JOURNAL OF CHEMOMETRICS, VOL. 11, 533–545 (1997)

DISCRIMINATION BY MEANS OF COMPONENTS THAT ARE ORTHOGONAL IN THE DATA SPACE

HENK A. L. KIERS
Department of Psychology, University of Groningen, Grote Kruisstraat 2/1, NL-9712 TS Groningen, Netherlands

SUMMARY

Krzanowski (*J. Chemometrics*, **9**, 509 (1995)) proposed a method for obtaining so-called orthogonal canonical variates (henceforth called components) for discrimination purposes. In contrast with ordinary discriminant analysis, this method employs components that are orthogonal in the original data space. These components are derived in a successive way, thus optimizing discrimination of a component given the previously extracted components. Two alternative procedures are proposed to extract the desired number of components simultaneously, yielding a better overall discrimination. The simultaneous approaches are applied to the same two data sets as analysed by Krzanowski, as well as to Anderson's Iris data, and a comparison of discriminatory quality of the solutions is presented. © 1997 John Wiley & Sons, Ltd.

Journal of Chemometrics, Vol. **11,** 533–545 (1997) (No. of Figures: 3 No. of Tables: 5 No. of References: 5)

KEY WORDS discriminant analysis; multivariate analysis

1. INTRODUCTION

For data where the (n) observation units belong to several different groups, ordinary discrimant analysis can be used for searching dimensions that discriminate the observation units well. In ordinary discriminant analysis these dimensions are the 'canonical variates' (henceforth called 'discriminant components') that successively maximize the ratio of between-group to pooled within-group sample variance. These ordinary discriminant components are computed as linear combinations of the (p) original variables. The successive discriminant components are chosen such that the scores on these components are uncorrelated.

When one is interested in the orientation of the discriminant components in a representation of the variables in $\mathbb{R}^n$, the uncorrelatedness of the discriminant components is useful, because it implies that these components are found as orthogonal directions in $\mathbb{R}^n$, which in turn implies that plotting of the variables with respect to the discriminant components (taken as orthogonal axes) is meaningful. However, when performing a discriminant analysis, there often is considerable interest in the 'reverse' representation of the n observation units ('objects' for short) in $\mathbb{R}^p$, where the axes refer to the p variables. This high-dimensional configuration visualizes the dissimilarities between the objects (in terms of their scores on the p variables) as distances in the configuration. The aim of discriminant analysis is to find directions in $\mathbb{R}^p$ along which the groups are optimally discriminated. In this space the directions found by the ordinary discriminant components are (usually) not orthogonal. This would only be the case if the vectors of weights that define the discriminant components were orthogonal, because in that case the discriminant components could be viewed as axes obtained by an orthogonal rotation of the original axes formed by the p variables. Because ordinary discriminant components are

Correspondence to: H. A. L. Kiers, Department of Psychology, University of Groningen, Grote Kruisstraat 2/1, NL-9712 TS Groningen, Netherlands.

CCC 0886–9383/97/060533–13 $17.50

Received November 1996
Accepted May 1997

(usually) non-orthogonal in $\mathbb{R}^p$, a plot of the scores of the objects on the discriminant components drawn as *orthogonal* axes (as one usually does when plotting scores) leads to a distorted view of the configuration of the objects.

To illustate the distortion of the configuration of the objects by a plot with respect to discriminant components, we consider discriminant analysis of a subset of the Iris data (first published by Anderson;[1] see e.g. Reference 2). The full data set (to be analysed later on) consists of scores of three groups of 50 irises on the four variables sepal length, sepal width, petal length and petal width. The three groups of irises belong to three different species: *Iris setosa, Iris versicolor* and *Iris virginica.* Here we consider only two variables (allowing us to plot the whole data space); we chose the first and the third variable to clearly demonstrate the distorting effects of a plot with respect to discriminant components. In Figure 1(a) the scores of the irises on the first variable (vertical axis) are plotted versus those on the third (horizontal axis). A discriminant analysis of these data led to two uncorrelated discriminant components (standardized to mean within-group variance equal to one), the scores on which are plotted in Figure 1(b). Some of the irises are labelled individually (by capitals) to facilitate comparison of the two plots; the others are labelled with respect to their species only. It can be seen that the original configuration (Figure 1(a)), with all irises along a relatively narrow band from lower left to upper right, is not recovered by the plot with respect to the discriminant components. In fact, the whole configuration is severely distorted, as can be seen, for instance, from the fact that in the original data space both M and X are approximately equally distant from A and B, whereas in the plot on the discriminant components both M and X are considerably closer to B than to A. Other such distortions can easily be found in terms of distances between the other individually labelled irises.

Krzanowski[3] proposed a procedure for finding discriminating components such that the vectors of weights defining the successive components are orthogonal and normalized to unit sums of squares. Therefore these discriminant components do refer to orthogonal axes in $\mathbb{R}^p$. As a consequence, when plotting objects with respect to these components, we get an undistorted picture of their positions in the original data space. In fact, a (two-dimensional) plot of the objects with respect to such axes comes down to a rotation of the full data space such that two axes coincide with these discriminating components, and the procedure of plotting the co-ordinates with respect to these components comes down to an orthogonal projection on the ensuing plane. Thus Krzanowski's procedure consists of searching planes (or higher-dimensional structures) in the *original data space*, along which the data are optimally discriminated (in a certain sense) after *orthogonally projecting* them on those planes. A particularly useful feature of his approach is that when data are correlated considerably, this correlation remains visible in the plot with respect to the discriminating axes, whereas in ordinary discriminant analysis such a correlation is deliberately removed from the data. This is exemplified by Krzanowski's (Reference 3, p. 516) first exemplary analysis.

