JOURNAL OF CHEMOMETRICS
J. Chemometrics 2003; **17**: 166–173
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.785

Partial least squares for discrimination

Matthew Barker[1] and William Rayens[2]*

[1]Biometrics and Statistical Sciences, The Procter and Gamble Company, Mason, OH 45040, USA
[2]Department of Statistics, University of Kentucky, Lexington, KY 40506, USA

Received 25 April 2000; Revised 29 October 2002; Accepted 18 November 2002

Partial least squares (PLS) was not originally designed as a tool for statistical discrimination. In spite of this, applied scientists routinely use PLS for classification and there is substantial empirical evidence to suggest that it performs well in that role. The interesting question is: why can a procedure that is principally designed for overdetermined regression problems locate and emphasize group structure? Using PLS in this manner has heurestic support owing to the relationship between PLS and canonical correlation analysis (CCA) and the relationship, in turn, between CCA and linear discriminant analysis (LDA). This paper replaces the heuristics with a formal statistical explanation. As a consequence, it will become clear that PLS is to be preferred over PCA when discrimination is the goal and dimension reduction is needed. Copyright © 2003 John Wiley & Sons, Ltd.

KEYWORDS: partial least squares; linear discrimination; dimension reduction

1. INTRODUCTION

Although partial least squares (PLS) was not inherently designed for problems of classification and discrimination, it is routinely used for those purposes. For example, PLS has been used to:

- distinguish Alzheimer's, senile dementia of the Alzheimer's type, and vascular dementia [1];
- discriminate between Arabica and Robusta coffee beans [2];
- classify waste water pollution [3];
- separate active and inactive compounds in a quantitative structure–activity relationship study [4];
- differentiate two types of hard red wheat using near-infrared analysis [5];
- distinguish transsexualism, borderline personality disorder and controls using a standard instrument [6];
- determine the year of vintage port wine [7];
- classify soy sauce by geographic region [8];
- determine emission sources in ambient aerosol studies [9].

Many other examples can be cited[10–14]. The way in which PLS facilitates the classification varies according to the user. Often, operating on the heuristic that one would like to effectively 'predict' group membership, a dummy matrix (**Y** block) that records this membership is paired with a training set (**X** block) and PLS is implemented in the usual way. The associated PLS scores are then calculated and plotted pairwise, allowing a visual assessment of group separation. If the separation is pronounced, then it is typical to classify an unknown into the group admitting the closest mean score. Rarely will PLS be followed by an actual discriminant analysis on the scores and rarely is the classification rule given a formal interpretation. Still, this method often produces nice separation. Since it is well known that PLS is related to canonical correlation analysis (CCA) and that CCA is, in turn, related to linear discriminant analysis (LDA), it is reasonable to expect that PLS might have some direct connection to LDA. This paper gives a formal statistical explanation of this connection. This connection would suggest that PLS, and not principal component analysis (PCA), should be used for dimension reduction aimed at discrimination when a training set is available.

*Correspondence to: W. Rayens, Department of Statistics, University of Kentucky, Lexington, KY 40506, USA.
E-mail: rayens@ms.uky.edu

2. BACKGROUND

2.1. Partial least squares

The origins of PLS are traced to Herman Wold's original non-linear iterative partial least squares (NIPALS) algorithm, an algorithm developed to linearize models which were non-linear in the parameters [15,16]. The NIPALS method was adapted for the overdetermined regression problem, mentioned above—a problem typically addressed with principal component regression (PCR)—and that extension was termed partial least squares [17]. Although many standard software packages still implement PLS via the original NIPALS algorithm, there are other useful perspectives on the same paradigm [18–22]. These latter views are more consistent with classical multivariate statistical theory and allow for the implementation of PLS

by way of solutions to well-posed eigenstructure problems that clearly exhibit the compromise PLS strikes between variance summary and score correlation. Unfortunately, this eigenstructure perspective is still not universally known within the community of users, and even more critically, the effect on the eigenvalue problem of different constraints on the PLS directions is less well known. In short, PLS does not correspond to just one eigenstructure problem, but rather to several. Although a full discussion of these issues is peripherial to this paper, it is important to be clear as to what is meant by 'PLS' in the material presented herein. In any case, if one adopts the general eigenstructure paradigm, the impact of constraints on the directions is irrelevant for the first direction. If the successive PLS directions are constrained to be orthogonal, then the corresponding eigenstructure problem is particularly simple, as reviewed below. The eigenstructure that defines PLS under the more popular uncorrelated scores constraint is generated from successively defined, asymmetric matrices (see the Appendix for more discussion on this). In either case the connection with classical LDA is clear when PLS is used for discrimination. For the sake of brevity, only the orthogonal constraints case will be dealt with explicitly, however.

Let $\mathbf{x} \in \mathbf{R}^p$, with dispersion matrix $(\mathbf{\Sigma}_x)_{p\times p}$; $\mathbf{y} \in \mathbf{R}^q$, with dispersion matrix $(\mathbf{\Sigma}_y)_{q\times q}$; and denote the covariance of $\mathbf{x}$ and $\mathbf{y}$ by $(\mathbf{\Sigma}_{xy})_{p\times q}$. The following defines the PLS directions constrained to be orthogonal in the $\mathbf{X}$ space.

Theorem 1

Suppose $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \ldots, \mathbf{a}_k]_{p\times k}$.

$$\arg\max_{\substack{\mathbf{a}\in\mathbf{R}^p \\ \mathbf{b}\in\mathbf{R}^q \\ \mathbf{a}^{\mathrm{T}}\mathbf{A}=\mathbf{0}^{\mathrm{T}}}} \left\{ \frac{[\mathrm{cov}(\mathbf{a}^{\mathrm{T}}\mathbf{x}, \mathbf{b}^{\mathrm{T}}\mathbf{y})]^2}{(\mathbf{a}^{\mathrm{T}}\mathbf{a})(\mathbf{b}^{\mathrm{T}}\mathbf{b})} \right\} = \{\mathbf{a}_{k+1}, \mathbf{b}_{k+1}\}$$

where $\mathbf{a}_{k+1}$ is the eigenvector of $\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yx}$ corresponding to the $(k+1)$th largest eigenvalue, and $\mathbf{b}_{k+1} = \mathbf{\Sigma}_{yx}\mathbf{a}_{k+1}$.

