

Once a model is built, it is then possible to predict the value of $\boldsymbol{c}$ both for the original data (autoprediction) and for future samples of unknown origins, or for test set samples of known origins, as follows.

The relationship between $\boldsymbol{X}$ and $\boldsymbol{c}$ can be expressed by

$$\boldsymbol{c} = \boldsymbol{Xb} + \boldsymbol{f} = \boldsymbol{Tq} + \boldsymbol{f}$$

where $\boldsymbol{b}$ is a regression coefficient vector of dimensions $J \times 1$; hence, an unknown sample value of c can be predicted by

$$\hat{c} = \boldsymbol{xb}$$

The estimation of $\boldsymbol{b}$ can be obtained as follows

$$\boldsymbol{b} = \boldsymbol{W}(\boldsymbol{PW})^{-1}\boldsymbol{q}$$

The class a sample belongs to is determined by its value of $\hat{c}$. The simplest decision rule is if the value is above 0, assign it to class A and below to class B. We will examine this decision rule later, but it is most usual to choose a value halfway between the numerical class labels, even though, as we show, this may not be the most appropriate value.

3. RELATIONSHIP BETWEEN PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS, EUCLIDEAN DISTANCE TO CENTROIDS, AND LINEAR DISCRIMINANT ANALYSIS

3.1. Euclidean distance to centroids

Euclidean distance to centroids is the simplest type of linear classifier that can be envisaged. The principle is that a sample is assigned to the group of samples whose centroid it is nearest using the squared Euclidean distance in the variable space. Formally,

$$d_{iA}^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_A)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_A)'$$

where d_{iA} is the Euclidean distance of sample i to the centroid of group A and $\overline{x}_A$ is the centroid of group A. Note that we use row vectors to represent the measurements for a sample. A similar equation can be obtained for group B. The group whose distance is smallest is the one that the sample is defined as belonging to.

The boundary between two classes occurs when

$$d_{iA}^2 = d_{iB}^2$$

We can call this the separator or boundary between two classes. If two classes are not linearly separable, then perfect prediction is not possible using a linear model.

We will consider an example of two classes consisting of equal numbers of samples characterised by two variables; extensions when group sizes are unequal and there are more than two groups are discussed in later sections. In Figure 3, two groups are illustrated. The contours of the squared Euclidean distances from the centroids of each group are presented in Figure 3(a). It can be observed that these are circular with a linear separator or decision function where the two distances are equal. In Figure 3(b), the contours for $d_{iB}^2 - d_{iA}^2$ are illustrated, representing the difference in squared distances. A positive value corresponds to membership of class A and a negative value to class B. A value of 0 represents the separator. We will see later that the position of this separator can be adjusted using Bayesian methods (Section 4).

When there are more than two variables, the separator becomes a plane in hyperspace. Although EDC is a common method, it has several drawbacks. The first is that it is assumed that each variable has approximately equal variance, which may not be the case if measurements are on very different scales or different in nature. The second is that it is assumed that the variance structure of each group is the same. In many real situations, these limitations mean that the simple Euclidean model is inadequate.

3.2. Partial least squares discriminant analysis with one component

For PLS-DA, instead of calculating the difference between two squared distances from the centroid, a value of c is estimated from training or test set or validation data. If c is set to 1 for all training set values of group A and c to -1 for group B, we can

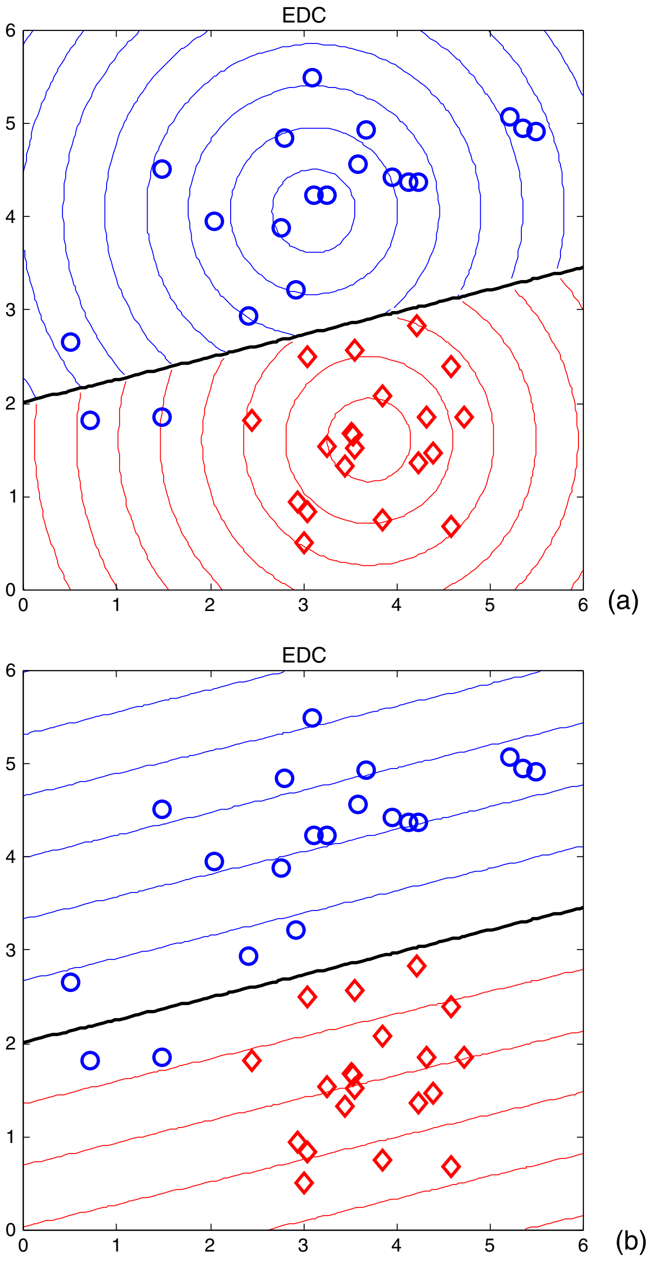


Figure 3. Illustration of Euclidean distance to centroids (EDC). (a) Contours of the squared distance from the centroid of two groups, with the equidistant separator indicated in bold and black; (b) contours of the difference between squared distances with the separator indicated as in (a).

estimate this for unknowns. A simple rule might be that if c is positive, assign to group A, otherwise to group B. The estimated values of c from the one-component PLS-DA model can be contoured as in Figure 4. It can be seen these are linearly related with the separator at $c = 0$. The position of the separator for a one-component PLS-DA model is the same (in this case) as for an EDC model.

Hence, using a single PLS component, the classification model is the same as for EDC and so has the same disadvantages as EDC. Of course, if variables are standardised in advance, then they are likely to be on a similar scale, and EDC or one-component PLS-DA may be appropriate.

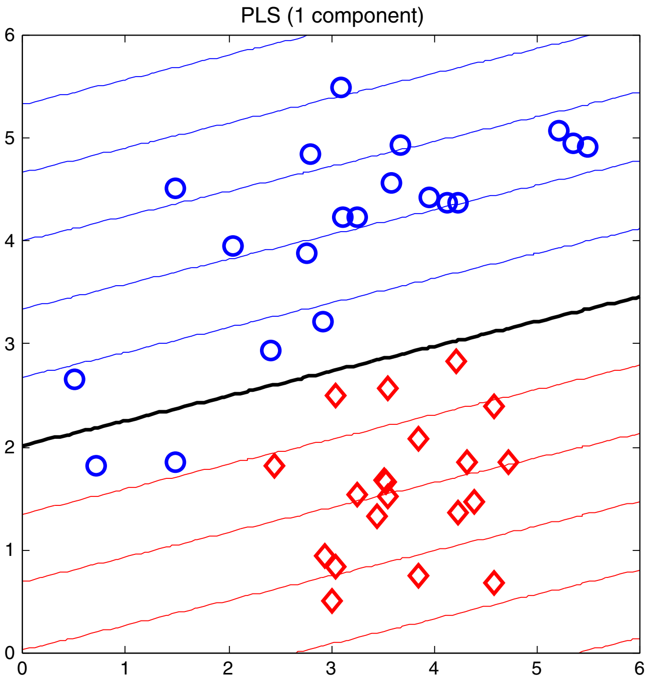


Figure 4. Estimated values of c for a one-component partial least squares (PLS) discriminant analysis model. Blue = positive and red = negative, with $c = 0$ as separator.

