



Projection-pursuit approach to robust linear discriminant analysis

Ana M. Pires*, João A. Branco

Department of Mathematics and CEMAT, IST, Technical University of Lisbon, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

ARTICLE INFO

Article history:

Received 26 May 2009

Available online 1 July 2010

AMS subject classifications:

62H30

62G35

62P10

Keywords:

Projection-pursuit

Robustness

Linear discriminant analysis

Influence function

Asymptotic variance

High-dimensional data

Microarray data

ABSTRACT

Discriminant analysis plays an important role in multivariate statistics as a prediction and classification method. It has been successfully applied in many fields of work and research. As it happens with other multivariate methods, discriminant analysis is highly vulnerable to the presence of outliers that commonly occur in many real world data sets. The lack of robustness of the classical estimators on which the linear discriminant function is based is a severe disadvantage and several authors have worked to find efficient ways to prevent the damage that outliers can cause. This paper focuses on the projection-pursuit approach to discriminant analysis. The projection-pursuit estimators are described and theoretical properties are deduced and their relevance is highlighted. These include Fisher consistency, affine equivariance, partial influence functions and asymptotic distributions. Application to real data and a simulation study reveal the robustness of the projection-pursuit approach. In both analyses the data relates to a large number of variables, a situation that is becoming common when new technology is applied to data gathering.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Discriminant analysis is a widely used multivariate statistical method. Interesting applications are found in almost every traditional scientific area, ranging from social sciences to health sciences, from industry to economy. It also has large intersections with recent interdisciplinary areas such as pattern recognition and data mining. A standard reference is [37].

The need of robust methods is obvious because most applications have a large number of variables and/or observations and data seldom come from nice theoretical models. Lachenbruch et al. [33] were the first authors pointing this necessity in the discriminant analysis setup.

Robust methods for discriminant analysis have been proposed in the past three decades, mostly for linear discrimination between two groups. One of the first references is [1], followed by Randles et al. [43]. Other authors addressing this problem were [5], and [50]. Quadratic discrimination was considered by Clarke et al. [8] and Broffitt et al. [3]. A great majority of this early work concentrated on replacing the classical mean vectors and covariance matrices by robust counterparts (this is usually called “the plug-in method”). The analogy with linear regression analysis was also explored, for instance in one of the proposals made by Ahmed and Lachenbruch [1] and by Todorov et al. [51]. All the methods considered in the above references have some drawbacks, mainly related to the low breakdown point or the lack of equivariance of the robust estimators used. Meanwhile the use of high breakdown equivariant estimators of multivariate location and scatter has been considered. Chork and Rousseeuw [7] applied the Minimum Volume Ellipsoid estimators (MVE, introduced by Rousseeuw [45]). Hawkins and McLachlan [24] developed a specific method for estimating a pooled covariance matrix based on the idea of Minimum Covariance Determinant (MCD, also introduced by Rousseeuw [45]), while He and Fung [25] and Croux and Dehon [10] did similarly but considering S-estimators ([16], [46, p. 174]). Along the same lines, Hubert and Van Driessen [31] investigated

* Corresponding author.

E-mail address: apires@math.ist.utl.pt (A.M. Pires).

the plug-in approach using fast-MCD estimators of location and shape [47] whereas Todorov and Pires [52] reviewed existing methods for (high breakdown) robust linear discriminant analysis and considered several new robust procedures (plug-in approach based on the constrained M-estimates as defined by Rocke [44], and on the pairwise estimator, OGK, of [36]).

Robust versions of non-standard discriminant analysis methods have also been proposed. For instance, Vanden Branden and Hubert [54] proposed a new robust method called RSIMCA, which consists in a robustification of the classification method entitled Soft Independent Modelling of Class Analogies (SIMCA), proposed by Wold [57]. The RSIMCA method has also been studied by Daszykowski et al. [17]. The RSIMCA method is more general than most of the above-mentioned methods in that it is very flexible (with respect to the number of groups, the number of variables and observations, and data configurations). It is therefore interesting for comparison with the methods proposed in this paper.

Other relevant publications to robust discriminant analysis have also been published in the last decade: Pires [40] discusses theoretical and practical issues related to the projection-pursuit approach to robust discriminant analysis; Croux and Joossens [12] look at the behavior of the total probability of misclassification of robust linear and quadratic discriminant analysis; Croux and Joossens [13] analyze the influence of observations on error rates in quadratic discriminant analysis; Filzmoser et al. [18] measure the performance of classical and robust Fisher discriminant analysis using the error rate as a performance criterion; Han and Jin [23] use robust linear discriminant analysis model to devise a face recognition technique with good recognition performance; Hubert et al. [29] focus on high-breakdown robust multivariate methods including discriminant analysis; Croux et al. [11] compute relative classification efficiencies of robust Fisher's linear discriminant analysis with respect to the classical method; Filzmoser et al. [19] provide an insight of a number of robust methods that can be used or extended to classification, including discriminant analysis.

Most of the robust methods proposed so far and referenced above are quite good, however there are still cases where they may fail, for instance when there are categorical variables, when the number of variables is large relatively to the number of observations or, in general, when the covariance estimates are close to singularity. It is therefore important to explore other possibilities for robustifying discriminant analysis.

Since the pioneering work of Friedman and Tukey [21] projection-pursuit (PP) methods have had a visible impact on multivariate analysis (see the review papers: [27,28,32]). Some of the advantages of PP methods are their ability to overcome the so called "curse of dimensionality" and the possibility of extending one-dimensional techniques to higher dimensions. The method is thus suitable for the robustification of multivariate methods that are already special cases of PP, such as principal components and linear discriminant analysis. This was originally pointed out by Huber [27]. The case of principal components was explored by Li and Chen [35] and revisited later by Croux and Ruiz-Gazen [15]. In this paper PP ideas are used to obtain a robust estimate of the linear discriminant function (l.d.f.) for two groups.

In general a PP technique searches for low-dimensional projections of higher-dimensional data where an objective function called projection index is maximized. Fisher's original idea for linear discriminant analysis between two groups [20] is indeed a PP procedure considering as projection index the squared standardized distance between the projected observations of the two groups. This is the approach adopted in this paper. A similar but more restrict idea was considered by Chen and Muirhead [6] and is also referred in Van Ness and Yang [53]. Other attempts with different projection indexes have also been made (two of the methods proposed by Randles et al. [43], Posse [41], Chen and Muirhead [6] and Van Ness and Yang [53]).

Section 2 of the paper describes the proposed estimators of the parameters involved in the linear discriminant analysis for two groups. Theoretical properties are derived in Section 3. These include Fisher consistency, equivariance, partial influence functions and asymptotic distributions. Section 4 is devoted to practical considerations, starting with a description of a numerical algorithm, showing the application to real data sets and ending with the results of a comparative Monte Carlo study. The main conclusions of this work are drawn in Section 5.

In order to establish relevant notation and terminology a brief introduction to discriminant analysis will be made at the beginning of the coming section.

2. Description of the method

2.1. Fisher linear discriminant analysis for two groups

It is well known that discriminant analysis can be applied to a set of g samples, $g \geq 2$. In this study we consider only $g = 2$ samples of m -variate observations known to come from each of two pre-specified groups (or populations, or classes, denoted by $G_i, i = 1, 2$): $\mathbf{x}_{ij}, i = 1, 2; j = 1, \dots, n_i$ (the dimensions n_i may be fixed *a priori*, in the case of separate sampling or may be determined from the sampling process itself, in the case of mixture sampling). It is assumed that each \mathbf{x}_{ij} is a realization of a random variable with distribution F_i (and density f_i) with location parameter μ_i (the condition $\mu_1 \neq \mu_2$ is assumed throughout the paper) and scatter matrix Σ_i . The objectives of the analysis may be

- (a) to describe the differences between the two groups in terms of the m variables (sometimes referred to as descriptive discriminant analysis);

or/and

(b) to find a rule for classifying a new observation of unknown origin, characterized by the m -dimensional vector \mathbf{x} , into one of the two groups (which some authors call predictive discriminant analysis).

Fisher [20] had the second idea in mind when, for two groups, he proposed the criterion of finding the linear combination of the original variables, characterized by an m -dimensional vector α , for which the squared standardized distance between the groups is maximized, assuming common dispersion, $\Sigma_1 = \Sigma_2 = \Sigma$. At the population level this amounts to

$$\alpha = \operatorname{argmax}_{\mathbf{a} \in \mathbb{R}^m} \frac{(\mathbf{a}^T \mu_1 - \mathbf{a}^T \mu_2)^2}{\mathbf{a}^T \Sigma \mathbf{a}}. \quad (1)$$

Any vector proportional to $\Sigma^{-1}(\mu_1 - \mu_2)$ is a solution. In order to define α uniquely it is necessary to impose restrictions.

The restrictions used in this paper are $\|\mathbf{a}\| = 1$ and $\mathbf{a}^T \mu_1 > \mathbf{a}^T \mu_2$ (Fisher forced the first component of \mathbf{a} to be equal to 1, and obtained a solution that automatically satisfies the second restriction). The second restriction is easy to apply since once a solution (verifying the first restriction) is obtained and if the second restriction is not verified, then multiplication by -1 yields the right solution. This remark applies to all similar restrictions and will not be repeated. To distinguish between a generic solution and the solution verifying the restrictions, the later will be denoted by α_N . Eq. (1) is now written as

$$\alpha_N = \operatorname{argmax}_{\mathbf{a} \in \mathbb{R}^m, \|\mathbf{a}\|=1, \mathbf{a}^T \mu_1 > \mathbf{a}^T \mu_2} \frac{(\mathbf{a}^T \mu_1 - \mathbf{a}^T \mu_2)^2}{\mathbf{a}^T \Sigma \mathbf{a}} \quad (2)$$

which has the unique solution

$$\alpha_N = \frac{\Sigma^{-1}(\mu_1 - \mu_2)}{\|\Sigma^{-1}(\mu_1 - \mu_2)\|}. \quad (3)$$

Note that normality is never assumed and that this criterion does not yield a classification rule but only a direction of maximum separation between the two groups (widely used in descriptive discriminant analysis). In order to obtain a classification rule Fisher proposed to use the midpoint between the projected means for separating the two groups.