PLS can also be equivalently derived by minimizing the appropriate sums-of-squared-residuals term [20,21,23]. For example:

Theorem 2

The PLS components minimize $E\{\|\mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{x}_i - \mathbf{x}_i\|^2 + \|\mathbf{a}^{\mathrm{T}}\mathbf{x}_i - \mathbf{b}^{\mathrm{T}}\mathbf{y}_i\|^2 + \|\mathbf{b}\mathbf{b}^{\mathrm{T}}\mathbf{y}_i - \mathbf{y}_i\|^2\}$ over all unit-length vectors $\mathbf{a}$ and $\mathbf{b}$.

Regardless, the plug-in sample solutions in the $\mathbf{X}$ space are eigensolutions of $\mathbf{S}_{xy}\mathbf{S}_{yx}\mathbf{a}_{k+1} = \lambda\mathbf{a}_{k+1}$.

2.2. Linear discriminant analysis

In this paper, 'LDA' will refer to the canonical discriminant procedure developed by Fisher [24] in 1936 and designed to maximize between-groups variability relative to a measure of pooled within-groups variability. It is to this paradigm that PLS is intimately related. Formally, suppose one has a training set consisting of n_i observervations on each of p feature variables $\mathbf{X} = (\mathbf{x}_{11}\ \mathbf{x}_{12}\ \ldots\ \mathbf{x}_{1n_1}\ \ldots\ \mathbf{x}_{g1}\mathbf{x}_{g2}\ \ldots\ \mathbf{x}_{gn_g})^{\mathrm{T}}$. Let $\mathbf{H} = \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}}$ denote the among-groups sums-of-squares and cross-products matrix and $\mathbf{E} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^{\mathrm{T}}$ the pooled within-groups sums-of-squares and cross-products matrix, where $\bar{\mathbf{x}}_i = (1/n_i)\sum_{j=1}^{n_i}\mathbf{x}_{ij}, \bar{\mathbf{x}} = (1/n)\times\sum_{i=1}^{g}\sum_{j=1}^{n_i}\mathbf{x}_{ij}, n = \sum_{i=1}^{g} n_i, \mathbf{x}_{ij}$ is the $p\times 1$ vector for the jth observation in the ith group, n_i is the number of observations in the ith group and g is the number of groups. Of course, $\mathbf{E} + \mathbf{H} = (n-1)\mathbf{S}_x$, where $\mathbf{S}_x$ is the usual consistent estimator for $\mathbf{\Sigma}_x$. Fisher was interested in the following optimization problem:

$$\arg\max_{\mathbf{a}\in\mathbf{R}^p}\left\{\frac{\mathbf{a}^{\mathrm{T}}\mathbf{H}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{E}\mathbf{a}}\right\}$$

So that a classification rule might be constructed, one often projects one's data orthogonally onto the subset of directions that emerge and classifies according to the usual closest mean rule. Of course, for $g = 2$ this 'discriminant rule' is the same as what emerges from the better-known formal perspectives that seek to minimize misclassification probabilities or to maximize the posterior probabilities of group membership. For $g > 2$ this is not the case unless the projection is on all p discriminant directions [25].

2.3. CCA, LDA and coding

Frank and Friedman [19] noted that since

$$\left[\mathrm{cov}(\mathbf{a}^{\mathrm{T}}\mathbf{x}, \mathbf{b}^{\mathrm{T}}\mathbf{y})\right]^2 = \mathrm{var}(\mathbf{a}^{\mathrm{T}}\mathbf{x})\left[\mathrm{corr}(\mathbf{a}^{\mathrm{T}}\mathbf{x}, \mathbf{b}^{\mathrm{T}}\mathbf{y})\right]^2\mathrm{var}(\mathbf{b}^{\mathrm{T}}\mathbf{y})$$

it was proper to think of PLS as penalized canonical correlation analysis (CCA), with basically a PCA in the $\mathbf{X}$ space and a PCA in the $\mathbf{Y}$ space providing the penalties. It is well known that when CCA is performed on a training set, $\mathbf{X}$, and a dummy matrix representing group membership, $\mathbf{Y}$, that the CCA directions to emerge are just Fisher's LDA directions. Bartlett [26] is generally credited with first recognizing this connection. Further, the coding of the dummy matrix, within reason, does not matter. Suppose $\mathbf{1}_k$ is a $k\times 1$ vector of all ones and, likewise, $\mathbf{0}_k$ is a $k\times 1$ vector of all zeros. Then one can choose to code the $\mathbf{Y}$ block two equally reasonable ways:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_g} & \mathbf{0}_{n_g} & \cdots & \mathbf{1}_{n_g} \end{pmatrix}_{n\times g}$$

or

$$\mathbf{Z} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \mathbf{1}_{n_{g-1}} \\ \mathbf{0}_{n_g} & \mathbf{0}_{n_g} & \cdots & \mathbf{0}_{n_g} \end{pmatrix}_{n\times(g-1)}$$

and the resulting CCA will be invariant. In the first (overdetermined) case a conditional or generalized inverse of $\mathbf{S}_y$, say $\mathbf{S}_y^c$, is required for the analysis, but it is not difficult to show that the eigenstructure problem that defines CCA is invariant to the choice of conditional inverse as well.