3.2.1. *Proof that Euclidean distance to centroids and one-component partial least squares discriminant analysis provide the same classification model*

To simplify this proof, we assume that there are an equal number of samples in each group and that the $\boldsymbol{X}$ matrix is mean centred. Under such circumstances, $\overline{\boldsymbol{x}}_A = -\overline{\boldsymbol{x}}_B$, that is, the mean of group B is negative to that of group A, and the overall mean is zero.

For EDC, define D, which is the difference between the squared distance to the mean of group B minus that of group A. A positive value of D is indicative of a member of group A as the distance to the mean of group B is greater than to the mean of group A, analogous to c.

$$\begin{aligned} D &= (\boldsymbol{x} - \overline{\boldsymbol{x}}_B)(\boldsymbol{x} - \overline{\boldsymbol{x}}_B)' - (\boldsymbol{x} - \overline{\boldsymbol{x}}_A)(\boldsymbol{x} - \overline{\boldsymbol{x}}_A)' \\ &= (\boldsymbol{x} + \overline{\boldsymbol{x}}_A)(\boldsymbol{x} + \overline{\boldsymbol{x}}_A)' - (\boldsymbol{x} - \overline{\boldsymbol{x}}_A)(\boldsymbol{x} - \overline{\boldsymbol{x}}_A)' \\ &= 4\boldsymbol{x}\overline{\boldsymbol{x}}_A' \end{aligned}$$

because other terms cancel out.

For PLS-DA using one component, we have the following.

Since the class sizes are equal $\bar{t}_A = -\bar{t}_B$

$$\begin{aligned} \boldsymbol{t}'\boldsymbol{c} &= \sum_A t_A - \sum_B t_B \\ &= (I/2)(\bar{t}_A - \bar{t}_B) = I\bar{t}_A \end{aligned}$$

where $\sum_A t_A$ represents the sum of scores for group A, and remembering that c has a value of +1 for group A and −1 for group B,

$$\text{so } \boldsymbol{c} = I\boldsymbol{t}\bar{t}_A / \left(\sum t^2\right)$$

We define $\boldsymbol{H}$, which contains the normalised weight vectors of the PLS components, i.e. $\boldsymbol{h} = \boldsymbol{w}/\sqrt{(\sum w^2)}$ for each component, which when there is one component is a vector.

So because $\boldsymbol{t} = \boldsymbol{Xh}$

$$\boldsymbol{c} = I\boldsymbol{Xh}\bar{\boldsymbol{x}}_A\boldsymbol{h}/\Sigma t^2$$

However, for a one-component centred model

$$\boldsymbol{w} = \left(\sum_A \boldsymbol{x}_A - \sum_B \boldsymbol{x}_B\right)' = (I/2)(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B)'$$
$$= I\bar{\boldsymbol{x}}_A'$$

so $\boldsymbol{h} \propto \bar{\boldsymbol{x}}_A'$

hence, $c \propto \boldsymbol{X}\bar{\boldsymbol{x}}_A'\bar{\boldsymbol{x}}_A\bar{\boldsymbol{x}}_A'$

so $\boldsymbol{c} \propto \boldsymbol{X}\bar{\boldsymbol{x}}_A'$

as $\bar{\boldsymbol{x}}_A\bar{\boldsymbol{x}}_A'$ is a scalar

or for a single sample $c \propto \boldsymbol{x}\bar{\boldsymbol{x}}_A'$

Hence, c obtained from PLS-DA is proportional to D for EDC, and when $c = 0$, $D = 0$ and the boundary or decision threshold is the same.

3.3. Linear discriminant analysis

Linear discriminant analysis can be expressed either in a Bayesian or non-Bayesian form. The former allows prior probabilities to be taken into account. For example, if one knows in advance that there is around 75% chance that a sample belongs to one group rather than the other, this is called the Bayesian prior and can be used as a starting probability for LDA, the experimental observations being used to improve this estimate. However, we will use the non-Bayesian and more classical form later. Section 4 will provide additional comments about Bayesian extensions. As before, we restrict discussions in this section to the case of two equal-sized groups.

Instead of using the Euclidean distance as a measure, the Mahalanobis distance [7,8] is employed. Formally,

$$d_{iA}^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_A)\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_A)'$$

where $\boldsymbol{S}$ is a variance covariance matrix. For LDA, if there are two groups, this is the pooled variance covariance matrix over all groups, i.e.

$$\boldsymbol{S} = ((I_A - 1)\boldsymbol{S}_A + (I_B - 1)\boldsymbol{S}_B)/(I_A + I_B - 2)$$

or for two equal-sized groups

$$\boldsymbol{S} = (\boldsymbol{S}_A + \boldsymbol{S}_B)/2$$

where $\boldsymbol{S}_A$ is the variance covariance matrix for group A and $\boldsymbol{S}_B$ for group B. Note that there are other definitions of $\boldsymbol{S}$, for example, for QDA, it is the variance covariance matrix of the relevant group and differs for each group to be modelled, rather than the pooled variance covariance matrix.

The corresponding LDA plots are presented in Figure 5. Note that the contours are no longer circular, but ellipsoidal. The separator represents the class decision function. The line separating the classes differs from that obtained using the EDC model. They would only be the same if both classes were uncorrelated and with equal variance in all directions.

An advantage of LDA over EDC is that it takes into account the different scales of and correlations between variables. Standardisation usually puts variables on a similar scale, and so there is less advantage under these circumstances. A traditionally cited disadvantage of LDA is that the number of variables needs to be less than the number of samples. However, identical results are obtained if LDA is performed on PC scores so long as all nonzero PCs are used in the model, as this just represents a rotation. Hence, if there are more variables than samples, it is simply necessary to perform PCA first and then retain all nonzero PCs. Hence, LDA in practice can be applied when the number of variables exceeds the number of samples.

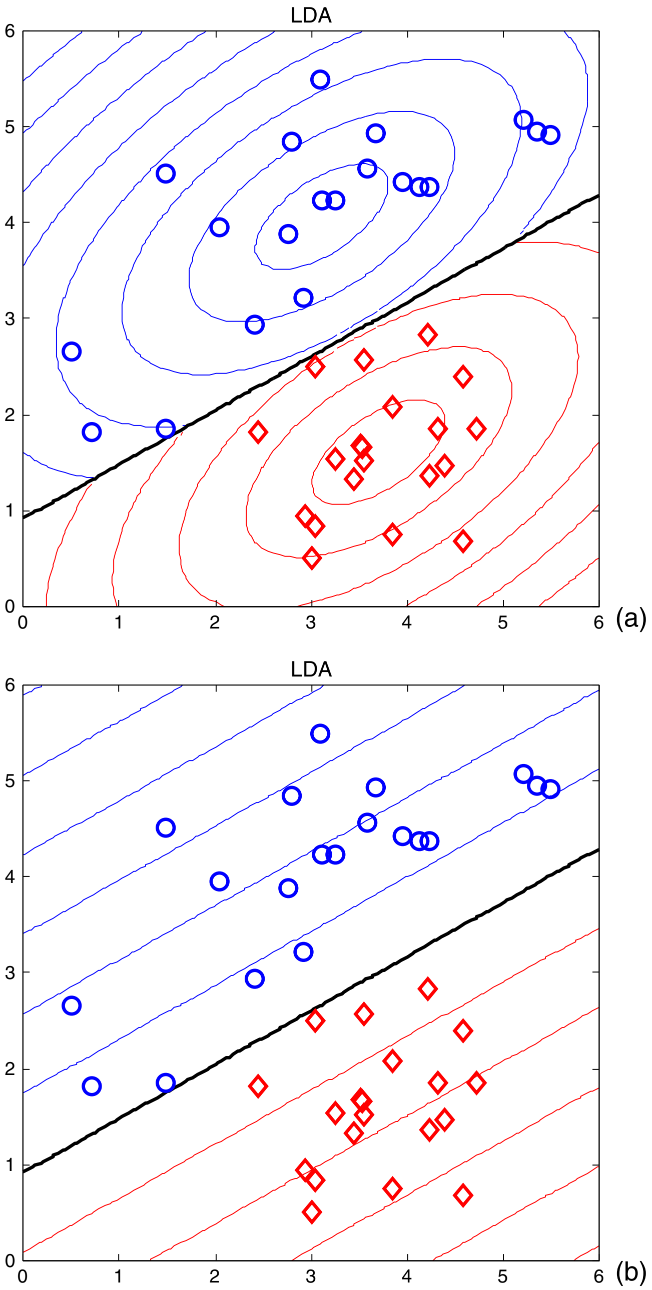


Figure 5. Illustration of linear discriminant analysis (LDA). (a) Contours of the squared distance from the centroid of two groups, with the equidistant separator indicated in bold and black; (b) contours of the difference between squared distances with the separator indicated as in (a).

