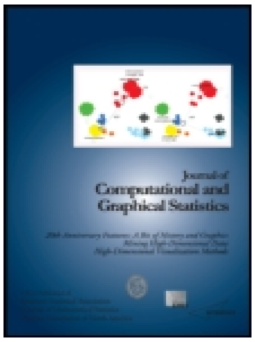

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

# Asymmetric Linear Dimension Reduction for Classification

Christian Hennig

**To cite this article:** Christian Hennig (2004) Asymmetric Linear Dimension Reduction for Classification, Journal of Computational and Graphical Statistics, 13:4, 930-945, DOI: 10.1198/106186004X12740

**To link to this article:** http://dx.doi.org/10.1198/106186004X12740

# Asymmetric Linear Dimension Reduction for Classification

Christian HENNIG

This article discusses methods to project a $p$-dimensional dataset with classified points from $s$ known classes onto a lower dimensional hyperplane so that the classes appear optimally separated. Such projections can be used, for example, for data visualization and classification in lower dimensions. New methods, which are asymmetric with respect to the numbering of the groups, are introduced for $s = 2$. They aim at generating data projections where one class is homogeneous and optimally separated from the other class, while the other class may be widespread. They are compared to classical discriminant coordinates and other symmetric methods from the literature by a simulation study, the application to a 12-dimensional dataset of 74,159 spectra of stellar objects, and to land snails distribution data. Neighborhood-based methods are also investigated, where local information about the separation of the classes is averaged. The use of robust MCD-covariance matrices is suggested.

**Key Words:** Canonical coordinates; Cluster validation; Discriminant coordinates; MCD estimator; Nearest neighbor; Projection pursuit; Quasars; Visualization.

## 1. INTRODUCTION

This article deals with the following problem: Given is a dataset $\mathbf{X}$ (rows are cases, columns are variables) of $n$ points from $I\!R^p$, which is partitioned into $s > 1$ known classes. The aim is to find a $p \times k$-matrix $\mathbf{C}$ leading to $k$-dimensional projected data $\mathbf{Y} = \mathbf{XC}$, in order to separate the classes as well as possible in $\mathbf{Y}$ according to various criteria. The columns of $\mathbf{C}$ might be interpreted as optimal separating projection directions.

There are many applications of this task. Many methods for supervised classification suffer from the so-called "curse of dimensionality" (see, e.g., Hastie, Tibshirani, and Friedman 2001, pp. 22–27) and may profit from performing a linear dimension reduction first. My main motivation has been the visual validation of the outcomes of clustering meth-

Christian Hennig is Research Assistant, Fachbereich Mathematik, Universität Hamburg, Bundesstrasse 55, D-20146, Hamburg, Germany (E-mail: hennig@math.uni-hamburg.de).

*Journal of Computational and Graphical Statistics, Volume 13, Number 4, Pages 930–945*
*DOI: 10.1198/106186004X12740*

ods (Hennig and Christlieb 2002; see also the land snails example in Section 6), and such a visualization may also be useful for an exploratory analysis of a supervised classification problem (see Seber 1984, pp. 269–273, and the spectra example in Section 6). In another application, Strecker (2002) used classification adapted linear dimension reduction to demonstrate that a group of fishes constitutes an up-to-now unknown species by showing that the group is separated strongly from any known candidate species. The entries of the matrix $\mathbf{C}$ may be used to interpret the separation between groups in terms of the variables.

In this article, I concentrate on data visualization. The assessment of the performance of particular classification rules combined with the proposed linear dimension reduction methods is left for future research.

The most widespread linear dimension reduction method is the method of discriminant coordinates (DCs; see Gnanadesikan 1977, pp. 84–90). DCs implicitly assume that all classes have the same covariance matrix. Fukunaga (1990) proposed an extension of this technique for $s = 2$, which can make visible the differences in the covariance structure as well. These two methods are discussed in Section 3. Both techniques are invariant with respect to the numbering of the classes. I call such methods "symmetric." The crucial new idea of the present article is that useful projection methods can be defined by violating this principle of symmetry.

The rationale for a reasonable "asymmetric" method is that in many applications there are two classes, one of which is very homogeneous because, for example, it consists of observations that share an important characteristic, while the observations of the other class have nothing more in common than the absence of this characteristic and may therefore be very heterogeneous. Such a situation may happen, for example, in medicine where patients in the homogeneous class may suffer from a particular disease. An example in Section 6 deals with the separation of quasars from other celestial objects. Although the definition of DCs forces both classes to appear simultaneously as homogeneous as possible, asymmetric methods are defined by the maximization of the ratio between measures for between-groups heterogeneity and measures for the projected variation of only the homogeneous class. Thus, in terms of visualization, the "homogeneous class" (called "H-class" from now on) should look homogeneous, the two classes should appear separated, but the second class may look as scattered as necessary to yield a good separation from the H-class.

The idea of asymmetry is not only useful when there are two groups, only one of which is homogeneous. $s > 2$ classes can be visualized by $s$ two-dimensional asymmetric projections, each of them defining one of the classes as the H-class. If unclassified points are present in the data, they can simply be included in the nonhomogeneous class ("N-class" from now on). This is illustrated by the land snails example in Section 6.

Asymmetric methods are introduced in Section 4. In Section 4.1, I suggest maximizing the ratio of the projected variation between points of the H-class and the N-class and the projected variance of the H-class. Sometimes this ratio can be dominated by extreme points of the N-class. In Section 4.2 and Section 4.3, robustness improvements are suggested.

Some mathematical preliminaries are given in Section 2. Section 3 discusses methods that implicitly assume an approximately elliptical shape of the classes. The same holds for the H-class in Section 4. Hastie and Tibshirani (1996) proposed a dimension-reduction procedure based on local differences between the groups. This procedure allows for more complicated distributional shapes of the classes, for example, skewness or a mixture structure. It is presented in Section 5.1. Again, the method may be affected strongly by outliers, and a robustification is proposed in Section 5.2. An asymmetric version is introduced in Section 5.3.

Two data examples are treated in Section 6. The results of a comprehensive simulation study are summarized along with a concluding discussion in Section 7.

## 1.1 FURTHER REFERENCES

The DC criterion may be interpreted as a particular projection pursuit index (see Huber 1985). More sophisticated projection pursuit indexes based on robust estimators (Pires 2003) or within-group density estimators (Polzehl 1995) can also be used for dimension reduction in classification. However, they are more difficult to compute than the eigenanalyses required for the methods of the present article. Röhl and Weihs (1999) proposed a computer-intensive dimension-reduction method to optimize the misclassification rate under the Normal assumption. Kiers and Krzanowski (2000) used a projection method to isolate the most extreme of more than two groups as well as possible. Fukunaga (1990, chap. 10) discussed the optimization of some alternative indexes to the DC-index based on the within-groups and between-groups covariance matrices. Young, Marco, and Odell (1987) proposed a linear dimension reduction technique where differences in mean and covariance matrices between classes are aggregated. The use of a robust minimum covariance determinant estimator in classification was proposed by Hawkins and McLachlan (1997). Ripley (1996, p. 105) computed robustified DCs based on minimum volume ellipsoid covariance matrix estimators.

