ROBUST CANONICAL DISCRIMINANT ANALYSIS

Peter Verboon
Ivo A. van der Lans

Department of Data Theory
University of Leiden

RR-93-02

ROBUST CANONICAL DISCRIMINANT ANALYSIS

Peter Verboon

Ivo A. van der Lans

Department of Data Theory

University of Leiden

This research was supported by a grant of the Netherlands Organization for Scientific Research (NWO) for the first author.
Requests for reprints should be sent to Peter Verboon, Department of Data Theory, Faculty of Social Sciences, University of Leiden, P.O. Box 9555, 2300 RB Leiden, The Netherlands.

Robust Canonical Discriminant Analysis

Abstract

A method for robust canonical discriminant analysis via two robust objective loss functions is discussed. These functions are useful to reduce the influence of outliers in the data. Majorization is used at several stages of the minimization procedure to obtain a monotonically convergent algorithm.

An advantage of the proposed method is that it allows for optimal scaling of the variables. In a simulation study it is shown that under the presence of outliers the robust functions outperform the ordinary least squares function, both when the underlying structure is linear in the variables as when it is nonlinear.

Key words: canonical discriminant analysis, robustness, outliers, optimal scaling, majorization.

1. Introduction

In classical canonical discriminant analysis (CDA) the data consist of scores on p observed variables for n objects, which are classified in m groups. In CDA this data set is usually called the training data set. The problem is to find r linear combinations of the observed variables that maximize the variation between the groups, relative to the variation within the groups (Fisher, 1936; Rao, 1948; 1952). The linear combinations are called canonical variates and the weights discriminant coefficients.

Classification is a major objective of CDA. The discriminant coefficients and group means are used to classify new observations (from the classification set), which may be difficult or impossible to classify directly. When CDA is defined in terms of the between versus within variance ratio, it has been shown by Campbell (1978; 1982) and Critchley and Vitiello (1991) that outliers can have a large influence on the CDA solution, and therefore on the classification of new observations. An outlier occurs when scores of an observation on one or more of the variables are atypical for the group to which it belongs, for instance due to misclassification.

Another objective of CDA is to map the high-dimensional observation space into a low-dimensional representation space, in which the scores of the objects on the canonical variates are displayed together with the group means and the variables (cf. Meulman, 1992). This low-dimensional representation may also be influenced by the presence of outliers in the data.

Robust procedures for CDA to decrease the disturbing influence of outliers were proposed by Randles et al. (1978), Broffitt et al. (1980), and Campbell (1982) . The general idea behind these procedures is to downweight observations that have a large Mahalanobis distance to their group mean. In section 4 a more detailed description is given of Campbell's procedure. In the present paper we propose an alternative

procedure. In contrast to Campbell's procedure, our procedure is based on the formulation of an explicit robust objective loss function for CDA. Two objective loss functions are used. The first function is based on the Huber estimator (Huber 1964), and the second on the biweight estimator (Beaton & Tukey, 1974). Both functions are well-known as robust criteria in multiple regression (Huber, 1981; Verboon, 1993). A monotonically converging algorithm is proposed to minimize these loss functions.

By formulating the problem in terms of a general objective loss function, it is possible to allow for optimal scaling of variables when the optimal canonical variates are not linear in the variables, for instance, by incorporating Kruskal's monotonic regression steps (Kruskal, 1964a; 1964b). It is hypothesized that the use of optimal scaling will yield better predictions and a more parsimonious description if there is some nonlinear structure in the data.

In a simulation study it is shown that CDA defined by robust objective loss functions performs better than least squares in terms of prediction of group membership and recovery of some initial group structure when the data contain outliers. In addition, it is shown that CDA defined by robust loss functions performs better than Campbell's robust procedure for CDA. Furthermore, the merits of optimal scaling are shown when the underlying structure is nonlinear.

2. Classical CDA

Let $\mathbf{X} = \{x_{ij}\}(i=1,..,n; j=1,...p)$ be a matrix with scores on p variables for n objects from m different groups. In classical CDA the aim is to find a linear combination of the observed variables (also called canonical variate), such that the variance of the group means of this canonical variate is maximized relative to the variance within the groups. The discriminant criterion is usually written down as the ratio

$$\psi = \frac{\text{SSQ(between)}}{\text{SSQ(within)}}, \tag{1}$$

where SSQ denotes sum of squares. The sum of SSQ(within) and SSQ(between) is equal to SSQ(total).

Let $\mathbf{G} = \{g_{ik}\}(i=1,..,n;\ k=1,...m)$ be a matrix that indicates which object belongs to which group, thus $g_{ik} = 1$ if the ith object belongs to group k, otherwise $g_{ik} = 0$. The canonical variate is given by $\mathbf{Xa}$, where $\mathbf{a}$ is the vector with discriminant coefficients. The group means of the canonical scores are given by

$$\mathbf{c} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{Xa}\ , \tag{2}$$

where $\mathbf{D}$ is defined as $\mathbf{G}'\mathbf{G}$, which is diagonal with the m group frequencies on the diagonal. Furthermore, $\mathbf{Gc}$ contains the group mean for each object . The total sum of squares of the canonical variate is $\mathbf{a}'\mathbf{X}'\mathbf{Xa}$, and the between sum of squares is $\mathbf{c}'\mathbf{G}'\mathbf{Gc}$. This last term is the same as $\mathbf{a}'\mathbf{X}'\mathbf{G}\mathbf{D}^{-1}\mathbf{G}'\mathbf{Xa}$, which is easily seen when expression (2) is substituted for $\mathbf{c}$. The within sum of squares is defined by the scores in deviation from their group means; thus by using (2):

$$(\mathbf{Xa} - \mathbf{Gc})'(\mathbf{Xa} - \mathbf{Gc}) = \mathbf{a}'\mathbf{X}'\mathbf{Xa} - \mathbf{a}'\mathbf{X}'\mathbf{G}\mathbf{D}^{-1}\mathbf{G}'\mathbf{Xa}, \tag{3}$$

which is the difference between the total and the between sum of squares. The within sum of squares is usually normalized to 1.

Having found a solution for $\mathbf{a}$ ($\mathbf{a}_1$) which maximizes the discriminant criterion ψ, a second canonical variate may be computed. The second canonical variate $\mathbf{Xa}_2$ also maximizes the discriminant criterion, subject to the restriction that it is orthogonal to the first one. In general, min $(m\text{-}1,p)$ canonical variates can be computed under the restriction $\mathbf{A}'\mathbf{X}'\mathbf{XA}$ = diagonal. In CDA the normalization restriction on the within sum of squares is $\mathbf{A}'\mathbf{X}'\mathbf{XA} - \mathbf{A}'\mathbf{X}'\mathbf{G}\mathbf{D}^{-1}\mathbf{G}'\mathbf{XA} = \mathbf{I}$. Henceforth, $\mathbf{XA}$ will be referred to as the matrix with object scores and $\mathbf{A}$ are the discriminant coefficients.

These discriminant coefficients can be found by an eigenvector decomposition of the matrix $\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}$, where **B** is the between-groups variance-covariance matrix of the variables, computed as $\mathbf{X'GD}^{-1}\mathbf{G'X}$, and **S** the within-groups variance-covariance matrix, computed as $\mathbf{X'X} - \mathbf{X'GD}^{-1}\mathbf{G'X}$. Subsequently, the group means of the canonical scores can be found via (2).