A significant disadvantage of LDA is that it does not take into account the differing variance structures of each group, for example, one group may be more dispersed than the other one. For a dispersed group, a relatively large distance from a mean may be less significant than for a compact group. LDA uses a single pooled variance covariance matrix and therefore is not always appropriate if the variance structure differs for two or more groups. Under these circumstances, it is often preferable to use nonlinear or multilinear methods, for example QDA, which allow different structures for each group; however, it is important to recognise that PLS-DA as usually implemented is a linear method and does not therefore take into account situations in which each group has very different structures.

3.4. Partial least squares discriminant analysis with all nonzero components

In our case, there can be no more than two nonzero components. The PLS-DA model with two components is illustrated in Figure 6. It can be seen that the separator is the same as for LDA; hence, PLS-DA with all nonzero components has the same disadvantages as LDA. The only simplification computationally is that there is no need to perform PCA prior to PLS if the number of variables exceeds the number of samples. However, classification performance is identical for both methods using the approach described in this article.

The relationship is illustrated in Figure 7. Note that for two variables, we can only have two PLS models, but when the number of variables is increased, there will be intermediate situations. For brevity, we do not review these as they are hard to visualise. However, EDC and LDA represent well-established statistical approaches with well-understood properties.

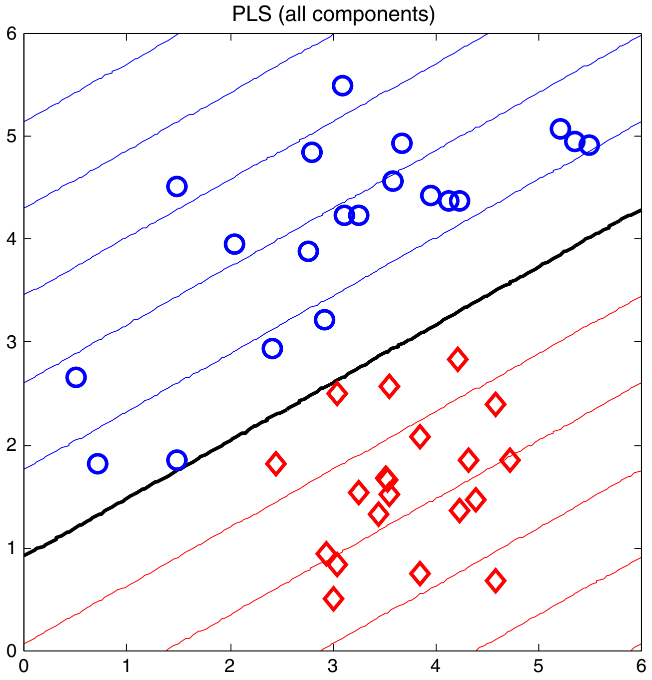


Figure 6. Estimated values of c for a two-component partial least squares (PLS) discriminant analysis model. Blue = positive and red = negative, with $c = 0$ as separator. When there are two variables, this represents all nonzero PLS components.

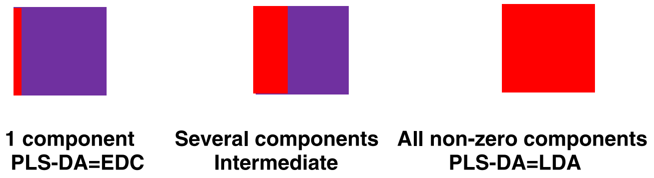


Figure 7. Relationship between partial least squares discriminant analysis (PLS-DA) and common statistical approaches when there are equal class sizes and the $\boldsymbol{X}$ matrix is centred. EDC, Euclidean distance to centroids; LDA, linear discriminant analysis.

The difference between the boundaries for a one-component PLS (or EDC) model and a full PLS (or LDA) model is illustrated in Figure 8. A model with all nonzero components is usually more appropriate to one with one component, especially if each variable has quite different characteristics. However, if each class has very different variance structure, it is sometimes preferable to stick with a simpler model, as LDA or PLS-DA with all nonzero components assumes that each class has a similar structure: this is inevitable as PLS is by origin a method for calibration and by default assumes all measurements are equally significant.

3.4.1. Proof that linear discriminant analysis and a full partial least squares discriminant analysis model provide the same classification model

For LDA, define D, which is the difference between the squared Mahalanobis distance using the pooled variance covariance matrix from the mean of group B minus that of group A. A positive value of D is indicative of a member of class A as the distance to the mean of group B is greater than to the mean of group A, analogous to c.

$$\begin{aligned} D &= (\boldsymbol{x} - \bar{\boldsymbol{x}}_B)\boldsymbol{S}_{\text{pooled}}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_B)' - (\boldsymbol{x} - \bar{\boldsymbol{x}}_A)\boldsymbol{S}_{\text{pooled}}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_A)' \\ &= (\boldsymbol{x} + \bar{\boldsymbol{x}}_A)\boldsymbol{S}_{\text{pooled}}^{-1}(\boldsymbol{x} + \bar{\boldsymbol{x}}_A)' - (\boldsymbol{x} - \bar{\boldsymbol{x}}_A)\boldsymbol{S}_{\text{pooled}}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_A)' \\ &= 4\boldsymbol{x}\boldsymbol{S}_{\text{pooled}}^{-1}\bar{\boldsymbol{x}}_A' \end{aligned}$$

However, the pooled variance covariance matrix can be simplified because $\bar{\boldsymbol{x}}_A = -\bar{\boldsymbol{x}}_B$, and $I_A = I_B = I/2$ and is the average of the variance covariance matrices of each group, providing the overall data matrix is centred for simplicity (this makes no

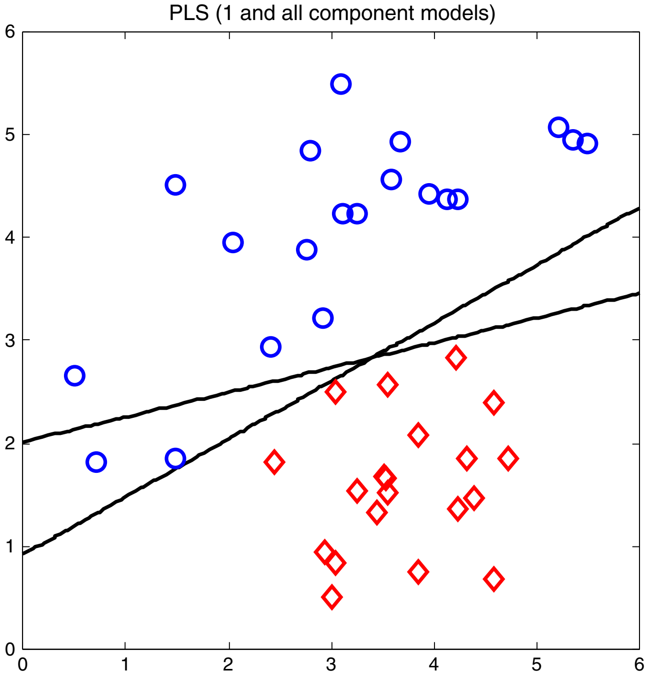


Figure 8. Difference between the one-component and two-component partial least squares (PLS) models.

difference but reduces the number of terms in the derivation).

The formal definition when both groups are of equal size is given by

$$\boldsymbol{S}_{\text{pooled}} = (\boldsymbol{S}_A + \boldsymbol{S}_B)/2 \\ = \left((\boldsymbol{X}_A - \boldsymbol{1}\overline{\boldsymbol{x}}_A)'(\boldsymbol{X}_A - \boldsymbol{1}\overline{\boldsymbol{x}}_A)\right) + \left((\boldsymbol{X}_B + \boldsymbol{1}\overline{\boldsymbol{x}}_A)'(\boldsymbol{X}_B + \boldsymbol{1}\overline{\boldsymbol{x}}_A)\right)/ \\ (2(I/2 - 1))$$

using the sample variance where $\boldsymbol{X}_A$ represents the samples from group A and $\boldsymbol{1}$ is a unit vector.