There is a large amount of literature on data visualization (e.g., Cleveland and McGill 1988; Rao 1993, Part VII). Much attention is paid to the problem of making high-dimensional patterns visible on a two-dimensional screen (see, e.g., Wegman and Carr 1993, and the references given therein). Nowadays, many statistical packages such as XGobi (Buja, Cook, and Swayne 1996), its recent version GGobi (www.ggobi.org) and ExplorN (Carr, Wegman, and Luo 1996) contain dynamic and interactive graphics where motion is used to produce informative views of the high-dimensional space. A prominent example is the grand tour (Asimov 1985; Wegman 1991), which generates a smooth sequence of two-dimensional projections and has been applied in classification and clustering (Wilhelm, Wegman, and Symanzik 1999; Hennig and Christlieb 2002). This article is devoted to the more traditional approach of defining informative, but static two-dimensional projections (Huber 1985). Such projections can be combined with dynamic and interactive graphics in a useful way. For example, they may provide target planes for smooth sequences (Hurley and Buja 1990; Cook, Buja, Cabrera, and Hurley 1995) or they may serve to validate findings from dynamic graphics in a reproducible manner (Hennig and Christlieb 2002).

## 2. MATHEMATICAL PRELIMINARIES

The methods treated in the present article are defined in the following way: Let $\mathbf{Q}$ and $\mathbf{R}$ be symmetric, positive definite $p \times p$-matrices ($\mathbf{Q}$ may be positive semidefinite). $\mathbf{Q}$ should measure the variance/covariance structure between the classes, and $\mathbf{R}$ measures the variance/covariance within the classes, within the H-class or within the whole dataset, depending on the projection method.

**Definition 1.** *The first $k$ projection vectors (with respect to the corresponding projection method, which will be defined by the choice of $\mathbf{Q}$ and $\mathbf{R}$) $\mathbf{c}_1, \ldots \mathbf{c}_k$ are defined as the vectors maximizing*

$$F_{\mathbf{c}} = \frac{\mathbf{c}'\mathbf{Q}\mathbf{c}}{\mathbf{c}'\mathbf{R}\mathbf{c}} \tag{2.1}$$

*subject to $\mathbf{c}_i'\mathbf{R}\mathbf{c}_j = \delta_{ij}$, where $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ else.*

**Corollary 1.** *The first $k$ projection vectors of $\mathbf{X}$ are the (suitably scaled) eigenvectors of $\mathbf{R}^{-1}\mathbf{Q}$ corresponding to the $k$ largest eigenvalues.*

All discussed methods are required to be affine equivariant, that is,

$$\mathbf{X} = \mathbf{Z}\mathbf{T} + \mathbf{v} \Rightarrow \mathbf{C}_k(\mathbf{X}) = \mathbf{T}^{-1}\mathbf{C}_k(\mathbf{Z}), \tag{2.2}$$

where $\mathbf{Z}$ and $\mathbf{X}$ are $n \times p$-data matrices, $\mathbf{C}_k(\mathbf{X})$ is the $p \times k$-matrix, where the $k \leq p$ columns are the projection vectors of the method under study, $\mathbf{T}$ is nonsingular $p \times p$, and $\mathbf{v} \in I\!R^p$. A justification of (2.2) is given in the Appendix. Most methods suggested here fulfill (2.2) because of the corresponding affine equivariance properties of $\mathbf{Q}$ and $\mathbf{R}$. The proofs of (2.2) for the resulting procedures are straightforward and omitted.

Exceptions are the methods proposed in the Sections 5.1 and 5.2, where $\mathbf{R} = \mathbf{I}_p$, being the $p$-dimensional unit matrix. To ensure affine equivariance for these procedures, I assume that the dataset $\mathbf{X}$ is sphered by some affine equivariant covariance matrix estimator. The resulting dimension-reduction method inherits the equivariance with respect to linear transformations from the equivariance of $\mathbf{S}(\mathbf{Z})$. We do not have to worry about centering, because all methods treated in the present article are based on translation invariant statistics.

In the present article, the minimum covariance determinant covariance matrix estimator (MCD, Rousseeuw 1984) is used for sphering, as implemented in the procedure `cov.rob` of the statistical package R according to Rousseeuw and van Driessen (1999). The MCD is based on the $h$ observations whose classical covariance matrix has the lowest determinant. $h$ defaults to $\lfloor (n + p + 1)/2 \rfloor$, the largest integer smaller or equal to $(n + p + 1)/2$.

Note that Krzanowski (1995) and Kiers and Krzanowski (2000) argued that in some applications projection directions should be optimized subject to an orthogonality constraint with respect to the Euclidean metric (as the projection vectors defined in the Sections 5.1 and 5.2 would be without sphering), which is incompatible with affine equivariance.

Some further notation: Let $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$ the $p$-dimensional points of group $i = 1, \ldots,$

$s$, $n = \sum_{i=1}^{s} n_i$. Let $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})', i = 1, \ldots, s$, and $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_s')'$. Let

$$\begin{aligned}
\mathbf{m}_i &= \frac{i}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij},\ \mathbf{m} = \frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \mathbf{x}_{ij}, \\
\mathbf{U}_i &= \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \mathbf{m}_i)(\mathbf{x}_{ij} - \mathbf{m}_i)',\ \mathbf{U} = \sum_{i=1}^{s} \mathbf{U}_i, \\
\mathbf{S}_i &= \frac{1}{n_i - 1} \mathbf{U}_i,\ \mathbf{W} = \frac{1}{n - s} \mathbf{U},\ \mathbf{B} = \frac{1}{n(s-1)} \sum_{i=1}^{s} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})',
\end{aligned}$$

that is, $\mathbf{S}_i$ is the covariance matrix of group $i$ with mean vector $\mathbf{m}_i$, $\mathbf{W}$ is the pooled within-groups scatter matrix and $\mathbf{B}$ is the between-groups scatter matrix.

## 3. A REVIEW OF SYMMETRIC METHODS

The method of discriminant coordinates is the most common approach for the projection of high-dimensional data with a given grouping to a lower-dimensional subspace. The term "discriminant coordinates" is used according to Gnanadesikan (1977). They are also known under the name "canonical variates." The approach goes back at least to Rao (1952), who developed discriminant coordinates as a generalization of Fisher's linear discriminant function to more than two groups.