For classification, objective (b), other criteria were proposed and are commonly accepted: (b1) to minimize the total misclassification probability [56]; (b2) to minimize the total expected cost of misclassification [55]. For using these criteria more ingredients are needed: the *a priori* probability, π_i , $i = 1, 2$, that the observation that is going to be classified belongs to G_i (note that these probabilities may differ from the proportions of the groups, either at the population or the sample level, and will be assumed known); and for (b2) also the costs $C(i|j)$, $i, j = 1, 2$, $i \neq j$, of misclassifying in G_i an observation actually coming from G_j . The criteria (b1) and (b2) are equivalent when $C(i|j) = c$ for all $i \neq j$. The optimal classification rule at the population level is

$$\text{classify } \mathbf{x} \text{ in } G_1 \text{ if } \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > k \quad (4)$$

and in G_2 otherwise, where $k = \pi_2 C(1|2) / (\pi_1 C(2|1))$. Note that it is assumed that k does not depend on the data. If $\mathbf{X}_i \sim \mathcal{N}_m(\mu_i, \Sigma)$, $i = 1, 2$, then rule (4) is equivalent to

$$\text{classify } \mathbf{x} \text{ in } G_1 \text{ if } \alpha^T \mathbf{x} + \alpha_0 > 0 \quad (5)$$

and in G_2 otherwise, where

$$\alpha = \Sigma^{-1}(\mu_1 - \mu_2) \quad \text{and} \quad \alpha_0 = -\alpha^T \frac{\mu_1 + \mu_2}{2} - \log k. \quad (6)$$

These are the usual representations, but they can be multiplied by any positive constant without changing rule (5). In particular α can be replaced by its normalized version (if α_0 is changed accordingly). This case will be denoted differently using

$$\alpha_N = \frac{\alpha}{\|\alpha\|} \quad \text{and} \quad \alpha_{0N} = -\alpha_N^T \frac{\mu_1 + \mu_2}{2} - \frac{\log k}{\|\alpha\|}. \quad (7)$$

That is, apart from a difference in the separation point (when $k \neq 1$), Fisher's solution is obtained. This suggests that one can still adopt Fisher's criterion, which does not assume multivariate normality, and define the separation point in order to include costs and/or *a priori* probabilities, assuming only approximate normality of the projected observations. There is, of course, no guarantee that the solution is optimal under one of the classification criteria, (b1) or (b2), except for multivariate homoscedastic normal populations.

2.2. The projection-pursuit estimators

At the sample level, in order to obtain estimators of α_N and α_{0N} , two approaches are possible:

- (I) to replace the unknown population parameters, μ_1 , μ_2 and Σ , by estimators of multivariate location and scatter (it is equivalent to do that either in the solutions, (6) or (7), or in the defining equations, (1) or (2));
- (II) to replace the unknown univariate population parameters along the projections, $\mathbf{a}^T \mu_1$, $\mathbf{a}^T \mu_2$ and $\mathbf{a}^T \Sigma \mathbf{a}$, in the defining equation, (2), by univariate estimators of location and scatter.

If one chooses the classical estimators, sample means and variances (multivariate in I and univariate in II) the two approaches are equivalent leading to the classical solutions. However, this choice has a very serious drawback: it is not robust, meaning that a single outlier (sometimes very hard to detect in multivariate samples) may distort the entire results.

For robust estimators the two approaches are no longer equivalent. As mentioned in the Introduction the first approach, that is, the plug-in approach, has received much more attention, but there are still cases presenting difficulties, namely when the estimate of the common covariance matrix is singular or near singularity. On the other hand, with this approach, observations which are not harmful for discriminant analysis may be downweighted and vice versa. This may lead to poor estimation.

Approach (II) is a projection-pursuit approach. As will be proved it may lead to affine equivariant high breakdown estimators of α_N and α_{0N} which are competitive in terms of efficiency to other robust estimators in any situation. Moreover, these estimators can be determined even in cases where robust estimators of the means and covariance matrices cannot. The PP estimators proposed are defined next.

Definition 1. Let T and S be univariate equivariant estimators (functionals) of location and scale, respectively, that is

$$T(b + cY) = b + cT(Y) \quad \text{and} \quad S(b + cY) = |c|S(Y) \quad (8)$$

for all real numbers b and c , where by convention $T(Y) \equiv T(F)$ and $S(Y) \equiv S(F)$ if $Y \sim F$.

Given a training sample, (\mathbf{x}_{ij}) , $i = 1, 2, j = 1, \dots, n_i$, with m -variate observations from group G_i , the PP estimator of α_N , the normalized discriminant vector, induced by the univariate estimator of location, T , and the univariate estimator of scale, S , is

$$\hat{\alpha}_{N,T,S} = \operatorname{argmax}_{\mathbf{a} \in \Omega_{(m-1)/2}} \frac{[T(\mathbf{a}^T \mathbf{x}_1) - T(\mathbf{a}^T \mathbf{x}_2)]^2}{a_1 S^2(\mathbf{a}^T \mathbf{x}_1) + a_2 S^2(\mathbf{a}^T \mathbf{x}_2)} \quad (9)$$

where $\mathbf{a}^T \mathbf{x}_i = (\mathbf{a}^T \mathbf{x}_{i1}, \dots, \mathbf{a}^T \mathbf{x}_{ini})$, $i = 1, 2$ and $\Omega_{(m-1)/2}$ denotes the collection of all unit vectors in R^m such that $T(\mathbf{a}^T \mathbf{x}_1) > T(\mathbf{a}^T \mathbf{x}_2)$ (that is, $\Omega_{(m-1)/2}$ is half of the unit radius hypersphere in R^m with center at the origin). \square

In order to simplify the notation, the explicit reference to T , S will be omitted from now on, $\hat{\alpha}_{N,T,S} \equiv \hat{\alpha}_N$. An estimator of the associated separation point can be obtained directly from the definition of α_{0N} in Eq. (7). $\alpha_N^T \mu_i$ may be estimated by $T(\hat{\alpha}_N^T \mathbf{x}_i)$, but if $k \neq 1$ it is also necessary to obtain an estimate of the norm of α , denoted by $\|\alpha\|$ (note that this does not mean “the norm of the estimate of α ”, which does not make sense here because α cannot be estimated by projection-pursuit):

$$\hat{\alpha}_{0N,T,S} \equiv \hat{\alpha}_{0N} = -\frac{T(\hat{\alpha}_N^T \mathbf{x}_1) + T(\hat{\alpha}_N^T \mathbf{x}_2)}{2} - \frac{\log k}{\|\alpha\|}. \quad (10)$$

The estimator of the norm of α proposed is

$$\widehat{\|\alpha\|}_{T,S} \equiv \widehat{\|\alpha\|} = \frac{T(\hat{\alpha}_N^T \mathbf{x}_1) - T(\hat{\alpha}_N^T \mathbf{x}_2)}{a_1 S^2(\hat{\alpha}_N^T \mathbf{x}_1) + a_2 S^2(\hat{\alpha}_N^T \mathbf{x}_2)}, \quad (11)$$

where $T(\hat{\alpha}_N^T \mathbf{x}_i)$ and $S(\hat{\alpha}_N^T \mathbf{x}_i)$ are simply the location and scale estimates, respectively, of the projection of the sample from the i -th group onto $\hat{\alpha}_N$. Eq. (11) is justified by the fact that, for $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ and $\alpha_N = \alpha/\|\alpha\|$,

$$\frac{\alpha_N^T (\mu_1 - \mu_2)}{\alpha_N^T \Sigma \alpha_N} = \|\alpha\|. \quad (12)$$

A natural choice for the constants a_1 and a_2 is

$$a_i = \frac{n_i}{n_1 + n_2}, \quad i = 1, 2. \quad (13)$$

Incidentally the method here discussed also allows to obtain a PP estimator of the Mahalanobis distance between the two groups, $\Delta = [(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)]^{1/2}$. For α_N given by Eq. (2) it is easy to verify that

$$\frac{[\alpha_N^T (\mu_1 - \mu_2)]^2}{\alpha_N^T \Sigma \alpha_N} = \Delta^2. \quad (14)$$

Therefore

$$\hat{\Delta}_{T,S} \equiv \hat{\Delta} = \left[\frac{\left(T(\hat{\alpha}_N^T \mathbf{x}_1) - T(\hat{\alpha}_N^T \mathbf{x}_2) \right)^2}{a_1 S^2(\hat{\alpha}_N^T \mathbf{x}_1) + a_2 S^2(\hat{\alpha}_N^T \mathbf{x}_2)} \right]^{1/2} \quad (15)$$

is a natural PP estimator of Δ , induced by the univariate estimators T and S . If those are the sample mean and standard deviation then the classical estimator

$$\hat{\Delta} = [(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^{1/2}$$

is obtained.