The object scores, together with the variables and group means can be represented in the *r*-dimensional space (cf. Ter Braak, 1990; Meulman, 1992). A new observation is classified by taking its scores $\mathbf{x}_i'$, and multiply them with **A**, yielding the co-ordinates of this new observation ($\mathbf{x}_i'\mathbf{A}$) in the objects space. By computing the Euclidean distances from the new object to the group means, the object can be classified in the group to which it is closest.

The discriminant criterion can be written as a canonical correlation problem, that is

$$\min \| \mathbf{GC} - \mathbf{XA} \|^2 \text{ over } \mathbf{A} \text{ and } \mathbf{C}, \tag{4}$$

where **G**, **C**, **X** and **A** are defined as before (Gittins, 1985). In canonical correlation analysis the normalization restrictions $\mathbf{A'X'XA} = n\mathbf{I}$ and $\mathbf{C'DC} = n\mathbf{I}$ are imposed to avoid degenerate solutions. Because of these normalizations **C** does not contain the group means anymore. The solution to (4) can be found by alternating least squares, thus by alternately solving one set of parameters while fixing the others, until the decrease in loss has become sufficiently small (Van der Burg, 1988) or by a singular value decomposition (e.g. Van de Geer, 1986).

It can be shown that the solution for canonical correlation analysis is identical to the CDA solution, except for a renormalization. In Van der Burg and De Leeuw (1983) the canonical correlation problem with optimal scaling is discussed .

3. Outliers in Canonical Discriminant Analysis

Like in most other MVA techniques, in CDA one has to be prepared for outliers in the data. In CDA, outliers are objects whose scores on the predictor variables are atypical for the group to which they belong. They may be due to (i) an only moderate relation between predictor variables and group classifications, in the sense that some objects do not follow the structure present in the majority of the data, (ii) errors in the scores on the predictor variables, (iii) misclassifications. No matter how they come about, the outliers can be expected to have different scores on the canonical variate compared to the scores of the other objects from the same group. Therefore, under CDA criteria that are sensitive to large residuals, like the least squares criterion in (4), outliers may lead to estimates for discriminant coefficients and group means that are distorted in the sense that they do not reflect the structure in the majority of the data. The influence of outliers on discriminant coefficients and group means depends on the structure of the outliers (Critchley and Vitiello, 1991; McLachlan, 1992, Section 2.5). For instance, discriminant coefficients and group means can be expected to be less influenced by random misclassifications in the training set, than by nonrandom misclassifications. In McLachlan (1992, p. 35-37) an overview is given of likelihood approaches to misclassifications in the training set.

When the means and the discriminant coefficients are distorted under the influence of outliers, the low-dimensional representation will not reflect the majority of the data. Thus, the Euclidean distances between the objects and their group means in the low-dimensional space will be distorted and, because these are the distances on which the classification is based, the prediction of the group membership for new observations can be expected to deteriorate. However, it should be noticed that under certain conditions it may also happen that outliers improve the classification of new observations, as Critchley and Vitiello (1991) demonstrated for two groups CDA.

4. Campbell's Robust Procedure for CDA

Campbell's robust procedure for CDA reduces the influence of outliers by downweighting them in the computation of the between groups variance-covariance matrix of the variables **B** and the within groups variance-covariance matrix of the variables **S**. The weights that are used for downweighting the outliers are determined iteratively, based on the maximum likelihood estimators for the functional relationship model for CDA under the assumption of elliptically symmetric densities (see Campbell, 1982). The functional relationship model assumes that the population group means of the variables lie in an r-dimensional hyperplane. In each iteration the matrices **B** and **S** are computed using the weights from the previous iteration. Then, discrimimant coefficients are obtained from the eigenvalue-eigenvector decomposition of $\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}$ and used to compute new estimates for the population group means and the population covariance matrix. Finally, new weights are computed as a monotonically decreasing function of the Mahalanobis distance (using the estimated population covariance matrix) between each object and its estimated (population) group mean, which concludes one iteration. Campbell suggest two different functions for the weights: one based on a non-descending influence function like the one suggested by Huber (1964) and one based on a redescending influence function like the one suggested by Hampel (1968). Clearly, Campbell's procedure is intuitively appealing. Unfortunately, however, there is no guarantee that it converges nor does Campbell give a criterion for when to stop the iterations.

5. The Huber and Biweight Functions

In the present paper we elaborate upon an approach which replaces the least squares criterion in (4) by a robust loss criterion, for which outliers are less harmful. The term

to be minimized in CDA can very generally be written as a summation over functions of residual elements

$$\sum_{i=1}^{n} f(r_i),$$

where the residuals are defined by the Euclidean distances between group means and canonical variate scores:

$$r_i = \sqrt{\sum_{s=1}^{r}\Big(\sum_{k=1}^{m} g_{ik}c_{sk} - \sum_{j=1}^{p} x_{ij}a_{js}\Big)^2} \tag{5}$$

Instead of choosing for $f(r_i)$ the least squares function, we will choose for it the Huber function (Huber, 1964) and Tukey's biweight function (Beaton & Tukey, 1974). The Huber function, $\phi(*)$, for a separate residual element, is defined as

$$\phi(r_i) = \begin{cases} 1/2\, r_i^2 & \text{if} \quad |r_i| < c \\ c|r_i| - 1/2c^2 & \text{if} \quad |r_i| \geq c. \end{cases} \tag{6}$$

where the constant c is called the tuning constant. For small residuals the ordinary least squares function is used, while for relatively large residuals, the least absolute residuals criterion is minimized. This means that with the Huber function the influence of large residuals is reduced compared to least squares. The biweight function, $\beta(*)$, is even more radical in this respect:

$$\beta(r_i) = \begin{cases} (c^2/6)(1 - (1 - (r_i/c)^2)^3) & \text{if} \quad |r_i| \leq c \\ c^2/6 & \text{if} \quad |r_i| > c. \end{cases} \tag{7}$$

Tukey's biweight is a hard redescending function, which means that its tolerance towards large residuals is large. It follows that outliers in the data may be arbitrarily far away from the 'model' and thus have large residuals, because beyond the tuning constant their contribution to the loss is constant. Consequently these points have no

further influence upon the solution, which is now entirely based on 'good' points only.

To minimize the Huber and biweight functions, the majorization method will be used. It can be proved (Verboon & Heiser, 1992) that majorization leads to a straightforward iteratively reweighted least squares (IRLS) algorithm, with two main steps. In one step, a weighted least squares problem is solved for a fixed set of weights, and in the other, the weights are chosen as a monotonically decreasing function of the residuals from the previous step.

It follows that in one step of the IRLS algorithm the following problem must be solved

$$\text{min tr } (\mathbf{GC} - \mathbf{XA})'\mathbf{V}(\mathbf{GC} - \mathbf{XA}), \tag{8}$$

over $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{C}$ with a fixed diagonal matrix $\mathbf{V}$ of order nxn, subject to the normalization restrictions $\mathbf{A'X'XA} = n\mathbf{I}$ and $\mathbf{C'G'GC} = n\mathbf{I}$. In the other step of the IRLS algorithm the weights v_i are adjusted. The algorithm alternates between these two steps until convergence. If the objective function is convex, like Huber's function with all variables treated as numerical, then majorization theory guarantees that the global minimum of the function will be attained. For non-convex functions like the biweight at least a local minimum will be attained.