Take note that

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}_A'\boldsymbol{X}_A + \boldsymbol{X}_B'\boldsymbol{X}_B$$

The equation can be simplified, in the special case that there are equal numbers of samples in each group and the overall data matrix is centred as follows

$$\boldsymbol{S}_{\text{pooled}} = \left(\boldsymbol{X}'\boldsymbol{X} - I\overline{\boldsymbol{x}}_A'\overline{\boldsymbol{x}}_A\right)/(2I - 2)$$

because other terms cancel out.

For PLS-DA using all nonzero components, we have the following derivation.

Because the class sizes are equal, $\overline{\boldsymbol{t}}_A = -\overline{\boldsymbol{t}}_B$, where these are vectors, as opposed to scalars for a one-component model, providing the overall $\boldsymbol{X}$ matrix has been centred first. Hence,

$$\boldsymbol{T}'\boldsymbol{c} = (I/2)(\overline{\boldsymbol{t}}_A - \overline{\boldsymbol{t}}_B) = I\overline{\boldsymbol{t}}_A'$$

$$\text{so } \boldsymbol{c} = I\boldsymbol{T}\left(\boldsymbol{T}'\boldsymbol{T}\right)^{-1}\overline{\boldsymbol{t}}_A'$$

But because $\boldsymbol{T} = \boldsymbol{XH}$

$$\boldsymbol{c} = I\boldsymbol{XH}\left(\boldsymbol{H}'\boldsymbol{X}'\boldsymbol{XH}\right)^{-1}(\overline{\boldsymbol{x}}_A\boldsymbol{H})'$$

However if all nonzero components have been determined and the number of components equals the number of variables, $\boldsymbol{H}$ is a square matrix. Note that if there are fewer nonzero components than variables, the dataset could be first reduced using PCA. Under such circumstances,

$$\boldsymbol{HH}^{-1} = \boldsymbol{I}$$

where $\boldsymbol{I}$ is the unit matrix: note that this is only valid when $\boldsymbol{H}$ is a square matrix and so all nonzero components have been found. This simplifies the preceding equation to

$$\boldsymbol{c} = I\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\overline{\boldsymbol{x}}_A'$$

or for a single sample

$$c = I\boldsymbol{x}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\overline{\boldsymbol{x}}_A'$$

However, the value of D for LDA is given by

$$D = 4\boldsymbol{x}\boldsymbol{S}_{\text{pooled}}^{-1}\overline{\boldsymbol{x}}_A' \\ = 4\boldsymbol{x}\left(\boldsymbol{X}'\boldsymbol{X} - I\overline{\boldsymbol{x}}_A'\overline{\boldsymbol{x}}_A\right)^{-1}(2I - 2)\overline{\boldsymbol{x}}_A' \\ \propto \boldsymbol{x}\left(\boldsymbol{X}'\boldsymbol{X} - I\overline{\boldsymbol{x}}_A'\overline{\boldsymbol{x}}_A\right)^{-1}\overline{\boldsymbol{x}}_A'$$

At this stage, we need a lemma [9] that states that if matrix $\boldsymbol{A}$ is invertible (as is $\boldsymbol{X'X}$) and matrix $\boldsymbol{B}$ is of rank 1 (which is the case for $\overline{\boldsymbol{x}}_A'\overline{\boldsymbol{x}}_A$), then

$$(\boldsymbol{A} + \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} - \frac{1}{1+g}\boldsymbol{A}^{-1}\boldsymbol{BA}^{-1} \\ = \boldsymbol{A}^{-1}\left(I - \frac{1}{1+g}\boldsymbol{BA}^{-1}\right) \\ \propto \boldsymbol{A}^{-1}$$

where g is the trace of $\boldsymbol{BA}^{-1}$.

Hence, it can be seen that c obtained from PLS-DA using all nonzero components is proportional to D obtained using LDA, and so the boundary corresponding to $D = 0$ and $c = 0$ is the same.

4. PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS FOR UNEQUAL CLASS SIZES

When the group sizes are unequal, using the PLS-DA method described earlier unchanged will not usually result in the most appropriate decision boundary. Figure 9(a) shows the boundary obtained when the $\boldsymbol{X}$ matrix is column centred for a model based on all nonzero components; a similar result is obtained using a one-component model. This boundary is shifted towards the larger group and so misclassifies many samples from this group. The reason is that centring shifts the centre of gravity towards the larger group and so a decision threshold of $c = 0$ is no longer appropriate. For LDA and EDC, this is not a problem as each centroid is of equal significance; however, many samples are in each group.

The solution to this is to weight centre the $\boldsymbol{X}$ matrix for PLS by subtracting the average of the means of the two groups, that is, $(\overline{\boldsymbol{x}}_A + \overline{\boldsymbol{x}}_B)/2$, from the columns. The columns are no longer centred; however, the centre of gravity for $\boldsymbol{c}$ is now the same as for $\boldsymbol{X}$. The result is presented in Figure 9(b). The new boundary is the same as that obtained for LDA or EDC (for one PLS component). Hence, by changing the column means, the position of the PLS-DA boundary is shifted as illustrated in Figure 10. There are several alternatives that give similar results, but it is always necessary to understand how the columns have been shifted and what decision threshold is used for c without which the results of PLS-DA may not be appropriate. Many users of packaged software do not understand this and therefore are in danger of using inappropriate models unless each group is of equal size and rarely state these details in presentations.

The relationship between the values of c for the mean-centred and weighted-centred models is as follows.

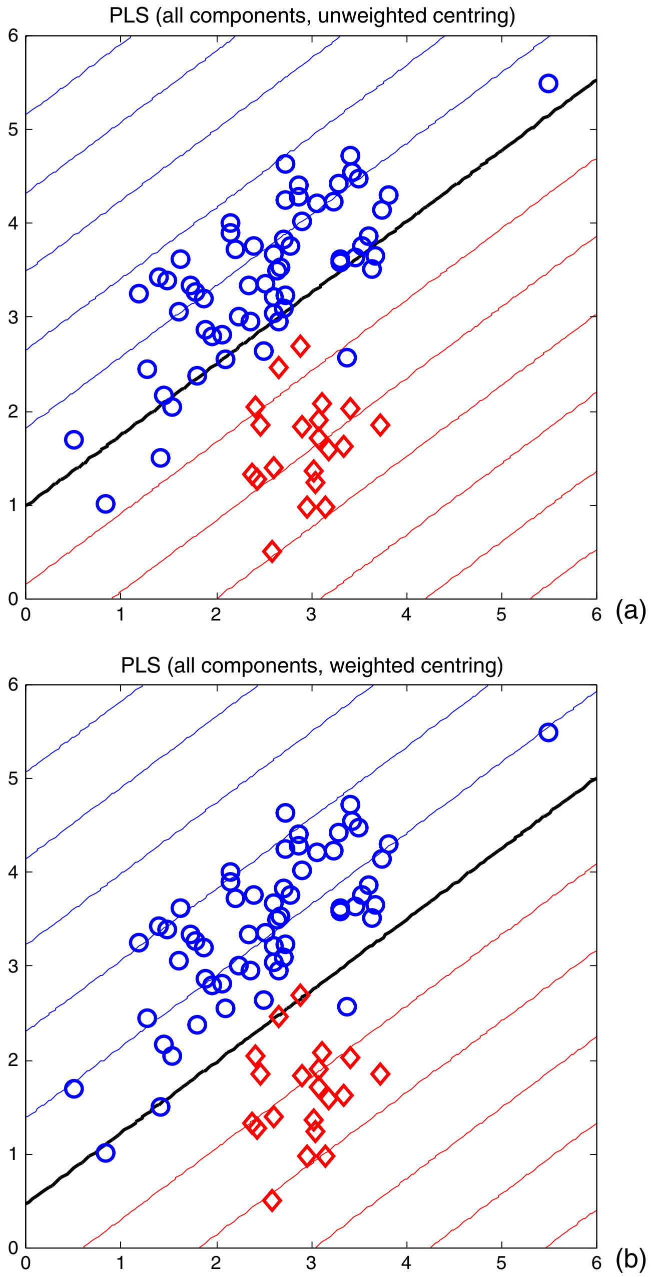
We will determine the value of c for the mean-centred model that corresponds to $c = 0$ for the weighted model.

Define $N = I_A/I_B$. For equal class sizes, this is 1.

For weighted centring, we note that $\overline{\boldsymbol{x}}_A = -\overline{\boldsymbol{x}}_B$ but that the overall mean $\overline{\boldsymbol{x}}$ is nonzero unless the group sizes are equal.