Finally, the actual total misclassification cost, that is the cost incurred when a new observation is misclassified, given the training sample (or the estimates $\hat{\alpha}_N$ and $\hat{\alpha}_{0N}$) is

$$\widehat{TMC}_{T,S} \equiv \widehat{TMC} = C(2|1)\pi_1 P(\hat{\alpha}_N^T \mathbf{x} + \hat{\alpha}_{0N} < 0 | G_1) + C(1|2)\pi_2 P(\hat{\alpha}_N^T \mathbf{x} + \hat{\alpha}_{0N} > 0 | G_2). \quad (16)$$

If the underlying distributions are elliptically symmetric, $\mathbf{X} \sim \mathcal{E}\mathcal{S}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ (that is $f_i(\mathbf{x}) = \det(\boldsymbol{\Sigma}^{-1/2})g[(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)]$, with $g: [0, +\infty[\rightarrow \mathbb{R}$, continuous) then, using Lemma 1 of [15],

$$\widehat{TMC} = C(2|1)\pi_1 F_0 \left(\frac{-\hat{\alpha}_{0N} - \hat{\alpha}_N^T \boldsymbol{\mu}_1}{\sqrt{\hat{\alpha}_N^T \boldsymbol{\Sigma} \hat{\alpha}_N}} \right) + C(1|2)\pi_2 \left[1 - F_0 \left(\frac{-\hat{\alpha}_{0N} - \hat{\alpha}_N^T \boldsymbol{\mu}_2}{\sqrt{\hat{\alpha}_N^T \boldsymbol{\Sigma} \hat{\alpha}_N}} \right) \right], \quad (17)$$

where the density of F_0 is given by $f_0(y) = \int \cdots \int g(y^2 + x_2^2 + \cdots + x_m^2) dx_2 \cdots dx_m$. If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ then $F_0 = \Phi$ (the standard normal distribution function). Expressions (16) and (17) give the actual misclassification probability when $C(2|1) = C(1|2) = 1$. Note that \widehat{TMC} is not really an estimate, since it still depends of unknown population parameters, and it is useful mostly for theoretical comparisons. The problem of estimating/fixing the misclassification costs or probabilities is beyond the scope of this paper.

2.3. The choice of univariate estimators

Now it remains to discuss the choice of T and S . Many robust univariate estimators of location and scale have been proposed in the literature. At one end, is the pair $T = \text{Median}$, $S = \text{Median Absolute Deviation}$ (or MAD), very well known and regarded as being very robust because of their maximal breakdown point. These are the estimators considered by Chen and Muirhead [6]. Although this choice is not discarded, it has two main drawbacks, the lack of differentiability and the low efficiency. At the other end there are differentiable estimators with controllable trade-off between efficiency and breakdown point, but with non-explicit definitions, like M or S estimators (of course other alternatives with similar properties can also be selected).

Recall that, given a random sample of univariate observations (X_1, \dots, X_n) , a robust M -estimator of location, T , is a solution of the equation

$$\sum_{i=1}^n \psi \left(\frac{X_i - T}{S_0} \right) = 0,$$

where S_0 denotes an auxiliary estimate of scale and ψ is a bounded odd function [26]. A robust M -estimator of scale, S , is a solution of the equation

$$\sum_{i=1}^n \rho \left(\frac{X_i - T_0}{S} \right) = n\beta,$$

where T_0 denotes an auxiliary estimate of location, ρ is a bounded even function, such that $\rho(0) = 0$, and $\beta = \int \rho(u) dF_0(u)$ (to ensure that S is a consistent estimate of the scale parameter of F_0). In this paper we will consider the particular case of Huber M -estimators, for which

$$\psi(x) = \begin{cases} -b, & x < -b \\ x, & |x| \leq b \\ b, & x > b \end{cases} \quad \rho(x) = \begin{cases} x^2, & |x| \leq c \\ c^2, & |x| > c \end{cases}$$

where b and c are tuning constants. The resulting estimators (using $T_0 = \text{Median}$ and $S_0 = \text{MAD}$) will be denoted by $H(b, c)$.

Simultaneous M -estimators of location and scale, (T, S) , are defined as the solution of the system

$$\sum_{i=1}^n \psi \left(\frac{X_i - T}{S} \right) = 0, \quad \sum_{i=1}^n \rho \left(\frac{X_i - T}{S} \right) = n\beta,$$

The S-estimators of location and scale are in turn defined as the solution of the constrained minimization problem

$$T = \operatorname{argmin}_{\theta \in R} \sum_{i=1}^n \rho \left(\frac{X_i - \theta}{S} \right) \quad \text{subject to} \quad \sum_{i=1}^n \rho \left(\frac{X_i - T}{S} \right) = n\beta,$$

where ρ has the characteristics mentioned above for the M-estimators of scale (in fact S-estimators are a special type of simultaneous M-estimators, with $\psi = \rho'$). A usual choice for ρ is

$$\rho(x) = \begin{cases} \left(\frac{x}{r}\right)^6 - 3\left(\frac{x}{r}\right)^4 + 3\left(\frac{x}{r}\right)^2, & |x| \leq r \\ 1, & |x| > r. \end{cases}$$

The corresponding ψ , $\rho'(x)$, is usually known as Tukey's biweight (or bisquare) function. The resulting estimators will be denoted by $S(r)$.

In the next section several theoretical properties of the PP estimators of the linear discriminant analysis parameters, defined in this section, are presented and studied. It will be shown that if T and S are robust estimators then the PP estimators $\hat{\alpha}_N$, $\hat{\alpha}_{0N}$ and $\hat{\Delta}$ will also have robustness properties, independently of the number of variables. The results of a Monte Carlo study given in Section 4 will also confirm, empirically, the robustness of the proposed method, when T and S are chosen from the classes just described.

3. Theoretical properties

In order to establish asymptotic properties of the estimators it is convenient to define the equivalent functionals. Let $\mathbf{X} \sim G$, where G is an arbitrary m -variate distribution and, for each $\mathbf{a} \in R^m$ such that $\|\mathbf{a}\| = 1$, let $G^{\mathbf{a}}$ be the (univariate) distribution of $\mathbf{a}^T \mathbf{X}$.

Definition 2. If F_i , $i = 1, 2$, denote the distribution of the i -th group ($\mathbf{X}_i \sim F_i$) then the PP functional for the normalized discriminant vector is

$$\alpha_N(F_1, F_2) = \operatorname{argmax}_{\mathbf{a} \in \Omega_{(m-1)/2}} \frac{[T(F_1^{\mathbf{a}}) - T(F_2^{\mathbf{a}})]^2}{a_1 S^2(F_1^{\mathbf{a}}) + a_2 S^2(F_2^{\mathbf{a}})}, \quad (18)$$

where now $\Omega_{(m-1)/2}$ denotes the collection of all unit vectors in R^m such that $T(F_1^{\mathbf{a}}) > T(F_2^{\mathbf{a}})$. \square

The functionals $\alpha_{0N}(F_1, F_2)$, $\|\alpha\|(F_1, F_2)$, $\Delta(F_1, F_2)$ and $TMC(F_1, F_2)$ are defined by expressions similar to (10), (11), (15) and (17), respectively, obtained by replacing $\hat{\alpha}_N^T \mathbf{x}_i$ by $F_i^{\alpha_N(F_1, F_2)}$ (in the first three) and $\hat{\alpha}_{0N}$ by $\alpha_{0N}(F_1, F_2)$ and $\hat{\alpha}_N$ by $\alpha_N(F_1, F_2)$, in the last one.

The estimators given in Section 2 can be obtained from the previous definitions simply by setting $F_i \equiv F_{n_i}$, where F_{n_i} denotes the empirical distribution function of the i -th sample ($i = 1, 2$).

If both F_i belong to an elliptically symmetric family of distributions with distinct locations μ_i and common scatter matrix Σ then Fisher consistency of the PP functionals is a straightaway conclusion if T and S are Fisher consistent for the corresponding univariate parameters. This happens for most robust univariate location and scale estimators, which are Fisher consistent for symmetric distributions, leading thus to Fisher consistent PP functionals. (In fact, for $\alpha_N(F_1, F_2)$ and $\alpha_{0N}(F_1, F_2)$ with $k = 1$, it is sufficient that S be Fisher consistent up to a constant).

3.1. Affine equivariance

Under any F_i , with common Σ (not necessarily elliptically symmetric), the parameters have certain affine equivariance (or invariance) properties. Consider a non-singular affine transformation: let F_i^* be the distribution of $\mathbf{X}_i^* = \mathbf{A}\mathbf{X}_i + \mathbf{b}$, where \mathbf{A} is a real non-singular $m \times m$ matrix and $\mathbf{b} \in R^m$. Then $\mu_i^* = \mathbf{A}\mu_i + \mathbf{b}$, $\Sigma^* = \mathbf{A}\Sigma\mathbf{A}^T$, and

$$\alpha^* = (\mathbf{A}^{-1})^T \alpha \quad \text{or} \quad \alpha = \mathbf{A}^T \alpha^*, \quad (19)$$

$$\alpha_N^* = \frac{(\mathbf{A}^{-1})^T \alpha_N}{\|(\mathbf{A}^{-1})^T \alpha_N\|} = \frac{(\mathbf{A}^{-1})^T \alpha}{\|(\mathbf{A}^{-1})^T \alpha\|}, \quad (20)$$

$$\|\alpha^*\| = \|(\mathbf{A}^{-1})^T \alpha\| = \|\alpha\| \|(\mathbf{A}^{-1})^T \alpha_N\|, \quad (21)$$

$$\alpha_{0N}^* = \frac{\alpha_{0N}}{\|(\mathbf{A}^{-1})^T \alpha_N\|} - \alpha_N^{*T} \mathbf{b}, \quad (22)$$

and $\Delta^* = \Delta$, $TMC^* = TMC$. It is convenient that the estimators behave similarly for simultaneous affine transformations of both samples. This is stated in the next proposition. Proofs of the propositions and theorems included in the text are given in the Appendix.