Different robust functions correspond to different choices for the weights function. For the Huber function the weights are computed as

$$v_i = \begin{cases} 1 & \text{if} \quad \mathrm{r}_i < c \\ \dfrac{c}{\mathrm{r}_i} & \text{if} \quad \mathrm{r}_i \geq c\,. \end{cases} \tag{9}$$

where r_i represents the residual found in the previous step of the algorithm. The weights for the biweight function are found by

$$v_i = \begin{cases} (1 - (r_i/c)^2)^2 & \text{if} \quad r_i \leq c \\ 0 & \text{if} \quad r_i > c. \end{cases} \tag{10}$$

In both cases we obtain a set of weights ($0 \leq v_i \leq 1$) that can be used as diagnostics. Weights close to 1 are assigned to data that fit the CDA model well, while badly fitting points (outliers) will have small weights.

The tuning constant (c) for these functions is usually chosen as a function of some resistant spread measure (S) of the absolute residuals obtained by the least squares method (Verboon, 1993). For the Huber function the tuning constant is then chosen as 2/3S and for the biweight as 2S. In this paper the measure S is defined as

$$S = \text{median } \mathbf{r} + 4 \text{ MAD } \mathbf{r}, \tag{11}$$

where $\mathbf{r}$ is the vector with absolute residuals from the least squares analysis and MAD stands for the median absolute deviation. This choice for S and c has given fairly good results in a variety of situations that were studied.

Notice that the weights are a function of the distances between the objects and their group means in the r-dimensional discriminant space. This feature distinguishes our method from Campbell's procedure, in which the weights are a function of the Mahalanobis distances in the p-dimensional variables space (between the objects and the r-dimensional population group means).

6. Minimizing Weighted CDA

Within each iteration of the IRLS algorithm, the problem in (8) has to be solved over all $\mathbf{A}$ and $\mathbf{C}$ satisfying $\mathbf{A'X'XA} = n\mathbf{I}$, and $\mathbf{C'G'GC} = n\mathbf{I}$. Due to the weight matrix $\mathbf{V}$ in (8), the problem can not be rewritten in terms of a singular value

decomposition anymore. Therefore, an alternating least squares algorithm will be used. This algorithm alternates between two steps. In the first step, a better estimate for **A** is found given some (previous) estimates for **A** and **C**. In the second step, a better estimate for **C** is found given some (previous) estimates for **A** and **C**. In the sequel, better estimates are indicated by a hat, e.g. $\hat{\mathbf{A}}$, and previous estimates by the corresponding characters in outline, e.g. $\mathbb{A}$. The algorithm starts with some $\mathbb{A}$ and $\mathbb{C}$ that satisfy the restrictions. After each step, the just found better values are used to substitute the corresponding values in either $\mathbb{A}$ or $\mathbb{C}$. In that way, the algorithm gives a series of monotonically decreasing values of the term in (8). This series is bounded from below by zero, and will therefore converge to a stationary point.

Since matrices $\hat{\mathbf{A}}$ and $\hat{\mathbf{C}}$ can both be found in a similar way, only the steps taken to find $\hat{\mathbf{A}}$ are described here. They involve a second majorization procedure, in which a quadratic function in the metric **V** is majorized by a quadratic function in the metric **I**, the identity matrix (for details about this particular majorization procedure, see Heiser, 1987). In addition, the problem of finding $\hat{\mathbf{A}}$ is reformulated by writing **Z** instead of **XA**, in which case $\hat{\mathbf{Z}}$ has to be found subject to the constraints that the columns of $\hat{\mathbf{Z}}$ lie in the hyperplane spanned by the columns of **X**, and that $\hat{\mathbf{Z}}'\hat{\mathbf{Z}} = n\mathbf{I}$. The majorization procedure finds $\hat{\mathbf{Z}}$ by minimizing

$$\xi(\mathbf{Z}) = (\mathbf{Z}^{+} - \mathbf{Z})'(\mathbf{Z}^{+} - \mathbf{Z}), \tag{12}$$

subject to the above constraints, where $\mathbf{Z}^{+}$ is defined as the adjusted unrestricted update for **Z**, equal to

$$\mathbf{Z}^{+} = \mathbb{Z} + \frac{\mathbf{V}}{\max(\mathbf{V})}(\mathbf{G}\mathbb{C} - \mathbb{Z}), \tag{13}$$

where max(**V**) is the largest element of the diagonal matrix **V**.

The matrix $\mathbf{G}\mathbb{C}$ is the optimal unrestricted update of **Z**. Clearly, the function in (12) is minimized subject to the constraints, by first projecting $\mathbf{Z}^{+}$ onto the hyperplane

spanned by the columns of $\mathbf{X}$, and then solving the orthogonal Procrustes problem for the projected $\mathbf{Z}^+$. Subsequently, $\hat{\mathbf{A}}$ is found as

$$\hat{\mathbf{A}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{Z}}. \tag{14}$$

One can either do one step for $\mathbf{A}$ and go on with a step that finds a better $\mathbf{C}$, or do more steps for $\mathbf{A}$. Interchanging the roles of $\mathbf{X}$ and $\mathbf{G}$, and $\mathbf{A}$ and $\mathbf{C}$ in the previous description, gives steps for finding a $\hat{\mathbf{C}}$. Thus, the overall IRLS algorithm consists of outer majorization loops, in which steps towards the minimization of a robust loss function are taken by minimizing a series of weighted least squares functions, and inner majorization loops, in which the minimization of weighted least squares loss functions is accomplished by minimizing a series of unweighted least squares functions.

7. Extension with Optimal Scaling

Because the current approach formulates robust CDA in terms of the minimization of robust objective loss functions, it can easily be generalized to CDA with optimal scaling features. The idea behind CDA with optimal scaling is that optimal scores for the objects in the variables are computed, in addition to the optimal discriminant coefficients and the group means (cf. Gifi, 1990; Young, 1981). Thus a solution for (robust) CDA with optimal scaling is found by minimizing the (robust) loss function over $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{X}$. The scores are estimated subject to *scaling level* restrictions that usually operate within the variables. Well-known restrictions are inequality and equality restrictions that imply that the scores of each variable should lie in a closed convex cone. In the sequel, it is assumed that the restrictions on the jth variable define a cone Ω_j, such that they can be written as $\mathbf{x}_j \in \Omega_j$. For CDA with optimal scaling

extra steps, in which updates of $\hat{\mathbf{x}}_j$ are found subject to $\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A} = n\mathbf{I}$ and $\mathbf{x}_j \in \Omega_j$, are incorporated in the alternating lower squares algorithm. To find $\hat{\mathbf{x}}_j$ a procedure similar to the one developed by Meulman (1986, Chapter 4) is used. Instead of working with $\mathbf{X}$ and $\mathbf{A}$, the algorithm works with orthogonal bases $\mathbf{X}^o$'s of $\mathbf{X}$, and orthogonal matrices $\mathbf{A}^o$ of discriminant coefficients. That is, instead of the term in (8) we minimize

$$\text{tr}\,(\,\mathbf{GC} - \mathbf{X}^o\mathbf{A}^o\,)'\mathbf{V}(\,\mathbf{GC} - \mathbf{X}^o\mathbf{A}^o\,). \tag{15}$$