Krzanowski's procedure for finding optimally discriminating orthogonal components can be described as follows. Let **B** denote the between-group covariance matrix for a particular data set and let **W** denote the pooled within-group covariance matrix. Then his method consists of first finding the vector $\mathbf{e}_1$ that maximizes $V_1=\mathbf{e}_1^T\mathbf{B}\mathbf{e}_1/\mathbf{e}_1^T\mathbf{W}\mathbf{e}_1$ subject to $\mathbf{e}_1^T\mathbf{e}_1=1$, next finding the vector $\mathbf{e}_2$ that maximizes $V_2=\mathbf{e}_2^T\mathbf{B}\mathbf{e}_2/\mathbf{e}_2^T\mathbf{W}\mathbf{e}_2$ subject to the constraints that $\mathbf{e}_2^T\mathbf{e}_2=1$ and $\mathbf{e}_2^T\mathbf{e}_1=0$, etc. That is, in his procedure the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_r$ (where r indicates the number of desired components) are obtained by successively maximizing $V_l=\mathbf{e}_l^T\mathbf{B}\mathbf{e}_l/\mathbf{e}_l^T\mathbf{W}\mathbf{e}_l$ subject to the constraints that $\mathbf{e}_l^T\mathbf{e}_l=1$ and $\mathbf{e}_l^T\mathbf{e}_j=0$ for $j=1, \ldots, l-1$. This procedure differs from ordinary discriminant analysis (DA) in that DA successively maximizes $V_l=\mathbf{e}_l^T\mathbf{B}\mathbf{e}_l/\mathbf{e}_l^T\mathbf{W}\mathbf{e}_l$ subject to the constraint that $\mathbf{e}_j^T\mathbf{W}\mathbf{e}_l=0$ for $j=1, \ldots, l-1$, and the DA solution for the weight vectors is given by the first r eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$.

In ordinary DA the successive maximization of V_l is equivalent to a simultaneous maximization of $\sum_{l=1}^{r}V_l$ subject to the constraint that $\mathbf{E}^T\mathbf{W}\mathbf{E}$ is diagonal (where $\mathbf{E}$ contains $\mathbf{e}_1, \ldots, \mathbf{e}_r$ as its columns). Therefore ordinary DA not only maximizes the discriminatory quality successively, given the previous

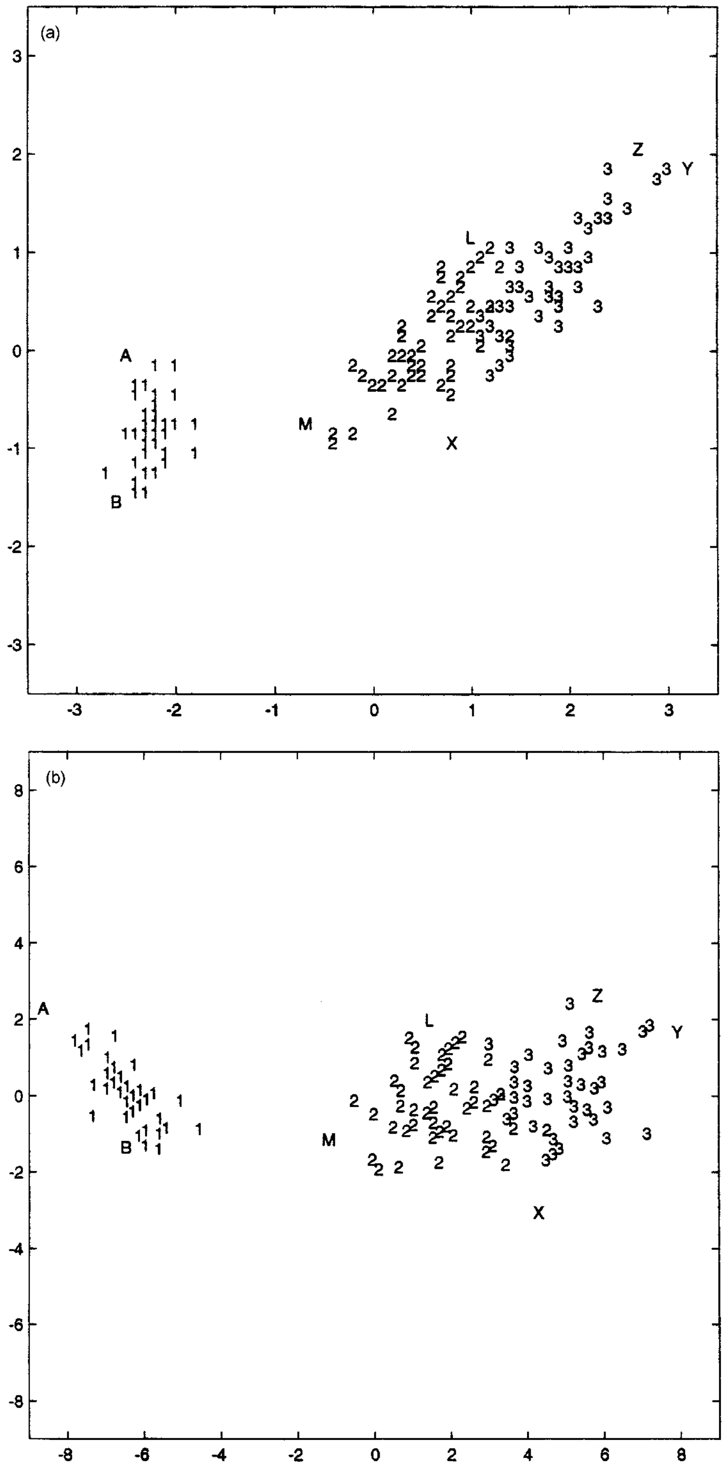

Figure 1. (a) Plot of Iris data with respect to third and first variables. (b) Plot of scores of irises on first two discriminant components. Legend: A, B, 1, *iris setosa*; L, M, 2, *iris versicolor*; X, Y, Z, 3, = *iris virginica*

components, but also the total discriminatory quality (expressed by $\sum_{l=1}^{r} V_l$) of the r components jointly. Krzanowski's discriminant analysis employing orthogonal components (abbreviated as ODA for 'orthogonal discriminant analysis') does not share this feature with ordinary DA: the successively obtained weights vectors do not maximize the total discriminatory quality $\sum_{l=1}^{r} V_l$.