Proposition 1. Let the samples for the two groups be simultaneously transformed according to $\mathbf{x}_{ij}^* = \mathbf{A}\mathbf{x}_{ij} + \mathbf{b}$, $i = 1, 2$, $j = 1, \dots, n_i$, where \mathbf{A} is a real non-singular $m \times m$ matrix and $\mathbf{b} \in R^m$. If, as assumed, the univariate estimators T and S are location-scale equivariant then the PP estimators defined in Section 2 are transformed in exactly the same way as the parameters. That is

$$\hat{\alpha}_N^* = \frac{(\mathbf{A}^{-1})^T \hat{\alpha}_N}{\|(\mathbf{A}^{-1})^T \hat{\alpha}_N\|}, \quad (23)$$

$$\|\widehat{\alpha}\|^* = \|\widehat{\alpha}\| \|(\mathbf{A}^{-1})^T \hat{\alpha}_N\|, \quad (24)$$

$$\hat{\alpha}_{0N}^* = \frac{\hat{\alpha}_{0N}}{\|(\mathbf{A}^{-1})^T \hat{\alpha}_N\|} - (\hat{\alpha}_N^*)^T \mathbf{b}, \quad (25)$$

$$\hat{\Delta}^* = \hat{\Delta} \quad \text{and} \quad \widehat{TMC}^* = \widehat{TMC}. \quad \square \quad (26)$$

The theoretical relevance of this result will become apparent in the next subsections. But the property is also important in practice, meaning that good or bad results do not depend for instance on the particular scale of the variables, and allows prior standardization which is important for badly scaled problems. Note that for plug-in estimators of the same parameters (designated by approach (I) in Section 2), this property is inherited from affine equivariance of the multivariate location and scatter estimators (recall that, using similar notation, $\hat{\mu}$ and $\hat{\Sigma}$ are affine equivariant estimators of multivariate location and scatter if the estimates for the transformed data, $\mathbf{x}_i^* = \mathbf{A}\mathbf{x}_i + \mathbf{b}$, $i = 1, \dots, n$, are, respectively, $\hat{\mu}^* = \mathbf{A}\hat{\mu} + \mathbf{b}$ and $\hat{\Sigma}^* = \mathbf{A}\hat{\Sigma}\mathbf{A}^T$).

3.2. Partial influence functions

The influence function [22] of an estimator (or of the equivalent functional) is an important tool in robustness studies. It measures the relative change in the functional induced by an infinitesimal contamination at the point \mathbf{x} :

$$IF(\mathbf{x}; T, F) = \lim_{\varepsilon \rightarrow 0} \frac{T((1 - \varepsilon)F + \varepsilon\Delta_{\mathbf{x}}) - T(F)}{\varepsilon}.$$

For functionals depending on more than one distribution, as the PP functionals considered in Definition 2, it makes sense to consider contamination of the distributions one at a time, leading to as many influence functions as the number of distributions involved. By analogy with derivatives and partial derivatives these influence functions may be called partial influence functions [39]. For instance, the first partial influence function of $\hat{\alpha}_N$ (which is a function from R^m onto R^m) at the pair (F_1, F_2) is

$$PIF_1(\mathbf{x}; \hat{\alpha}_N, F_1, F_2) = \lim_{\varepsilon \rightarrow 0} \frac{\alpha_N((1 - \varepsilon)F_1 + \varepsilon\Delta_{\mathbf{x}}, F_2) - \alpha_N(F_1, F_2)}{\varepsilon}.$$

and so on. Given the equivariance properties in Proposition 1 it is sufficient to obtain the partial influence functions (PIF) at a chosen “central” pair of distributions for which all the others can be reduced by a suitable affine transformation. The “central” pair of distributions used in this paper is characterized by $\mu_1^* = (\Delta/2, 0, \dots, 0)^T$, $\mu_2^* = (-\Delta/2, 0, \dots, 0)^T$ and $\Sigma^* = \mathbf{I}_m$ and denoted by $\mathbf{F}^* = (F_1^*, F_2^*)$. It is easy to verify that a transformation relating this distribution to a general (F_1, F_2) , in the same family, characterized by μ_1, μ_2 and Σ , is $\mathbf{x}^* = \mathbf{A}\mathbf{x} + \mathbf{b}$ (or $\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}^* - \mathbf{A}^{-1}\mathbf{b}$), with $\mathbf{A} = \mathbf{U}\Sigma^{-1/2}$ and $\mathbf{b} = -\mathbf{A}(\mu_1 + \mu_2)/2$, where $\Sigma^{-1/2}$ is the unique symmetric positive definite matrix such that $\Sigma^{-1/2}\Sigma^{-1/2} = \Sigma^{-1}$ and \mathbf{U} is a non-unique orthogonal matrix such that its first row is $(\mu_1 - \mu_2)^T \Sigma^{-1/2} / \Delta$. The linear discriminant parameters at the “central” distribution are simply $\alpha^* = (\Delta, 0, \dots, 0)^T$, $\alpha_N^* = (1, 0, \dots, 0)^T$ and $\alpha_{0N}^* = -\log k / \Delta$. The relations between the PIF of the PP functionals at (F_1, F_2) and the corresponding PIF at (F_1^*, F_2^*) are given in the next proposition.

Proposition 2. Let $\mathbf{F} = (F_1, F_2)$ and $\mathbf{F}^* = (F_1^*, F_2^*)$ be related as above, then

$$PIF_i(\mathbf{x}; \hat{\alpha}_N, \mathbf{F}) = (\mathbf{I}_m - \alpha_N \alpha_N^T) \frac{\Delta}{\|\alpha\|} \mathbf{A}^T PIF_i(\mathbf{x}^*; \hat{\alpha}_N, \mathbf{F}^*), \quad (27)$$

$$PIF_i(\mathbf{x}; \|\widehat{\alpha}\|, \mathbf{F}) = \frac{\|\alpha\|}{\Delta} PIF_i(\mathbf{x}^*; \|\widehat{\alpha}\|, \mathbf{F}^*) + \Delta \alpha_N^T \mathbf{A}^T PIF_i(\mathbf{x}^*; \hat{\alpha}_N, \mathbf{F}^*), \quad (28)$$

$$PIF_i(\mathbf{x}; \alpha_{0N}, \mathbf{F}) = \frac{\Delta}{\|\alpha\|} [PIF_i(\mathbf{x}^*; \hat{\alpha}_{0N}, \mathbf{F}^*) + (\mathbf{b}^T - \alpha_{0N} \alpha_N^T \mathbf{A}^T) PIF_i(\mathbf{x}^*; \hat{\alpha}_N, \mathbf{F}^*)], \quad (29)$$

$$PIF_i(\mathbf{x}; \hat{\Delta}, \mathbf{F}) = PIF_i(\mathbf{x}^*; \hat{\Delta}, \mathbf{F}^*) \quad \text{and} \quad PIF_i(\mathbf{x}; \widehat{TMC}, \mathbf{F}) = PIF_i(\mathbf{x}^*; \widehat{TMC}, \mathbf{F}^*). \quad \square \quad (30)$$

The PIF of the PP estimators at the “central” distribution are stated in the next theorem.

Theorem 1. If the following regularity conditions hold,

- (C1) F_i^* are elliptically symmetric distributions with distinct location parameters $\mu_i = ((-1)^{i+1} \Delta/2, 0, \dots, 0)^T$ and common scatter matrices $\Sigma = \mathbf{I}_m$;
- (C2) The functions $(\varepsilon, y) \rightarrow S[(1 - \varepsilon)F_0 + \varepsilon\Delta_y]$, $T[(1 - \varepsilon)F_0 + \varepsilon\Delta_y]$, with F_0 defined as in (17), are twice continuously differentiable at all points $(0, y)$;
- (C3) T and S are location–scale equivariant and Fisher consistent for the corresponding univariate parameters, which means that $S(F_0) = 1$;

then the partial influence functions of the projection-pursuit functionals at $\mathbf{F}^* = (F_1^*, F_2^*)$ are given by

$$PIF_i(\mathbf{x}; \hat{\alpha}_N, \mathbf{F}^*) = (\mathbf{I}_m - \alpha_N^* \alpha_N^{*T}) \mathbf{x} \left(\frac{(-1)^{i+1}}{\Delta} IF'(x_1 - \mu_{i1}^*; T, F_0) - a_i IF'(x_1 - \mu_{i1}^*; S, F_0) \right), \quad (31)$$

where $IF'(z; W, F_0) = \frac{d}{dz} IF(z; W, F_0)$ with $W = T$ or $W = S$.

$$PIF_i(\mathbf{x}; \widehat{\|\alpha\|}, \mathbf{F}^*) = (-1)^{i+1} IF(x_1 - \mu_{i1}^*; T, F_0) - 2a_i \Delta IF(x_1 - \mu_{i1}^*; S, F_0), \quad (32)$$

$$PIF_i(\mathbf{x}; \hat{\alpha}_{0N}, \mathbf{F}^*) = -\frac{1}{2} IF(x_1 - \mu_{i1}^*; T, F_0) + \frac{\log k}{\Delta^2} IF_i(\mathbf{x}; \widehat{\|\alpha\|}, \mathbf{F}^*), \quad (33)$$

$$PIF_i(\mathbf{x}; \hat{\Delta}, \mathbf{F}^*) = (-1)^{i+1} IF(x_1 - \mu_{i1}^*; T, F_0) - a_i \Delta IF(x_1 - \mu_{i1}^*; S, F_0), \quad (34)$$

$$PIF_i(\mathbf{x}; \widehat{TMC}, \mathbf{F}^*) = IF_i(\mathbf{x}; \hat{\alpha}_{0N}, \mathbf{F}^*) \left[C(1|2) \pi_2 f_0 \left(\frac{\log k}{\Delta} + \frac{\Delta}{2} \right) - C(2|1) \pi_1 f_0 \left(\frac{\log k}{\Delta} - \frac{\Delta}{2} \right) \right]. \quad \square \quad (35)$$

Remarks. (1) The PIF of the PP functionals at $\mathbf{F} = (F_1, F_2)$, where F_i is elliptically symmetric, with generic but distinct locations μ_i and common scatter matrices Σ , may then be obtained applying Proposition 2. Note that it is not necessary to compute the matrix \mathbf{A} explicitly, since it is easy to verify that $x_1^* - \mu_{i1}^* = \alpha^T(\mathbf{x} - \mu_i)/\Delta$, and that

$$\mathbf{A}^T (\mathbf{I}_m - \alpha_N^* \alpha_N^{*T}) \mathbf{x}^* = \left(\Sigma^{-1} - \frac{\|\alpha\|^2}{\Delta^2} \alpha_N \alpha_N^T \right) \left(\mathbf{x} - \frac{\mu_1 + \mu_2}{2} \right).$$

(2) Theorem 1 is not valid for $T = \text{Median}$, $S = \text{MAD}$, because for these estimators regularity condition (C2) is not verified.

