Special issue article

Journal of CHEMOMETRICS

Received: 10 December 2015, Revised: 31 December 2015, Accepted: 06 January 2016, Published online in Wiley Online Library: 25 February 2016

(wileyonlinelibrary.com) DOI: 10.1002/cem.2779

Re-evaluating the role of the Mahalanobis distance measure

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

It is shown that the sum of squares of the standardised scores of all non-zero principal components (PCs) equals the squared Mahalanobis distance. A new distance measure, the reduced Mahalanobis distance, is explored in which the number of PCs retained is less than the full rank model. It is illustrated by both one-class and two-class classifiers. Linear discriminant analysis can be employed as a soft model, and principal component analysis using the pooled variance-covariance matrix is introduced as an intermediate view between conjoint and disjoint models allowing linear discriminant analysis to be used on these reduced rank models. By choosing the most discriminatory PCs, it can be shown that the reduced Mahalanobis distance has superior performance over the full rank model for discriminating via soft models. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: Mahalanobis distance; pooled variance–covariance matrix; linear discriminant analysis; soft models; one-class classifiers; discrimination; principal component analysis

1. INTRODUCTION

The Mahalanobis distance was first introduced in 1936 [1] by the Indian mathematician P.C. Mahalanobis, in a paper that has been cited over 4000 times according to Google Scholar. He was part of a group of statisticians working in the 1920s and 1930s, including R.A. Fisher and H. Hotelling who pioneered many of the concepts in multivariate statistics that we continue to use nowadays. There are, in fact, few concepts in chemometrics that cannot be traced back to that era. The pioneering paper by R.A. Fisher on classification using Iris data [2] also uses what is now recognised as the Mahalanobis distance to define how far an object is from the centre of a distribution.

The key to the Mahalanobis distance is that it takes into account correlations and scales of variables. If we were to take measurements of a series of humans, some of those measurements may be on different scales to others, such as weight and height. Under such circumstances it is advisable to scale these measurements to ensure that their variance is the same. In addition, some measurements may be correlated. For example, the length of the 10 fingers will usually be correlated, and the Mahalanobis distance ensures that new uncorrelated variables are used in the distance measure. The Mahalanobis distance differs from the Euclidean distance [3] in that the distances of objects from the centre of a distribution are along the major axis of variation, as illustrated in Figure 1. There are two objects, which are at equal distances using the Mahalanobis criterion but at different distances using the Euclidean criterion.

The Mahalanobis distance is usually preferred to the Euclidean distance, in that it can cope with different structures of data. It is also directly related to the standard deviation, so when there is one variable, an object one Mahalanobis distance unit from the centre of a group is also one standard deviation away. This can be extended to multivariate distributions so the Mahalanobis distance is the analogy of the standard deviation in more than one dimension and is also called the standard distance [4].

In univariate statistics, we can usually compare distributions and calculate statistics, such as the probability of belonging to a distribution (confidence level) or rejecting the null hypothesis (p value), by computing how many standard deviations an object is from the centre of a distribution, and then using the normal or t distributions [5–7]. When there is more than one dimension, the analogous Mahalonobis distance can be used, for example, for similar purposes, using F or χ^2 statistics [8,9] and the related Hotelling T^2 [10]. Hence, the concept of the Mahalonobis distance is deeply buried within chemometrics [11]. There are, however, generally considered, some limitations to using this distance measure, primarily because the number of objects is usually assumed to be more than the number of variables. For many modern datasets where the possibilities for measuring variables are huge, such as in spectroscopy or chromatography, and sampling is expensive, the object to variable ratio is often less than one. Traditionally, this limitation has been overcome in a variety of ways, most often via variable selection, that is, removing uninformative variables and only using the most diagnostic ones, for example, the MS or NMR peaks that show best discrimination between groups; however, many of these approaches can result in overfitting, as the criteria for variable selection in themselves make assumptions about the dataset.

It can, however, be shown that the squared Mahalanobis distance is the same as of the sum of the squares of all non-zero standardised principal components [12], as we will show

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

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

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

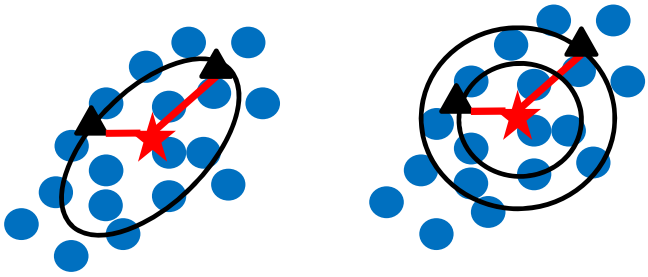

Figure 1. Two objects (black triangles) and the centroid (red star) of a distribution. The left-hand diagram illustrates the Mahalanobis distance to the ellipsoid in the direction of the variance, the right-hand diagram the Euclidean distances of the objects.

algebraically later on. Statisticians well recognised the relationship between principal component analysis (PCA) and the Mahalanobis distance [13,14] but although a few authors have commented on this relationship in the chemometrics literature [11,15,16], this has rarely been widely recognised or exploited in chemometrics. In particular, nearly all papers in the chemometrics literature state that the Mahalanobis distance cannot be calculated when the number of variables exceeds the number of objects, yet this is not correct, as one can perform PCA on a dataset even if the number of variables exceeds the number of objects many fold, so long as there are no perfectly correlated variables, and meaningful results are obtained providing the determinant is not too small numerically. Many traditional statistical applications did not suffer from the problems of modern chemometrics, where variables were expensive to measure, so datasets where the number of variables exceeds the sample size were probably non-existent to statisticians in the formative decades of multivariate analysis and pose a problem very characteristic of chemometrics.

Another consequence of the Mahalanobis distance being related to PCA is that it is not necessary to use a full PC model for determining distances. In almost all other areas of multivariate analysis within chemometrics such as PCA [17] and partial least squares [18], there has been a significant focus over several decades on estimating the appropriate number of components (or dimensionality) for the model. Yet when it comes to the Mahalanobis distance measure, there is no chemometrics literature relating to how many components are appropriate for this type of distance, and which components are most significant. In this paper, we demonstrate the relationship between the Mahalanobis distance and PCA and propose a new measure of the reduced Mahalanobis distance that involves selecting a reduced principal components (PC) model.

2. THEORETICAL BASIS

2.1. Relationship between Mahalanobis distance and principal components scores.

We will define a data matrix $\boldsymbol{X}$ of dimensions $I \times J$ to represent an experimental dataset. We will assume $\boldsymbol{X}$ is mean centred down the columns. This dataset may either be of a single class or of several classes. In this paper, we will discuss both one-class and two-class problems. In this section, we will assume that all samples belong to a single group. The extension to two groups is described in Section 2.3.

The squared Mahalanobis distance from the mean in variable space is defined by

$$\Delta^2 = (\boldsymbol{x} - \bar{\boldsymbol{x}})\boldsymbol{S}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}})' \tag{1}$$

where because $\boldsymbol{X}$ is centred, the variance–covariance matrix is

$$\boldsymbol{S} = \frac{\boldsymbol{X}'\boldsymbol{X}}{I} \tag{2}$$

Note that we use the traditional definition of the population variance rather than the sample variance, as we are using this for calculation rather than estimation, although in practice there is not a large difference.