Hence,

$$\overline{\boldsymbol{x}} = \left(I_A\,\overline{\boldsymbol{x}}_A + I_B\overline{\boldsymbol{x}}_B\right)/(I_A + I_B) \\ = \left(NI_B\overline{\boldsymbol{x}}_A + I_B\overline{\boldsymbol{x}}_B\right)/(I_A + I_B) \\ = \left(NI_B - I_B\right)\overline{\boldsymbol{x}}_A/(NI_B + I_B) = \frac{(N-1)}{(N+1)}\overline{\boldsymbol{x}}_A$$


PLS (all components, unweighted centring)

PLS (all components, weighted centring)

Figure 9. Partial least squares (PLS) discriminant analysis boundaries (bold black line) for two groups with unequal class sizes (a) centring the ***X*** matrix or (b) using weighted centring on the ***X*** matrix and full-component models.

The predicted ***c*** for the mean of the dataset is given by

$$\hat{\hat{c}} = \bar{\boldsymbol{x}}\boldsymbol{b} = \frac{(N-1)}{(N+1)}\bar{\boldsymbol{x}}_A\boldsymbol{b} = \frac{(N-1)}{(N+1)}\hat{\hat{c}}_A$$

However, we would like the decision threshold to be 0, so that positive values of ***c*** correspond to members of group A and negative values to group B, so the threshold is shifted for the unweighted model relative to the weighted model, by $\frac{(N-1)}{(N+1)}\hat{\hat{c}}_{\mathbf{A}}$. If $N = 1$, there is no shift, and the two values are equal. As the proportion of samples in group A increases, the unweighted threshold is shifted away from the weighted threshold towards the centre of class B.

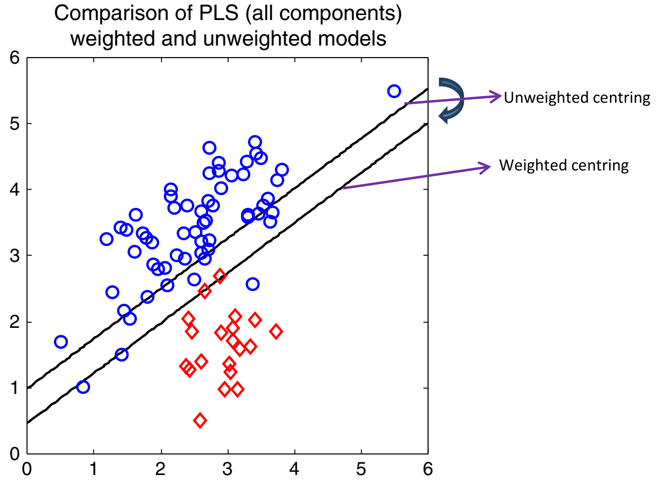

Comparison of PLS (all components)
weighted and unweighted models

Figure 10. Change in the boundary $c = 0$ when the ***X*** matrix is weight centred. PLS, partial least squares.

Hence, when there are unequal class sizes, it is necessary to adjust the column means for ***X*** to obtain an appropriate separator. There are of course other ways, for example by mean centring ***c***, which is equivalent, but the cut-off threshold is no longer 0 and should be halfway between the means of the two groups. Hence, when the group sizes are unequal, it is essential to understand how the columns of the data matrices have been processed and what decision threshold is suitable. If this has not been performed correctly, the results of prediction may be misleading unless the groups are very well separated, in which case almost any method will work and it is best to use a simpler approach.

For EDC, there is no such problem as the mean of each group is of equal importance, no matter how many samples characterise the group.

For LDA, many statisticians also include an additional probabilistic term [8] in addition to the classical equation. It is not the purpose of this paper to discuss probabilistic models in detail, but this is often called a Bayesian approach. The additional term relates to the probability that a sample is a member of each group. The default is that the probability of membership of each class is 0.5 when there are two groups, and under such circumstances, the results described earlier are derived. The probabilistic term is usually called a Bayesian prior and represents the prior probability of class membership. For example, we may know that only a 10th of samples belong to one of two categories; therefore, before we perform the measurement, the probability of belonging to group A is 0.1 and that to group B is 0.9.

In our example, we may want to investigate what happens if we use the relative group sizes as our prior probabilities. If one third of the samples belong to group A and two thirds to group B, then we could construct the model using prior probabilities of 0.33 and 0.67. This has the effect of shifting the position of the separator in LDA (Figure 11). The separator is shifted away from the centroid of the largest group, rather than towards it (as is the case for unweighted centring). To increase the probability of a sample belonging to a group, the relative size of the groups should increase rather than decrease. PLS-DA with a centred ***X*** block has the opposite effect of moving the separator closer to the centroid of the largest group because the centre of gravity of the data is closer to this group and is intuitively incorrect.

An advantage of LDA over PLS-DA is that prior probabilities can easily be incorporated, if required, so prior knowledge can

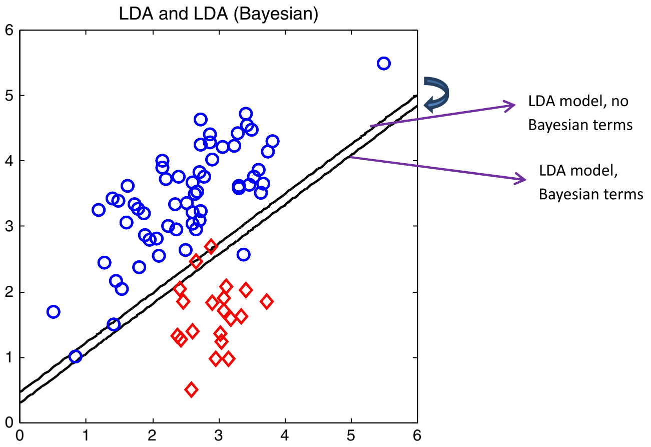


Figure 11. Bayesian example: the prior probabilities are equal to the relative group sizes. LDA, linear discriminant analysis.

be added to the model where necessary. This is not easy in PLS, which is usually expressed algorithmically.

5. MULTIGROUP PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS AND COMPARISON WITH LINEAR DISCRIMINANT ANALYSIS AND EUCLIDEAN DISTANCE TO CENTROIDS

When there are more than two groups, it is traditional to extend the PLS-DA model so that ***c*** instead of being a vector becomes a matrix ***C***. Each column represents a group or class. Each sample is considered to be a member of the relevant class ($c = +1$) or not ($c = -1$). Hence, if there are three groups A, B, and C, the second column of ***C*** would represent class membership of group B; samples from groups A and C will have values of $c = -1$, and samples from group B values of $c = +1$. We will call the class denoted by $c = +1$ as the 'in-group' representing one of the original groups and the class denoted by $c = -1$ as the 'out-group'. The principle is illustrated in Figure 12.

There are two fundamental approaches. The most usual is to perform three PLS1-DA models, one for each column, in what is often called a one-versus-all approach. For brevity, below we illustrate only the models involving all nonzero components corresponding to LDA. Similar results can be obtained for one-component PLS-DA models corresponding to EDC.

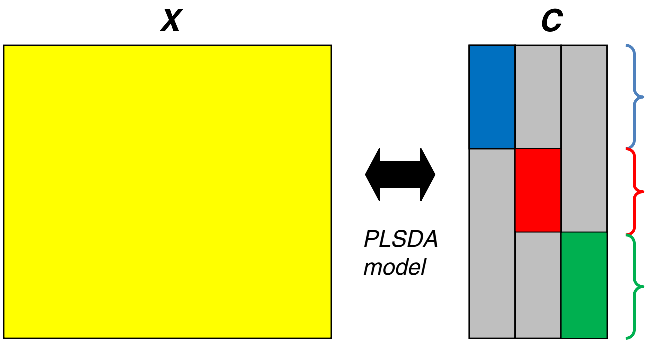


Figure 12. Partial least squares discriminant analysis (PLS-DA) model for three groups.

For three groups, there will be three PLS1-DA models. The aim is to perform a PLS-DA model between the corresponding column of ***C*** and the ***X*** matrix. The relative group sizes will be unequal in at least two out of the three models and, usually, all three (for three classes). If the number of samples in each of the original groups is equal, then in all cases, the ratio of samples in the in-group to that in the out-group will be 1: 2. To reflect this, we perform weighted centring (Section 4) for each of the PLS-DA models. Other combinations could be envisaged, but broadly similar conclusions will be obtained.