The crux of Meulman's procedure (see Appendix A) consists of a change between conveniently chosen orthogonal bases $\mathbf{X}^o$. Each new orthogonal basis $\mathbf{X}^o$ is a rotation of the previous orthogonal basis $\mathbf{X}^o$. With each change of basis, $\mathbf{A}^o$ is simultaneously premultiplied with the transpose of the particular rotation matrix, and therefore the loss will not change by the rotation. The steps in which $\hat{\mathbf{A}}$ is found are replaced by steps in which $\hat{\mathbf{A}}^o$ is found. In these steps the function in (12) is minimized again, where $\mathbf{Z}$ is now equal to $\mathbf{X}^o\mathbf{A}^o$. Obviously, $\hat{\mathbf{A}}^o$ is found directly by solving the orthogonal Procrustes problem for $\mathbb{X}^{o\prime}\mathbf{Z}^+/n$. After convergence of the overall IRLS algorithm, $\mathbf{A}$ is obtained as $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}^o\mathbf{A}^o$, and the constraint $\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A} = n\mathbf{I}$ is automatically satisfied.

At convergence of the overall algorithm, at least a local minimum of the loss function for robust CDA is found, subject to the CCA normalization. This solution is unique up to a rotation. A unique solution is identified by postmultiplying $\mathbf{A}$ and $\mathbf{C}$ by the rotation matrix that is obtained by taking, in reverse order, all r eigenvectors of the matrix $(\,\mathbf{GC} - \mathbf{XA}\,)'\mathbf{V}(\,\mathbf{GC} - \mathbf{XA}\,)$. This procedure assures that the first dimension of the rotated CCA solution has the lowest loss, followed by the second dimension, and so on. For robust CDA, the solution is renormalized in a way that resembles the renormalization in the least squares case. The idea here is to renormalize the solution in such a way that the weighted SSQ(within)'s are equal to 1. For that purpose, each

column $\mathbf{c}_r$ of **C** is first multiplied by the weight from the regression of $\mathbf{Xa}_r$ on $\mathbf{Gc}_r$. Secondly, corresponding columns of **A** and **C** are renormalized by multiplying them with the same factor such that $(\mathbf{Xa}_r - \mathbf{Gc}_r)'\mathbf{V}(\mathbf{Xa}_r - \mathbf{Gc}_r) = 1$ for each dimension. When all weights are equal to 1, the renormalization will yield the usual CDA solution, with the usual additivity of the SSQ's, equivalence of SSQ's and variances, and diagonality of the total, between, and within variance-covariance matrices.

Some of these properties are lost, when there are weights that are smaller than 1. Because of the CCA constraint $\mathbf{C}'\mathbf{G}'\mathbf{GC} = n\mathbf{I}$ and the weight matrix **V**, **C** will not contain the group means of **XA** neither its (weighted) group means, and the columns of **GC** and **XA** - **GC** will not have zero mean.

8. A Simulation

In a simulation study it has been studied how the three loss criteria, least squares (LSQ), Huber, and biweight perform under various experimental conditions.

The data

The data have been constructed such that the following model holds

$$y_{ks} = \mathbf{x}_{ik}' \mathbf{b}_s, \tag{16}$$

where $\mathbf{x}_{ik}'$ ($\mathbf{x}_{ik}' \neq \mathbf{x}_{jk}'$ if $i \neq j$) represents the scores of the ith object belonging to the kth group (i=1,...,42; k=1,...,7) on five variables and $\mathbf{b}_s$ a vector with 'true' discriminant coefficients for the sth dimension (s=1, 2). The matrix **B** with 'true' discriminant coefficients is fixed throughout the experiment. The y_{ks} is the mean of the kth group, which are gathered in the matrix **Y** (7 by 2) representing an ideal group

structure in two dimensions. The seven points are positioned on the vertices of an hexagon with one point in the centroid.

Two situations are distinguished: in the first there is a linear structure of the variables, and in the second the structure is nonlinear. In the linear situation, uniformly distributed random error has been added to the columns of **X**, with variance proportional to the variances of the corresponding columns of **X**. The error proportions (ε) are 0%, 1%, 5%, and 10%. Finally, outliers have been added by randomly selecting α percent of the data and randomly assigning these data points to different groups. The levels of outlier contamination (α) are 0%, 5%, 10%, and 20%. These data were also analyzed with Campbell's procedure (Campbell, 1982), using the Huber weights, to see if the procedure that we propose, performs better with respect to group prediction in the linear situation. The number of replications in each condition is chosen to be 50.

In the nonlinear situation, four variables have been transformed as follows:

$$\mathbf{x}_1 <= \sqrt[3]{\mathbf{x}_1} ,$$

$$\mathbf{x}_2 <= e^{\mathbf{x}_2} ,$$

$$\mathbf{x}_3 <= (\mathbf{x}_3)^5 ,$$

$$\mathbf{x}_4 <= \log (\mathbf{x}_4 + \min \{\mathbf{x}_4\}).$$

Two conditions have been studied in the nonlinear situation, in which two proportions of random error have been added to the transformed variables and two percentage of outlier contamination have been considered. The two extra conditions are (i) $\varepsilon = 1\%$, $\alpha = 5\%$, and (ii) $\varepsilon = 5\%$, $\alpha = 10\%$. These conditions were replicated 10 times.

Analysis

In the linear situation the variables are treated as numerical, while in the nonlinear situation the variables are treated both numerically and ordinally. In order to compare the three loss functions, first the number of correctly predicted observations is computed by the resubstitution method, that is, take the Euclidean distances between the object scores and the centroids, and assign the observation to the group to which centroid it has the smallest distance. The proportion of correctly predicted group memberships (PRED) is taken as the first external measure, where correct is evaluated with respect to the true group (i.e. the classifications before outliers were added).

As a second measure of performance, the group means (**C**) are compared with the ideal group structure (**Y**). The comparison is achieved by orthogonally rotating the two structures to each other. The remaining least squares loss, defined as

$$\text{LOSS} = 1/n \text{ tr } (\mathbf{Y} - \mathbf{CR})'(\mathbf{Y} - \mathbf{CR}), \tag{17}$$

where **R** is a (2 by 2) rotation matrix, is used as a measure that indicates how close the analysis recovered the original group structure.

Finally, the weights obtained by the robust loss functions are examined. For each analysis an average weight has been computed over all constructed outliers and also over all other points. Small weights should ideally be assigned to outliers and large ones to the other points.

Results

First, results are given for the analyses in the linear conditions, in which the variables have not been transformed. In Table 1 the three output measures are given for the three loss functions averaged over all replications, all error levels, and over all proportions of outliers.

Table 1. *Results Simulation Study, Averaged over Error Levels and Proportion Outliers*

	PRED	LOSS	Weights for outliers	Weights for others
LSQ	48.8	.807	1.000	1.000
HUB	50.0	.801	.76	.98
BIW	50.5	.800	.62	.93

With respect to group prediction (percentage correctly classified) the biweight function performs slightly better than the Huber function, and both robust ones are doing better than least squares. The loss due to the lack of fit of the recovered group structure with the ideal structure, is very similar for all three functions: it is largest for least squares, while this loss for the Huber and biweight functions is approximately equal. For the robust functions weights assigned to outliers are much smaller than the ones assigned to other points.