In the present paper an alternative to Krzanowski's method is proposed that, just as ordinary DA, maximizes $\sum_{l=1}^{r} V_l$ and, just like Krzanowski's method, yields mutually orthogonal and normalized weights vectors. It will be shown how such components can be obtained (using an algorithm proposed by Kiers[4]) and it will be illustrated on the data sets analysed by Krzanowski[3] as well as on Anderson's[1] Iris data set to what extent this simultaneous approach gives different results. In fact, in Krzanowski's second data set, $\mathbf{W}$ is singular, a situation for which Krzanowski made a special adjustment of his procedure. As will be shown, the same adjustment can be used with the simultaneous approach proposed here.

Krzanowski's approach is described as a method for successively maximizing V_l. However, as is readily verified, maximization of V_l is equivalent to maximization of $\eta_l^2 = \mathbf{e}_l^{\mathrm{T}}\mathbf{B}\mathbf{e}_l/\mathbf{e}_l^{\mathrm{T}}\mathbf{T}\mathbf{e}_l$, where $\mathbf{T}=\mathbf{B}+\mathbf{W}$ denotes the 'total' covariance matrix. The measure is denoted by η^2 because it equals the correlation ratio for the component defined by $\mathbf{e}_l$ with respect to the group structure in the data. The correlation ratio is related to V as follows: $\eta_l^2 = \mathbf{e}_l^{\mathrm{T}}\mathbf{B}\mathbf{e}_l/(\mathbf{e}_l^{\mathrm{T}}\mathbf{B}\mathbf{e}_l+\mathbf{e}_l^{\mathrm{T}}\mathbf{W}\mathbf{e}_l) = (\mathbf{e}_l^{\mathrm{T}}\mathbf{B}\mathbf{e}_l/\mathbf{e}_l^{\mathrm{T}}\mathbf{W}\mathbf{e}_l)/(\mathbf{e}_l^{\mathrm{T}}\mathbf{B}\mathbf{e}_l/\mathbf{e}_l^{\mathrm{T}}\mathbf{W}\mathbf{e}_l+1) = V_l/(V_l+1)$. The correlation ratio is somewhat more convenient than the V measure, because it is normed between zero and one. Moreover, the interpretation of this measure as the proportion of variation accounted for by the between-group variation may give this measure more intuitive appeal. Krzanowski's method maximizes the η_l^2-values successively, but does not maximize the sum of η_l^2-values over a set of normalized, mutually orthogonal vectors $\mathbf{e}_l$, $l=1, \ldots, r$. The problem of maximizing $\sum_l \eta_l^2$ over columnwise orthonormal $\mathbf{E}$ is also different from the problem of maximizing $\sum_l V_l$ over columnwise orthonormal $\mathbf{E}$, because maximizing $\sum_l [V_l/(V_l+1)]$ is not equivalent to maximizing $\sum_l V_l$. Therefore in the present paper the maximization of $\sum_l \eta_l^2$ over columnwise orthonormal $\mathbf{E}$ is handled separately. The algorithm proposed for this maximization problem is applied to the earlier mentioned data, and the results will again be compared with the outcomes of Krzanowski's method.

2. OBTAINING ORTHOGONAL DISCRIMINATING DIRECTIONS SIMULTANEOUSLY

The problems of maximizing

$$f(\mathbf{E}) = \sum_{l=1}^{r} V_l = \sum_{l=1}^{r} \frac{\mathbf{e}_l^{\mathrm{T}}\mathbf{B}\mathbf{e}_l}{\mathbf{e}_l^{\mathrm{T}}\mathbf{W}\mathbf{e}_l} \tag{1}$$

and

$$g(\mathbf{E}) = \sum_{l=1}^{r} \eta_l^2 = \sum_{l=1}^{r} \frac{\mathbf{e}_l^{\mathrm{T}}\mathbf{B}\mathbf{e}_l}{\mathbf{e}_l^{\mathrm{T}}\mathbf{T}\mathbf{e}_l} \tag{2}$$

subject to $\mathbf{E}$ being columnwise orthonormal can be solved as follows. The problems of maximizing (1) and (2) subject to $\mathbf{E}^{\mathrm{T}}\mathbf{E}=\mathbf{I}$ can be seen as special cases of the more general problem of maximizing

$$h(\mathbf{E}) = \sum_{l=1}^{r} \sum_{k=1}^{K} \frac{\mathbf{e}_l^{\mathrm{T}}\mathbf{A}_k\mathbf{e}_l}{\mathbf{e}_l^{\mathrm{T}}\mathbf{C}_k\mathbf{e}_l} \tag{3}$$

subject to $\mathbf{E}^{\mathrm{T}}\mathbf{E}=\mathbf{I}$, for which Kiers[4] has proposed an iterative algorithm. This algorithm can be summarized as follows.

Step 0. Initialize $\mathbf{E}$ (e.g. as an orthonormalized random matrix);
—compute $h(\mathbf{E})$;
—compute $\rho_k=\lambda_1(\mathbf{C}_k)$, the first eigenvalue of $\mathbf{C}_k$, $k=1, \ldots, K$.