**Definition 2.** *DCs are defined by Definition 1 with* $\mathbf{Q} = \mathbf{B}$ *and* $\mathbf{R} = \mathbf{W}$.

**Corollary 2.** *Only s – 1 eigenvalues of* $\mathbf{W}^{-1}\mathbf{B}$ *are larger than 0. The whole information about the mean differences can be displayed in* $s - 1$ *dimensions (see Gnanadesikan 1977).*

A reasonable application of DCs requires that $\mathbf{W}$ is an adequate measure of the covariance structure within all classes, which holds at least under approximate equality of their covariance matrices. The differences between the classes are measured as differences between the class means. Let $\mathbf{C}_k = (\mathbf{c}_1, \ldots, \mathbf{c}_k)$. Because $\mathbf{c}_i'\mathbf{W}\mathbf{c}_j = \delta_{ij}$, the within-groups covariance matrix of the projected data $\mathbf{X}\mathbf{C}_k$ is $\mathbf{I}_k$, that is, the projected groups appear spherical under the assumption of equality of the classes' covariance matrices.

Because of Corollary 2, in the two-class-setup (including more than two classes when one class is declared as H-class and the others are merged), only one DC is informative. In this setup, the first and only DC can be usefully complemented by an idea of Fukunaga (1990), who suggested to choose further projection directions orthogonal to the DC in order to maximize the difference in the projected covariance matrices.

Fukunaga (1990) proposed the following expression to measure the dissimilarity between two covariance matrices $\Sigma_1$ and $\Sigma_2$:

$$D(\Sigma_1, \Sigma_2) = \frac{1}{2} \log \frac{\det\left(\frac{\Sigma_1 + \Sigma_2}{2}\right)}{\sqrt{\det(\Sigma_1)\det(\Sigma_2)}}.$$

Fukunaga replaced $\mathbf{W}$ by $\mathbf{W}_D = \frac{1}{2}(\mathbf{S}_1 + \mathbf{S}_2)$ in the DC, which corresponds to Definition 2 only if $n_1 = n_2$ or $\mathbf{S}_1 = \mathbf{S}_2$. Further details and motivation are given in Fukunaga (1990, p. 99, p. 455 f).

**Definition 3.** *Under $s = 2$, the first Bhattacharyya coordinate (BC) is defined by Definition 1 with $\mathbf{Q} = \mathbf{B}$ and $\mathbf{R} = \mathbf{W}_D$.*

*The second to $k$th BCs $\mathbf{c}_2, \ldots, \mathbf{c}_k$ maximize $D(\mathbf{c}'\mathbf{S}_1\mathbf{c}, \mathbf{c}'\mathbf{S}_2\mathbf{c})$ subject to $\mathbf{c}_i'\mathbf{W}_D\mathbf{c}_1 = 0$, $\mathbf{c}_i'\mathbf{S}_1\mathbf{c}_j = \delta_{ij}$, $i, j = 2, \ldots, k$.*

The second to $k$th BCs are computed as the eigenvectors corresponding to the largest values of $\lambda + \frac{1}{\lambda}$, $\lambda$ denoting the eigenvalues of $\mathbf{S}_{Y1}^{-1}\mathbf{S}_{Y2}$, where $\mathbf{S}_{Yi}$ is the within-group covariance matrix of the $i$th group projected onto the 2nd to $k$th eigenvector of $\mathbf{W}_D^{-1}\mathbf{B}$.

# 4. ASYMMETRIC METHODS

As far as I know, the explicit distinction of a homogeneous (H-) class and a less homogeneous (N-) class did not appear in the literature up to now. However, as mentioned in the Introduction, there are many applications where such a distinction suggests itself.

The term "homogeneity" may not have a unique meaning. A class can qualify as a H-class by being less scattered, clean of atypical and extreme observations, or by being of a more homogeneous distributional shape, for example, bell-shaped as opposed to Normal mixture-shaped. The term "H-class" is used technically here to denote the class which is treated as homogeneous by an asymmetric projection method.

## 4.1 Basic Approach

Asymmetric discriminant coordinates (ADCs) are the most simple direct "asymmetrization" of classical DCs. For ADCs, the within-groups covariance matrix $\mathbf{W}$ is replaced by the covariance matrix of the H-class. Thus, the denominator of the criterion to maximize does no longer become small by reducing the projected variance of the N-class. The between-groups covariance matrix $\mathbf{Q}$ no longer takes only mean differences into account, but is now based on squared differences between points of different classes. Let

$$\mathbf{B}^* = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (\mathbf{x}_{1i} - \mathbf{x}_{2j})(\mathbf{x}_{1i} - \mathbf{x}_{2j})'.$$

Note that the computation of $\mathbf{B}^*$ (and of $\mathbf{B}^{**}$ and $\mathbf{B}^{***}$ in the Sections 4.2 and 4.3) needs only $n_1 + n_2 + 1$ vector products instead of $n_1 n_2$ if it is multiplied out.

**Definition 4.** *ADCs are defined by Definition 1 with $\mathbf{Q} = \mathbf{B}^*$ and $\mathbf{R} = \mathbf{S}_1$.*

The projected points of the H-class have a unit covariance matrix.

Here is some motivation for the definition of $\mathbf{B}^*$. Assume $s = 2$ and recall $\mathbf{S}_i = (n_i - 1)\mathbf{U}_i$, $i = 1, 2$. The matrices $n\mathbf{B}$ and $(n-1)\mathbf{W}$ from the definition of DCs sum up to the total sum of squares, that is,

$$\mathbf{U} = \mathbf{U}_1 + \mathbf{U}_2 + n\mathbf{B}. \tag{4.1}$$

DCs require that the projected $\mathbf{B}$ gets large and the projected $\mathbf{U}_1$ and $\mathbf{U}_2$ get small. The basic idea of asymmetry is to drop the demand that the second class should appear homogeneous, that is, to drop its contribution to the denominator of the criterion function. Thus, the simplest idea would be to maximize $(\mathbf{c}'\mathbf{B}\mathbf{c})/(\mathbf{c}'\mathbf{S}_1\mathbf{c})$, but this is not much better than DCs because $\dim(\mathbf{B}) = 1$ and the points of the second class remain represented only by their mean, around which they may scatter extremely. It is more informative to consider the decomposition

$$\begin{aligned}\mathbf{U} &= \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{2}\sum_{l=1}^{n_k}(\mathbf{x}_{ij}-\mathbf{x}_{kl})(\mathbf{x}_{ij}-\mathbf{x}_{kl})' \\ &= \frac{1}{2n}\left[\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{l=1}^{n_i}(\mathbf{x}_{ij}-\mathbf{x}_{il})(\mathbf{x}_{ij}-\mathbf{x}_{il})' + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\mathbf{x}_{1i}-\mathbf{x}_{2j})(\mathbf{x}_{1i}-\mathbf{x}_{2j})'\right] \\ &= \frac{1}{2n}\left[2n_1\mathbf{U}_1 + 2n_2\mathbf{U}_2 + n_1n_2\mathbf{B}^*\right]. \qquad (4.2)\end{aligned}$$