For PCA, if all non-zero components are used, then

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P} \tag{3}$$

The mean of the dataset is given by the row vector

$$\bar{\boldsymbol{x}} = \bar{\boldsymbol{t}}\boldsymbol{P} \tag{4}$$

The squared Mahalanobis distance of a single object to the mean of the dataset becomes

$$\Delta^2 = (\boldsymbol{t}\boldsymbol{P} - \bar{\boldsymbol{t}}\boldsymbol{P})\boldsymbol{S}^{-1}(\boldsymbol{t}\boldsymbol{P} - \bar{\boldsymbol{t}}\boldsymbol{P})' \quad = (\boldsymbol{t} - \bar{\boldsymbol{t}})\boldsymbol{P}\boldsymbol{S}^{-1}\boldsymbol{P}'(\boldsymbol{t} - \bar{\boldsymbol{t}})' \tag{5}$$

where $\boldsymbol{t}$ is a row vector consisting of the scores of the object over all non-zero components. The variance–covariance is (as $\boldsymbol{X}$ is centred and this is a full component model)

$$\boldsymbol{S} = \frac{\boldsymbol{P}'\boldsymbol{T}'\boldsymbol{T}\boldsymbol{P}}{I} \tag{6}$$

Substitute $\boldsymbol{S}$ into the equation for the Mahalanobis distance:

$$\Delta^2 = (\boldsymbol{t} - \bar{\boldsymbol{t}})\boldsymbol{P}\left(\frac{\boldsymbol{P}'\boldsymbol{T}'\boldsymbol{T}\boldsymbol{P}}{I}\right)^{-1}\boldsymbol{P}'(\boldsymbol{t} - \bar{\boldsymbol{t}})' \tag{7}$$

As all non-zero components (and assuming $I > J$; the case where $I < J$ will be discussed later in this section) are retained, $\boldsymbol{P}$ is square matrix, and because $(\boldsymbol{ABC})^{-1} = \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ (if $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are invertible)

$$\Delta^2 = (\boldsymbol{t} - \bar{\boldsymbol{t}})\boldsymbol{P}\boldsymbol{P}^{-1}\left(\frac{\boldsymbol{T}'\boldsymbol{T}}{I}\right)^{-1}(\boldsymbol{P}')^{-1}\boldsymbol{P}'(\boldsymbol{t} - \bar{\boldsymbol{t}})' \tag{8}$$

The loadings all cancel to give

$$\begin{aligned}\Delta^2 &= (\boldsymbol{t} - \bar{\boldsymbol{t}})\left(\frac{\boldsymbol{T}'\boldsymbol{T}}{I}\right)^{-1}(\boldsymbol{t} - \bar{\boldsymbol{t}})' \\ &= (\boldsymbol{t} - \bar{\boldsymbol{t}})\boldsymbol{S}_{\boldsymbol{T}}{}^{-1}(\boldsymbol{t} - \bar{\boldsymbol{t}})'\end{aligned} \tag{9}$$

where $\boldsymbol{S}_{\boldsymbol{T}} = \left(\frac{\boldsymbol{T}'\boldsymbol{T}}{I}\right)$ is the variance covariance matrix of the principal component scores. Because $\boldsymbol{X}$ is centred, $\bar{\boldsymbol{t}} = \boldsymbol{0}$ so

$$\Delta^2 = \boldsymbol{t}\boldsymbol{S}_{\boldsymbol{T}}{}^{-1}\boldsymbol{t}' \tag{10}$$

For PCA, the scores are orthogonal by definition and so $\boldsymbol{S}_{\boldsymbol{T}}$ is a diagonal matrix (i.e. is all zeros apart from the diagonal

elements). The diagonal elements are the variance of each component:

$$v_a = \frac{\sum_{i=1}^{I}(t_{ia} - \overline{t_a})^2}{I} \quad (11)$$

Hence, the squared Mahalanobis distance becomes

$$\Delta^2 = \sum_{a=1}^{A} \frac{{t_a}^2}{v_a} \quad (12)$$

where $t_{ia}/\sqrt{v_a}$ are standardised (or autoscaled) principal component scores and A is the total number of non-zero PCs.

If a reduced number of PC scores is used, then a reduced Mahalanobis distance is obtained simply by including fewer components in the model. This provides an additional route to calculate the equivalent Mahalanobis distance in variable space when $I < J$; by using all I non-zero components, the PCA residuals are zero and so the reduced distance using all I components is equivalent to the Mahalanobis distance in variable space, which could not be calculated otherwise. It is also possible to select the PCs to include in the model so there is no requirement to take the largest PCs. Note that because all PCs are standardised, each PC has equal numerical influence on the resultant distance, so there is no overwhelming numerical reason for selecting the largest PCs. In linear discriminant analysis (LDA) for example, there is no guarantee that the larger components are the most discriminatory, and including non-discriminatory components may be detrimental to the model.

2.2. Hotelling's *T* squared

When there is more than one variable, the Mahalanobis distance to the centre of a distribution has to be scaled in order to be comparable with the one-dimensional distance (or standard deviation). This allows statistics such as using the F distribution to be employed to determine the significance of a given distance from the average of the data. There are historic reasons for why this scaling is required. Essentially, the F statistic was first most used in hypothesis tests, such as ANOVA [19] to determine, for example, which factors are significant when analysing the results of a designed experiment, and there needs to be a numerical correction when applying to multivariate distance measures. Had history been different, the critical values of F would have been scaled differently. However, the development of multivariate distances in multidimensional space was around two decades later than the use of F statistics and ANOVA for hypothesising whether experimental factors have a significant effect on a response and so tables of F values retain the traditional scaling, requiring a correction factor when used in multivariate space.

This correction transforms the Mahalanobis distance to the Hotelling's T squared statistic [10,20]. By scaling the squared Mahalanobis distance by $(I-A)/(A(I-1))$, where A is the number of PCs in the model, a new statistic H can be defined that follows an F distribution:

$$H = \frac{(I-A)}{A(I-1)}\Delta^2 \approx F(A, I-A) \quad (13)$$

Note that there are several closely similar equations in the literature all of which converge if there are sufficient sample sizes or numbers of variables, and in this paper, we employ just one for brevity. The one chosen is often used, for example, in process monitoring and one-class classifiers [21]. If it is assumed that the underlying distribution of objects follows a multinormal pattern, a given percentage level of F can be used to define cut-offs at specific p values. In Figure 2, a p value of 0.1 is illustrated. In statistical terms, this represents the Mahalanobis distance from the mean above which we would expect 10% of the data to lie, that is, a 10% probability of rejecting the null hypothesis that an object is from the parent population.

These critical values can be used to determine whether a measurement represents an outlier (e.g. a fault in a process) or is part of a predefined group. They depend in part on the assumption that the group has an underlying normal distribution but are often relied upon in numerous statistical and chemometric packages to make decisions at varying levels of confidence and depends on transforming the distance to give Hotelling's T squared statistic so that it can be compared with critical values of the F distribution for a given number of degrees of freedom (related to the sample size and number of variables) .