The following results can be used to formally (and clearly) reprove Bartlett's connection between CCA and LDA, but, more importantly, will provide the essential links between LDA and PLS. Let $\mathbf{S}_x = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c\mathbf{X}, \mathbf{S}_y = \frac{1}{n-1}\mathbf{Y}^{\mathrm{T}}\mathbf{P}_c\mathbf{Y}, \mathbf{S}_z =$

$\frac{1}{n-1}\mathbf{Z}^{\mathrm{T}}\mathbf{P}_c\mathbf{Z}$, $\mathbf{S}_{xy} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c\mathbf{Y}$ and $\mathbf{S}_{xz} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c\mathbf{Z}$, where $\mathbf{P}_c = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ is the usual centering projector ($\mathbf{I}_n$ is the $n \times n$ identity matrix). The first two derivations can be found in the Appendix; the third is not repeated here.

Theorem 3

$$\mathbf{S}_z^{-1} = (n-1)\left(\frac{1}{n_g}\mathbf{1}_{g-1}\mathbf{1}_{g-1}^{\mathrm{T}} + \mathbf{M}^{-1}\right)$$

where

$$\mathbf{M} = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_{g-1} \end{pmatrix}_{(g-1)\times(g-1)}$$

and

$$\mathbf{S}_y^c = \begin{pmatrix} \mathbf{S}_z^{-1} & \mathbf{0}_{g-1} \\ \mathbf{0}_{g-1}^{\mathrm{T}} & 0 \end{pmatrix}_{g\times g}$$

Theorem 4

$$\mathbf{S}_{xy}\mathbf{S}_y^c\mathbf{S}_{yx} = \mathbf{S}_{xz}\mathbf{S}_z^{-1}\mathbf{S}_{zx} = \frac{1}{n-1}\mathbf{H}$$

In the context of this paper, Barlett's 1938 result could then be restated as:

Theorem 5
$\mathbf{S}_x^{-1}\mathbf{S}_{xy}\mathbf{S}_y^c\mathbf{S}_{yx}\mathbf{a} = \mathbf{S}_x^{-1}\mathbf{S}_{xz}\mathbf{S}_z^{-1}\mathbf{S}_{zx}\mathbf{a} = \lambda\mathbf{a}$ iff $\mathbf{E}^{-1}\mathbf{H}\mathbf{a} = \mu\mathbf{a}$, where $\mu = \lambda/(1-\lambda)$.

3. PLS FOR DISCRIMINATION

3.1. Connections between PLS and LDA

As mentioned above, sample-based PLS is essentially concerned with the eigenstructure of $\mathbf{S}_{xy}\mathbf{S}_{yx}$, or perhaps $\mathbf{S}_{xz}\mathbf{S}_{zx}$ if the application is discrimination and the coding has been with $\mathbf{Z}$ instead of $\mathbf{Y}$. In the spirit of the above results it is easy to show the following (proofs in the Appendix).

Theorem 6
$\mathbf{S}_{xy}\mathbf{S}_{yx} = \mathbf{H}^*$, where

$$\mathbf{H}^* = \frac{1}{(n-1)^2}\sum_{i=1}^{g} n_i^2(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}}$$

Theorem 7
$\mathbf{S}_{xz}\mathbf{S}_{zx} = \mathbf{H}^{**}$, where

$$\mathbf{H}^{**} = \frac{1}{(n-1)^2}\sum_{i=1}^{g-1} n_i^2(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}}$$

The following assessments are immediate.

- The eigenstructure problem essential to discriminant PLS coded with $\mathbf{Y}$ basically just depends on a slightly altered version of the usual among-groups sums-of-squares and cross-products matrix $\mathbf{H}$, wherein the weights have been altered.
- The eigenstructure problem essential to discriminant PLS coded with $\mathbf{Z}$ depends, as well, on a slightly altered version of the usual among-groups sums-of-squares and cross-products matrix $\mathbf{H}$, but the form of this eigenstructure problem depends arbitrarily on what one designates to be the gth group.
- Coding matters, but is probably not a critical problem; in fact, our experiences have been that the classification results using $\mathbf{H}^*$ and $\mathbf{H}^{**}$ are almost identical.

The slightly illogical form of the among-groups cross-products matrix $\mathbf{H}^*$ can be traced to what might be arguably seen as a slightly illogical posing of the usual PLS problem when discrimination is the goal. Recall that PLS can be thought of as penalized canonical correlation analysis, with basically a PCA in the $\mathbf{X}$ space and a PCA in the $\mathbf{Y}$ space providing the penalties. When the goal of the analysis is discrimination and the $\mathbf{Y}$ block is coded with dummy variables, the $\mathbf{Y}$ space penalty is perhaps not meaningful. Removing this $\mathbf{Y}$ space penalty from the original objective function leaves

$$\left[\mathrm{corr}(\mathbf{a}^{\mathrm{T}}\mathbf{x}, \mathbf{b}^{\mathrm{T}}\mathbf{y})\right]^2\mathrm{var}(\mathbf{a}^{\mathrm{T}}\mathbf{x}) = \frac{\left[\mathrm{cov}(\mathbf{a}^{\mathrm{T}}\mathbf{x}, \mathbf{b}^{\mathrm{T}}\mathbf{y})\right]^2}{\mathrm{var}(\mathbf{b}^{\mathrm{T}}\mathbf{y})}$$

As the following results show, when this new objective function is employed and the $\mathbf{X}$ space directions are constrained to be orthogonal, the eigenstructure problem reduces to that of the ordinary $\mathbf{H}$ (proofs in the Appendix).