The three PLS-DA models together with the corresponding LDA models are illustrated in Figure 13 and are as expected, the same. However, a problem occurs when we superimpose the PLS-DA or LDA separators, we find that they do not divide the data space into three clear regions, and so there is ambiguity for many samples, as illustrated in Figure 14.

This is because the three separators do not intersect, which is a consequence of the weighted centring. The average of the mean in-group and out-group is different for each of the three models. Even if ***X*** were mean centred, the models would intersect at the average of the overall dataset, which in itself would not be satisfactory.

The problem of ambiguous models in multiclass PLS1-DA is usually overcome by developing elaborate decision rules as to class membership of each sample. For example, if sample A has a predicted value of $c = 0.8$ for a model of group A against the rest, $c = 1.4$ of a model of group B against the rest and $c = -0.5$ for a model of group C against the rest, it is usually assigned to group B as this represents the most positive value, but some methods might assign it to group A because it is closest to 1. These assignments are ambiguous especially because relative group sizes may differ for each model and so will the pooled variance covariance matrix, so there is in practice no really satisfactory agreed universal decision rule. It is important to recognise that classification methods differ from each other in performance where the answer is not completely straightforward. If all three groups were well separated, almost any simple method would work. Hence, there are many difficulties in employing a one-versus-all PLS1-DA model and comparing it with other approaches.

For LDA or EDC, this is not a serious limitation, as it is easy to obtain a three (or more)-group model, the LDA three-group model for our example is illustrated in Figure 15. It is simply necessary to determine the Mahalanobis distance to the centroid of each group and choose the group that a sample is nearest to. This results in an unambiguous classification and so overcomes the limitations of one-versus-all PLS1-DA.

An alternative to PLS1 is to use PLS2 [5,6]. It is not the purpose of this paper to expand on the PLS2 algorithm, but in brief, the *c* block is treated as a single matrix, so that there is a single calculation. There are many difficulties with PLS2-DA, which is rarely successful. An important problem relates to the difficulty of satisfactorily transforming the columns of ***X***, because each column of ***C*** would require a different type of weighted centring, but ***X*** can be transformed only once. Furthermore, PLS2-DA assumes that there are interactions between columns of the ***C*** matrix, which can cause difficulties. If two columns of ***C*** are known, the third is dependent on these columns. The problems are too numerous to discuss in detail in this paper. Yet there are very regular reports of the use of PLS-DA (either PLS1 or PLS2) in the literature when there are more than two groups in the data, often with a very detailed comparison of performance against other methods. It is unlikely that the majority of authors of these

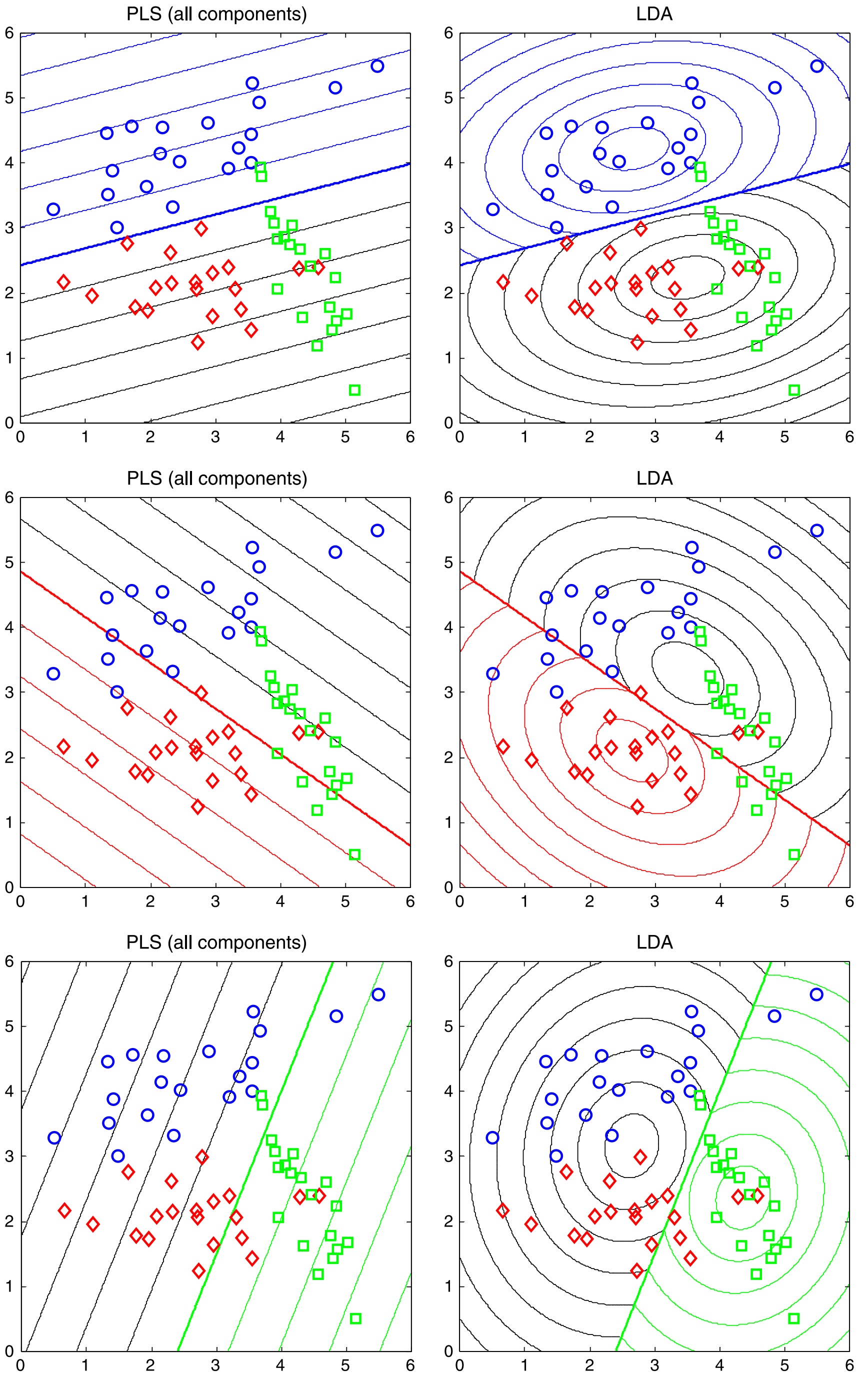


Figure 13. Partial least squares discriminant analysis (PLS-DA) and linear discriminant analysis (LDA) models for three groups characterised by two variables using weighted centring and all nonzero partial least squares discriminant (PLS) components (left PLS-DA models and right one-vs-all LDA models).

papers have much understanding of the pitfalls in their results, and indeed, many do not even distinguish between PLS1 and PLS2 in their papers.

6. PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS SCORES PLOTS AND PREDICTIONS FOR TRAINING SETS

It is very common to present PLS-DA scores plots, usually by plotting the scores of PLS component 2 versus component 1. PCA scores plots are well established, and there are a very large numbers of papers published reporting data this way: PCA [10] is a valuable approach for data visualisation, especially in cases where there are many more variables than samples and can be used to look at an entire set of samples.

Partial least squares discriminant analysis scores plots on an entire set of samples can, in contrast, be highly misleading, especially if the number of variables far exceeds the number of samples. We will use a simple example consisting of a matrix of dimensions 40 × 200 consisting of random numbers generated using a uniform distribution between −1 and +1: for the purpose

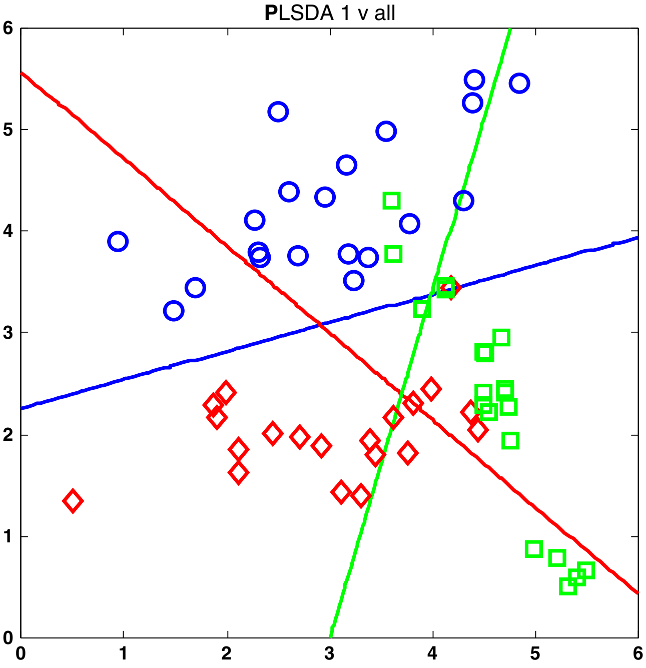


Figure 14. Superimposing the three partial least squares discriminant analysis (PLS1-DA) separators obtained in Figure 13.