Step 1. Compute $\mathbf{D}_k=\mathrm{diag}(\mathbf{E}^{\mathrm{T}}\mathbf{A}_k\mathbf{E})[\mathrm{diag}(\mathbf{E}^{\mathrm{T}}\mathbf{C}_k\mathbf{E})]^{-2}$, $k=1, \ldots, K$.

Step 2. Compute $\mathbf{M}=\sum_{k=1}^{K} \{\rho_k\mathbf{E}\mathbf{D}_k-\mathbf{C}_k\mathbf{E}\mathbf{D}_k+\mathbf{A}_k\mathbf{E}[\mathrm{diag}(\mathbf{E}^{\mathrm{T}}\mathbf{C}_k\mathbf{E})]^{-1}\}$.

Step 3. Compute the singular value decomposition $\mathbf{M}=\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathrm{T}}$, with $\mathbf{P}^{\mathrm{T}}\mathbf{P}=\mathbf{Q}^{\mathrm{T}}\mathbf{Q}=\mathbf{I}$ and $\boldsymbol{\Delta}$ a diagonal matrix with the singular values on the diagonal.

Step 4. Update $\mathbf{E}$ by $\mathbf{E}^{\mathrm{u}}=\mathbf{P}\mathbf{Q}^{\mathrm{T}}$.

Step 5. Evaluate $h(\mathbf{E}^{\mathrm{u}})$; if $h(\mathbf{E}^{\mathrm{u}})-h(\mathbf{E})>h(\mathbf{E}^{\mathrm{u}})*\varepsilon$ for some prespecified small value ε, go to Step 1, else consider the algorithm converged.

Kiers has shown that this algorithm converges monotonically. As with most iterative algorithms, this does not imply that the global optimum will always be attained. In a simulation study, Kiers reported that the global optimum was found in at least 50% of the cases, in various conditions, and usually much more frequently. This suggests that, employing a multistart procedure with, for instance, five random starts, one can be confident that the global maximum is found.

The above general algorithm can be used for maximizing $f(\mathbf{E})$ by taking $K=1$, $\mathbf{A}_1=\mathbf{B}$ and $\mathbf{C}_1=\mathbf{W}$. Taking $K=1$ clearly simplifies the algorithm, but the set-up of the algorithm remains as above and is therefore not repeated. For the maximization of $g(\mathbf{E})$ we take $K=1$, $\mathbf{A}_1=\mathbf{B}$ and $\mathbf{C}_1=\mathbf{T}$.

Krzanowski[3] described a modification of his algorithm to handle cases where the number of variables exceeds the number of observations, or more in general, for cases where $\mathbf{W}$ is singular. For that case, to avoid indefinite values of V_l, Krzanowski proposed to constrain solutions of $\mathbf{E}$ to the column space of $\mathbf{W}$. Let $\mathbf{F}$ denote a basis for the column space of $\mathbf{W}$. Then, for the singular case, his procedure comes down to applying his original procedure to matrices $\tilde{\mathbf{B}}=\mathbf{F}^{\mathrm{T}}\mathbf{B}\mathbf{F}$ and $\tilde{\mathbf{W}}=\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F}$ instead of $\mathbf{B}$ and $\mathbf{W}$ respectively. The simultaneous maximization procedures proposed here can readily be adapted to that situation by taking $\mathbf{A}_1=\mathbf{F}^{\mathrm{T}}\mathbf{B}\mathbf{F}$ and $\mathbf{C}_1=\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F}$ (maximization of $\sum_l V_l$) or $\mathbf{C}_1=\mathbf{F}^{\mathrm{T}}\mathbf{T}\mathbf{F}$ (maximization of $\sum_l \eta_l^2$).

3. ANALYSIS OF KRZANOWSKI'S FIRST DATA SET

Krzanowski's first data set* consists of scores of 87 students on seven physical variables. The students belonged to six different cohorts (consisting of 17, 17, 17, ten, eleven and 15 students respectively). The successive V-values obtained by Krzanowski's successive ODA analysis are reported in the first column of Table 1. (It should be noted that, for computing these, we used the eigenvector-based procedure given by Krzanowski[3] in passing (first paragraph of p. 516) rather than the iterative approach mentioned on p. 511.) The same data set was first analysed by our procedure for simultaneously maximizing $\sum_l V_l$. For comparative purposes we took $r=2, \ldots, 7$. Each analysis was based on five randomly started runs and the convergence criterion ε was taken as $\varepsilon=10^{-10}$. (In fact, these choices have been used for all analyses reported in the present paper.) It was found for all values of r that the five differently started runs led to the same solution, so we are confident that indeed the global maximum has been found in each case. The resulting function values (sums of r V-values) are reported in Table 1, as well as the individual V-values (in decreasing order). The latter are only

* The author thanks Wojtek Krzanowski for kindly making these data available.

Table 1. Values of V from successive and simultaneous ODA for Krzanowski's first data set

Successive	Simultaneous $r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$
4·54	4·44	4·41	4·38	4·36	4·35	4·34
3·48	3·70	3·68	3·65	3·61	3·60	3·59
2·92		3·08	2·87	2·70	2·66	2·61
1·79			2·68	2·43	2·39	2·33
1·22				2·40	2·35	2·29
0·23					1·58	1·43
0·06						1·10
$\sum_l^r V_l$	8·14	11·17	13·58	15·50	16·92	17·69
$\sum_l^r V_l$ (successive)	8·02	10·95	12·74	13·96	14·19	14·26

meaningful if the solution is (rotationally) unique. Although we could not prove this uniqueness, the conjecture that the solution is unique is corroborated by the finding that for each r all runs led to matrices **E** that differed only by permutations.