Theorem 8
Suppose $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \ldots, \mathbf{a}_k]_{p\times k}$.

$$\arg\max_{\substack{\mathbf{a}\in\mathbf{R}^p \\ \mathbf{b}\in\mathbf{R}^q \\ \mathbf{a}^{\mathrm{T}}\mathbf{A}=\mathbf{0}^{\mathrm{T}}}}\left\{\frac{\left[\mathrm{cov}(\mathbf{a}^{\mathrm{T}}\mathbf{x}, \mathbf{b}^{\mathrm{T}}\mathbf{y})\right]^2}{\mathrm{var}(\mathbf{b}^{\mathrm{T}}\mathbf{y})(\mathbf{a}^{\mathrm{T}}\mathbf{a})}\right\} = \{\mathbf{a}_{k+1}, \mathbf{b}_{k+1}\}$$

where $\mathbf{a}_{k+1}$ is the eigenvector of $\mathbf{\Sigma}_{xy}\,\mathbf{\Sigma}_y^{-1}\,\mathbf{\Sigma}_{yx}$ corresponding to the $(k+1)$th largest eigenvalue, and $\mathbf{b}_{k+1} = \mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}\mathbf{a}_{k+1}$. Of course, only $\mathbf{a}_{k+1}$ is of interest when classification is the goal.

Using Theorem 4, it is clear that when PLS is posed as in Theorem 8, then the plug-in sample solutions in the $\mathbf{X}$ space are eigensolutions of $\mathbf{H}\mathbf{a}_{k+1} = \lambda\mathbf{a}_{k+1}$. It follows immediately that:

- with the slightly redefined, perhaps more reasonably redefined PLS for discrimination, the essential eigenstructure is exactly that of the usual among-groups sums-of-squares and cross-products matrix;
- coding does not matter, since

$$\mathbf{S}_{xy}\mathbf{S}_y^c\mathbf{S}_{yx} = \mathbf{S}_{xz}\mathbf{S}_z^{-1}\mathbf{S}_{zx} = \frac{1}{n-1}\mathbf{H}$$

It should be noted that when $g = 2$,

$$\mathbf{H}^* = 2\mathbf{H}^{**} = \frac{n-1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{H}$$

and hence the eigenstructure problems are equivalent and all proportional to $(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}})(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}})^{\mathrm{T}}$.

Whether this new definition of PLS is employed or not, the relationship between PLS when used for discrimination and Fisher's version of LDA is now clear. In fact, the results suggest with equal clarity that PLS should be superior to principal component analysis (PCA) for reducing dimension with the goal of achieving class separation. That is, the dimension reduction provided by PLS in a discriminant application is guided explicitly by among-groups variability, while the dimension reduction provided by PCA is guided only by total variability. Although the above results make this point quite clear, it is perhaps useful to see an illustration of the phenomenon.

3.2. Illustration: PLS vs PCA in discrimination

In spite of not being a tool that can inherently identify group structure, PCA has long been used for discrimination [27]. Often the rationale is as follows.

- A PCA is performed on the original training set, and a subset of scores (components) is extracted.
- If the groups appear well separated in pairwise score plots, then one might construct a rule that will classify an unknown into the group with the closest average score (in the sense of Mahalanobis distance, perhaps).

Of course, this is not a particularly reliable approach, since PCA is only capable of identifying gross variability and is not capable of distinguishing 'among-groups' and 'within-groups' variability, as is the explicit goal of the simple LDA paradigm. LDA would clearly be the better thing to do if an LDA were possible. This is typically the problem. That is, surely one of the reasons for the persistent use of PCA instead of LDA is that often a formal LDA cannot be performed owing to the large number of variables in the available training set relative to the number of observations. Hence dimension reduction is necessary and PCA is a ready vehicle for that purpose. Further, PCA has enjoyed considerable success as a 'classification' methodology, because in many, if not most, of the reported applications the among-groups variability soundly dominates the within-groups variability. Therefore, when PCA locates a direction of 'maximum gross variability', PCA has in fact found a direction that is consistent with group separation.

What if this is not the case? That is, what if one is in a situation wherein dimension reduction is necessary and it is not at all clear if the resulting group-to-group differences will dominate the total variability as measured by the sample variance/covariance matrix? Surely PCA will no longer perform well as a classification tool. The mathematical results stated above suggest that PLS is the obvious alternative. Indeed, as an alternative to PCA, PLS will surely be no worse than PCA owing to its direct relationship to structures that are essentially Fisher's 'among-groups' sums-of-squares and cross-products matrix. In fact, when the within-groups variability dominates the among-groups variability, PLS will necessarily perform better. Hence, when discrimination is ultimately the goal, and dimension reduction is needed, PLS is to be preferred to PCA. This is illustrated in the following simple simulation.

We generated 100 observations on each of two groups, distributed bivariate normal, with means $(-2, 0)$ and $(2, 0)$ and dispersion

$$\begin{pmatrix} 1 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

respectively. The variance σ^2 in the vertical direction was varied from one to six. Notice that the between-groups variability is fully summarized by the distance between the two group means, both of which are arranged on the horizontal axis. As the variability in the vertical direction increases, PCA will lose sight of this discriminating information, but PLS for discrimination will not.

In Figure 1 we have plotted the first component extracted by PCA (full line) and by PLS (broken line) for $\sigma^2 \in \{1, 4, 6\}$. We see that when $\sigma^2 = 1$, PCA is aligning itself along the axis of maximum total variability (horizontal axis), while PLS aligns itself along the axis of maximum among-groups variability (also the horizontal axis). However, as σ^2 increases, we see that PCA (full line) aligns itself along the vertical axis, while PLS (broken line) remains along the horizontal axis. PLS continues to find the group structure in the data even though PCA has been led astray. This is not surprising, of course, given the form of PLS for discrimination derived earlier in this paper. PLS depends on the between-groups sums-of-squares and cross-products matrix for information on reducing dimension, while PCA relies on the sample (total) variance/covariance matrix.