ADCs are defined by applying the above idea to this decomposition. Get, by combination of (4.1) and (4.2),

$$\frac{\mathbf{c}'\mathbf{B}^*\mathbf{c}}{\mathbf{c}'\mathbf{S}_1\mathbf{c}} = \frac{2}{n_1n_2}\left[n^2\frac{\mathbf{c}'\mathbf{B}\mathbf{c}}{\mathbf{c}'\mathbf{S}_1\mathbf{c}} + (n_1-1)n_2\frac{\mathbf{c}'\mathbf{S}_2\mathbf{c}}{\mathbf{c}'\mathbf{S}_1\mathbf{c}} + d\right],$$

$d$ depending only on $n_1$ and $n_2$. Thus, under $\mathbf{S}_1 = \mathbf{S}_2 = \mathbf{W}$, the first and only informative ADC is the DC, while under $\mathbf{S}_1 \neq \mathbf{S}_2$, the ADCs optimize a weighted mean of the asymmetrized DC criterion function and an analogous measure of the projected ratio between the covariance matrices.

## 4.2 Adequately Accounting for Nonhomogeneity

This section introduces asymmetric weighted discriminant coordinates (AWCs). They improve ADCs with respect to objects in the N-class lying extremely far from the H-class. Such objects can even occur if there are no gross outliers in the dataset, namely when there is a subclass of the N-class that is very far from the H-class. In many applications, where the H-class is really homogeneous, gross outliers occur only in the N-class. The problem of ADCs is that in such a situation a projection direction can be chosen that mainly shows the difference between the H-class and the extreme objects. If the difference between the H-class and the majority of the N-class lies in another direction, this direction will be obscured. The principle of AWCs is to weight the points of the N-class according to their Mahalanobis distance to the H-class. Thus, the resulting objective function weights up the differences between the H-class and the points from the N-class that are close to the H-class and therefore more difficult to separate.

Let

$$\mathbf{B}^{**} = \frac{1}{n_1\sum_{j=1}^{n_2}w_j}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}w_j(\mathbf{x}_{1i}-\mathbf{x}_{2j})(\mathbf{x}_{1i}-\mathbf{x}_{2j})',$$

where

$$w_j = \min\left(1, \frac{d}{(\mathbf{x}_{2j}-\mathbf{m}_1)'\mathbf{S}_1^{-1}(\mathbf{x}_{2j}-\mathbf{m}_1)}\right), \quad j=1,\ldots,n_2, \qquad (4.3)$$

$d>0$ being some constant, for example, the .99-quantile of the $\chi_p^2$-distribution.

**Definition 5.** *AWCs are defined by Definition 1 with* $\mathbf{Q}=\mathbf{B}^{**}$ *and* $\mathbf{R}=\mathbf{S}_1$.

The choice of the weights is motivated as follows: Consider some point $\mathbf{x}_{2j}=\mathbf{m}_1+q\mathbf{v}$ from the N-class, where $\mathbf{v}$ is a unit vector with respect to $\mathbf{S}_1$ giving the direction of the deviation of $x_{2j}$ from the mean $\mathbf{m}_1$ of the H-class and $q>0$ is the amount of deviation. The contribution of $\mathbf{x}_{2j}$ to $\mathbf{B}^{**}$ is, for $q$ large enough,

$$\sum_{i=1}^{n_1} \frac{d}{(\mathbf{x}_{2j}-\mathbf{m}_1)'\mathbf{S}_1^{-1}(\mathbf{x}_{2j}-\mathbf{m}_1)}(\mathbf{x}_{1i}-\mathbf{x}_{2j})(\mathbf{x}_{1i}-\mathbf{x}_{2j})',$$

which converges to $n_1 d\frac{\mathbf{v}\mathbf{v}'}{\mathbf{v}'\mathbf{S}_1^{-1}\mathbf{v}}$ for $q\to\infty$. Thus, the information about the direction of the deviation is maintained, but the information about the amount of deviation vanishes.

### 4.3 Robustifying the Estimation of the Homogeneous Class

The idea of asymmetric robustified discriminant coordinates (ARCs) is to replace the mean and covariance matrix estimator of the H-class by robust alternatives. Points from the H-class that are not consistent with its main part, that is, that have a too large robust Mahalanobis distance from the H-class, are then weighted down in the same manner as points from the N-class. The effect of such a downweighting scheme is that the majority of the H-class can be projected more homogeneously and better separated from the N-class in presence of outliers in the H-class. Downweighted points from the H-class contribute fewer to the target criterion and can be projected far away from the core of the H-class.

Let $\mathbf{m}_1^*$ be the location estimator associated with the MCD-estimator of the first group $\mathbf{S}_{\mathrm{MCD}}(\mathbf{X}_1)$. Here I choose $h=\lfloor 3(n+p+1)/4\rfloor$ instead of $\lfloor (n+p+1)/2\rfloor$ as in Section 2. The reason is that there should be 75% or more points belonging together in an adequately chosen H-class, and a covariance estimator should take all these points into account, even if a majority of them would look even more homogeneous. Let

$$\mathbf{B}^{***} = \frac{1}{\sum_{i,j} w_{1i}w_{2j}} \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{1i}w_{2j}(\mathbf{x}_{1i}-\mathbf{x}_{2j})(\mathbf{x}_{1i}-\mathbf{x}_{2j})',$$

where

$$w_{ij} = \min\left(1, \frac{d}{(\mathbf{x}_{ij}-\mathbf{m}_1^*)'\mathbf{S}_{\mathrm{MCD}}(\mathbf{X}_1)^{-1}(\mathbf{x}_{ij}-\mathbf{m}_1^*)}\right), \quad i=1,2,\ j=1,\ldots,n_i,$$

with $d>0$ as in (4.3).

**Definition 6.** *ARCs are defined by Definition 1 with* $\mathbf{Q}=\mathbf{B}^{***}$ *and* $\mathbf{R}=\mathbf{S}_{\mathrm{MCD}}(\mathbf{X}_1)$.

The motivation for the weights from the previous subsection continues to hold if either $\mathbf{x}_{1i}$ or $\mathbf{x}_{2j}$ have a small enough robust Mahalanobis distance to the H-class. If this distance converges to $\infty$ for both of them, the corresponding term in $\mathbf{B}^{***}$ vanishes. Pairs consisting of a central point of the H-class and a point of the N-class lying near to the H-class are weighted up.