It can be concluded that the results from the successive and the simultaneous approach do not differ very much for $r=2$ and 3, although it is already apparent that the successive approach gives a smaller sum of V-values and that in the successive approach only the first component has better discriminatory quality than its counterpart in the simultaneous approach. From $r=4$ on, differences become considerable. In the simultaneous approach with $r=4$ the fourth component is almost as good as the third, whereas in the successive approach the fourth V-value is much smaller than the third. For $r=5$, 6, 7 a similar difference can be observed, culminating in the fact that the sixth and seventh components still have appreciable V-values (still larger than one), whereas those of the successive approach practically vanished.

The same data were next analysed by the method maximizing $\sum_l \eta_l^2$. Again in each analysis the five runs led to identical solutions. The values for η_l^2 are reported in Table 2, in the same format as in Table 1. It should be noted that the results for the successive approach are directly related to those in Table 1, since $\eta_l^2 = V_l/(V_l+1)$. The conclusions to be drawn from this analysis are almost identical with those

Table 2. Values of η^2 from successive and simultaneous ODA for Krzanowski's first data set

Successive	Simultaneous $r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$
0·82	0·81	0·81	0·81	0·81	0·81	0·80
0·78	0·79	0·79	0·78	0·78	0·78	0·78
0·74		0·76	0·74	0·73	0·72	0·71
0·64			0·73	0·71	0·71	0·69
0·55				0·71	0·70	0·69
0·19					0·63	0·61
0·06						0·57
$\sum_l^r \eta_l^2$	1·60	2·36	3·07	3·74	4·35	4·86
$\sum_l^r \eta_l^2$ (successive)	1·60	2·34	2·98	3·53	3·72	3·78

from the former analysis. For $r=2$ and 3 there are hardly any differences, but from $r=4$ on, differences become considerable, and even the sixth and seventh dimensions lead to appreciable discrimination as expressed by η^2=values over 0·5 (which imply that still more than 50% of the variation in these directions is accounted for by between-group variation).

The practical implications of these results on discrimination values (V and η^2) are clear. When, for these data, discrimination in only a few dimensions is desired, both methods lead to almost equally good discriminant functions. However, when discrimination in terms of many directions of the original data space is desired (up to as many as the dimensionality of the data space), the simultaneous approach does, overall, a considerably better job than the successive approach. The use of many components may seem unfeasible and uninteresting in practice. However, it should be borne in mind that these components, in fact, pertain to rotations of the original data space such that the main separations are found along such discriminating axes. As a consequence, when searching for important distinguishing directions, one can study such a solution component-by-component and one does not have to aim at grasping the full space spanned by such components. Hence, in case one wishes to cover much of the information in the data, it is useful and feasible to consider many discriminating components.

We also compared the components resulting from the successive and the simultaneous approaches. The first three dimensions were very similar. Correlations between co-ordinates obtained from the successive approach with corresponding co-ordinates from the simultaneous approach were over 0·85 for the V-based analyses and over 0·83 for the η^2-based analyses. As soon as four dimensions were considered, these correlations dropped considerably (to as low as 0·58 and 0·55 respectively). Taking more than four dimensions, correlations became even lower. To illustrate such differences, in Figure 2(a) we have plotted the data with respect to the sixth and seventh components obtained by successive ODA and in Figure 2(b) the data are plotted with respect to the sixth and seventh components from the simultaneous approach maximizing $\sum_{l=1}^{7}\eta_l^2$. It can be seen that, with respect to these dimensions, all groups overlap almost completely in the successive approach, whereas in the simultaneous approach the major distinction between groups 1, 2, 3 and 4, 5, 6 (males versus females) is still clearly visible, and in the male cohorts, some distinction between cohorts 1 and 2 can be observed.

4. ANALYSIS OF KRZANOWSKI'S SECOND DATA SET

The second data set* analysed by Krzanowski[3] concerned 24 meat products which were judged on 31 descriptors of texture. The meat products consisted of four groups ($n_1=6$ reformed meats, $n_2=5$ sausages, $n_3=7$ whole meats, $n_4=6$ beefburgers). This example served to illustrate Krzanowski's approach for cases where the number of variables exceeds the number of cases, thus leading to a singular matrix $\mathbf{W}$. In this case, Krzanowski's approach successively maximizes $V_l=\tilde{\mathbf{e}}_l^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{BF}\tilde{\mathbf{e}}_l/\tilde{\mathbf{e}}_l^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{WF}\tilde{\mathbf{e}}_l)$ over $\tilde{\mathbf{e}}_l$ subject to the constraints that $\tilde{\mathbf{e}}_l^{\mathrm{T}}\tilde{\mathbf{e}}_l=1$ and $\tilde{\mathbf{e}}_l^{\mathrm{T}}\tilde{\mathbf{e}}_j=0$ with $l=1, \ldots, r$ and $j=1, \ldots, l-1$, where $\mathbf{F}$ denotes a basis for the column space of $\mathbf{W}$. Our simultaneous variants of this approach consist of maximizing $\sum_l V_l=\sum_l(\tilde{\mathbf{e}}_l^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{BF}\tilde{\mathbf{e}}_l/\tilde{\mathbf{e}}_l^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{WF}\tilde{\mathbf{e}}_l)$ and $\sum_l\eta_i^2=\Sigma_l(\tilde{\mathbf{e}}_l^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{BF}\tilde{\mathbf{e}}_l/\tilde{\mathbf{e}}_l^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{TF}\tilde{\mathbf{e}}_l)$ respectively.