2.3. Two or more groups. Linear discriminant analysis and principal components analysis using the pooled covariance

Previously, we have discussed the situation when all samples are part of a single group, but many chemometric problems involve classifying samples into two or more predefined groups. LDA is one common method for such discrimination [2,3,22,23]. In this paper, for brevity, we will restrict the discussion to LDA although the approach we use can be extended to quadratic discriminant analysis (QDA). We will also use the non-Bayesian form of the classifier.

In LDA, the Mahalanobis distance of an object to the centroid of group g is calculated:

$$\Delta_g{}^2 = (\boldsymbol{x} - \overline{\boldsymbol{x}}_g)\boldsymbol{S}_P{}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}_g)' \quad (14)$$

where Δ_g is the Mahalanobis distance, $\overline{\boldsymbol{x}}_g$ is the mean of group g and $\boldsymbol{S}_P$ is the pooled variance–covariance matrix over all groups.

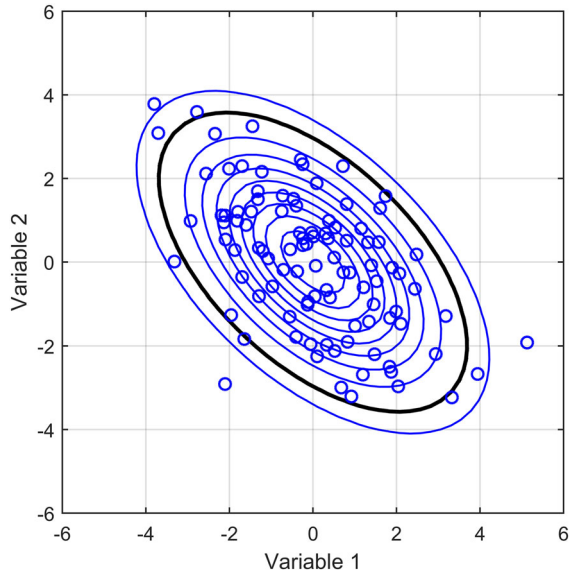

Figure 2. Different p values as the Mahalanobis distance of a distribution increases from the centre. A p value of 0.1 (90% confidence) is indicated in black.

The pooled variance–covariance matrix is not the same as the overall variance–covariance matrix, and in effect is the weighted average of the variance–covariance matrix for each group. In this paper, we will restrict to two groups for brevity.

In Section 2.1, we showed that the Mahalanobis distance can be calculated using the standardised scores of the principal components of all objects. However, in that derivation, we use the overall variance–covariance matrix $\boldsymbol{S}$ and not the pooled variance–covariance matrix $\boldsymbol{S}_P$, which is required for LDA. To calculate the Mahalanobis distance for LDA using PCA-based methods, we therefore have to make some small adjustments when there is more than one group in the data.

It is well known that the principal component loadings of $\boldsymbol{X}'\boldsymbol{X}$ are the same as the loadings for $\boldsymbol{X}$ [24] providing $\boldsymbol{X}$ is mean centred using the overall mean:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P} \tag{15}$$

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{T}_{\boldsymbol{X}}\boldsymbol{P} \tag{16}$$

where $\boldsymbol{T}_{\boldsymbol{X}}$ are the scores of $\boldsymbol{X}'\boldsymbol{X}$ and differ from $\boldsymbol{T}$, the scores for $\boldsymbol{X}$.

Note that the loadings are normalised and are insensitive to scaling of $\boldsymbol{X}$. Thus, the same loadings are obtained if PCA is calculated using the overall variance–covariance matrix $\boldsymbol{S}$ instead of $\boldsymbol{X}'\boldsymbol{X}$. The scores of $\boldsymbol{X}$ can then be calculated by multiplying $\boldsymbol{X}$ by the transpose of the loadings.

$$\boldsymbol{S} = \boldsymbol{T}_{\boldsymbol{X}}\boldsymbol{P} \tag{17}$$

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{P}' \tag{18}$$

To calculate the equivalent LDA Mahalanobis distance to the centroid of group g using pooled covariance, we can calculate a new set of PCs characterised by $\boldsymbol{T}_P$ and $\boldsymbol{P}_P$ using the pooled covariance matrix $\boldsymbol{S}_P$ instead of $\boldsymbol{S}$ and scale the scores using the pooled standard deviation:

$$\boldsymbol{S}_P = \boldsymbol{T}_P\boldsymbol{P}_P \tag{19}$$

$$\Delta_g{}^2 = \sum_{a=1}^{A}\left(\frac{t_{Pa} - \bar{t}_{Pa,g}}{\sqrt{v_{Pa}}}\right)^2 \tag{20}$$

where $\bar{\boldsymbol{t}}_{Pa,g}$ is the mean of the scores of component a for group g.

The pooled variance for component a when there are two groups is

$$v_{Pa} = \frac{(I_1 - 1)v_{1a} + (I_2 - 1)v_{2a}}{I_1 + I_2 - 2} \tag{21}$$

with a similar equation for the pooled covariance. For equal sample sizes, this involves averaging the variance–covariance matrix over the two groups.

The principal components of $\boldsymbol{S}_P$ maximise the pooled variance, rather than the variance, and the pooled variance–covariance matrix of the scores are a diagonal matrix. The Mahalanobis distance for LDA can therefore be calculated using the PCA scores of $\boldsymbol{S}_P$ using a similar approach to that described in Section 2.1.

Chemometricians do not normally calculate PCA using the pooled variance, although as discussed previously, this is analogous to using LDA for the Mahalanobis distance, and for the full component model gives exactly the same result. As discussed in Section 4, this is an intermediate view between conjoint and disjoint PCA, the latter being used for soft independent modeling of class analogy [25] for example.

2.4. Hard and soft models for two groups

When using the Mahalanobis distance to discriminate between two groups, there are two possible ways of using it.

The first is as a hard model. If the Mahalanobis distance to group A is less than that to group B, then an object is assigned to group A and *vice versa*; there is no ambiguity.

The second is as a soft model. Under such circumstance, there can be ambiguous solutions. A confidence level or p value is established for the distance to each group. For example, we can calculate the Mahalanobis distance to a group centroid, and if it is below a defined level, it is assigned to that group. In this paper, we will use a p value of 0.1 (sometimes called a confidence value, or α, of 90%), although of course other levels could be employed. If an object has a p value of less than 0.1 of belonging to both groups A and B, it is considered an outlier, whereas if it is greater than 0.1 of belonging to both groups, it is considered ambiguous; note that in a dataset with no outliers, this means that 10% of samples that genuinely belong to a group will be assigned as outliers. Uniquely assigned samples have a p value of belonging to one group of greater than 0.1 and of the other group of less than 0.1. Using a soft model allows for overlap between groups, and so ambiguous samples are found in this region.

Traditionally, soft modelling is used for one-class classifiers [21], where each group is modelled independently. LDA requires a form of conjoint modelling because the variance–covariance matrix is calculated using both groups together. However, there is no requirement that soft modelling is restricted to one class classifiers and no requirement that LDA models are used to form a hard boundary between two groups. Although it would be possible to discuss both hard and soft classifiers, in this paper, we restrict the discussion to soft models.