(3) When T is the sample mean and S is the sample standard deviation $IF(z; T, F_0) = z$ and $IF(z; S, F_0) = (z^2 - 1)/2$, and the PIF of the classical discriminant analysis estimators are obtained (some of these influence functions were first obtained by Campbell [5]).

(4) For the classical estimators all the PIF are unbounded (which is not surprising because these estimators are not robust) except for \widehat{TMC} when $k = 1$ (see remark 7).

(5) For robust T and S the PIF of $\hat{\alpha}_N$ are also unbounded, meaning that this estimator may be very sensitive to small amounts of contamination. At \mathbf{F}^* , very large values of the PIF are obtained for small x_1 combined with large values on one of the other variables. For very large x_1 this will not happen, in general, because for common robust location and scale estimators $IF(z; \cdot, \cdot)$ is either constant or rapidly approaching a constant. This is apparently an intriguing result but it is not unique, since there are other cases of robust estimators with unbounded influence function. In particular, the same happens for the PP estimators of principal components (cf. [15]). It would be interesting to investigate if this is a general characteristic of PP estimators of directions.

(6) $PIF_i(\mathbf{x}; \hat{\alpha}_{N1}, \mathbf{F}^*) \equiv 0$ and $\hat{\alpha}_N^T PIF_i(\mathbf{x}; \hat{\alpha}_N, \mathbf{F}) \equiv 0$, meaning that the relevant changes are orthogonal to the discriminant direction (which is not surprising due to the unit norm constrain).

(7) When $k = 1$, $PIF_i(\mathbf{x}; \widehat{TMC}, \mathbf{F}) \equiv 0$, because of the symmetry of f_0 .

Fig. 1 shows plots of $PIF_1(\mathbf{x}; \hat{\alpha}_N, \mathbf{F}^*)$ at a bivariate normal distribution with $a_1 = a_2 = 0.5$ and $\Delta = 2$ for PP estimators with four different choices for T and S : (a) classical estimators, that is, T is the mean and S is the standard deviation; (b) separate Huber M-estimators with different tuning constants for the location and scale components, $H(1.5, 2)$; (c) separate Huber M-estimators with common tuning constants for the location and scale components, $H(3, 3)$; (d) S-estimators with the maximal asymptotic breakdown point of 50%, $S(1.548)$. The plots illustrate the different behavior of the estimators: with the classical method the influence is unbounded in every direction whereas with the robust estimators it is bounded in almost every direction (it is unbounded only in some directions parallel to the x_2 axis). On the other hand, all the plots are of similar shape near the origin, but (c) is the most similar to (a) and (d) is the most different (it is thus expected that the corresponding estimators present different trade-off between robustness and efficiency).

3.3. Asymptotic distributions

Pires and Branco [39] show that under regularity conditions the asymptotic variances of a general estimator depending on two samples at given populations (F_1 and F_2) can be obtained from the partial influence functions of the corresponding

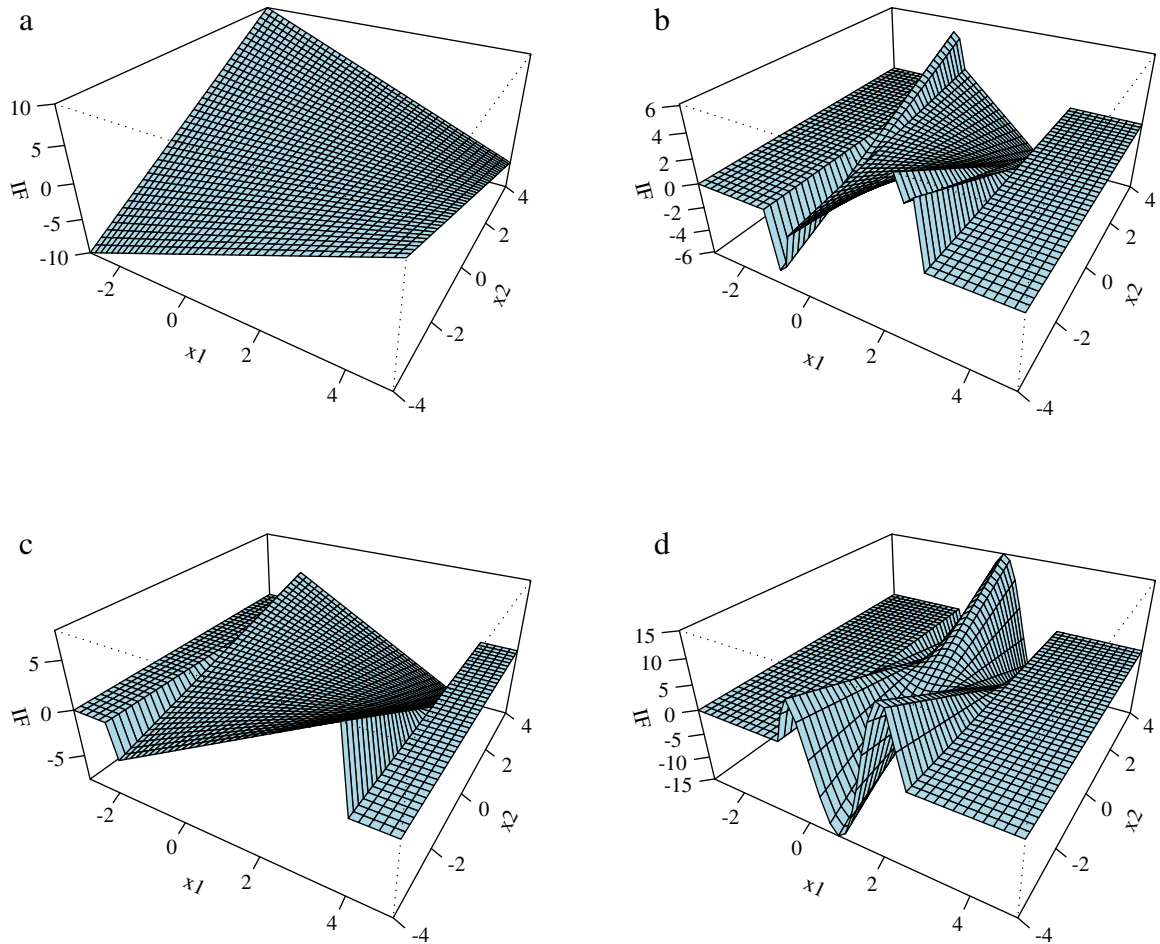


Fig. 1. $PIF_1(\mathbf{x}; \hat{\alpha}_N, \mathbf{F}^*)$ at a bivariate normal distribution with $a_1 = a_2 = 0.5$ and $\Delta = 2$ for PP estimators with four different choices for T and S : (a) classical (Mean, Standard deviation), (b) $H(1.5, 2)$, (c) $H(3, 3)$, and (d) $S(1.548)$.

functional by

$$\begin{aligned} V(\mathbf{T}, F_1, F_2) &= \lim_{n_1+n_2 \rightarrow \infty} (n_1 + n_2) \text{var } \mathbf{T}(F_{n_1}, F_{n_2}) \\ &= \frac{1}{a_1} V_1(\mathbf{T}, F_1, F_2) + \frac{1}{a_2} V_2(\mathbf{T}, F_1, F_2), \end{aligned} \quad (36)$$

where $a_i = \lim_{n_i \rightarrow \infty} n_i / (n_1 + n_2)$ and

$$V_i(\mathbf{T}, F_1, F_2) = \int PIF_i(\mathbf{x}; \mathbf{T}, F_1, F_2) PIF_i(\mathbf{x}; \mathbf{T}, F_1, F_2)^T dF_i(\mathbf{x}). \quad (37)$$

The next theorem states the conditions under which the PP estimators of the normalized discriminant vector are asymptotically normal and gives the expression of its asymptotic variance.

Theorem 2. If the partial influence functions of $\hat{\alpha}_N^*$ at \mathbf{F}^* are given by (31), that is, if the regularity conditions of Theorem 1 hold, and the remainder of the first-order von Mises expansion of the empirical distribution (F_{n_1}, F_{n_2}) around $\mathbf{F}^* = (F_1^*, F_2^*)$ converges in probability to zero (see [39], Theorem 2.1) then, for the corresponding PP estimators of the normalized discriminant vector, it holds that

$$\sqrt{n_1 + n_2} (\hat{\alpha}_N^* - \alpha_N^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}_m, V(\hat{\alpha}_N, \mathbf{F}^*)),$$

where $V(\hat{\alpha}_N, \mathbf{F}^*) = (\mathbf{I}_m - \alpha_N^* \alpha_N^{*T}) C = \text{diag}[0, C, \dots, C]$, and the scalar C is given by

$$C = C_S + \frac{C_T}{a_1 a_2 \Delta^2}, \quad (38)$$

Table 1

Values of the constant C (asymptotic variance of the i th component of $\hat{\alpha}_N$, $i = 2, \dots, m$) at normal (\mathcal{N}), t-Student (t_5) and symmetric contaminated normal (\mathcal{SCN} , with $c = 9$ and $\varepsilon = 0.1$) distributions with $a_1 = a_2 = 0.5$ and $\Delta = 1, 2, 4$ for some PP and PI estimators.