In Figure 1 the predictions are shown as function of the proportion outliers for all levels of random error. Note that the scale values of the first plot (0% random error) differ from the others. Without random error the robust functions perform perfectly: all observations, including the outliers, are correctly classified. For LSQ there is a slight decrease in proportion correctly predicted, but even with 20% outliers PRED is still about 95%. The improvement of the robust functions over LSQ is rather consistent over the random error levels. Of course, prediction becomes worse with increasing error. The differences in the percentage correctly predicted increase with proportion outliers involved. Huber predicts up to 1.5% better, and the biweight up to 3% for the highest level of α.

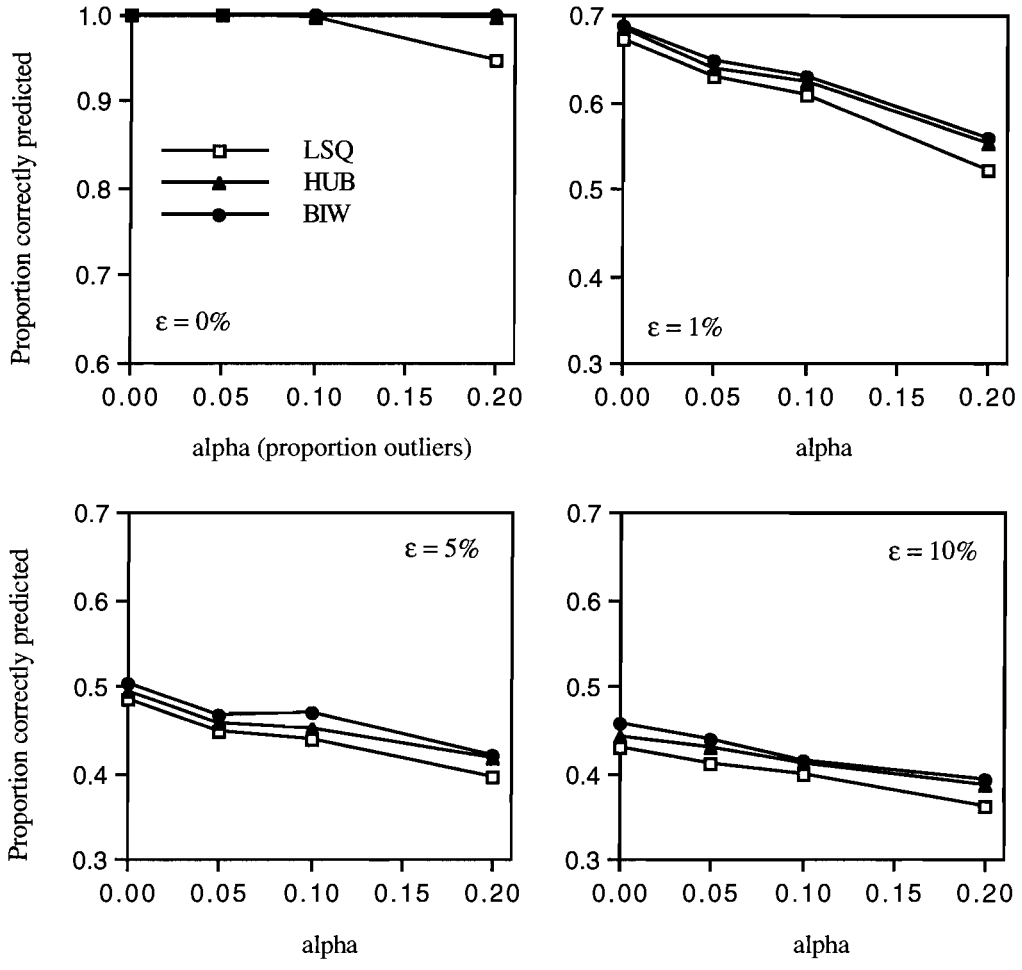

Figure 1. *Proportion correctly predicted by LSQ, HUB, and BIW for four levels of α and four levels of ε.*

The loss values obtained by comparing the computed group structures with the initial ones are given in Figure 2. Note again that the scale values of the 0% error plot differ from the others. Without random error the group structure obtained by the robust functions is closer to the original one than least squares is. However, with increasing error the differences between the three functions become smaller, and sometimes LSQ is even performing somewhat better. It appears that the effect of the outliers is rather mixed for all three functions, as is shown by the non-monotonic results in Figure 2.

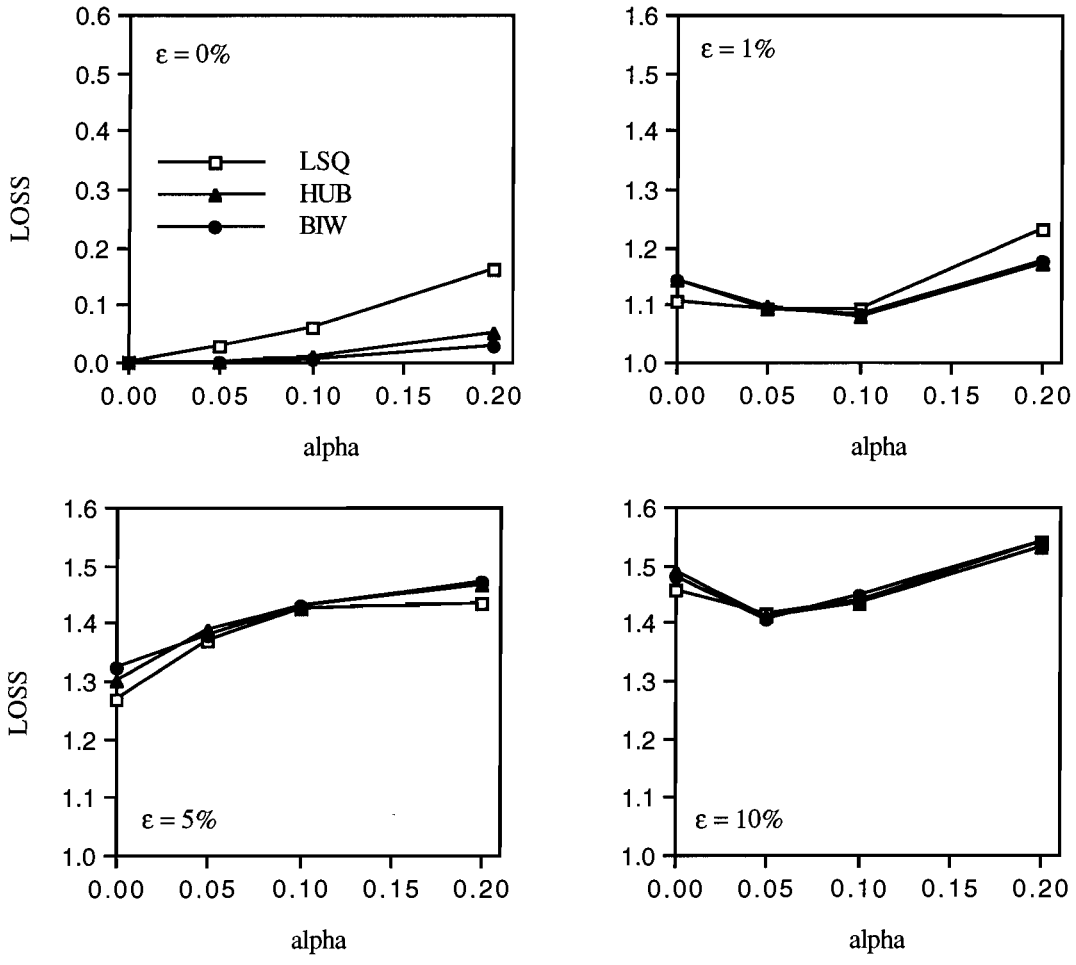

Figure 2. *Loss after matching recovered structure with initial structure by LSQ, HUB, and BIW for four levels of* α *and four levels of* ε.