The distances from the centroids of each group can be represented by a class distance or Coomans plot [26], in which each axis represents the distance to the centroid of a class, and the confidence limits at a given p value (in our case 0.1 or 90% confidence) are drawn.

3. ONE-CLASS DATASETS

3.1. Three-variable illustration

A sample size of 100 was used to generate a simple normally distributed three-dimensional simulation of dimensions 100×3. Some correlation between the first two variables is introduced through stretching and rotation. We use this simulation to graphically illustrate the calculation of the Mahalanobis distance when a reduced number of principal components are used.

The reduced distances using the equations discussed previously for a 90% cut-off ($p = 0.10$) are 1.66, 2.18 and 2.56 units from the centre using a 1, 2 and 3 PC model, respectively. In Figure 3, we see the three models, both in the original variable space and in the reduced PC space. The shape of the boundaries change considerably from planar (1 PC) to cylindrical (2 PCs) and a hyper-ellipsoid (3 PCs), with corresponding projections into PC space. For this simulation which is quite simple, there is little difference in the number of samples rejected as the complexity

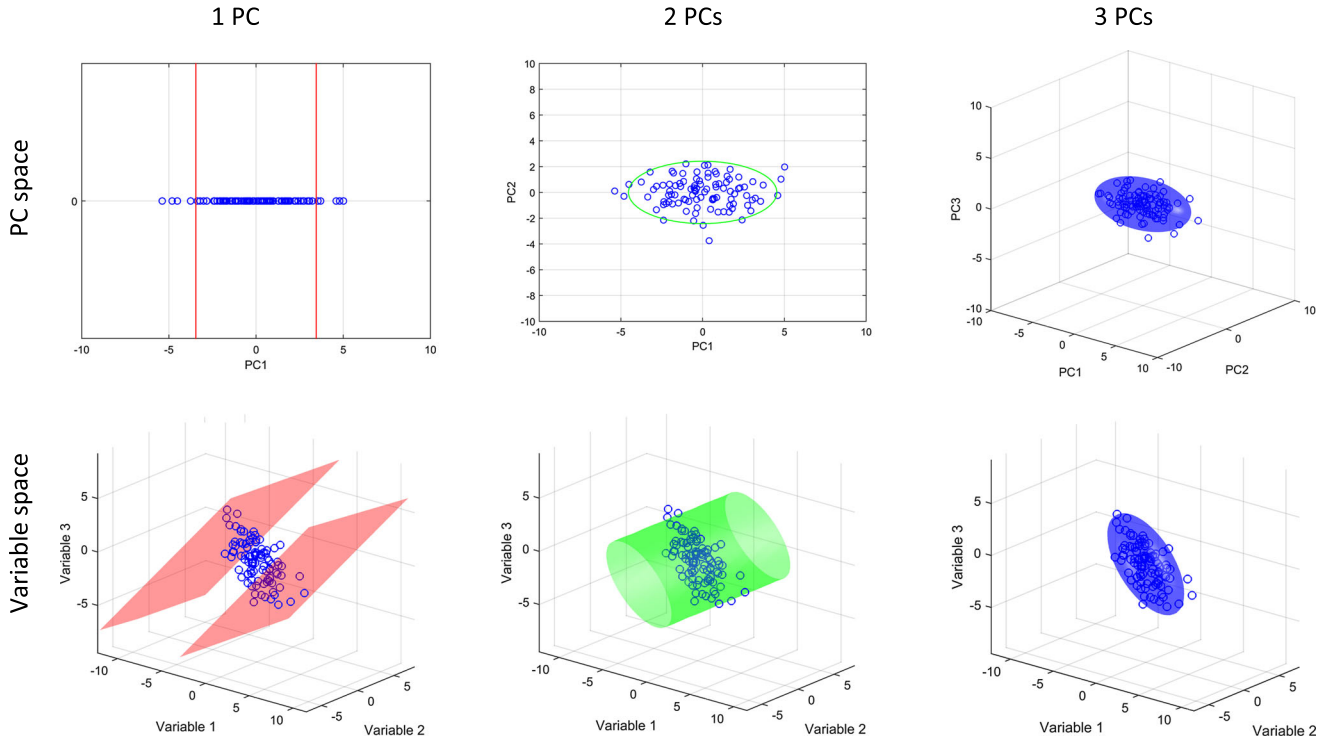

Figure 3. Boundaries using reduced Mahalanobis distance formed using 1, 2 and 3 principal component models and 90% cut-off ($p = 0.1$) for the data of Section 3.1.

of the model changes, but this is because the dataset is multinormal and all 3 PCs have some significance. However, the Mahalanobis distance clearly changes when calculated with differing numbers of PCs.

3.2. Multivariate one-class simulation

We simulate a 100-object, 10-variable matrix to demonstrate the benefit of changing the number of PCs in the reduced Mahalanobis distance model.

A matrix $\boldsymbol{X}$ of dimensions 100×10 was generated by

$$\boldsymbol{X} = \boldsymbol{R}\,\boldsymbol{C} + \boldsymbol{N} \tag{22}$$

where $\boldsymbol{R}$ is a 100×5 matrix of random numbers with population mean of 0 and standard deviation 1, using a normal random number generator, $\boldsymbol{C}$ is a 5×10 matrix of random numbers with population mean of 0 and standard deviation 1, using a normal random number generator, and $\boldsymbol{N}$ is a 100×10 matrix of random numbers with population mean of 0 and standard deviation 0.1, using a normal random number generator. Finally, the matrix $\boldsymbol{X}$ was centred.

The simulation was designed so that there approximately five factors, with the matrix $\boldsymbol{C}$ designed to introduce correlation between 10 composite variables. The precise nature of the simulation is not important for our purpose.

In order to obtain an overview of the main trends, this simulation is repeated 100 times. The number of objects rejected using the Hotelling T squared criterion and three different p values is presented in Table I for different numbers of components. For brevity, we consider only models with the largest components, so, for example, a two-component model always consists of PCs 1 and 2; for one-class models, it is normally expected that the earlier components are the most characteristic of the group; this differs to multiclass models, where larger components may not necessarily have the largest discriminatory power. The aim of this set of simulations is to show the effect of increasing the number of components in the model, rather than choosing which components to retain.

Table I. Average number of objects (100 iterations) rejected using reduced Mahalanobis distance measures of 1 to 10 components for the simulations of Section 3.2 by the Hotellings T squared criterion and three p values

	p	0.5	0.25	0.1
# PCs	1	50.2	24.7	9.9
	2	49.6	24.5	9.0
	3	49.8	23.8	8.3
	4	49.2	23.0	7.7
	5	48.3	22.0	7.2
	6	47.3	20.8	6.4
	7	46.4	19.5	5.8
	8	44.7	18.8	5.1
	9	43.2	17.6	4.3
	10	41.4	15.8	3.8

As there were 100 iterations, these numbers are also percentages.

It can be seen clearly that the quality of model degrades as the number of components increases. This is due to most components representing noise, a problem that is amplified by standardisation, which is giving equal significance to all components in the Mahalanobis distance model.