Estimator		$\Delta = 1$			$\Delta = 2$			$\Delta = 4$		
Type	Name	\mathcal{N}	\mathcal{SCN}	t_5	\mathcal{N}	\mathcal{SCN}	t_5	\mathcal{N}	\mathcal{SCN}	t_5
PP	S(1.548)	66.90	21.99	25.39	19.34	7.71	8.84	7.45	4.14	4.70
	S(2)	26.26	11.51	13.65	8.17	4.51	5.18	3.65	2.77	3.06
	S(3)	9.18	6.69	7.33	3.27	3.04	3.11	1.79	2.13	2.06
	S(4.685)	5.69	5.71	5.57	2.21	2.87	2.56	1.34	2.16	1.81
PP	H(1.5, 1.5)	6.71	5.82	6.39	3.25	3.19	3.47	2.38	2.53	2.74
	H(1.5, 2)	5.97	5.67	5.87	2.51	3.04	2.96	1.64	2.38	2.23
	H(2, 2)	5.54	5.71	5.79	2.40	3.05	2.94	1.62	2.38	2.22
	H(3, 3)	5.04	6.36	5.67	2.03	3.53	2.78	1.28	2.83	2.05
PI	($m = 2$)									
	MCD _{25%}	16.04	9.82	9.46	8.61	6.04	5.79	6.75	5.10	4.87
	RMCD _{25%}	5.92	4.79	4.89	2.50	2.45	2.30	1.64	1.87	1.65
PI	MultS _{25%}	5.56	4.07	4.61	2.27	2.02	2.25	1.45	1.50	1.66
	($m = 5$)									
	MCD _{25%}	10.62	6.36	6.38	4.97	3.48	3.49	3.56	2.77	2.77
PI	RMCD _{25%}	5.44	5.32	4.92	2.20	2.83	2.29	1.39	2.21	1.64
	MultS _{25%}	5.13	3.83	4.33	2.06	1.89	2.04	1.29	1.40	1.47
	($m = 10$)									
PI	MCD _{25%}	8.98	5.47	5.52	3.96	2.85	2.89	2.70	2.20	2.23
	RMCD _{25%}	5.31	5.67	4.98	2.13	3.10	2.34	1.34	2.45	1.68
	MultS _{25%}	5.05	3.75	4.28	2.02	1.83	2.01	1.26	1.36	1.44
Classical		5.00	6.78	7.00	2.00	3.78	4.00	1.25	3.03	3.25

with $C_S = E_{F_1^*} \left[x_2^2 (IF'(x_1 - \mu_{i1}^*; S, F_0))^2 \right]$ and $C_T = E_{F_1^*} \left[x_2^2 (IF'(x_1 - \mu_{i1}^*; T, F_0))^2 \right]$. \square

Given the model, F^* , the values of C_T and C_S , necessary to compute C , can be obtained by numerical integration. Table 1 presents some results for three different types of elliptically symmetric distributions with discrimination parameters determined by $a_1 = a_2 = 0.5$ and $\Delta = 1, 2, 4$: normal (\mathcal{N}), t-Student with five degrees of freedom (t_5) and a symmetric contaminated normal (\mathcal{SCN} , with $c = 9$ and $\varepsilon = 0.1$). For the t and \mathcal{SCN} a rescaled form, with identity covariance matrix, was adopted (see [39]), which means that the standard densities are given by

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{m+v}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\pi^{m/2}(\nu-2)^{m/2}} \left[1 + \frac{\mathbf{x}^T \mathbf{x}}{\nu-2}\right]^{-\frac{m+v}{2}}$$

and

$$f(\mathbf{x}) = (1 - \varepsilon) \left(\frac{a}{2\pi}\right)^{m/2} \exp\left[-\frac{a\mathbf{x}^T \mathbf{x}}{2}\right] + \varepsilon \left(\frac{a}{2\pi c}\right)^{m/2} \exp\left[-\frac{a\mathbf{x}^T \mathbf{x}}{2c}\right] \quad (\text{with } a = 1 + (c - 1)\varepsilon),$$

for the t_ν and $\mathcal{SCN}_{c,\varepsilon}$ respectively. $S(r)$ and $H(b, c)$ denote the pair of univariate estimators, (T, S) , used in the definition of the projection index (see Section 2.3).

The table includes the asymptotic variances for some plug-in estimators (PI) using multivariate robust covariance–location estimators. The methods considered were: the Minimum Covariance Determinant [45] with a 25% breakdown point, denoted by MCD_{25%}; the one-step Reweighted Minimum Covariance Determinant [45] with a 25% breakdown point, denoted by RMCD_{25%}¹; and the multivariate S-estimators [48] also with a 25% breakdown point, denoted by MultS_{25%}. Unlike the PP case, for this type of estimators the asymptotic variance depends on m , the number of variables. The figures reported were obtained as described in [39].

The last line shows the asymptotic variance of the classical estimators, which can be classified, as mentioned previously, both as PP and PI. This information allows the comparison of the various methods in terms of their asymptotic relative efficiency to the classical method (ARE_{cla}). Almost all estimators are less efficient than the classical at the normal distribution but more efficient at the \mathcal{SCN} or the t_5 (the exceptions are the first two $S(r)$ and MCD_{25%} when $m = 2$). It is also possible to conclude that, for the cases considered, the asymptotic efficiency of PP estimators is in general worse than that of PI estimators. This means that for large samples following these model distributions the PI estimators are to be preferred, instead of the PP estimators, on grounds of efficiency. However the PP estimators may still be of value for smaller sample sizes, especially when the number of observations is close to the number of variables, or even smaller, a situation where most PI estimators cannot be computed at all.

Remark. The breakdown point is another important theoretical concept. However, adaptation of the usual definition to the discriminant analysis setup is not straightforward. Van Ness and Yang [53] have presented a proposal but there are other possibilities needing further research. Intuitively what can be said about the breakdown point of linear discrimination procedures is that it must be smaller or equal than the minimum of the breakdown points of all the location and scale estimators involved, either univariate or multivariate. Moreover, it must depend on Δ^2 and approach 0 as $\Delta^2 \rightarrow 0$.

4. Applications and comparison with other methods

In this section we present the results of the application of projection-pursuit methodology to two real data sets as well as the results of a small scale simulation study. Other appropriate discrimination methods were selected for comparison. We start by describing the algorithms used in those applications to compute the PP estimates.

4.1. Algorithms

The definition of the PP estimate of α_N given by Eq. (9) in Definition 1, poses a non-trivial optimization problem, except for non-interesting (and non-robust) cases like the classical, or in situations with at most 3 variables, for which it is possible to visually inspect the optimizing surface. For all the other cases it is necessary to devise a strategy to obtain a good approximation of the true estimate (i.e., of the global maximum).

Following an idea proposed by Croux and Ruiz-Gazen [14] for projection-pursuit principal components, we define the first approximation to the solution as in (9) but considering a finite number of candidate directions,

$$\hat{\alpha}_N^{(1)} = \operatorname{argmax}_{\mathbf{a} \in \mathcal{D}} \frac{[T(\mathbf{a}^T \mathbf{x}_1) - T(\mathbf{a}^T \mathbf{x}_2)]^2}{a_1 S^2(\mathbf{a}^T \mathbf{x}_1) + a_2 S^2(\mathbf{a}^T \mathbf{x}_2)},$$

where \mathcal{D} contains the $n_1 \times n_2$ unit vectors defined by all pairs of points such that one belongs to G_1 and the other to G_2 :

$$\mathcal{D} = \left\{ \frac{\mathbf{x}_{1i} - \mathbf{x}_{2j}}{\|\mathbf{x}_{1i} - \mathbf{x}_{2j}\|}, i = 1, \dots, n_1, j = 1, \dots, n_2 \right\}.$$

As a second approximation we consider the solution returned by a numerical optimization algorithm using $\hat{\alpha}_N^{(1)}$ as initial solution. Taking into account that the nature of the function to be optimized is not well known, we have chosen a flexible method, the Nelder and Mead [38] simplex algorithm, which uses only function values and is implemented by function `optim(stats)` in R [42]. This approximation will be denoted $\hat{\alpha}_N^{(2)}$.

It is important to remark that the first algorithm, by considering a finite number of directions, loses the equivariance properties of the theoretical method. More specifically, equivariance no longer holds for general affine transformations and it is only possible to guarantee orthogonal equivariance. However, if the number of candidate directions is large and the approximate solution is very close to the true solution, the resulting estimates will be approximately affine equivariant, that is, the expressions given in Proposition 1 will be approximately verified, as long as the affine transformation is not very close to a singular transformation. On the contrary, the estimates obtained with the second algorithm will be affine equivariant whenever the numerical optimization converges to the global maximum of the objective function.

In the remainder of this section we will use eight PP “methods” denoted PPX_i , with $X = H$, for $(T, S) \equiv H(1.5, 2)$, $X = M$, for $(T, S) \equiv (\text{Median}, \text{MAD})$, $X = S$, for $(T, S) \equiv S(1.548)$, $X = C$, for $(T, S) \equiv (\text{Mean}, \text{Standard deviation})$; and $i = 1, 2$ referring to $\hat{\alpha}_N^{(i)}$. The particular H and S representatives were chosen because they have different efficiency and breakdown properties. For H, the ARE_{cla} is around 80% for $1 \leq \Delta \leq 4$, and the maximal breakdown point is $\approx 23\%$. For S, the ARE_{cla} is around 10% for $1 \leq \Delta \leq 4$, and the maximal breakdown point is $\approx 50\%$. The pair $(\text{Median}, \text{MAD})$ (M) is included because it has different properties and it is thus expected to behave differently. Finally the pair $(\text{Mean}, \text{Standard deviation})$ (C) is considered because the exact solution is in this case known, and thus it is possible to have an indirect evaluation of the contribution of the algorithm to the performance of the methods.

4.2. Ceramic data set

This data set was first published in [49] and is reproduced in [9] who use it to introduce and illustrate discriminant analysis. The main data set has 27 observations (from fragments of Greek pottery), which were classified into two groups (G_1 : Attic origin, with 13 observations; G_2 : Eritrean origin, with 14 observations) and measured on six chemical variables (metallic oxide constituents – Si, Al, Fe, Mg, Ca, Ti). There is a separate data set with 13 observations, the first 4 classified as “probably Attic” and the remaining 9 as “probably Eritrean”. This data set can be used as a separate test set.

Let us consider first only two variables, Mg and Ca, as done by Cooper and Weekes [9], in order to gain insight from graphical presentations. The eight PP methods described in the previous subsection were applied to the main data set. For comparison we also applied the classical linear discriminant (LDA) and two PI robust methods using 50% breakdown point reweighted MCD estimates of the group means and of the pooled covariance matrix obtained from two different algorithms: the minimum within covariance determinant based on the “feasible solution algorithm”, proposed by Hawkins and McLachlan [24], denoted fsa-MWCD , and the minimum within covariance determinant based on the “fast-MCD algorithm”, proposed by Hubert and Van Driessen [31], denoted fast-MWCD (for implementation details see [52], and the R package `rrcov`).