The values of the weights that are assigned to the outliers and to the other observations are closer with increasing values of ε. In other words, in the situation with much random error it is harder to distinguish between outliers and other points by inspecting the weights. This can best be seen if the weights are plotted against the level of random error (Figure 3).

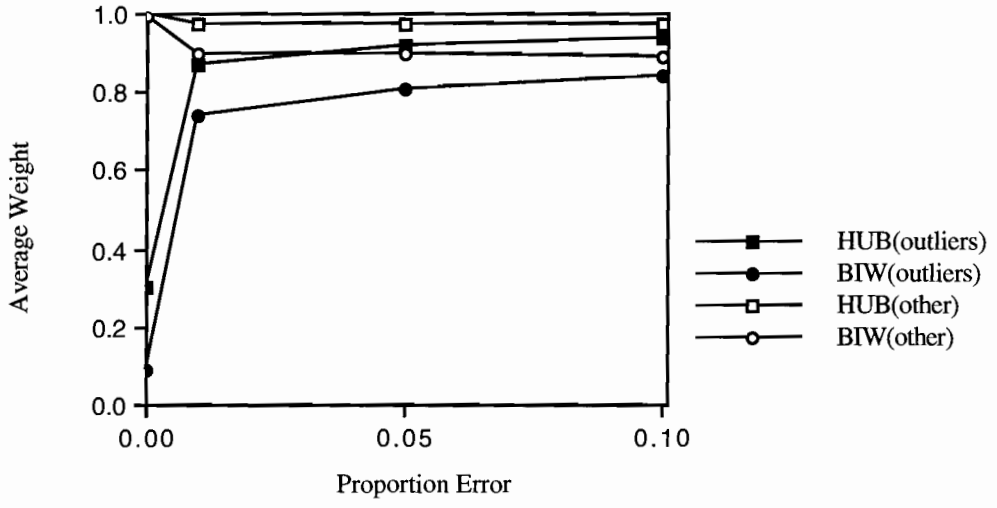

Figure 3. *Weights assigned to outliers and other observations for HUB and BIW.*

When there is no random error almost zero weights are assigned to the outliers for the biweight. The magnitude of the weights assigned to the outliers rapidly grows when more error is added. For the Huber function this effect is also present but the weights are larger. The weights assigned to the other points are well above .90 for the biweight and about .97 for the Huber function.

The results of the analyses of these data by Campbell's procedure are given in Table 2. For Campbell only the Huber weighting scheme has been used. The number of correctly predicted observations is a few percent smaller in Campbell's procedure. The differences between Campbell's procedure and the one (with Huber weights) that we propose are consistent over all levels of outlier contamination and random error. Campell's procedure is performing slightly better than LSQ.

The weights assigned to the observations in Campell's procedure do not distinguish very well between outliers and other points. The average weight for outliers was .98 (cf. .76 for HUB in Table 1), and all other points were assigned weights equal to one.

Table 2. Results Simulation Study: Proportion Correctly Predicted with Campbell's Procedure using Huber Weights

	α			
ε	0	5	10	20
.00	1.00	.99	.99	.96
.01	.66	.63	.59	.53
.05	.47	.46	.45	.40
.10	.42	.40	.39	.37

In the conditions with transformed variables it is found that treating the variables at an ordinal scaling level, thus applying monotonic regression, outperforms the numerical treatment (see Figure 4 and 5). However, contrary to the numerical situation the weights do not distinguish very well between outliers and other points. The robust functions are still predicting a little better than least squares (Huber 4% ordinal and 0.3% numerical; biweight 4.5% and 2.0%, respectively). Also, least squares gives a slightly higher loss after matching the group structure with the original structure than the other functions. Curiously, the numerical solutions seems to be closer to the original structure than the ordinal solutions.

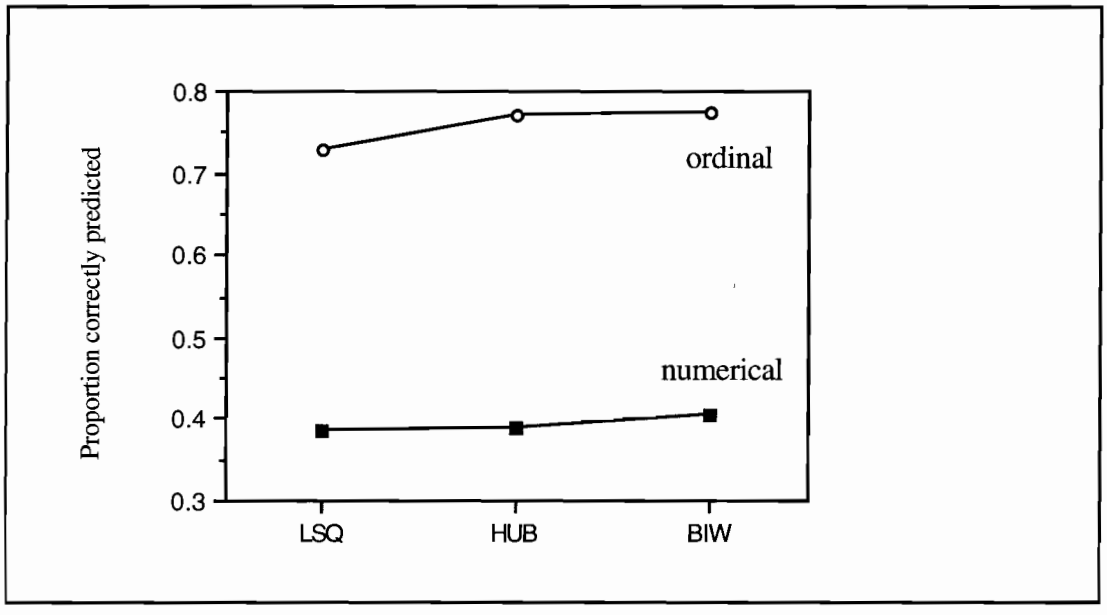

Figure 4. *Proportion correctly predicted by LSQ, HUB, and BIW in ordinal and numerical situation, after transformation of the variables.*