To compare the classification performance of PCA and PLS, we adopted the usual rule based on projecting an 'unknown' orthogonally onto the appropriate subspace and classifying it into the group exhibiting the closest sample mean, as measured by Mahalanobis distance. Lachenbruch's 'hold-one-out' method was employed for evaluating the misclassification rate of each rule [28]. These individual rates were then averaged over 100 randomly generated training sets for each σ^2 configuration. The bar chart in Figure 1 summarizes the results. Notice that as σ^2 increases, the misclassification rate for PCA increases to over 40%, while that for PLS remains at approximately 2·4%.

Much more elaborate simulations were performed for more than two groups with similar results. These are not reported here, since the mathematical results presented earlier in the paper make it clear that the simulations had to turn out this way.

4. DISCUSSION

The results in this paper bring to the surface the underlying statistical constructs that are resident when PLS is used as a discrimination procedure. The essential construct being manipulated is for all practical purposes the between-groups sums-of-squares and cross-products matrix from Fisher's original LDA problem that was first discussed in 1936. No new classification rule has been developed, nor has a new use for PLS been suggested. Rather, the practical upshot of this work is that one now can have a full understanding of what really goes on when PLS is used as a classification tool; and it is now clear why PLS can be expected to perform reasonably well in this role.

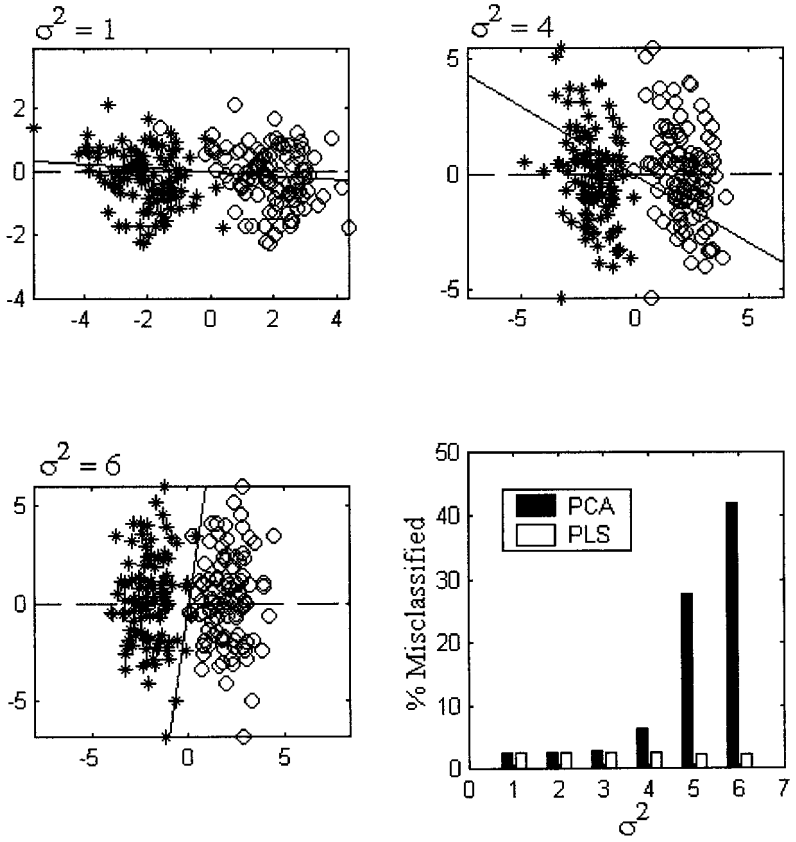

Figure 1. Two groups.

It was also noted that Fisher's between-groups matrix emerged unaltered in the associated eigenstructure problem if one would only slightly and much more logically repose the PLS optimization problem when LDA is the goal. This is a very interesting result, although it is not intended to influence practice. Indeed, in practice, neither the coding (version of **H**) nor the way in which PLS was posed seems to make that much difference with respect to performance. In all cases the eigenstructure of **H** is essentially what is being manipulated, and that recognition is the chief contribution herein.

Of course, PLS users should note that the class structure embodied in **E** is ignored when using PLS for the type of discrimination described in this paper. The most immediate consequence of this is that there is no claim to minimizing misclassification probabilities as there is for $g = 2$ in the Fisher paradigm. Hence there is every reason to believe that LDA will typically outperform PLS when LDA can actually be implemented. However, it could very well be that this gap would decrease or even reverse direction in the presence of a high degree of collinearity. This is a matter for future research. The within-class structure enters this problem naturally if one employs the procedure 'oriented partial least squares' (OrPLS) as proposed by Rayens and Andersen [29]. Specifically, if OrPLS is used for the purpose of discrimination wherein one chooses to orient away from the within-class covariability in the presence of orthogonal constraints, then the eigenstructure of $\mathbf{E}^{-1}\,\mathbf{H}$ surfaces.

Since PCA is the most common 'discriminant' tool used by chemometricians in the presence of high-dimensional data, a simple simulation was presented just to show how using PLS in such a case would have to be better. However, it should be noted, perhaps, that this point is already made clear by the more theoretical results in the paper that articulate that PLS for discrimination is using an LDA construct, while PCA, of course, is not. Hence the example is intended to simply affirm what the theory makes clear has to happen. One can imagine alternatives to PCA—such as principal component regression (PCR)—that could perhaps be adapted and focused toward a discrimination goal. Such alternatives may, in fact, outperform PLS in this capacity, or at least perform better than PCA. However, the purpose of this paper has not been to develop a new classification rule, but rather to point out—for the first time—the mathematical structure being manipulated by PLS in such a context. The reader is referred to Reference [18] for more details.