4. TWO-CLASS LINEAR DISCRIMINANT ANALYSIS

4.1. Two variable simulation

As an example, we will illustrate the case where there are two variables. Each group consists of 50 objects, and there is some overlap between the groups. PCA is performed using the pooled variance–covariance matrix as described in Section 2.3.

Figure 4 illustrates the 90% ($p = 0.1$) confidence bands for the model based on both PCs, on PC1 and on PC2, together with the corresponding Coomans plots. Note that when there is only one PC in the model the points will lie on a straight line. The first important feature to note is that the PC1 model is very poor. This is because PC1 is in the direction of maximum pooled variance of the two groups, which are approximately parallel to each other and loses most of the discrimination, whereas PC2 is the direction of discrimination and as such results in a much better model. The discriminatory power of successive PCs can be calculated by the t value of their scores given by

$$\frac{|\bar{t}_{Aa} - \bar{t}_{Ba}|}{s_{Pa}} \tag{23}$$

where s_{Pa} is the pooled standard deviation of the scores of PC a over groups A and B and $\bar{t}_{Aa}$ is the mean scores of group A and PC a. The same result is obtained whether the scores are standardised or not. The higher the t value the more discriminatory the PC is. For PC1 the t value is 0.21, whereas for PC2 it is 3.04, suggesting PC2 is very much more discriminatory.

Using soft models as described previously, the contingency tables can be obtained as in Table II. It can be seen that the model using just PC2 provides an improvement in the number of objects unambiguously classified into their correct groups, with 88% correctly assigned as compared with 77% for the two-component model. This is because the influence of PC1, which in this case is noise, is removed, resulting in improved discrimination.

Table II. Contingency tables for soft models at $p = 0.1$ (90% confidence) and dataset of Section 4.1

Two components ($p = 0.1$)

Predicted group

		One only	Two only	One and Two	Neither
True group	A	37	1	9	3
	B	1	40	4	5
1st component only ($p = 0.1$)					
True group	A	0	1	47	2
	B	1	1	44	4
2nd component only ($p = 0.1$)					
True group	A	41	2	4	3
	B	1	47	0	2

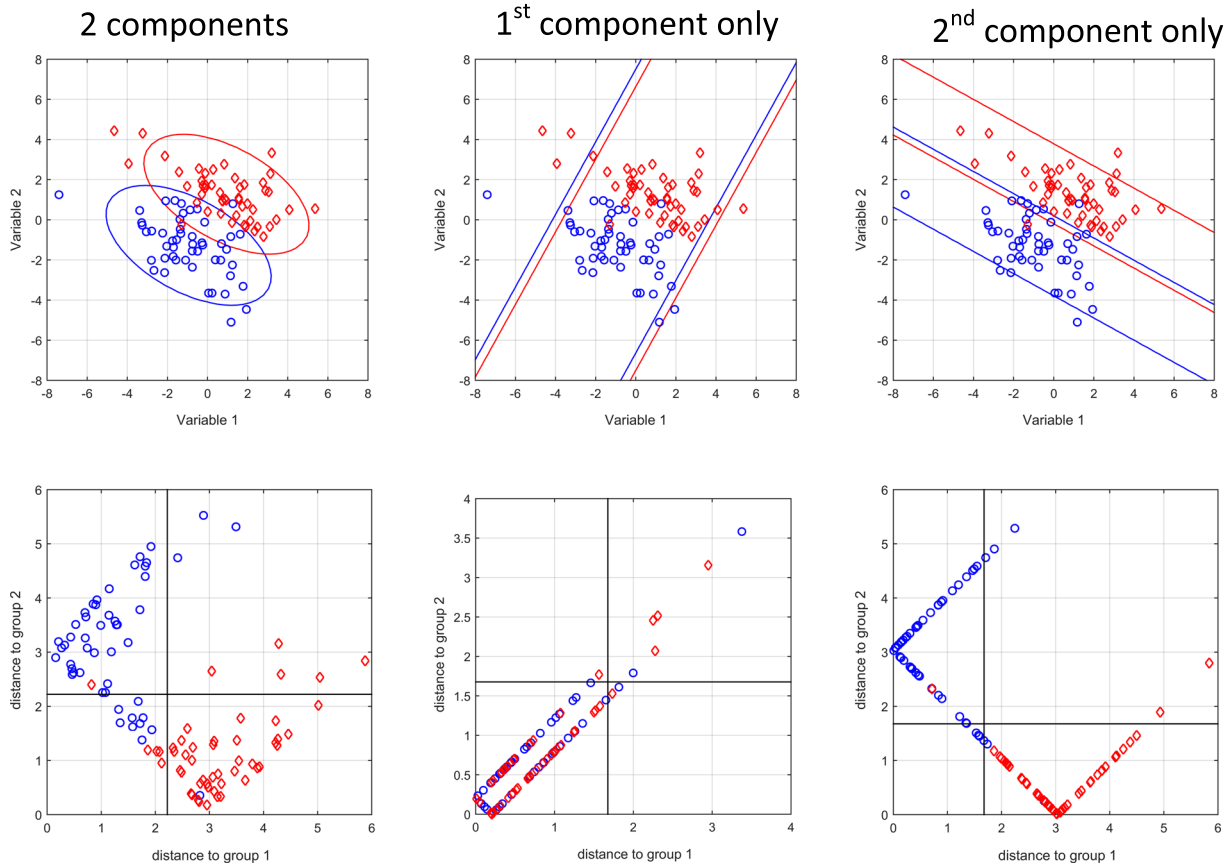

Figure 4. Ninety per cent ($p = 0.1$) confidence bands for two component models and the two one-component models for a two-variable, two-class simulation and the corresponding Coomans plots with $p = 0.1$ limits indicated using linear discriminant analysis and the reduced Mahalanobis distance measure.

It is important to remember that PC1 is sometimes very useful for discrimination, and the relative *t* values for different components will depend on how the groups are oriented towards each other on their main axes.

4.2. Multivariate simulation

A simulation with 100 samples (50 belonging to class A and 50 to class B) and characterised by 10 variables was generated.

The results of PCA using the traditional conjoint method and using the pooled variance–covariance matrix as discussed in Section 2.3 are presented in Figure 5. There is a significant difference between the two types of PC plots. As discussed previously, PCA using a variance–covariance matrix is an intermediate view between conjoint and disjoint PCA. In Figure 6, we show why there is such a difference between performing PCA using each method. If two groups are oriented in parallel to each other, their pooled variance–covariance matrix loses discrimination, although it is a good approximation to the variance of each individual group.