## 5. NEIGHBORHOOD-BASED DIMENSION REDUCTION

The principle of neighborhood-based dimension reduction was introduced by Hastie and Tibshirani (1996). The difference to the methods discussed earlier is that the classes are no longer characterized by location and scatter estimators. Instead, mean differences between the classes are computed locally in the neighborhoods of the points, and the dimension-reduction criteria combine this local information about the directions of class separation. The approach provides a useful supplementation to the methods described above in particular for datasets where the separation between the classes cannot be adequately assessed by comparing locations and covariance matrices. See the simulation study cited in Section 7 for examples.

### 5.1 The Basic Approach

For every point $\mathbf{x}_{ij},\ i = 1, \ldots, s,\ j = 1, \ldots, n_i$, let $\mathbf{X}(i,j)$ denote the dataset consisting of its $K$ nearest neighbors (with respect to the Euclidean distance and including $\mathbf{x}_{ij}$ itself). Let $\mathbf{m}(i,j)$ be the $p$-dimensional mean of $\mathbf{X}(i,j)$ and $\mathbf{m}_k(i,j)$ the mean of the points of group $k,\ k = 1, \ldots, s$, in $\mathbf{X}(i,j)$. $n_k(i,j)$ denotes the number of such points. $K$ is chosen as $\max(50, n/5)$ according to Hastie and Tibshirani (1996). $n$ must be much larger than $K$. $K$ should be so large that the number of points from different classes occurring in neighborhoods of points of each class is not too small.

Let

$$\begin{aligned} \mathbf{B}(i,j) &= \frac{1}{K}\sum_{k=1}^{s} n_k(i,j)(\mathbf{m}_k(i,j) - \mathbf{m}(i,j))(\mathbf{m}_k(i,j) - \mathbf{m}(i,j))', \\ \tilde{\mathbf{B}} &= \frac{K}{n}\sum_{i,j} \mathbf{B}(i,j). \end{aligned}$$

**Definition 7.** *(Hastie and Tibshirani 1996) NCs are defined by Definition 1 with* $\mathbf{Q} = \tilde{\mathbf{B}}$ *and* $\mathbf{R} = \mathbf{I}_p$.

A maximizer of $\mathbf{c}'\tilde{\mathbf{B}}\mathbf{c}$ is the projection direction in which the squared mean difference of the projected points between the groups, averaged over the neighborhoods of all points, is maximal.

This definition yields basis vectors that are orthonormal with respect to the Euclidean metric. Therefore, the data is assumed to be sphered by the MCD as explained in Section 2.

$\tilde{\mathbf{B}}$ may be strongly dominated by gross outliers, which may occur in many datasets of interest. Therefore I propose a robustification.

### 5.2 Robustifying the Basic Approach

The principle of weighted neighborhood-based coordinates (WNCs) is to weight the neighborhood-based between-groups matrices $\mathbf{B}(i,j)$ so that the amount of contribution to the averaged matrix is no longer determined by the size of the squared mean difference between the groups. This size is reduced in the matrix $\mathbf{B}(i,j)/\text{trace}(\mathbf{B}(i,j))$, while the direction information is maintained. This matrix is then multiplied by $\prod_{k=1}^{s} n_k(i,j)$ to represent the reliability of the neighborhood belonging to $(i,j)$ for the estimation of the class means, which is maximal if all $n_k(i,j)$ are approximately equal. Again, the data are assumed to be sphered by the MCD.

Let

$$w(i,j) = \prod_{k=1}^{s} n_k(i,j),\ w = \sum_{i,j} w(i,j), \quad \tilde{\mathbf{B}}^* = \frac{1}{w}\sum_{i,j}\frac{w(i,j)}{\text{trace}(\mathbf{B}(i,j))}\mathbf{B}(i,j).$$

**Definition 8.** *WNCs are defined by Definition 1 with* $\mathbf{Q} = \tilde{\mathbf{B}}^*$ *and* $\mathbf{R} = \mathbf{I}_p$.

### 5.3 Asymmetric Neighborhood-Based Coordinates

The neighborhood-based projection principle can also be asymmetrized. By this I mean that only the neighborhoods of the points of the H-class are considered, and that the resulting projection vectors are required to be orthonormal with respect to a scale estimator of the H-class, namely $\mathbf{S}_{\text{MCD}}(\mathbf{X}_1)$.

As all the asymmetric methods introduced in the present article, asymmetric neighborhood-based coordinates (ANCs) imply an internal sphering of the data with respect to the homogeneous group. Thus, the data do not need to be sphered initially.

Let

$$\begin{aligned}
\mathbf{B}_1(i) &= \frac{1}{K}\sum_{j=1}^{2} n_j(i,1)(\mathbf{m}_j(i,1) - \mathbf{m}(i,1))(\mathbf{m}_j(i,1) - \mathbf{m}(i,1))', \\
\tilde{\mathbf{B}}_1^* &= \frac{1}{\sum_{i=1}^{n_1} w(i,1)}\sum_{i=1}^{n_1}\frac{w(i,1)}{\text{trace}(\mathbf{B}_1(i))}\mathbf{B}_1(i).
\end{aligned}$$

**Definition 9.** *ANCs are defined by Definition 1 with* $\mathbf{Q} = \tilde{\mathbf{B}}_1^*$ *and* $\mathbf{R} = \mathbf{S}_{\text{MCD}}(\mathbf{X}_1)$.

## 6. DATA EXAMPLES

The Hamburg/ESO survey (HES; Wisotzki et al. 2000) was carried out at the European Southern Observatory (ESO) in Chile. About half of the sky visible from southern hemisphere has been imaged on photographic plates. By mounting a prism in front of the telescope, the images of celestial objects are converted into spectra. From these spectra, a set of 16 spectral features (variables) has been computed. A description of these features was given by Christlieb et al. (2001). The main aim of the HES is to find new quasars. Recent statistical methods for finding quasars were discussed by Feigelson and Babu (2003).

The dataset used in the work presented here consists of 74,159 spectra (out of 4 million in the HES). It includes 344 spectra of quasars selected from all HES plates by "classical" criteria (related to the spectral features 13–16, see Wisotzki et al. 2000), and confirmed by follow-up observations, a sample of 2,856 spectra of known class, and 70,959 unclassified spectra from 10 HES plates. Because the confirmation of quasar candidates needs cumbersome additional observations, it is important to keep the number of quasar candidates from the unclassified spectra as low as possible.

It is of interest if new quasar candidates can be found by use of the features 1–12, which are used in the present paper, representing information distinct from the classical criteria.

The scatterplot in the upper left side of Figure 1 shows the first two ADCs where the spectra of known class have been the H-class and the 70,959 unclassified spectra have been the N-class. The H-class is represented by black full circles. It can clearly be seen that the sample of spectra of known class does not properly represent at least a vast majority of all unclassified spectra, contrary to the intentions of the astronomers. Therefore, for all subsequent plots of Figure 1, the quasars have been chosen as the H-class and all other objects have been declared as N-class. This is reasonable because the probability for an unclassified object to be a quasar is very low. Because the plot of the whole value range is dominated by a few gross outliers, all scatterplots show only the most interesting central part.