APPENDIX

It should be noted that not everyone would agree that it is most useful to implement PLS with orthogonal constraints. Indeed, perhaps the most commonly used eigenstructure implementation for PLS requires uncorrelated scores instead of orthogonal constraints (e.g. the SIMPLS option in the SAS procedure PROC PLS). When uncorrelated scores are required in the **X** space, then the $(r+1)$th PLS direction in the **X** space, $\mathbf{a}_{(r+1)1}$, is determined by finding the eigenvector corresponding to the largest eigenvalue, $\lambda_{(r+1)1}$, of $\mathbf{P}^{\mathbf{X}}_{(r)}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yx}$,

where

$$\mathbf{P}^{\mathrm{X}}_{(r)} = \mathbf{I} - \boldsymbol{\Sigma}_x \mathbf{A}^{\mathrm{T}}_{(r)} \left(\mathbf{A}^{\mathrm{T}}_{(r)} \boldsymbol{\Sigma}^2_x \mathbf{A}_{(r)} \right) \mathbf{A}^{\mathrm{T}}_{(r)} \boldsymbol{\Sigma}_x$$

$$\mathbf{A}_{(r)} = [\mathbf{a}_{(1)1}, \mathbf{a}_{(2)1}, \cdots, \mathbf{a}_{(r)1}]$$

The notation $\mathbf{a}_{(r+1)k}$ is to be interpreted as the eigenvector of $\mathbf{P}^{\mathrm{X}}_{(r)}\ \boldsymbol{\Sigma}_{xy}\ \boldsymbol{\Sigma}_{yx}$ corresponding to the kth largest eigenvalue. When $r = 0, \mathbf{P}^{\mathrm{X}}_{(r)}$ is taken to be the appropriate identity matrix. This result can be proved without the use of Lagrange multipliers [30]. Note that solutions are derived from successively defined eigenstructure problems. Hence one is forced to adopt a more cumbersome nested double-subscript notation. Still, when $\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yx}$ is replaced by $\mathbf{S}_{xy}\mathbf{S}_{yx}$, or by $\mathbf{S}_{xz}\mathbf{S}_{zx}$, depending on the coding, Theorems 6 and 7 articulate the roles of **H*** and **H****. Orthogonal directions were chosen in this paper so that the notation might stay simple and the point might emerge unobscured.

Proof of Theorem 3

$$\begin{aligned}\mathbf{S}_y &= \frac{1}{n-1}\mathbf{Y}^{\mathrm{T}}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right)\mathbf{Y} \\ &= \frac{1}{n-1}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathbf{Y}\right) \\ &= \frac{1}{n-1}\left(\mathbf{N} - \frac{1}{n}\mathbf{N}\mathbf{1}_g\mathbf{1}_g^{\mathrm{T}}\mathbf{N}\right)\end{aligned}$$

and

$$\begin{aligned}\mathbf{S}_z &= \frac{1}{n-1}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} - \frac{1}{n}\mathbf{Z}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathbf{Z}\right) \\ &= \frac{1}{n-1}\left(\mathbf{M} - \frac{1}{n}\mathbf{M}\mathbf{1}_{g-1}\mathbf{1}_{g-1}^{\mathrm{T}}\mathbf{M}\right)\end{aligned}$$

where

$$\mathbf{N} = \begin{pmatrix} \mathbf{M} & 0 \\ 0 & n_g \end{pmatrix}_{g\times g}$$

The claim for $\mathbf{S}_z^{-1}$ can be checked by simple matrix multiplication. The other claim follows once one notices that $\mathbf{S}_y$ can be written as

$$\mathbf{S}_y = \begin{pmatrix} \mathbf{S}_z & \mathbf{v} \\ \mathbf{v}^{\mathrm{T}} & \alpha \end{pmatrix}_{g\times g}$$

where $\mathbf{S}_z$ is as above,

$$\mathbf{v} = \frac{-n_g}{n(n-1)}\mathbf{M}^{\mathrm{T}}\mathbf{1}_{g-1}$$

and

$$\alpha = \frac{n_g(n-n_g)}{n(n-1)}$$

Proof of Theorem 4

First note that by definition

$$\mathbf{S}_{xy}\mathbf{S}_y^c\mathbf{S}_{yx} = \frac{1}{(n-1)^2}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c\mathbf{Y}\mathbf{S}_y^c\mathbf{Y}^{\mathrm{T}}\mathbf{P}_c\mathbf{X}$$

where $\mathbf{P}_c = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ is the usual (idempotent) centering matrix. It is easy to check that

$$\begin{aligned}\mathbf{Y}\mathbf{S}_y^c\mathbf{Y}^{\mathrm{T}} &= \begin{pmatrix}\mathbf{Z}\mathbf{S}_z^{-1} & \mathbf{0}_n\end{pmatrix}\begin{pmatrix}\mathbf{Z}^{\mathrm{T}} \\ \mathbf{w}^{\mathrm{T}}\end{pmatrix} \\ &= \mathbf{Z}\mathbf{S}_z^{-1}\mathbf{Z}^{\mathrm{T}} + \mathbf{0}_n\mathbf{w}^{\mathrm{T}} \\ &= \mathbf{Z}\mathbf{S}_z^{-1}\mathbf{Z}^{\mathrm{T}}\end{aligned}$$

where

$$\mathbf{w} = \begin{pmatrix}\mathbf{0}_m \\ \mathbf{1}_{n_g}\end{pmatrix}, \qquad m = n - n_g$$