The *t*-statistic for successive components using PCA based on the pooled variance–covariance matrix is presented in Figure 7. It can be seen PCs 4 and 5 are by far the most discriminatory. The relevant Coomans plots using all components, the best component and the best two components for calculating the reduced Mahalanobis distance are presented in Figure 8 for the 90% confidence limits (p=0.1). Note that in this paper we restrict the discussion to LDA rather than QDA for brevity, although similar conclusions could be obtained from QDA. When performing QDA, it is common to use the standard prediction error (SPE) or Q statistic [3,21] in addition to what is often called the D statistic, but when using the reduced Mahalanobis distance, only a single indicator is necessary, providing the correct components are chosen for the model, as we can decide to use only the most discriminatory PCs, which may be of minor importance for a

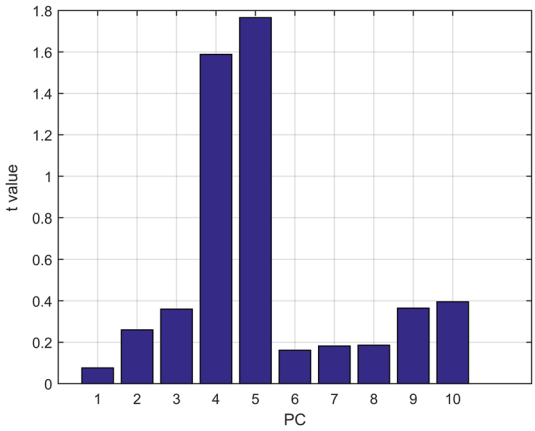

Figure 7. *T* statistic for successive principal components using the pooled variance–covariance matrix for the simulation of Section 4.2.

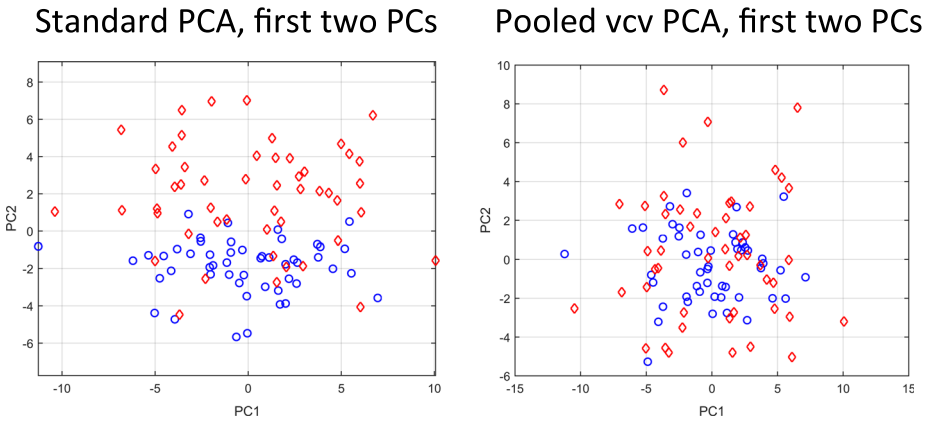

Figure 5. Simulation for section 4.2. The first two principal components (PCs) using the overall variance–covariance matrix and the pooled variance–covariance matrix.

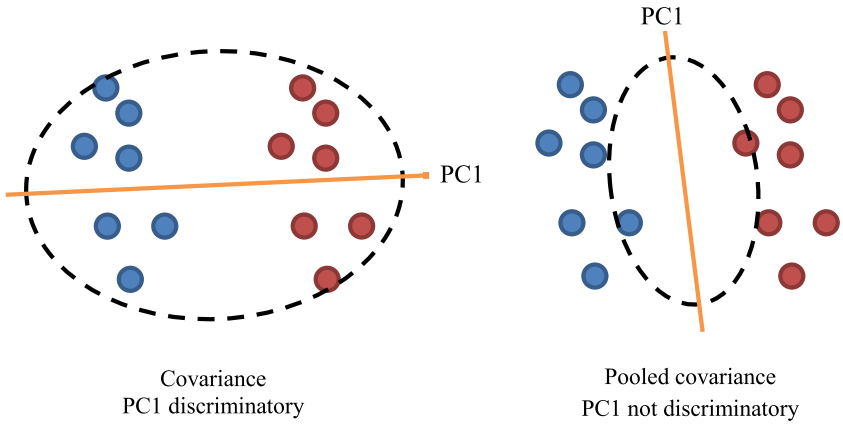

Figure 6. Difference between principal component analysis using the overall variance–covariance matrix and the pooled variance–covariance matrix for two groups that are oriented in parallel to each other.

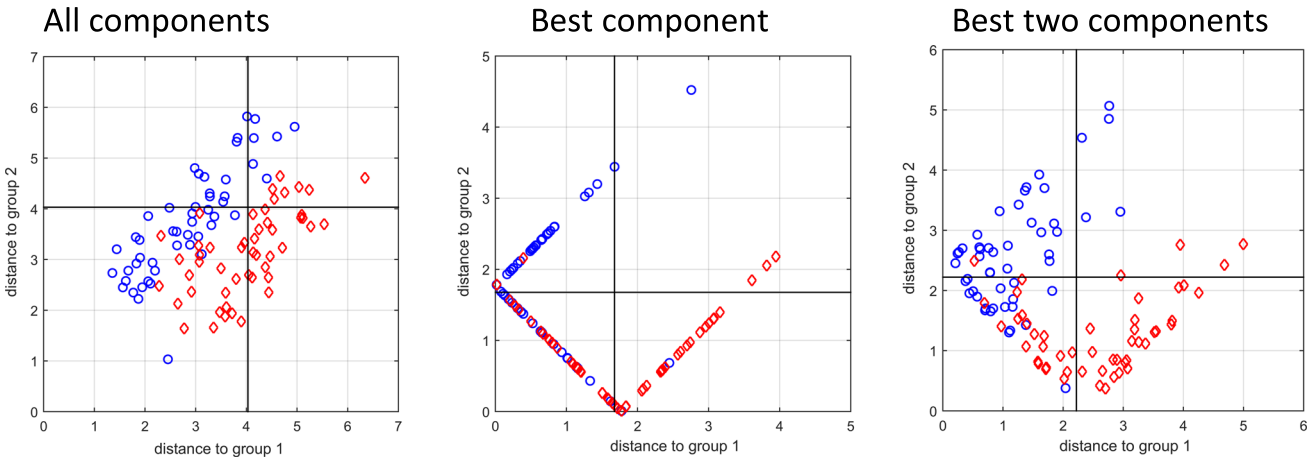

Figure 8. Coomans plot for simulation of Section 4.2 using various components, $p = 0.1$ (90% confidence limit) and reduced Mahalanobis distance using the pooled variance–covariance matrix.

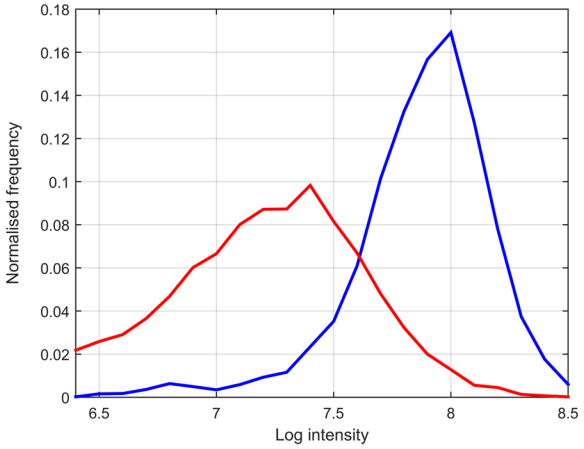

Figure 9. Intensity profile of a set of contaminated banknotes (blue) and background banknotes (red).