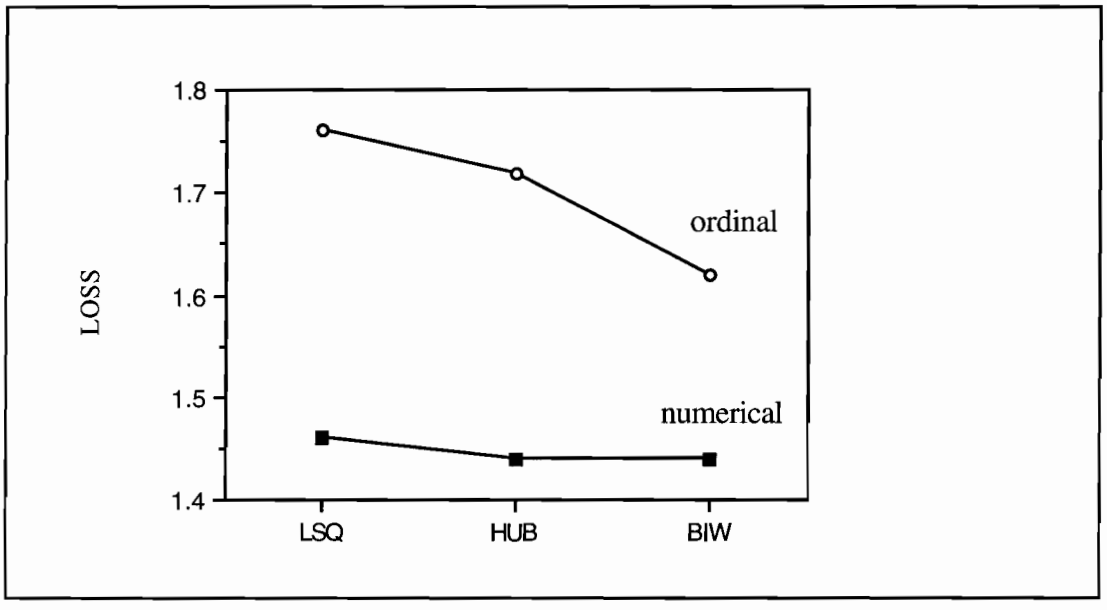

Figure 5. *Loss after matching recovered structure with initial structure by LSQ, HUB, and BIW in ordinal and numerical situation, after transformation of the variables.*

9. Discussion and Conclusions

In the present paper we have shown that the robust functions, used in this study as alternatives for least squares, were able to improve the solution when outliers were present, under various conditions of random error. With larger proportions of random error the robust approaches do not seem to be able to distinguish between outliers and the random error. The value of the weights assigned to the outliers then comes close to the value of the weights for the other points.

Especially with respect to the group prediction criterion the robust functions scored a few percent higher than least squares. Although a few percent does not seem much, one could easily think of real life situations in which a few percent improvement of prediction could have an enormous practical impact. For this reason, a robust approach to canonical discriminant analysis can be very useful.

Comparison of our Huber approach with the robust (Huber weights) procedure of Campbell yielded somewhat better results of our procedure with respect to group prediction. Also the weights discriminated better between the outliers and the other points. Furthermore, we found that Campbell's procedure occasionally yielded degenerate solutions.

Allowing for ordinal transformations of the variables may yield better predictions as was shown in the simulation study. In the nonlinear situation, the outliers were not clearly distinguished from the other points, and therefore the performance of the robust functions was only slightly better than least squares.

It is puzzling why the group means were further away from the original structure in the ordinal analyses than in the numerical ones, despite their much better predictions. It seems the ordinal analyses have much freedom to make group structures which may be better in predicting classifications than the constructed one.

In the simulation study, type (iii) outliers (see Section 3) were constructed by misclassifying a proportion of the objects. Clearly, the study showed that in that case

robust loss functions yield a better CDA-based classification of (new) objects than the least squares loss function does. In practice, we will also encounter type (i) and type (ii) outliers. The most important difference between type (i) outliers on the one hand, and type (ii) and type (iii) outliers on the other hand, is that type (i) outliers will also occur among new objects in the classification set. Therefore, to compare the loss functions on their ability to predict group membership under the presence of type (i) outliers, the predicted group memberships should be compared with the perturbed classification. The differences between the three loss functions on these percentages correct classifications can be expected to be smaller then. On the other hand, objects were randomly misclassified and the difference between the percentages can be expected to be larger in a simulation study in which the objects are non randomly misclassified. The difference between type (ii) and type (iii) outliers is that the scores of type (ii) outliers are atypical in the predictor space, not only for their own group, but also for the whole set of predictor scores (which was not the case in the simulation study). In the presence of type (ii) outliers, the improvement in classifications by using robust loss functions will be larger when the errors in the scores of new objects on the predictor variables can be kept to a minimum. As *a priori* classifications do not play a role in the classification of new objects, the most straightforward improvement in classification is obtained under type (iii) outliers.

A disadvantage of the robust loss functions is that the IRLS-algorithm for minimizing them requires a number of nested iterative procedures. Especially, the procedure for the optimal scaling of the variables under the orthogonality restrictions on the canonical variates is very time-consuming. It would therefore be worthwhile to consider different procedures for the optimal scaling step, or alternative loss functions, that have a comparable performance in terms of finding a low-dimensional representation for the majority of the data, and in terms of the ability to predict group membership under the presence of outliers.

Appendix

Let $\mathbf{G}$ be a given $n \times m$-matrix, $\mathbb{C}$ a given $m \times r$-matrix, $\mathbb{A}$ a fixed $p \times r$-matrix, and $\mathbf{V}$ a positive semi-definite $n \times n$-matrix. Furthermore, let $\mathbf{X}$ be the $n \times p$ matrix to be estimated. Consider the quadratic minimization problem

$$\min \sigma(\mathbf{X}) = \text{tr}\,(\mathbf{G}\mathbb{C} - \mathbf{X}\mathbb{A})'\mathbf{V}(\mathbf{G}\mathbb{C} - \mathbf{X}\mathbb{A}), \tag{A1}$$

$$\text{subject to } \mathbb{A}'\mathbf{X}'\mathbf{X}\mathbb{A} = n\mathbf{I} \text{ and } \mathbf{x}_j \in \Omega_j, \forall\, j \in \{1,\ldots, p\}.$$

Here, $\mathbf{I}$ denotes the $r \times r$ identity matrix, $\mathbf{x}_j$ denotes the j^{th} column of $\mathbf{X}$, and Ω_j is a cone. Clearly, at least a local minimum of the function in (A1) can be found by updating $\mathbf{X}$ column-wise in an alternating lower squares algorithm.

Let $\mathbb{X}$ be a given matrix $\mathbf{X}$ that satisfies the above constraints. To find an update for $\mathbf{x}_j$, a procedure similar to the one developed by Meulman (1986, Chapter 4) is used. That is, let $\mathbb{X}^o$ be an orthogonal basis of $\mathbb{X}$ ($\mathbb{X}^{o\prime}\mathbb{X}^o = n\mathbf{I}$). Because $\mathbb{A}'\mathbb{X}'\mathbb{X}\mathbb{A} = n\mathbf{I}$, there exists an orthogonal matrix $\mathbb{A}^o$ ($\mathbb{A}^{o\prime}\mathbb{A}^o = \mathbf{I}$), such that $\mathbb{X}^o\mathbb{A}^o = \mathbb{X}\mathbb{A}$. In particular, let $\mathbb{X}^p$ be a matrix, obtained by a column-wise permutation of $\mathbb{X}$, with $\mathbb{x}_j$ in the last column. Furthermore, let $\mathbb{X}^o$ be the orthogonal basis that is obtained by a back-permutation of the Gram-Schmidt decomposition of $\mathbb{X}^p$. Then we can write $\mathbb{x}^o_j$ as $\alpha\Psi\mathbb{x}_j$, in which Ψ is the anti-projector $\mathbf{I} - \mathbb{X}^o_{(-j)}\mathbb{X}^o_{(-j)}'/n$, where $\mathbb{X}^o_{(-j)}$ denotes $\mathbb{X}^o$ without its j^{th} column, and α is a normalization factor, such that $\mathbb{x}^o_j{}'\mathbb{x}^o_j = n$. Because of the particular Gram-Schmidt decomposition, $\mathbb{X}^o_{(-j)}$ is a function of the columns of $\mathbb{X}_{(-j)}$ only. Let $\mathbf{Z}$ be equal to $\mathbf{G}\mathbb{C} - \mathbb{X}^o_{(-j)}\mathbb{A}^o_{(-j)}$, in which $\mathbb{A}^o_{(-j)}$ denotes $\mathbb{A}^o$ without its j^{th} row. Then, minimizing