Hence

$$\begin{aligned}\mathbf{S}_{xy}\mathbf{S}_y^c\mathbf{S}_{yx} &= \frac{1}{(n-1)^2}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c\mathbf{Y}\mathbf{S}_y^c\mathbf{Y}^{\mathrm{T}}\mathbf{P}_c\mathbf{X} \\ &= \frac{1}{(n-1)^2}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c^{\mathrm{T}}\mathbf{Z}\mathbf{S}_z^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{P}_c\mathbf{X} \\ &= \mathbf{S}_{xz}\mathbf{S}_z^{-1}\mathbf{S}_{zx}\end{aligned}$$

and the first part is proved. To see that the common expression is a scalar multiple of **H**, note that

$$\begin{aligned}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c\mathbf{Z} &= \mathbf{X}^{\mathrm{T}}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right)\mathbf{Z} \\ &= \mathbf{X}^{\mathrm{T}}\mathbf{Z} - \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathbf{Z} \\ &= \begin{pmatrix}\bar{\mathbf{x}}_1 & \bar{\mathbf{x}}_2 & \cdots & \bar{\mathbf{x}}_{g-1}\end{pmatrix}\mathbf{M} - \bar{\mathbf{x}}\mathbf{1}_{g-1}^{\mathrm{T}}\mathbf{M} \\ &= \begin{pmatrix}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}) & (\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}) & \cdots & (\bar{\mathbf{x}}_{g-1} - \bar{\mathbf{x}})\end{pmatrix}\mathbf{M} \\ &= \mathbf{U}\mathbf{M}\end{aligned}$$

where **M** is as defined in the statement of Theorem 3. It follows that

$$\begin{aligned}\mathbf{S}_{xy}\mathbf{S}_y^c\mathbf{S}_{yx} &= \frac{1}{(n-1)^2}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c\mathbf{Z}\mathbf{S}_z^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{P}_c\mathbf{X} \\ &= \frac{1}{n-1}\mathbf{U}\mathbf{M}\left(\frac{1}{n_g}\mathbf{1}_{g-1}\mathbf{1}_{g-1}^{\mathrm{T}} + \mathbf{M}^{-1}\right)\mathbf{M}\mathbf{U}^{\mathrm{T}} \\ &= \frac{1}{n-1}\left(\frac{1}{n_g}\mathbf{U}\mathbf{M}\mathbf{1}_{g-1}\mathbf{1}_{g-1}^{\mathrm{T}}\mathbf{M}\mathbf{U}^{\mathrm{T}} + \mathbf{U}\mathbf{M}\mathbf{U}^{\mathrm{T}}\right) \\ &= \frac{1}{n-1}\left(n_g(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})^{\mathrm{T}} + \sum_{i=1}^{g-1} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}}\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}} \\ &= \frac{1}{n-1}\mathbf{H}\end{aligned}$$

Proof of Theorem 6

Notice

$$\begin{aligned}\mathbf{S}_{xy}\mathbf{S}_{yx} &= \frac{1}{(n-1)^2}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{P}_c\mathbf{X} \\ &= \frac{1}{(n-1)^2}\begin{pmatrix}n_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}) & n_2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}) & \cdots\end{pmatrix}

$$n_g(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})) \begin{pmatrix} n_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}) \\ n_2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}) \\ \vdots \\ n_g(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}) \end{pmatrix} = \frac{1}{(n-1)^2} \sum_{i=1}^{g} n_i^2 (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}} \equiv \mathbf{H}^*$$

Proof of Theorem 7

Notice

$$\mathbf{S}_{xz}\mathbf{S}_{zx} = \frac{1}{(n-1)^2}\mathbf{X}^{\mathrm{T}}\mathbf{P}_c\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{P}_c\mathbf{X}$$

$$= \frac{1}{(n-1)^2}(n_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}) \quad n_2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}) \quad \cdots$$

$$n_{g-1}(\bar{\mathbf{x}}_{g-1} - \bar{\mathbf{x}})) \begin{pmatrix} n_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}) \\ n_2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}) \\ \vdots \\ n_{g-1}(\bar{\mathbf{x}}_{g-1} - \bar{\mathbf{x}}) \end{pmatrix} = \frac{1}{(n-1)^2} \sum_{i=1}^{g-1} n_i^2 (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}} \equiv \mathbf{H}^{**}$$

Proof of Theorem 8

Assuming the PLS directions are constrained to be orthogonal,

$$\max_{\substack{\mathbf{a}\in\mathbf{R}^p \\ \mathbf{b}\in\mathbf{R}^q \\ \mathbf{a}^{\mathrm{T}}\mathbf{A}=\mathbf{0}^{\mathrm{T}}}} \left\{ \frac{[\mathrm{cov}(\mathbf{a}^{\mathrm{T}}\mathbf{x}, \mathbf{b}^{\mathrm{T}}\mathbf{y})]^2}{(\mathbf{a}^{\mathrm{T}}\mathbf{a})(\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_y\mathbf{b})} \right\} = \max_{\substack{\mathbf{a}\in\mathbf{R}^p \\ \mathbf{b}\in\mathbf{R}^q \\ \mathbf{a}^{\mathrm{T}}\mathbf{A}=\mathbf{0}^{\mathrm{T}}}} \left\{ \frac{(\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{xy}\mathbf{b})^2}{(\mathbf{a}^{\mathrm{T}}\mathbf{a})(\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_y\mathbf{b})} \right\}$$

$$= \max_{\substack{\mathbf{a}\in\mathbf{R}^p \\ \mathbf{a}^{\mathrm{T}}\mathbf{A}=\mathbf{0}^{\mathrm{T}}}} \left\{ \frac{1}{\mathbf{a}^{\mathrm{T}}\mathbf{a}} \max_{\mathbf{b}\in\mathbf{R}^q} \left\{ \frac{[\mathbf{b}^{\mathrm{T}}(\boldsymbol{\Sigma}_{yx}\mathbf{a})]^2}{\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_y\mathbf{b}} \right\} \right\}$$

$$= \max_{\substack{\mathbf{a}\in\mathbf{R}^p \\ \mathbf{a}^{\mathrm{T}}\mathbf{A}=\mathbf{0}^{\mathrm{T}}}} \left\{ \frac{1}{\mathbf{a}^{\mathrm{T}}\mathbf{a}} \max_{\mathbf{b}\in\mathbf{R}^q} \left\{ \frac{\mathbf{b}^{\mathrm{T}}(\boldsymbol{\Sigma}_{yx}\mathbf{a})(\boldsymbol{\Sigma}_{yx}\mathbf{a})^{\mathrm{T}}\mathbf{b}}{\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_y\mathbf{b}} \right\} \right\}$$

$$= \max_{\substack{\mathbf{a}\in\mathbf{R}^p \\ \mathbf{a}^{\mathrm{T}}\mathbf{A}=\mathbf{0}^{\mathrm{T}}}} \left\{ \frac{\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}} \right\} = \lambda_{k+1}$$

the largest eigenvalue of $\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}$, achieved at the corresponding eigenvector, $\mathbf{a}_{k+1}$.