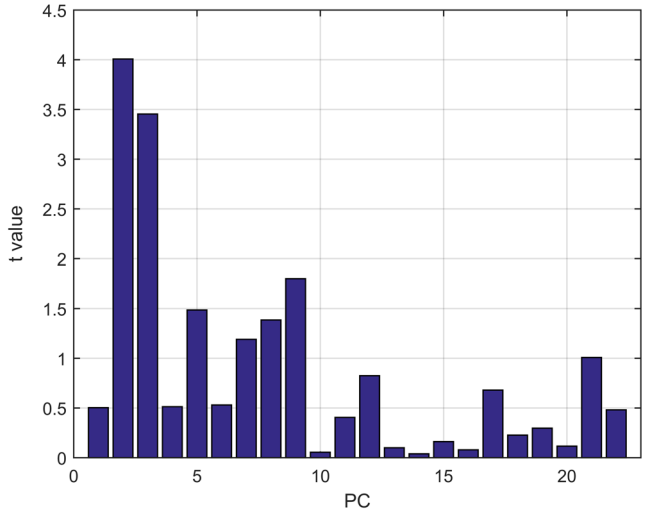

Figure 11. *T* statistic for the 22 non-zero principal components (PCs) using the pooled variance–covariance matrix of the dataset of Section 4.3.

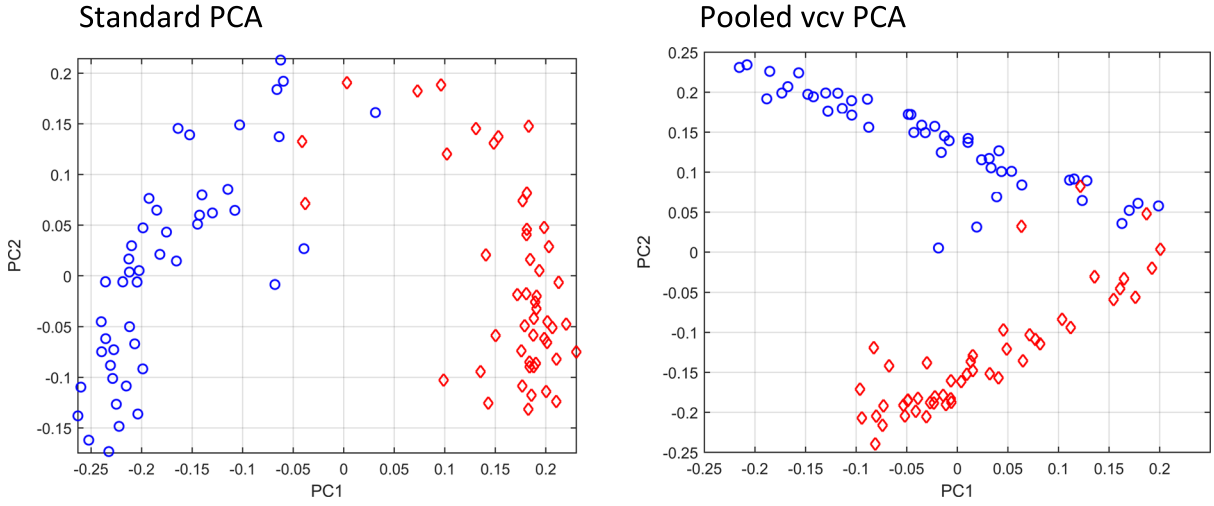

Figure 10. The first two principal components for the case study of Section 4.3, blue, contaminated; red, background. The left subfigure involves using the overall variance–covariance matrix, and the right subfigure involves the pooled variance–covariance matrix.

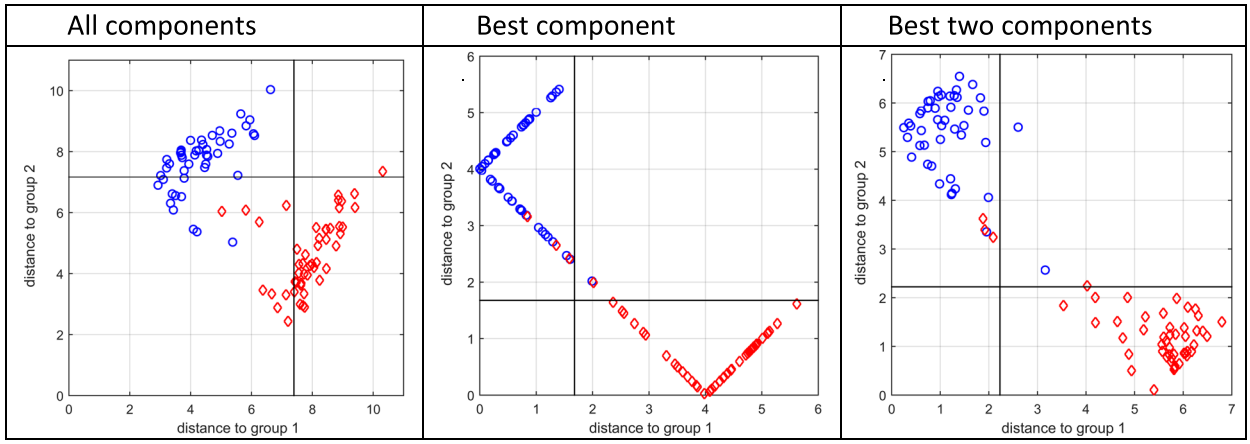

Figure 12. Coomans plot for simulation of Section 4.3 using various components, $p = 0.1$ (90% confidence limit) and reduced Mahalanobis distance using the pooled variance–covariance matrix.

full-rank model using the D statistic. For the full component LDA model, a large number of samples are ambiguous due to overlap, but if choosing the correct components, the number of ambiguous samples is reduced. Because 90% confidence limits are chosen we do not expect complete separation, but there is definite improvement using the reduced rather than the full Mahalanobis distance measure using soft models.

4.3. Real case study: detection of contamination by cocaine in banknotes

In order to illustrate the method using a real case study, we discuss the use of pattern recognition to determine whether a set of banknotes is contaminated by cocaine. The case study has been described in detail elsewhere [3,27] so just a summary is provided below. Most banknotes in the UK are contaminated with cocaine. In order to use this evidence to prosecute suspected drug dealers in court, it is necessary to show that the distribution of contamination by cocaine of a set of banknotes seized from possible defendants is substantially higher than those from innocent sources (in our case the controls are taken from banks). The amount of cocaine distribution is measured by tandem mass spectrometry and each sample consists of the peak heights of a characteristic peak representing cocaine from a set of banknotes. The distribution of intensities can be digitised and the proportion plotted against the logarithm of intensity, as represented in Figure 9, for the average signal from both groups. Note that there is considerable variation within each group, but we do not illustrate the full dataset for brevity. There were 46 sets of banknotes from defendants (suspected drug dealers) and 49 sets of controls (banks), each of which resulted in an intensity profile. We used 22 informative variables (equivalent to log10 intensities) so a matrix $\boldsymbol{X}$ of dimensions 95×22 was generated. All details of scaling were reported elsewhere. The aim is to distinguish samples from the background (control) to those from suspected drug dealers by the mass spectrometric intensity profiles because of cocaine contamination. It should be noted that although most banknotes from background sources have a low level of contamination, the occasional one will be highly contaminated due to the banknotes originally handled by drug dealers mixing in with the general background, so measurements on individual notes cannot with confidence establish their provenance. For example, an innocent person will occasionally carry a very contaminated note in their pocket, but if a set of several notes is highly contaminated, this is good evidence that they may originate from illegal activity.