On the upper right side, it can be seen that the difference in means is not suitable to separate the quasars properly from the N-class, because this difference is dominated by the spread of the unclassified spectra to the left side in the direction of the first BC. There is also a large difference in variances along this direction, and therefore the orthogonal variance maximizing second projection direction does also not lead to a strong separation. The ADC plot on the middle left side shows the quasars more homogeneously and separates them from a larger part of the unclassified spectra. The differences between ADCs and AWCs are negligible in this dataset.

The AWC plot on the middle right side demonstrates a diagnostic application of such plots. Additional to the quasars, the stars with known classes are indicated by a "+". A simple nearest neighbor classification has been applied, where the quasars and the classified stars have made up the training set. Three thousand two hundred ninety-three of the unclassified objects have been classified as quasar candidates, which would need far too much effort to confirm. The quasar candidates are denoted by darker gray full circles, and the AWC plot shows the reason why there are so many of them: Lots of them occur in an area which is not represented in the training sample, neither by quasars, nor by classified stars. This led us to the idea that a classification where the whole N-class is taken as nonquasars could generate much fewer quasar candidates and would therefore be more useful.

In the ARC solution (lower left side), the majority of quasars looks even more packed, while much more quasars can be recognized as outlying with respect to their own class. This is a consequence of the weighting scheme of ARCs, where points from the H-class are weighted down if they appear outlying with respect to the MCD. An ANC scatterplot

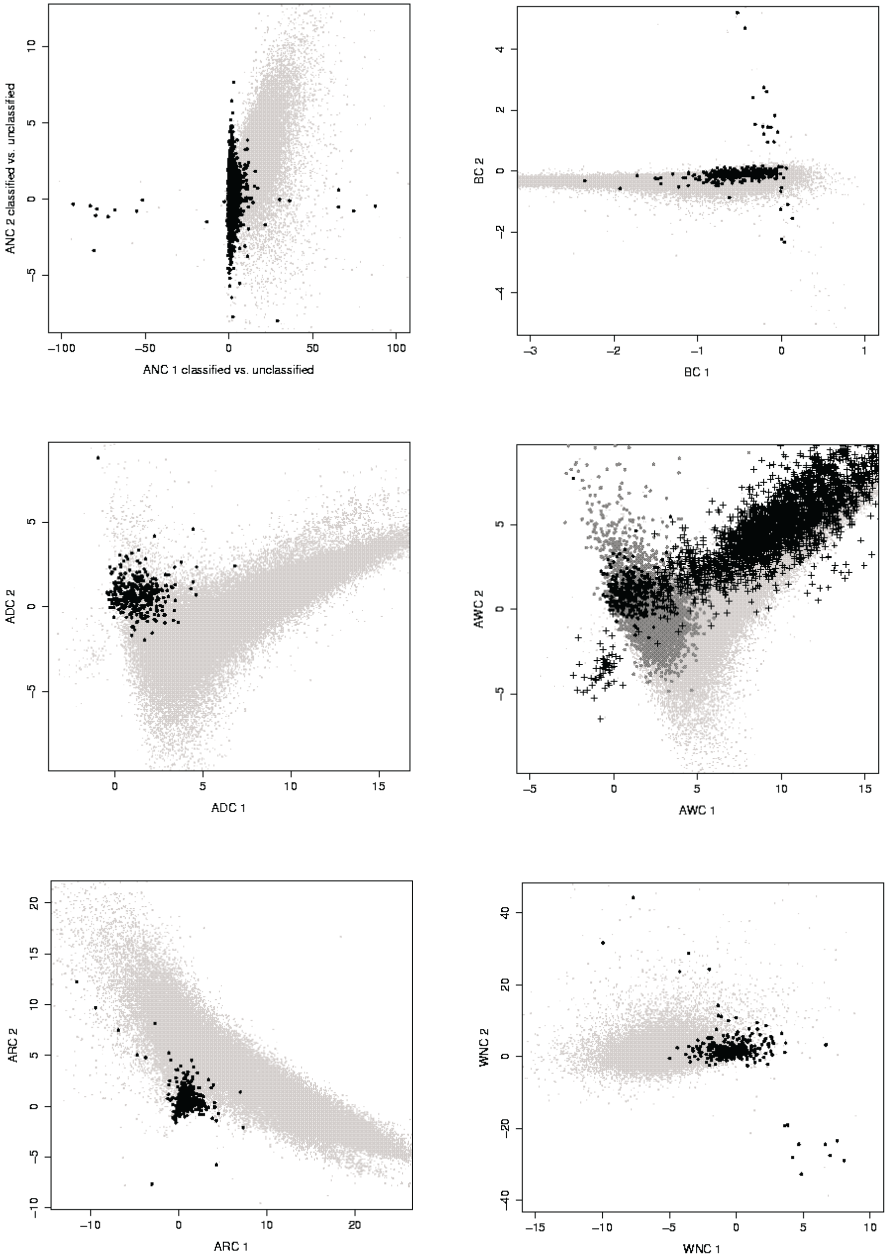

*Figure 1. Upper left side: First two ANCs for separating classified (black) from unclassified spectra. Upper right side: First two BCs. Quasars are black full circles, other objects are small gray points. Middle left side: ADCs, middle right side: AWCs with classified stars denoted by "+" and quasar candidates chosen by nearest neighbor denoted by dark gray circles. Lower left side: ARCs. Lower right side: WNCs.*

Table 1. Computing Time in Seconds for Spectra Data (74,159 × 12).

| *BC* | *ADC* | *AWC* | *ARC* | *NC* * | *WNC* * | *ANC* | *MCD* |
|---|---|---|---|---|---|---|---|
| 4.32 | 2.86 | 18.75 | 22.53 | 83426.30 | 87287.00 | 496.01 | 1476.51 |

NOTE: (*) Times for NC and WNC do not include the time for the computation of the MCD sphering matrix, which is given under "MCD."

is not shown, because it is almost identical to the ARC solution for this dataset (rotated by about 45 degrees). The NC (not shown) and WNC (lower right side) projections also resemble each other. The main difference is that mainly the quasar outliers are separated from the main part of the unclassified spectra along the first NC ($x$-axis). The second NC separates the majority of the quasars better from the majority of the N-class. The first WNC corresponds to the second NC and reversely, because the influence of outliers is weaker for WNCs. Furthermore, the first WNC is slightly better for separating the majority of quasars than the second NC.

To summarize, it can be seen that the quasars can indeed be distinguished to a satisfactory extent from the N-class. The asymmetric plots show clearly the nonhomogeneity of the unclassified points with respect to the quasars. ARCs and neighborhood-based methods show also some outliers among the quasars. As input for a nearest neighbor-type classification method to select quasar candidates from the unclassified points, presumably ARCs and ANCs are most useful.