We reanalysed this data set with our simultaneous approach taking $r=2 \ldots 6$. For all analyses we checked whether the five different runs led to the same solution. This was the case in all analyses maximizing $\sum_l V_l$ and in all but the one with $r=6$ of the analyses maximizing $\sum_l\eta_l^2$; the latter analyses led to three local optima and two (identical) solutions that we considered to be the global optimum. The resulting V-values, as well as those obtained by Krzanowski,[3] are reported in Table 3. The η^2-values showed a similar pattern and are therefore not reported here. It can be seen that differences

* The author thanks Wojtek Krzanowski for kindly making these data available.

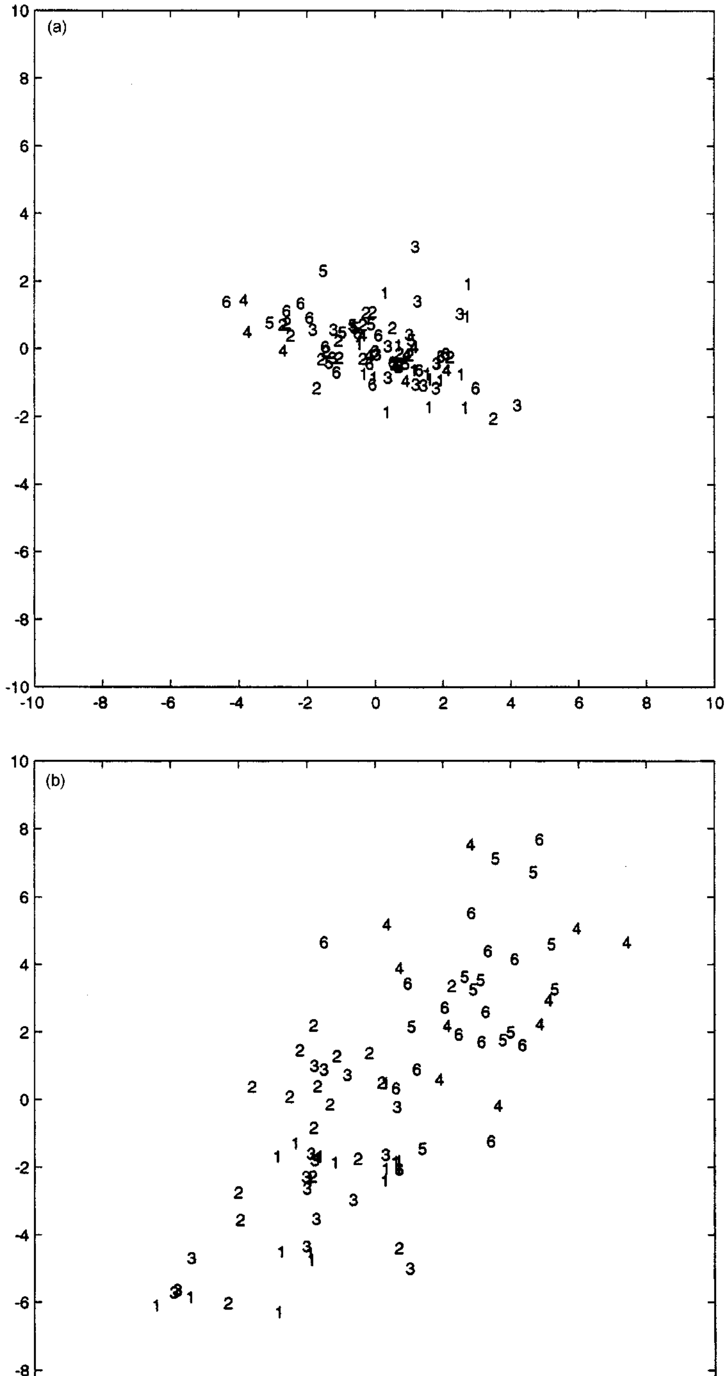

Figure 2. (a) Plot of sixth against seventh dimension of *successive* approach applied to first data set. (b) Plot of sixth against seventh dimension of *simultaneous* approach applied to first data set. Legend: 1, 2, 3, subjects from three male cohorts; 4, 5, 6, subjects from three female cohorts

Table 3. Values of V from successive and simultaneous ODA for Krzanowski's second data set

	Simultaneous				
Successive	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$
46·58	46·15	46·00	45·81	45·76	45·75
25·31	26·20	24·85	23·65	23·27	23·16
14·60		18·47	16·32	15·67	15·52
5·12			13·27	12·40	12·38
3·39				6·78	6·62
3·17					4·35
$\sum_l^r V_l$	72·35	89·31	99·06	103·89	107·78
$\sum_l^r V_l$ (successive)	71·90	86·50	91·62	95·01	98·17

become appreciable from $r=3$ and especially from $r=4$ on. It can be seen that the fourth dimension in the simultaneous solution has a considerably higher V-value than the fourth successive dimension. It can also be seen that there are cases where the first two dimensions of simultaneous solutions are poorer than the first two successive dimensions, but the third and higher ones are always better than their successive counterparts. A comparison of the meat product co-ordinates resulting from the different analyses is not pursued here, because the above reported results already demonstrate the main issues, i.e. the possibility and utility of the simultaneous variants of Krzanowski's approach.

5. ANALYSES OF IRIS DATA

In the above analyses, differences between the successive and simultaneous approaches only showed up from the third dimension on. To illustrate that this need not always be the case, we also report the results of the successive and simultaneous approaches (with $r=2$, 3, 4) applied to the Iris data set of Anderson[1] (see Section 1).