Principal component scores plots of the first two PCs are illustrated in Figure 10. As expected, for this case, using the pooled variance–covariance matrix changes the appearance of the graph with PC2 becoming the most discriminatory when using the pooled matrix. The *t*-statistic of all non-zero components is presented in Figure 11. It can be seen PCs 2 and 3 are the most discriminatory in this case.

Coomans plots are illustrated in Figure 12. It can be seen that much better separation is obtained when using the reduced Mahalanobis distance calculated by the most discriminatory PCs, rather than the full model. This suggests that when using all the PCs for calculating the Mahalanobis distance there is considerable overlap between the groups, but when reducing the number of PCs in the model, good separation can be obtained. Because all PCs have equal influence on the full PC model for the Mahalanobis distance, uninformative PCs dominate and degrade the discriminatory power, so by reducing the number of PCs, we can obtain better separation.

Similar types of observations can be used when determining the Mahalanobis distance for QDA using disjoint PCA (often called the D statistic), but we restrict this paper to LDA models for brevity. It is, however, well known that the D statistic loses discrimination, and it is usual to have to combine with the Q statistic or SPE in order to see whether samples belong to a given group when using QDA-based one-class classifiers such as soft independent modeling of class analogy [21], which would be unnecessary if the reduced Mahalanobis distance were employed.

5. CONCLUSION

In this paper, we show that the squared Mahalanobis distance equals the sum of squares of scores of non-zero principal components in a model. Although the relationship between Mahalanobis distance and PCA has been noted in the literature, this relationship has not been widely exploited by chemometricians. One of the limits of early statistical applications is that the sample size usually far exceeds the number of variables, whereas in modern chemometric practice, we often find the reverse.

There are numerous consequences of this relationship. The first is obviously that the Mahalanobis distance can be calculated even if the number of columns (variables) in a matrix exceeds the number of rows (sample size). There has been a huge amount of effort in the chemometric literature trying to overcome this limitation often by various elaborate methods for feature or variable selection, which can lead to overfitting.

The second is that we can use reduced Mahalanobis distance models by either choosing the largest or for two or more class models, the most discriminatory, PCs and form a model using just these PCs. For one group models, the advantage is that the smaller PCs, often dominated by noise, are removed from the model. However, there is a major advantage when using two group models either via the pooled variance–covariance matrix (LDA) or disjoint models (QDA) as typically the larger PCs are not discriminatory under such circumstances and often limit these models considerably: usually the Q statistic or SPE is employed as a second statistic because projection onto the new PC space loses discrimination; however, discrimination is still there for smaller PCs, which with the reduced Mahalanobis distance model can be retained, rejecting the less discriminatory PCs.

In this paper, we have also described the use of the pooled variance–covariance matrix for PCA, resulting in an intermediate view between conjoint and disjoint PCA, allowing the possibility of using the reduced Mahalanobis distance when performing LDA.

This paper has limited the assessment of the method for purposes of brevity, but we propose that the relationship between the Mahalanobis distance and PCA is not adequately exploited in the chemometrics literature and has significant potential. In traditional papers on partial least squares and PCA, it is usual to select the most important latent variables or components, but when using Mahalanobis distance-based measures, this is not normally done, even though, as demonstrated here, it can be advantageous to do so.

REFERENCES

1. Mahalanobis PC. On the generalised distance in statistics. *Proc Natl Inst Sci India Phys Sci.* 1936; **2**: 49–55.
2. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 1936; **7**: 179–188.
3. Brereton RG. *Chemometrics for Pattern Recognition*. Wiley: Chichester, 2009.
4. Flury BK, Reldwyl H. Standard distance in univariate and multivariate analysis. *Am Stat* 1986; **40**: 249–251.
5. Gauss CF. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, available. ABC Books: Lowfield Heath, Crawley, United Kingdom, 1809.
6. Student. The probable error of the mean. *Biometrika* 1908; **6**: 1–25.
7. Brereton RG. The t distribution and its relationship to the normal distribution. *J. Chemometr.* 2015; **29**: 481–483.
8. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Phil Mag* 1900; **50**: 157–175.
9. Brereton RG. The F distribution and its relationship to the chi squared and t distribution. *J. Chemometr.* 2015; **29**: 582–586.
10. Hotelling H. The generalization of Student's ratio. *Ann Math Stat* 1931; **2**: 360–378.
11. De Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. *Chemometr Intell Lab Syst.* 2000; **50**: 1–18.
12. Brereton RG. The Mahalanobis distance and its relationship to principal component scores. *J. Chemometr.* 2015; **29**: 143–145.
13. Rao CR. The use and application of principal components analysis in applied research. *Indian J.Stat: Series A* 1964; **26**: 329–358.
14. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 1966; **53**: 325–338.
15. Pomeranstev AL. Acceptance areas for multivariate classification derived by projection methods. *J. Chemometr.* 2008; **22**: 601–609.
16. Hubert M, Rousseeuw PJ. ROBPCA: a new approach to robust principal component analysis. *Technometrics* 2005; **47**: 64–79.
17. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intell Lab Syst* 1987; **2**: 37–52.
18. Geladi P, Kowalski BR. Partial least squares : a tutorial. *Anal. Chim. Acta* 1986; **185**: 1–17.
19. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Philos Trans R Soc* 1918; **52**: 399–433.
20. Brereton RG. Hotelling's T squared distribution, its relationship to the F distribution and its use in multivariate space. *J. Chemometr.* 2016; **30**: 18–21.
21. Brereton RG. One class classifiers. *J. Chemometr.* 2011; **25**: 225–246.
22. 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. *Chemometr Intell Lab Syst* 2009; **95**: 1–17.
23. Frank IE, Friedman JH. Classification: oldtimers and newcomers. *J. Chemometr.* 1989; **3**: 463–475.
24. Geladi P, Isaksson H, Lindqvist L, Wold S, Esbensen K. Principal component analysis of multivariate images. *Chemometr Intell Lab Syst* 1989; **5**: 209–220.
25. Wold S. Pattern-recognition by means of disjoint principal components models. *Pattern Recogn* 1976; **8**: 127–139.
26. Coomans D, Broeckaert I, Derde MP, Tassin A, Massart DL, Wold S. Use of a microcomputer for the definition of multivariate confidence regions in medical diagnosis based on clinical laboratory profiles. *Comput. Biomed. Res.* 1984; **17**: 1–14.
27. Dixon SJ, Brereton RG, Carter JF, Sleeman R. Determination of cocaine contamination on banknotes using tandem mass spectrometry and pattern recognition. *Anal. Chim. Acta* 2006; **559**: 54–63.