Table 1 gives the computing times of my implementations of the methods for the statistical software R for the spectra dataset on a Pentium III/928 MHz processor. The advantage of ANC compared to NC and WNC is due to the fact that mean difference matrices have to be computed only for the points of the H-class, which is very small in this dataset (.5%) and will often be larger.

To illustrate the variety of possible applications, I briefly discuss another example, where the objects are six-dimensional scores from a multidimensional scaling, applied to data about distribution areas of 366 species of northwest European land snails. Details and background were given by Hennig and Hausdorf (2004). It has been of interest to find clusters of species of snails according to their distribution areas. A cluster analysis based on a Normal mixture model with noise component (uniform on the convex hull of the data; points denoted by "N" in Figure 2) as explained by Fraley and Raftery (1998) yielded nine clusters including a noise component. The left side of Figure 2 shows the first two discriminant coordinates (omission of noise point and inclusion of them as additional class lead to almost identical plots). Asymmetric dimension reduction may help to decide if the estimated clusters can be interpreted as real groups of species. The right side of Figure 2 shows the first two ANCs with cluster 1 chosen as H-class and all other points chosen as N-class. As opposed to the DC plot, the ANC plot shows the points of cluster 1 as a noticeable pattern of the data. This is a valuable validation of this cluster, which can also be reasonably interpreted from a bio-geographical point of view. Note that some experience is needed to assess such plots, because the combination of a cluster analysis and an asymmetric plot to visualize the clusters may result in some small subsets which appear clustered by chance.

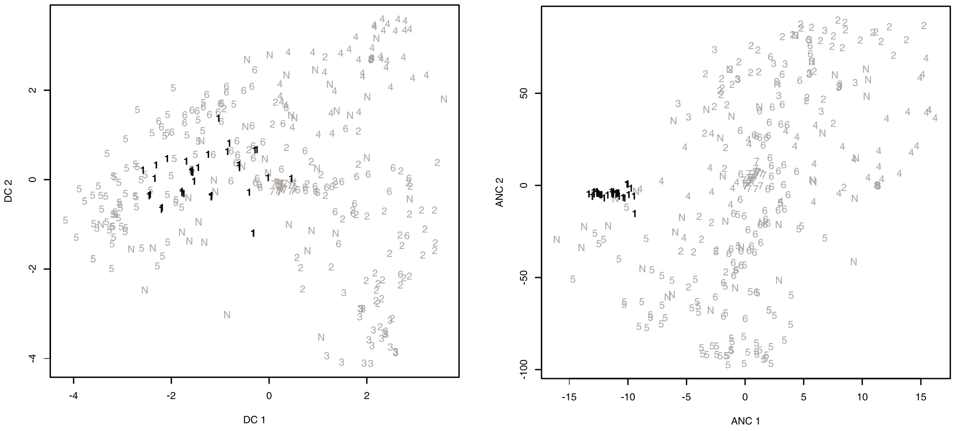

*Figure 2. First two dimensions of projected land snails data, plotting symbols indicating clusters. Left side: DCs, right side: ANCs for cluster 1.*

## 7. SIMULATION RESULTS AND DISCUSSION

A simulation study has been carried out to compare the behavior of the proposed methods in 10 different setups. These setups are defined by the distributions of the H-class and the N-class in the first two (independent) dimensions, which contain the whole information to separate the classes, and the noise distribution for the further dimensions ($p = 4$ for eight setups, $p = 10$ for two setups). The quality of the recovery of the informative first two dimensions by two-dimensional projections has been measured. A full description of the simulation study can be obtained from http://www.amstat.org/publications/jcgs/ftp.html.

Here are some main results:

- In all situations where the H-class is really more homogeneous than the N-class (normally or Cauchy distributed classes with smaller variation of the H-class), the asymmetric methods outperform the symmetric ones.
- If the differences between the classes do not manifest themselves in different means or covariance matrices (e.g., exponentially distributed classes with skewness in opposite directions), neighborhood-based methods are to be preferred. In presence of outliers (Cauchy distributed classes and noise), the robustified methods are superior.
- ARC is outperformed either by ANC or by AWC in all setups. It may be chosen only in large datasets with outliers, where it is faster to compute than ANC (compare Table 1 in Section 6).
- The correct specification of the H-class is crucial for the asymmetric methods.
- The neighborhood-based methods lose some quality in 10 dimensions. The problem may be prevented by choosing neighborhoods with more than 50 points if $n$ is large enough.

Although for data visualization in principle more than a single plot can be considered, the simulations and examples showed that ANC can be recommended in situations with outliers (assuming a homogeneous core of, say, 75% of the data of the H-class), and AWC is a good (and fast) choice if the H-class can be assumed to be free of outliers.

Further research is needed to assess the use of the methods for dimension reduction in combination with the application of particular classification procedures.

The R-package `fpc` to compute all methods treated in the present article is available at the Comprehensive R Archive Network via http://www.R-project.org.

## APPENDIX

The following proposition justifies (2.2): It ensures that the (optimal Bayes) classification problem is the same for $\mathbf{Z}\mathbf{C}_k(\mathbf{Z})$ and for $\mathbf{X}\mathbf{C}_k(\mathbf{X})$.

**Proposition A.1.** *Let $b_{\pi_1 g_1,\ldots,\pi_s g_s}$ denote the Bayes classifier for $s$ classes with assumed densities $g_1,\ldots,g_s$ on $I\!R^q$ and class probabilities $\pi_1,\ldots,\pi_s$, that is, for $\mathbf{y}\in I\!R^q$,*

$$b_{\pi_1 g_1,\ldots,\pi_s g_s}(\mathbf{y}) = \operatorname*{arg\,max}_{i=1,\ldots,s} \pi_i g_i(\mathbf{y}).$$

*Let $\mathbf{Z},\mathbf{T},\mathbf{v}$ and $\mathbf{C}_k$ be defined as for (2.2), $\mathbf{z}\in I\!R^p,\ \mathbf{x}=\mathbf{T}'\mathbf{z}+\mathbf{v}$. Let $g_i^*,\ i=1,\ldots,s$, be the density of $\mathbf{C}_k(\mathbf{X})'\mathbf{x}$ ($\mathbf{C}_k(\mathbf{X})$ is interpreted as fixed, $\mathbf{x}$ and $\mathbf{z}$ are interpreted as random variables) under the assumption that $\mathbf{C}_k(\mathbf{Z})'\mathbf{z}$ has the density $g_i$. Then, under (2.2),*

$$b_{\pi_1 g_1^*,\ldots,\pi_s g_s^*}(\mathbf{C}_k(\mathbf{X})'\mathbf{x}) = b_{\pi_1 g_1,\ldots,\pi_s g_s}(\mathbf{C}_k(\mathbf{Z})'\mathbf{z}).$$