In all six analyses for simultaneous maximization the five different runs led to identical solutions. The set of solutions maximizing $\sum_l V_l$ gave a similar pattern of results as the set of solutions maximizing $\sum_l \eta_l^2$, hence it made little sense to report both. This time, only the η^2-values are reported (see Table 4). It can be seen that for these data the simultaneous solutions already start differing appreciably from $r=2$ on, and even bigger differences are observed for $r=3$ and 4. We inspected the

Table 4. Values of η^2 from successive and simultaneous ODA for Iris data set

	Simultaneous		
Successive	$r=2$	$r=3$	$r=4$
0·97	0·96	0·95	0·95
0·91	0·96	0·94	0·93
0·79		0·92	0·91
0·67			0·81
$\sum_l^r \eta_l^2$	1·92	2·82	3·60
$\sum_l^r \eta_l^2$ (successive)	1·88	2·67	3·33

co-ordinate plots for the analyses with $r=2$ and 3. In all cases, with respect to all dimensions, the first group (*setosa*) was well distinguished from the other two groups. The latter two groups showed some overlap, the amount of which differed from analysis to analysis. Upon comparing the successive and simultaneous (η^2-based) analyses with $r=2$, we found that the distinction between *versicolor* and *virginica* was somewhat clearer for the simultaneous approach than for the successive approach. To avoid giving too many plots, we merely describe our findings here. In the simultaneous solution there was a visible gap between the *versicolor* cluster and the *virginica* cluster (ignoring for the moment that five *versicolors* seemed to have stuck to the outside of the *virginica* cluster). In the successive solution (where again five *versicolors* stuck to the *virginica* cluster) such a gap was much harder to discern. The comparison of the $r=3$ solutions shows bigger differences, as can be seen from Figures 3(a) and 3(b), which plot the second against the third dimension for the successive solution and the simultaneous solution respectively. The differences between the solutions are clear. In the successive solution the *versicolor* and *virginica* clusters overlap to a large extent, whereas in the simultaneous solution these clusters show overlap only in terms of one displaced *versicolor*, and the clusters merely touch, but are otherwise perfectly distinguishable. This example clearly shows that the results of a successive approach may be too limited and that dimensions with higher discriminatory quality can be obtained by using the simultaneous approach instead.

6. DISCUSSION

Krzanowski[3] has proposed a method for rotating the data space such that the ensuing dimensions optimally discriminate the data according to a given group structure. Specifically, he proposed a method for searching orthogonal dimensions in the original data space along which the data are successively discriminated optimally. The present paper offers two simultaneous variants for the ODA approach proposed by Krzanowski. Thus methods are proposed for searching orthogonal dimensions which yield the highest *joint* discrimination. The choice between the two depends on how one wishes to measure discriminatory quality: in terms of the V-value or in terms of η^2. Because the latter is normed between zero and one, we have a slight preference for the latter, but as seen from the three analyses reported here, differences tend to be small.

Although the procedures described in the present paper are meant as descriptive tools, one may nevertheless wonder to what extent they can be used for predictive purposes as well. To give some insight into this question, a simple cross-validation was carried out on the analysis of the first data set by means of three methods: standard discriminant analysis, Krzanowski's successive approach and the present simultaneous approach (employing η^2-values). The data set was split into a training and a validation sample by alternately assigning the objects in the full sample to the training sample and to the validation sample. We used four dimensions in all three analyses of the training sample and applied the obtained weights to the validation sample; consequently, we computed the η^2-values for the validation sample and summed these as a measure for overall discriminatory quality of the weights resulting from the training sample, and compared these with the maximal η^2-values (i.e. obtained when analysing the validation sample itself by the respective methods). The results are reported in Table 5. It can be seen that all three methods cross-validated reasonably well, considering that the $\sum_l \eta_l^2$-values turned out to be 90%, 93% and 95% of the respective maximum values for the validation sample. It can be concluded that, at least for this data set, the ODA methods, even though they have not been designed for predictive purposes, can well be used in prediction and even turn out to perform slightly better than those for ordinary DA (as follows from the fact that for the ODA methods the summed cross-validation values are relatively closer to the optimal values than they are for DA). It can also be seen that the simultaneous approach not only yields a higher value of $\sum_l \eta_l^2$ than the successive approach does, but, at least for this data set, gives better relative cross-validatory results as well. For