Fig. 2 shows the bivariate scatter plot of the data, together with the six distinct separation lines ($\hat{\alpha}_N^T \mathbf{x} + \hat{\alpha}_{0N} = 0$) that were obtained. Although PPX_1 and PPX_2 have produced slightly different numerical results, the differences cannot be observed graphically. The plot shows that several observations from both groups appear to be outlying relatively to the respective group (outlier detection using Mahalanobis distances computed with $\text{RMCD}_{50\%}$, for each group, detected the observations

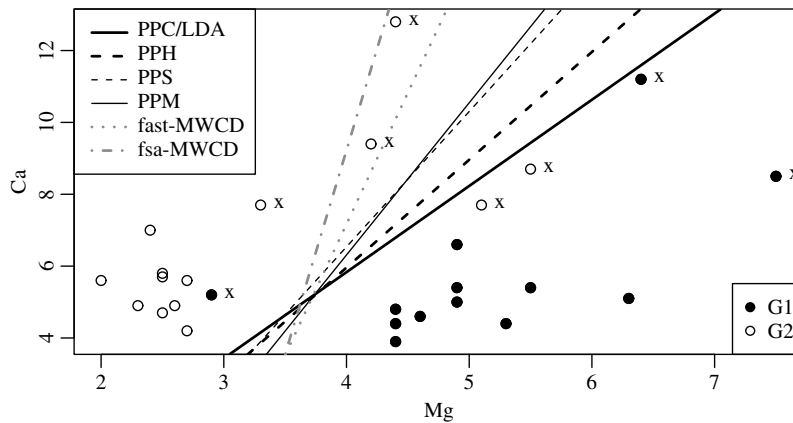


Fig. 2. Scatter plot of Ca versus Mg for the ceramic data, showing the separation lines estimated by six different methods (PPC/LDA, PPH, PPS, PPM, fast-MWCD, fsa-MWCD). The symbol “x” marks the observations that can be considered outliers, according to Mahalanobis distance, relatively to the respective group.

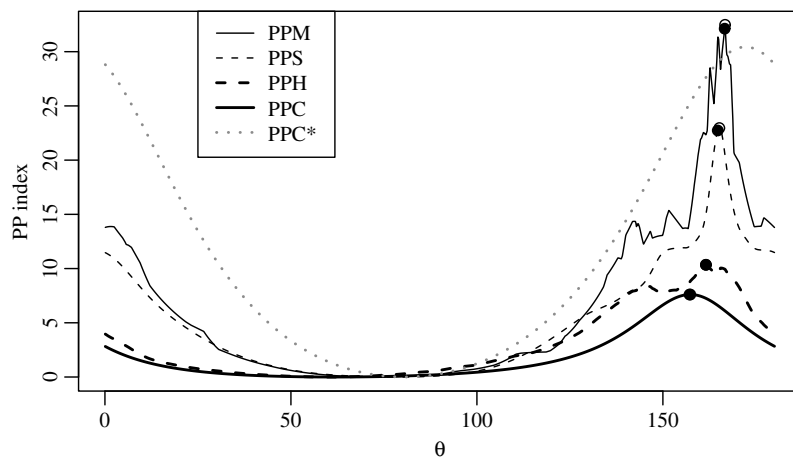


Fig. 3. Plots of the PP index with the M, S, H and C univariate estimators. The label “PPC*” refers to the PPC method computed without the observations marked with “x” in Fig. 2.

Table 2

Number of misclassified observations from the ceramic data sets obtained using several classification strategies.

N. of Variables	Type of Classif.	Estimators		PPH ₁	PPH ₂	PPM ₁	PPM ₂	PPS ₁	PPS ₂	PPC ₁	PPC ₂	LDA	fast-MWCD	fsa-MWCD
		3	3											
2	Apparent	3	3	3	3	3	3	3	3	3	3	3	3	5
	CV	4	4	3	3	3	3	3	3	4	4	4	3	3
	Test	4	4	3	3	3	3	3	3	4	4	4	2	3
6	Apparent	1	3	3	3	3	3	3	3	1	1	1	3	3
	CV	1	3	3	3	3	3	3	3	1	2	2	4	4
	Test	2	2	2	2	2	2	2	2	2	2	2	2	2

marked with “x”, 3 in G_1 and 5 in G_2). It is quite clear that the estimates are disturbed by those outlying observations in different degrees. The classical LDA (or the equivalent PPC) is the most affected, while the fsa-MWCD is the least affected.

In what concerns the PP methods, their different behavior can also be appreciated in Fig. 3 which shows plots of five projection-pursuit indexes, PPH, PPM, PPS, PPC and PPC* (the last one denoting the PPC method computed without the outliers shown in Fig. 2). It is clear that the M index is the most irregular and therefore the most difficult to maximize. The S index appears to be the second most difficult. On the other hand the Huber index seems to be smoother but it is also closer to the C index, so it must be less robust and simultaneously more efficient with regular/clean data. In this dimension the two estimating algorithms give very close solutions. The solid circles on the plots correspond to the solutions obtained using only data directions, whereas the empty circles mark the solutions obtained from the optimization algorithm. The empty circles are barely visible on the M and S curves and completely invisible on the H and C curves.

To conclude the analysis of these data we have used the different estimates obtained from the main data set to classify those same observations as well as the observations in the test data set. The results are given in Table 2. The first three lines

were obtained using only two variables (Mg and Ca), while the last three were obtained using the six original variables. The number of misclassified observations was computed in three different ways: apparent (number misclassified in the main data set), CV (cross-validation or leaving-one-out for the main data set), and test (number misclassified in the test data set). With two variables the best results are obtained with fast-MWCD, but when the six variables are used PPH₁ and PPC₁ lead to the best results. This fact is a little intriguing at first sight, but we think that it can be explained by the sparsity of the data (13 and 14 observations, respectively for G₁ and G₂, in 6 variables), something that may be better understood with the next example.

4.3. Colon cancer data set

This data set is the result of a microarray experiment [2]. It contains 62 observations on subjects classified into two groups (G₁: subjects with colon cancer, with 40 observations; G₂: healthy subjects, with 22 observations) and measured on 2000 variables (gene expression levels). The aim is to predict, as accurately as possible, the disease status from the gene expression levels.

This is a well known data set in the modern classification literature (e.g., [34,4,58]) and the original version is available in the `colonCA` R package from Bioconductor. The raw data is not normalized/preprocessed, which may lead to very bad classification results. Therefore a simple normalization procedure was applied: the data were log-transformed and after that each row was individually centered using its median.

When applying discriminant analysis to a data set like this, with many more variables than observations, it is convenient to select a smaller number of variables, presumably relevant to the classification task, rather than to blindly use all the variables. As the variable selection procedure is not the main focus of this paper, it was decided to use, just for illustration, two simple dimensionality reduction methods:

- (i) compute a univariate t -statistic to test the hypothesis of equality between the two group means of each variable, t_{0i} , $i = 1, \dots, m$; rank the m variables by decreasing order of $|t_{0i}|$ and choose the top p ;
- (ii) obtain the principal components of the complete data matrix ignoring the group structure, and use the first q principal component scores as input to the discriminant analysis procedure.

Both classical and robust variants were used for each of the two methods. The t -statistic in (i) was computed using the classical Welch's t -test,

$$t_{0i}^c = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sqrt{\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}}}, \quad i = 1, \dots, m$$

and an “ad hoc” robustified version of it, t_{0i}^r , obtained using, in the previous expression, medians in place of means and median absolute deviations in lieu of standard deviations. The resulting selections will be denoted by TC and TR, respectively.

The principal components in (ii) were obtained using the classical method, i.e., the eigendecomposition of the covariance matrix (CPC) and the ROBPCA method of [30] as implemented in the R package `rrcov` (with the default options: 25% breakdown point and 250 directions for computing the outlyingness), which will be denoted by RPC.

It was decided to use, for illustration, $p = 15$, $p = 75$, $q = 15$ and all the variables, $p = 2000$ (with 15 variables all the covariance matrices are non-singular but the opposite happens with 75 or 2000 variables; 15 principal components explain 76% of the total variance). Thus, the various discriminant methods will be applied to seven data sets: 15TR, 75TR, 15TC, 75TC, 15RPC, 15CPC and All. Note that if the aim of this study was to find a good classification method for this data set it would be necessary to consider much more values of p and q .

The discrimination methods selected for comparison were the eight PP variants described in Section 4.1, and the five following alternative methods chosen among those that can be used in problems with more variables than observations:

- LDA: Fisher's linear discriminant as described in Section 2, using the Moore–Penrose pseudoinverse when the pooled covariance matrix is singular;
- DDA: diagonal discriminant analysis, i.e., Fisher's linear discriminant but using only the diagonal of the pooled covariance matrix, $\hat{\Sigma} = \text{diag}(\mathbf{S})$; in this way $\hat{\Sigma}$ is never singular;
- OGK₁: similar to LDA but using the OGK multivariate location and scatter estimators [36] in place of the sample mean vectors and covariance matrices (the OGK estimates were obtained using function `covOGK` from the R package `robustbase`);
- OGK₂: DDA with OGK estimates;
- RSIMCA: this method is described in [54]; briefly it consists on performing a robust principal component analysis separately for each group, and to classify each observation using a certain function of the derived orthogonal and Mahalanobis distances to the group centers, so it can be regarded as a kind of robust generalized quadratic discriminant; the ROBPCA method from the R package `rrcov` (with the options described as above, plus an automatic rule for selecting the number of components to retain in each group, given by the smallest k such that $\hat{\lambda}_k/\hat{\lambda}_1 \leq 10^{-3}$ or $\sum_{i=1}^k \hat{\lambda}_i / \sum_{i=1}^r \hat{\lambda}_i \geq 0.8$, where r is the rank of an intermediate estimate of the covariance matrix) was also

Table 3

Colon data set. Cross-validation total error rates: mean (standard deviation) over 200 replications.