REFERENCES

1. Gottfries J, Blennow K, Wallin A and Gottfries CG. Diagnosis of dementias using partial least squares discriminant analysis. *Dementia* 1995; **6**: 83–88.
2. Briandet R, Kemsley E and Wilson R. Discrimination of Arabica and Robusta in instant coffee by Fourier transform infrared spectroscopy and chemometrics. *J. Agric. Food Chem.* 1996; **44**: 170–174.
3. Saaksjarvi E, Khalighi M and Minkkinen P. Waste water pollution modelling in the southern area of Lake Saimaa, Finland, by the SIMCA pattern recognition method. *Chemometrics Intell. Lab. Syst.* 1989; **7**: 171–180.
4. Berntsson P and Wold S. Comparison between x-ray crystallographic data and physicochemical parameters with respect to their information about the calcium channel antagonist activity of 4-phenyl-1,4-dihydropyridines. *Quant. Struct.–Activ. Relat.* 1986; **5**: 45–50.
5. Delwiche S, Chen Y-R and Hruschka W. Differentiation of hard red wheat by near-infrared analysis of bulk samples. *Grain Qual.* 1972; **3**: 243–247.
6. Sundbom E, Bodlund O and Hojerback T. Object relation and defensive operations in transsexuals and borderline patients as measured by the Defense Mechanism Test. *Nordic J. Psychiatry* 1995; **49**: 379–388.
7. Ortiz M, Sarabia L, Symington C, Santamaria F and Iniguez M. Analysis of ageing and typification of vintage ports by partial least squares and soft independent modelling class analogy. *Analyst* 1996; **121**: 1009–1013.
8. Iizuka K and Aishima T. Soy sauce classification by geographic region based on NIR spectra and chemometrics pattern recognition. *J. Food Sci.* 1997; **62**: 101–104.
9. Vong R, Geladi P, Wold S and Esbensen K. Source contributions to ambient aerosol calculated by discriminant partial least squares regression. *J. Chemometrics* 1988; **2**: 281–296.
10. Geladi P and Kowalski B. Partial least squares regression: a tutorial. *Anal. Chim. Acta* 1986; **185**: 1–17.
11. Helland IS. On the structure of partial least squares regression. *Commun. Statist. B – Simul. Comput.* 1988; **17**: 581–607.
12. Helland IS. Maximum likelihood regression on relevant components. *J. R. Statist. Soc. B* 1992; **54**: 637–645.
13. Hoskuldsson A. PLS regression methods. *J. Chemometrics* 1988; **2**: 211–228.
14. Kettaneh-Wold N. Analysis of mixture data with partial least squares. *Chemometrics Intell. Lab. Syst.* 1992; **14**: 57–69.
15. Wold H. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, Krishnaiah PR (ed.). Academic Press: New York, 1966; 391–420.
16. Wold H. Soft modeling: the basic design and some extensions. In *Systems under Indirect Observation, Causality–Structure–Prediction*, Joreskog KG, Wold H (eds). North-Holland: Amsterdam, 1981; 1–54.
17. Wold S, Martens H and Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In *Proceedings from the Conference on Matrix Pencils*, Ruhe A, Kagstrom B (eds). Springer: Heidelberg, 1983; 286–298.
18. Barker M. *Partial least squares for discrimination*. *PhD Dissertation*, University of Kentucky, 2000.
19. Frank I and Friedman J. A statistical view of some chemometrics regression tools. *Technometrics* 1993; **35**: 109–148.
20. Hinkle J and Rayens WS. Partial least squares and compositional data: problems and alternatives. *Chemometrics Intell. Lab. Syst.* 1995; **30**: 159–172.
21. Hinkle J and Rayens WS. *Partial least squares, reciprocal components and reciprocal curves*. *Proceedings from the 1998 Joint Statistical Meetings*, Dallas, TX, 1998; 126–130.
22. Stone M and Brooks RJ. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares, and principal component regression. *J. R. Statist. Soc. B* 1990; **52**: 237–269.
23. Hinkle J. *Reciprocal components, reciprocal curves, and partial least squares*. *PhD Dissertation*, University of Kentucky, 1995.
24. Fisher RA. The use of multiple measurement in taxonomic problems. *Ann. Eugen.* 1936; **7**: 179–188.
25. Kshiragar AM and Arseven E. A note on the equivalency

of two discrimination procedures. *Am. Statist.* 1975; **29**: 38–39.
26. Bartlett MS. Further aspects of the theory of multiple regression. *Proc. Camb. Philos. Soc.* 1938; **34**: 33–40.
27. McLachlan GJ. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley: New York, 1992.
28. Lachenbruch PA and Mickey MR. Estimation of error rates in discriminant analysis. *Technometrics* 1968; **10**: 1–11.
29. Rayens WS and Andersen A. Oriented partial least squares. *Ital. J. Appl. Statist.* 2003.
30. Rayens WS. *The art of maximizing covariance*. *University of Kentucky Technical Report 383*, 2000.