***Proof:***

$$\begin{aligned}
\mathbf{C}_k(\mathbf{X})'\mathbf{x} &= \mathbf{C}_k(\mathbf{Z})'\mathbf{z} + \mathbf{C}_k(\mathbf{Z})'(\mathbf{T}^{-1})'\mathbf{v} \Rightarrow g_i^*(\mathbf{y})\\
&= g_i(\mathbf{y} - \mathbf{C}_k(\mathbf{Z})'(\mathbf{T}^{-1})'\mathbf{v}), \quad i=1,\ldots,s,\\
&\Rightarrow \operatorname*{arg\,max}_{i=1,\ldots,s} \pi_i g_i^*(\mathbf{C}_k(\mathbf{X})'\mathbf{x}) = \operatorname*{arg\,max}_{i=1,\ldots,s} \pi_i g_i(\mathbf{C}_k(\mathbf{X})'\mathbf{x} - \mathbf{C}_k(\mathbf{Z})'(\mathbf{T}^{-1})'\mathbf{v})\\
&= \operatorname*{arg\,max}_{i=1,\ldots,s} \pi_i g_i(\mathbf{C}_k(\mathbf{Z})'\mathbf{z}).
\end{aligned}$$

□

## REFERENCES

Asimov, D. (1985), "The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM Journal of Statistical Computing*, 6, 128–143.

Buja, A., Cook, D., and Swayne, D. (1996), "Interactive High-Dimensional Data Visualization," *Journal of Computational and Graphical Statistics*, 5, 78–99.

Carr, D. B., Wegman, E. J., and Luo, Q. (1996), "ExplorN: Design Considerations Past and Present," Technical Report 129, Center for Computational Statistics, George Mason University, Fairfax, VA.

Christlieb, N., Wisotzki, L., Reimers, D., Homeier, D., Koester, D., and Heber, U. (2001), "The Stellar Content of the Hamburg/ESO Survey. I. Automated Selection of DA White Dwarfs," *Astronomy and Astrophysics*, 366, 898–912.

Cleveland, W. S., and McGill, M. E. (eds.) (1988), *Dynamic Graphics for Statistics*, Monterey, CA: Wadsworth.

Cook, D., Buja, A., Cabrera, J., and Hurley, C. (1995), “Grand Tour and Projection Pursuit,” *Journal of Computational and Graphical Statistics*, 4, 155–172.

Feigelson, E. D., and Babu, G. J. (eds.) (2003), “Statistical Challenges of Astronomy,” in *Proceedings of SCMA III*, Wiley, New York.

Fraley, C., and Raftery, A. E. (1998), “How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis,” *Computer Journal*, 41, 578–588.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition* (2nd Ed.), Boston: Academic Press.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: Wiley.

Hastie, T., and Tibshirani, R. (1996), “Discriminant Adaptive Nearest Neighbor Classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18, 607–616.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer.

Hawkins, D. M., and McLachlan, G. J. (1997), “High-Breakdown Linear Discriminant Analysis,” *Journal of the American Statistical Association*, 92, 136–143.

Hennig, C., and Christlieb, N. (2002), “Validating Visual Clusters in Large Datasets: Fixed Point Clusters of Spectral Features,” *Computational Statistics and Data Analysis*, 40, 723–739.

Hennig, C., and Hausdorf, B. (2004), “Distance-Based Parametric Bootstrap Tests for Clustering of Species Ranges,” *Computational Statistics and Data Analysis*, 45, 875–896.

Huber, P. J. (1985), “Projection Pursuit” (with discussion), *The Annals of Statistics*, 13, 435–475.

Hurley, C., and Buja, A. (1990), “Analyzing High-Dimensional Data with Motion Graphics,” *SIAM Journal of Scientific and Statistic Computation*, 11, 1193–1211.

Kiers, H. A. L., and Krzanowski, W. J. (2000), “Projections Distinguishing Isolated Groups in Multivariate Data Spaces,” in *Data Analysis*, eds. W. Gaul, O. Opitz, and M. Schader, Berlin: Springer, pp. 207–218.

Krzanowski, W. J. (1995), “Orthogonal Canonical Variates for Discrimination and Classification,” *Journal of Chemometrics*, 9, 509–520.

Pires, A. M. (2003), “Robust Discriminant Analysis and the Projection Pursuit Approach, Practical Aspects,” in *Developments in Robust Statistics*, eds. R. Dutter, P. Filzmoser, U. Gather, and P. J. Rousseeuw, Heidelberg: Physica, pp. 317–329.

Polzehl, J. (1995), “Projection Pursuit Discriminant Analysis,” *Computational Statistics and Data Analysis*, 20, 141–157.

Rao, C. R. (1952), *Advanced Statistical Methods in Biometric Research*, New York: Wiley.

——— (ed.) (1993), *Handbook of Statistics*, (vol. 9), Amsterdam: Elsevier.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.

Röhl, M. C., and Weihs, C. (1999), “Optimal vs. Classical Linear Dimension Reduction,” in *Classification in the Information Age*, eds. W. Gaul and H. Locarek-Junge, Berlin: Springer, pp. 252–259.

Rousseeuw, P. J. (1984), “Least Median of Squares Regression,” *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J., and van Driessen, K. (1999), “A Fast Algorithm for the Minimum Covariance Determinant Estimator,” *Technometrics*, 41, 212–223.

Seber, G. A. F. (1984), *Multivariate Observations*, New York: Wiley.

Strecker, U. (2002), “*Cyprinodum esconditus*, A New Pupfish from Laguna Chichancanab, Yucatan, Mexico (Cyprinodontidae),” *Cybium*, 26, 301–307.

Wegman, E. J. (1991), “The Grand Tour in $k$-Dimensions,” in *Computer Science and Statistics: Proceedings of the 22nd Symposium on the Interface*, Fairfax, VA: Interface Foundation, pp. 127–136.

Wegman, E. J., and Carr, D. B. (1993), “Statistical Graphics and Visualization,” in *Handbook of Statistics* (vol. 9), ed. C. R. Rao, Amsterdam: Elsevier.

Wilhelm, A. F. X., Wegman, E. J., and Symanzik, J. (1999), “Visual Clustering and Classification: The Oronsay Particle Size Dataset Revisited,” *Computational Statistics*, 14, 109–146.

Wisotzki, L., Christlieb, N., Bade, N., Beckmann, V., Köhler, T., Vanelle, C., and Reimers, D. (2000), “The Hamburg/ESO Survey for Bright QSOs. III. A Large Flux-Limited Sample of QSOs,” *Astronomy and Astrophysics*, 358, 77–87.

Young, D. M., Marco, V. R., and Odell, P. L. (1987), “Quadratic Discrimination: Some Results on Optimal Low-Dimensional Representation,” *Journal of Statistical Planning and Inference*, 17, 307–319.