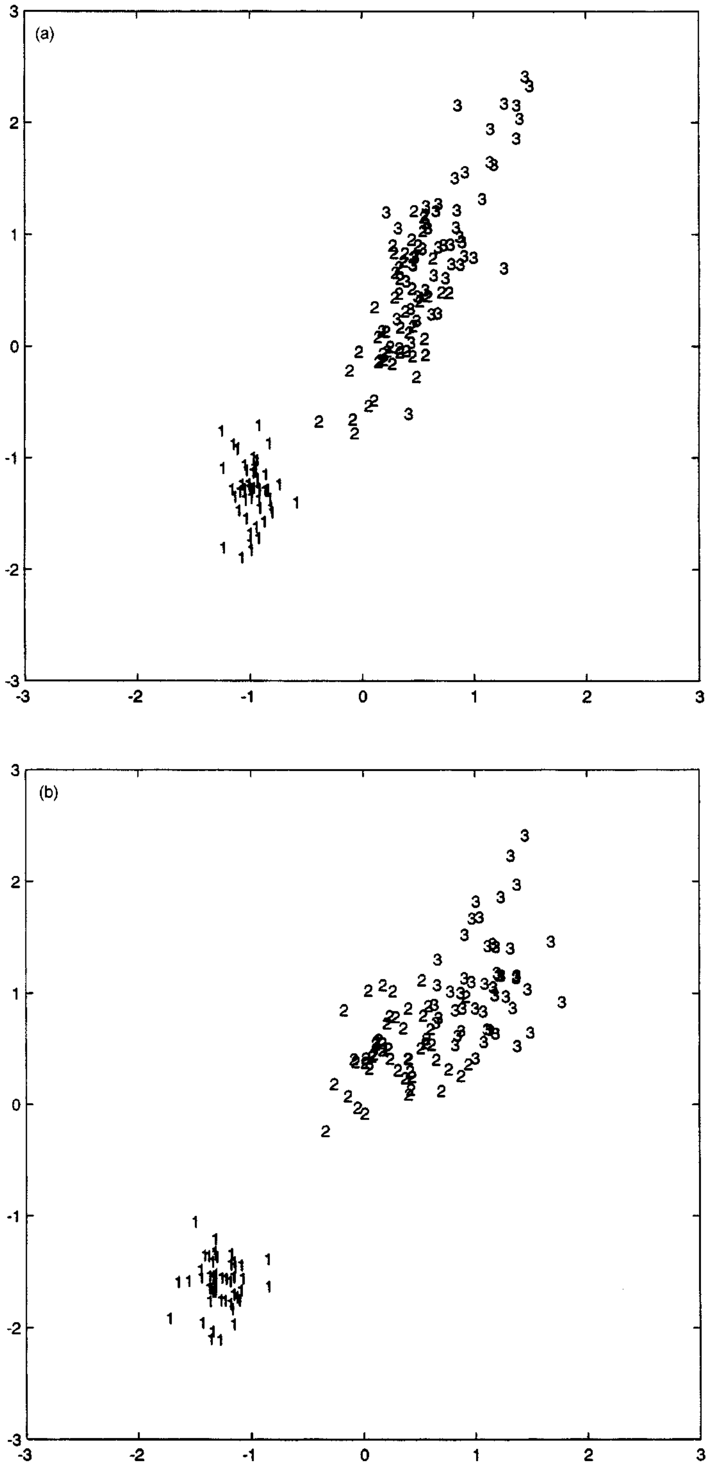

Figure 3. (a) Plot of second against third dimension of *successive* approach applied to Iris data set. (b) Plot of second against third dimension of *simultaneous* approach applied to Iris data set. Legend: 1, *iris setosa*; 2, *iris versicolor*; 3, *Iris virginica*

Table 5. Cross-validated (CV) and optimal (Opt.) values of η^2 from DA and successive and simultaneous ODA ($r=4$) for a validation sample derived from Krzanowski's first data set

	DA		Successive ODA		Simultaneous ODA	
	CV	Opt.	CV	Opt.	CV	Opt.
	0·80	0·83	0·80	0·83	0·79	0·82
	0·60	0·69	0·75	0·78	0·77	0·80
	0·09	0·17	0·68	0·74	0·72	0·75
	0·18	0·15	0·58	0·66	0·66	0·75
$\sum_l^r \eta_l^2$	1·67	1·85	2·81	3·01	2·95	3·11

the two ODA methods this difference turned out to be even larger when using seven dimensions, where in the successive approach the cross-validated $\sum_l \eta_l^2$ was 90% of its maximum value, whereas the simultaneous approach led to 96% of the maximal value; for ordinary DA, taking $r=7$ is impossible (since r is limited by the number of groups).

The simultaneous approaches proposed in the present paper are based on simply adding the discriminatory quality measures. Of course, simple addition is not the only possibility. Instead, one could take sums of squares of such measures, to mention just one possibility. This choice would lead to other maximization problems that can be handled by the general maximization technique proposed by Kiers.[4]

Following Krzanowski,[3] the approach for handling situations with more cases than variables was based on the analysis $\mathbf{F}^T\mathbf{BF}$ and $\mathbf{F}^T\mathbf{WF}$ instead of $\mathbf{B}$ and $\mathbf{W}$. These approaches, however, are by no means the only viable ways to handle such data. Other approaches that also give orthogonal components might, for instance, be constructed on the basis of the variety of methods discussed by Krzanowski *et al*.[5]

The use of orthogonal discriminating components was motivated by the fact that it does not distort the original data space. An obvious implication of this is that the ODA methods are sensitive to differential changes in scale of the variables. Therefore the method as such can only fruitfully be applied if the variables are measured at comparable scales or are first standardized to comparable scales. Another implication of the desire not to distort the original data space is that, if several variables discriminate the data in largely the same way, the scores on the ODA components will be strongly correlated as well, as was already found by Krzanowski[3] (p. 516). In fact, this demonstrates that the ODA methods, just as ordinary DA, do not aim at finding a limited number of components that optimally account for the variation in the data (as principal component analysis does), but merely at finding optimally discriminating components. A reviewer questioned the importance of the discriminatory measures being optimized by ODA methods in case of data with highly correlated variables. Indeed, when one analyses several variables that are correlated considerably, one may thus find rather uninteresting solutions where the same distinction is found repeatedly. To avoid such solutions, one may, prior to the search for discriminating directions, perform a 'summarizing' projection. One could follow a two-step approach in which first the data are optimally summarized by a limited number of components; the ensuing plot of the data will closely resemble the actual configuration in the full data space. Next, ODA methods can be used to find a (complete) rotation of this configuration such that the axes are the ones along which the data are best discriminated.

ACKNOWLEDGEMENTS

The author is obliged to Jos ten Berge and Wojtek Krzanowski for their helpful comments on an earlier version of this paper.

REFERENCES

1. E. Anderson, *Bull Am. Iris Soc.* **59**, 2 (1935).
2. B. Flury, *J. Am. Stat. Assoc.* **79**, 892 (1984).
3. W. J. Krzanowski, *J. Chemometrics*, **9**, 509 (1995).
4. H. A. L. Kiers, *Psychometrika*, **60**, 221 (1995).
5. W. J. Krzanowski, P. Jonathan, W. V. McCarthy and M. R. Thomas, *Appl. Stat.* **44**, 101 (1995).