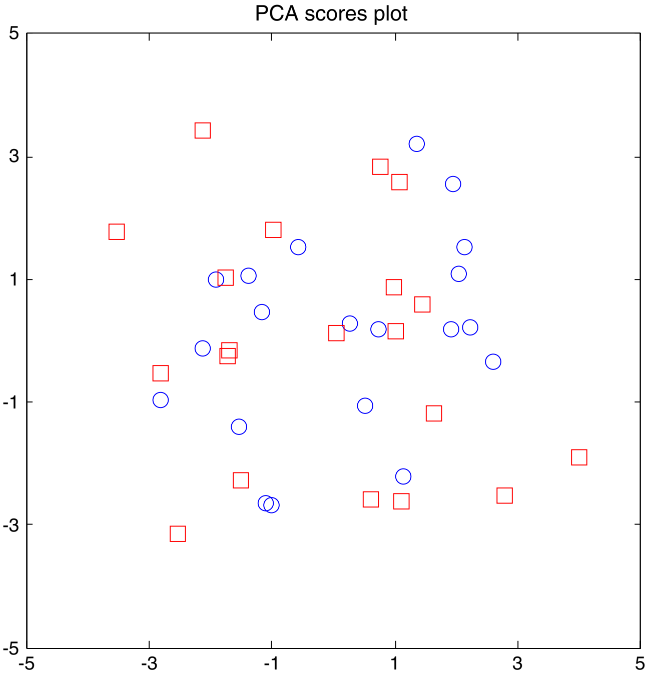


Figure 16. Principal component analysis (PCA) scores plot, PC1 horizontal and PC2 vertical of a randomly generated dataset consisting of 40 samples each characterised by 200 variables, divided arbitrarily into two groups.

of this paper, the precise method of generating these random numbers is not important. Of the 40 rows, 20 are assigned to group A and 20 to group B. There should be no significant difference between these groups. These wide datasets are quite common especially in metabolomics studies where samples are often difficult to obtain but a large number of variables such as metabolites, gas chromatography–mass spectrometry peaks, or nuclear magnetic resonance signals can be measured.

The PCA scores plot of the first two PCs (mean-centred data) is presented in Figure 16. There is no particular distinction between the two groups, although some clumping of samples can be observed: it is important to understand that randomness is not the same as uniformity. If we tossed an unbiased coin several times and we obtained a sequence HTHTHTHT (where H = heads and T = tails), this sequence is extremely unlikely to occur randomly.

When we perform PLS-DA and plot the scores of the first two PLS components against each other, we obtain the pattern shown in Figure 17, which falsely suggests there is an excellent

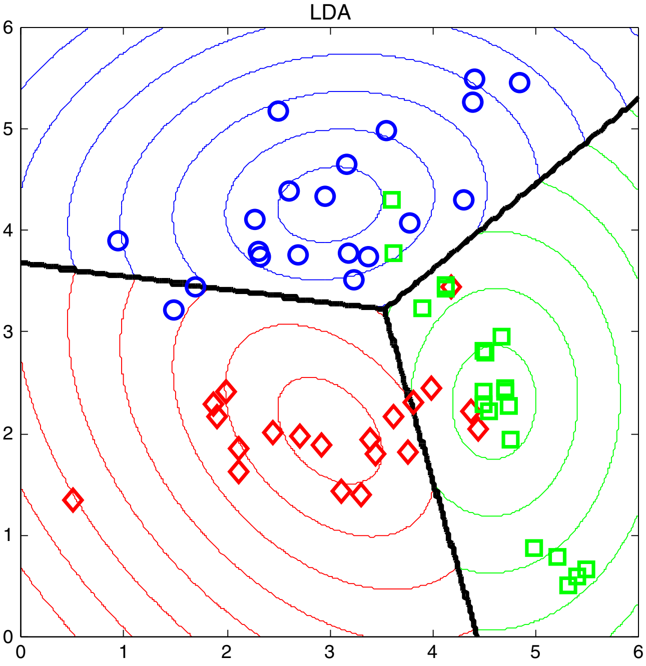


Figure 15. Linear discriminant analysis (LDA) three-group model.

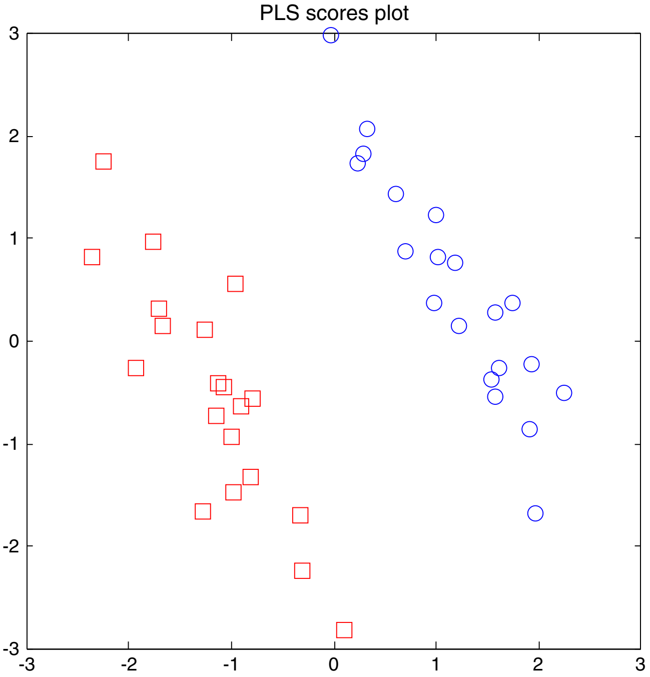


Figure 17. Partial least squares (PLS) discriminant analysis scores plot, centred data, corresponding to the dataset of Figure 16.

separation between the groups. The problem is that because there are so many variables, there are some correlations just by chance. Consider tossing an unbiased coin 10 times. Sometimes, there will be 9H or 2H as well as 5H, as illustrated in Figure 18. If this experiment were repeated 1000 times, there will be many cases of seven or more Hs. A clever classification algorithm might pick those cases where there are, for example, seven or more Hs and ignore or reduce the significance of the other situations and then might conclude that the coin is biased. PLS-DA by analogy looks for the variables that correlate best with the classifier. Even if, for example, a set of metabolites has an underlying even distribution between two groups of samples, if 1000 metabolites are analysed and there are, for example, just 10 samples in each of two groups, by chance, there will appear to be an uneven distribution of a few metabolites, just as an unbiased coin will sometimes show more Hs than Ts. It is therefore possible to find these, which would have a high weight or loading for the more significant PLS components and form a model for which it appears that the two groups are separated. When the number of variables strongly exceeds the number of samples, this is easy. For randomly generated data, we expect it, and it would be surprising if this did not happen. A consequence is that the predictions appear to be very good also, with samples classified correctly into their respective groups using the most common criteria as illustrated in Figure 19.

When classification methods are used as black boxes, or part of a very complex protocol, it is easy to make the mistake of producing over-optimistic models or overfitting. An expert would know that a way to avoid this is to split samples into training and test sets, but even then, there can be problems, for example, variable reduction might be performed on the raw data, leading to 'bad variables' being removed from the model and the test set appearing to show a separation. There are a number of solutions. The first is to create a null or random dataset and follow all the methods through using this dataset [3,11], to see if there is any separation, and if this looks significant, the methodology almost certainly overfits the model. The second is to permute the classifier so that a group label is randomly attached to each sample to try to destroy the structure [12]. Usually, a large number of permutations are necessary, and the results from the unpermuted dataset are compared with those from the ensemble of permutations.