$$\sigma(\mathbf{x}_j) = \text{tr}\,(\mathbf{Z} - \alpha\Psi\mathbf{x}_j\mathbb{a}^o_j{}')'\mathbf{V}(\mathbf{Z} - \alpha\Psi\mathbf{x}_j\mathbb{a}^o_j{}'), \tag{A2}$$

subject to the restrictions, for given $\mathbb{X}^o$, $\mathbb{C}$, and $\mathbb{A}^o$, finds an update for $\mathbf{x}_j$. This minimization problem is rather complicated because unrestricted updates for $\mathbf{x}_j$ have to

be projected onto Ω_j in the metric $\Psi\mathbf{V}\Psi$, whereas projected $\mathbf{x}_j$'s have to be normalized in the metric Ψ. The problem is solved by using two nested iterative majorization procedures. In the outer procedure, we bring the minimization of (A2) back to successive minimizations of

$$\xi(\mathbf{x}_j) = (\mathbf{q}^+ - \alpha\Psi\mathbf{x}_j)'(\mathbf{q}^+ - \alpha\Psi\mathbf{x}_j), \tag{A3}$$

in which $\mathbf{q}^+$ is the adjusted unrestricted update for $\alpha\Psi\mathbf{x}_j$ analogous to the majorization procedure in Heiser (1987). That is,

$$\mathbf{q}^+ = \alpha\Psi\mathbb{x}_j + \frac{\mathbf{V}}{\beta(\mathbf{V})}\left(\frac{\Sigma_i\, a^o_{ij}\mathbf{z}_i}{\Sigma_i\, a^o_{ij}{}^2} - \alpha\Psi\mathbb{x}_j\right), \tag{A4}$$

in which $\beta(\mathbf{V})$ is (an upper bound for) the largest eigenvalue of $\mathbf{V}$. Of course, this largest eigenvalue equals max($\mathbf{V}$) when $\mathbf{V}$ is diagonal. It follows from the results of De Leeuw (1977) on normalized cone regression, that (A3) can be minimized by minimizing

$$\xi(\mathbf{x}_j) = (\mathbf{q}^+ - \Psi\mathbf{x}_j)'(\mathbf{q}^+ - \Psi\mathbf{x}_j), \tag{A5}$$

and by computing a new α afterwards. In the inner majorization procedure, we bring the minimization of (A5) back to successive minimizations of

$$\xi(\mathbf{x}_j) = (\mathbf{x}^+_j - \mathbf{x}_j)'(\mathbf{x}^+_j - \mathbf{x}_j), \tag{A6}$$

where, the adjusted unrestricted update $\mathbf{x}^+_j$ is given as

$$\mathbf{x}^+_j = \mathbb{x}_j + \Psi(\mathbf{x}^+_j - \mathbb{x}_j). \tag{A7}$$

In each loop of the inner majorization procedure, $\mathbf{x}^+_j$ is projected onto the intersection of the hyperplane that is orthogonal to $\mathbf{1}$, the n vector of ones, and Ω_j. That is, the projected $\mathbf{x}^+_j$ has to have zero mean and to lie in Ω_j. These zero means are

required to avoid degenerate solutions with one canonical variate $\mathbf{Xa}_r$ that tends to be equal to **1**. In order to apply the results of De Leeuw (1977) on normalized cone regression, we have to obtain the minimum for (A5), and therefore we have to cycle through the inner majorization loop until convergence in theory. In practice, we can perform a limited number of inner majorization iterations and check whether they are sufficient to decrease the value of (A3). The algorithm can be made faster, by renormalizing $\mathbb{x}_j$ such that the renormalized $\mathbb{x}_j$ minimizes (A5) over all renormalized $\mathbb{x}_j$'s, before starting the inner majorization loop.

References

Beaton A. E. & Tukey J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics, 16*, 147-185.

Broffitt, J. B., Clarke, W. R., & Lachenbruch, P. A. (1980). The effect of Huberizing and trimming the quadratic discriminant function. *Comm. Stat. Theor. Meth., A9 (1)*, 13-25.

Campbell, N. A. (1978). The influence function as an aid in outlier detection in discriminant Analysis. *Applied Statistics, 27,* 251-258.

Campbell, N. A. (1982). Robust procedures in multivariate analysis. II: Robust canonical variate analysis. *Applied Statistics, 31,* 1-8.

Critchley, F. & Vitiello C. (1991). The influence of observations on misclassification probability estimates in linear discriminant analysis. *Biometrika, 78,* 677-690.

De Leeuw, J. (1977). *Normalized cone regression* (research report). Leiden: Department of Data Theory.

Fisher, R. A. (1936). The use of multiple measurement in taxonomic problems. *Ann. Eugen.,7*, 179-188.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester: Wiley.

Gittins, R. (1985). *Canonical Analysis.* Berlin: Springer-Verlag.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation.* Ph.D. thesis, University of California, Berkeley.

Heiser, W. J. (1987). Correspondence analysis with least absolute residuals. *Computational Statistics and Data Analysis, 5*, 337-356.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist., 35*, 73-101.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1-27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*, 115-129.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition.* New York: Wiley.

Meulman, J. J. (1986). *A distance approach to nonlinear multivariate analysis.* Leiden: DSWO-Press.

Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika, 57*, 539-565.

Randles, R. H., Broffitt, J. B., Ramberg, J. S. & Hogg, R. V. (1978). Generalized linear and quadratic discriminant functions using robust estimates. *Journal of the American Statistical Association, 73*, 564-568.

Rao, C. R. (1948). The utilization of multiple measurement in problems of biological classification. *Journal of the Royal Statistical Society, B, 10*, 159-203

Rao, C. R. (1952). *Advanced statistical methods in biometric research.* Wiley: New York.

Ter Braak, C. J. F. (1990). Interpreting canonical correlation analysis through biplots of structure and weights. *Psychometrika, 55*, 519-532.

Van de Geer, J. P. (1986) *Introduction to linear multivariate data analysis* (2 Vols.) Leiden: DSWO Press.

Van der Burg, E. (1988). *Nonlinear canonical correlation and some related techniques.* Leiden: DSWO-press.

Van der Burg, E. & De Leeuw, J. (1983). Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology, 36*, 54-80.

Verboon, P. & Heiser, W. J. (1992). Resistant orthogonal Procrustes analysis. *Journal of Classification, 9*, 237-256.

Verboon, P. (1993). Robust nonlinear regression analysis. *British Journal of Mathematical and Statistical Psychology* , (in press).

Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika, 46*, 347-388.