But above all, the danger is the presentation of PLS-DA scores plots on training sets: yet there is no software that prevents this; indeed, users of packages would not purchase the software if they were not allowed to show PLS-DA scores plots on training sets. And many investigators who are nonexperts in the chemometrics field often prefer to show desired patterns, even if they are false ones, and will usually choose the PLS training set scores plot over the PCA scores plot, if it suits their preconception. It is hard for the data analysis expert to withhold graphs and often even harder to explain to their colleagues that the graphs they are showing have little meaning and persuade their colleagues to publish graphs that appear inconclusive instead.

Hence, PLS-DA scores plots, while widely used, can in the wrong hands be very dangerous.

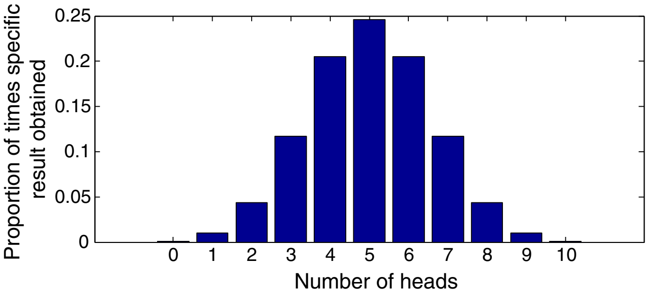


Figure 18. Expected spread of results for the toss of an unbiased coin 10 times.

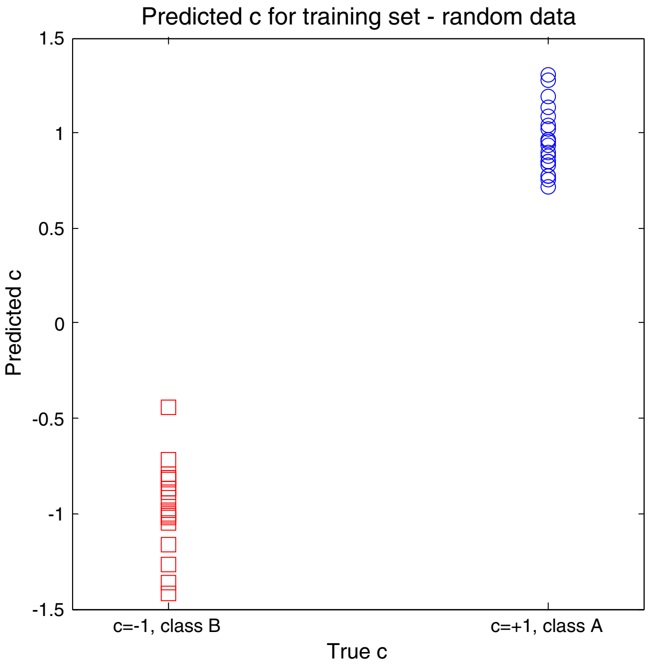


Figure 19. Predicted values of *c* for the two groups illustrated in Figure 16.

7. CONCLUSIONS

Partial least squares discriminant analysis is often regarded as a method in its own right. There are numerous papers and conference presentations where its use is described with no further elaboration. However, as we have shown, even in the simplest of cases, PLS-DA can perform like EDC or like LDA, both classical methods but regarded as quite distinct approaches with different underlying assumptions about the structure of the data and well-recognised statistical properties. As a minimum, the PLS-DA algorithm should be regarded just as one building block in a series of steps used to develop a classification procedure, including for example, validation, preprocessing, variable selection, and so on. Most equations contain several terms, and all are necessary to get to an answer, so by analogy PLS-DA should be regarded as just one step in a full classification procedure. All presentations and papers should at the minimum describe how the data had been preprocessed and what decision rules have been used.

But a danger is that PLS-DA is now strongly engrained in most commercial chemometrics software. Although expert users will understand the pitfalls, it is likely that only a very small fraction of nonexperts do. Many methods currently in use in chemometrics were first developed by experts in data analysis, who would safely evaluate the use of their procedures. But with the widespread availability of software, very few users have this knowledge base. There are many inherent decisions needed to get to an answer using PLS-DA such as the choice of threshold, the use of PLS1 or PLS2 when there are several classes and column centring that are critical steps in the analysis and can make radical differences to the result. Yet there are numerous papers in the literature that compare the performance of, for example, PLS-DA, without specifying adequate details, with other classification approaches: most of these papers are of limited value.

Partial least squares discriminant analysis is sold as a technique because it can provide what appears to be a very effective solution to classification, especially graphical presentations such as scores plots, whereas more traditional statistics do not have this flexibility. However, with the vast increase in the use of such approaches in areas such as medicine and biology, the possibilities of major disasters are high, particularly when many studies involve small sample sizes and large numbers of variables. PLS-DA is also often presented as a method that can cope with large variable-to-sample ratios, yet LDA with prior reduction of dimensionality can perform just as well. Simpler statistical approaches are more robust in that there are less decisions to be made and so less opportunities of overfitting or of misleading classification results and, furthermore, have well-known properties that can be related back to the data.

Hence, for classification, there are few reasons to use PLS-DA, despite its widespread availability. However, this paper has focussed on the use of PLS methods for classification: they were originally developed for calibration, and the extension to classification methods was perhaps misunderstood. An advantage of PLS methods is that they can provide insight into the variables via weights and loadings. LDA and EDC do not easily relate the classifier to the underlying variables. They originated in the work of Fisher and colleagues [13] in the 1930s, where typically two or three variables might be measured for a dataset: the classic iris dataset contains just three variables. In modern applications such as metabolomics, we do not only want to decide whether a sample belongs to one of a number of known groups, but which variables, often related to chemicals, are best discriminators. Here, PLS can be advantageous as direct information about the variables is available. As an exploratory graphical method that suggests which variables are most likely to be responsible for discrimination, used often in exploratory studies, PLS-DA has a significant role to play.

However, PLS-DA has little advantages over existing approaches as a method for discrimination, and it is often alarmingly misused by nonexperts who encounter this as an option in a variety of common chemometrics packages, without understanding its underlying strengths and weaknesses. In a previous era where the majority of users of chemometrics techniques were experts, the method would have been used with care, but current usage has put this widely available method into serious disrepute. A method that was developed as a good exploratory tool for experts is not necessarily appropriate as a widespread approach for nonexperts. Would we trust a competent car driver to pilot an airplane without specialist training? The amount of experience an expert in data analysis has may be greater than that required for an airplane pilot, or even a professional magician. Amateur magicians may not get their tricks right and need extensive experience and knowledge if they wish to practice as professionals.

Many methods that have since become incorporated into packages were first developed by specialists who had substantial expertise and understanding of the underlying maths. The advocacy of these methods has made some such as PLS widespread, but with the unintended consequences that the user often has limited understanding of how to safely apply the methods, and hence, an uncontrolled literature can build up in which quite dubious applications are reported and widely believed.

Acknowledgements

G. R. L. gratefully acknowledges the Gloucestershire Hospitals NHS Foundation Trust for funding.

REFERENCES

1. Barker W, Rayens W. Partial least squares for discrimination. *J. Chemom.* 2003; **17**: 166–173.
2. Gottfries J, Blennow K, Wallin A, Gottfries CG. Diagnosis of dementias using partial least squares discriminant analysis. *Dementia* 1995; **6**: 83–88.
3. Brereton RG. *Chemometrics for Pattern Recognition*. John Wiley and Sons: Chichester, 2009.
4. Dixon SJ, Brereton RG. Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure. *Chemom. Intell. Lab. Syst.* 2009; **95**: 1–17.
5. Geladi P, Kowalski BR. Partial least squares: a tutorial. *Anal. Chim. Acta*, 1986; **185**: 1–17.
6. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 2001; **58**: 109–130.
7. Mahalanobis PC. On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India* 1936; **2**: 49–55.
8. Friedman JH. Regularized discriminant-analysis. *J. Am. Stat. Assoc.*. 1989; **84**: 165–175.
9. Miller KS. On the inverse of the sum of matrices. *Math. Mag.* 1981; **54**: 67–72.
10. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom. Intell. Lab. Syst.* 1987; **2**: 37–57.
11. Brereton RG. Consequences of sample sizes, variable selection, model validation and optimisation for predicting classification ability from analytical data. *TrAC* 2006; **25**: 1103–1111.
12. Wongravee K, Lloyd GR, Hall J, Holmboe M, Schaefer ML, Reed RR, Trevejo J, Brereton RG. Monte-Carlo methods for determining optimal number of significant variables. Application to mouse urinary profiles. *Metabolomics* 2009; **5**: 307–406.
13. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 1936; **7**: 179–188.