Var.	Estimators												
(Meth.)	PPH ₁	PPH ₂	PPM ₁	PPM ₂	PPS ₁	PPS ₂	PPC ₁	PPC ₂	LDA	DDA	OGK ₁	OGK ₂	RSIMCA
15	7.83	9.08	<u>8.46</u>	<u>8.29</u>	<u>8.04</u>	8.88	9.96	13.75	14.54	10.46	13.42	10.46	13.04
(TR)	(0.47)	(0.50)	(0.50)	(0.48)	(0.50)	(0.47)	(0.57)	(0.60)	(0.62)	(0.56)	(0.65)	(0.55)	(0.62)
75	9.42	8.46	9.42	8.96	9.25	9.92	9.58	10.62	36.79	26.00	23.46	25.08	23.42
(TR)	(0.52)	(0.47)	(0.55)	(0.55)	(0.49)	(0.52)	(0.58)	(0.55)	(1.43)	(1.02)	(1.08)	(1.18)	(0.94)
15	7.92	8.08	8.75	<u>8.75</u>	<u>8.29</u>	10.46	<u>8.21</u>	12.21	13.62	11.08	12.12	11.00	13.12
(TC)	(0.49)	(0.47)	(0.49)	(0.49)	(0.49)	(0.59)	(0.50)	(0.52)	(0.62)	(0.54)	(0.55)	(0.54)	(0.70)
75	8.88	9.71	9.50	9.25	9.17	8.79	13.33	19.04	36.58	27.12	25.79	26.54	25.75
(TC)	(0.50)	(0.49)	(0.49)	(0.49)	(0.51)	(0.53)	(0.61)	(0.71)	(1.34)	(0.89)	(1.23)	(1.00)	(0.97)
15	13.25	8.67	14.96	13.58	14.96	11.29	14.00	9.87	9.83	11.67	13.71	14.08	17.42
(RPC)	(0.64)	(0.51)	(0.71)	(0.70)	(0.66)	(0.60)	(0.65)	(0.54)	(0.53)	(0.60)	(0.81)	(0.75)	(0.81)
15	14.25	8.38	22.83	22.88	24.38	17.50	15.17	11.62	11.42	12.83	18.42	14.29	31.54
(CPC)	(0.76)	(0.54)	(1.14)	(1.07)	(1.24)	(0.97)	(0.75)	(0.61)	(0.59)	(0.63)	(0.87)	(0.73)	(1.22)
2000	15.67	12.21	29.25	29.92	33.25	33.21	14.88	13.33	12.17	18.96	15.67	28.21	38.29
(All)	(0.92)	(0.85)	(1.29)	(1.27)	(1.44)	(1.45)	(0.76)	(0.70)	(0.68)	(0.80)	(0.80)	(1.14)	(1.25)

used in our implementation of this method. Note that although RSIMCA is primarily designed to work with all the variables, because it already comprises a dimension reduction, it is not incompatible with our preliminary principal components selection of variables which is performed globally, ignoring the groups, whereas RSIMCA works with PCA within each group.

The reasons for choosing these five methods were the following: LDA has to be considered because it represents the classical counterpart of the robust PP methods under evaluation; DDA has performed well in many classification studies when there are more variables than observations (see, e.g., [34]); the OGK₁ and OGK₂ are robust plug-in versions of the previous two, therefore they are direct competitors of the proposed PP methods, moreover, among the robust multivariate location and scatter estimators proposed so far, only those can be computed in singular situations and in a reasonable time (taking into account that the estimates have to be computed thousands of times). The RSIMCA method was suggested by a referee because it is a different robust classification method (in particular, the boundaries of the classification regions are not linear) and it is also suited to cases with a large number of variables for a small or large number of observations. There is a huge number of other classification methods (like, for instance, regularized linear discriminant analysis, penalized logistic regression, nearest neighbors, classification trees, neural networks) that could have been selected. However, besides the need to keep the size of this study feasible, we must remark that the majority of those methods have not yet proved to be better, for this data set, than the methods considered [34].

In order to evaluate the performance of the 13 discrimination methods, combined with the seven variable selection strategies, it is crucial to compute unbiased, or nearly unbiased, estimates of the error rates. We have used cross-validation in the following way: the original data set (62 obs. \times 2000 variables) is randomly divided into a training set with 50 observations and a test set with 12 observations. After that, the selection of variables is performed in the training set, leading to seven training data sets. The 13 discriminant methods are then applied to each of those sets. Finally, for each of the 13×7 combinations of methods, the observations in the test set are classified, and the proportion of errors, per group and total, recorded (these are called cross-validation error rates). The training sets were also classified leading to apparent error rates. The whole procedure was repeated 200 times. Tables 3 and 4 report the mean and standard deviation of the 200 total rates observed, cross-validation and apparent, respectively. The figures within a box are the overall minimum values obtained. The figures in boldface indicate the minimum within the respective row. The figures underlined correspond to cases that are not significantly different from the overall minimum (paired Wilcoxon signed rank test with Bonferroni correction, 90 tests, for a global significance level of 5%).

The best results in Table 3 show very clearly the good performance of the PP methods, and also that selection of variables is essential to achieve that performance. However, the particular selection method is not so important. The differences between the PP methods are not so clear, but point to PPH₂, though PPH₁ may also be a good choice if one does not consider the use of principal components. The fact that PPC₁ can in some cases produce good classification results, unlike PPC₂ and LDA, which were all supposed to lead to similar estimates, is an indication that the first algorithm performs a kind of regularization of the estimates and contributes to the good performance of the PP methods.

The not so good results produced by the OGK and the RSIMCA methods came out as a surprise. For RSIMCA the explanation may be related to the fact that quadratic discrimination is not a good choice for this data set, either because there is not a large difference between the two group variances, or because of the quite different number of observations per group, a situation that is usually unfavorable to quadratic discrimination.

As expected, the apparent error rates in Table 4 are in general very optimistic, especially if one looks at the six right columns. This is a strong signal of overfitting, which clearly appears to be happening with LDA and DDA, and to a lesser extent, with OGK₁, OGK₂ and RSIMCA. On the contrary, the apparent rates are similar to the cross-validation rates (sometimes larger, sometimes smaller), for all the 7 first PP methods on the 5 top rows, which means that the overfitting effect is small, and in this sense, we may say that there is a regularization of the estimates.

Table 4

Colon data set. Apparent total error rates: mean (standard deviation) over 200 replications.

Var. (Meth.)	Estimators												
	PPH ₁	PPH ₂	PPM ₁	PPM ₂	PPS ₁	PPS ₂	PPC ₁	PPC ₂	LDA	DDA	OGK ₁	OGK ₂	RSIMCA
15 (TR)	8.62 (0.11)	9.83 (0.12)	9.64 (0.16)	9.65 (0.16)	9.23 (0.15)	10.19 (0.14)	8.96 (0.13)	4.99 (0.13)	5.88 (0.15)	9.69 (0.16)	10.00 (0.19)	9.98 (0.17)	9.88 (0.20)
75 (TR)	8.73 (0.12)	8.74 (0.12)	10.00 (0.15)	10.02 (0.15)	10.81 (0.13)	10.70 (0.13)	8.64 (0.12)	8.17 (0.12)	0.52 (0.07)	1.51 (0.12)	8.06 (0.17)	8.55 (0.18)	11.60 (0.33)
15 (TC)	8.46 (0.11)	8.18 (0.12)	9.71 (0.18)	9.64 (0.17)	9.49 (0.18)	11.16 (0.22)	8.61 (0.12)	4.32 (0.12)	5.50 (0.14)	9.60 (0.14)	9.03 (0.17)	9.25 (0.14)	9.20 (0.17)
75 (TC)	8.63 (0.12)	8.66 (0.12)	9.72 (0.16)	9.78 (0.16)	9.79 (0.17)	9.66 (0.16)	8.43 (0.14)	7.44 (0.12)	0.61 (0.07)	1.49 (0.14)	7.79 (0.16)	8.04 (0.18)	11.36 (0.26)
15 (RPC)	12.98 (0.17)	9.04 (0.13)	15.15 (0.21)	14.84 (0.21)	15.55 (0.22)	12.60 (0.22)	13.72 (0.21)	8.44 (0.13)	8.38 (0.13)	9.19 (0.17)	11.49 (0.30)	11.69 (0.29)	13.18 (0.44)
15 (CPC)	12.46 (0.15)	8.37 (0.11)	14.72 (0.21)	14.10 (0.21)	15.35 (0.23)	11.07 (0.22)	13.11 (0.18)	7.30 (0.13)	7.32 (0.13)	8.50 (0.13)	10.64 (0.20)	11.18 (0.20)	17.61 (0.53)
2000 (All)	12.17 (0.15)	8.61 (0.12)	14.73 (0.21)	14.73 (0.21)	15.58 (0.21)	13.79 (0.21)	12.99 (0.20)	7.95 (0.12)	6.47 (0.12)	0.03 (0.02)	11.08 (0.17)	12.14 (0.24)	14.43 (0.64)

Table 5

Simulation results (uncontaminated case). Actual total error rates: mean (standard deviation) over 50 simulation runs.

Var. (Meth.)	Estimators												
	PPH ₁	PPH ₂	PPM ₁	PPM ₂	PPS ₁	PPS ₂	PPC ₁	PPC ₂	LDA	DDA	OGK ₁	OGK ₂	RSIMCA
20 (TR)	12.41 (0.24)	12.89 (0.33)	14.13 (0.34)	14.26 (0.34)	14.35 (0.35)	16.11 (0.45)	11.70 (0.23)	11.75 (0.29)	11.76 (0.30)	9.28 (0.19)	13.23 (0.34)	9.50 (0.20)	15.45 (0.62)
100 (TR)	25.13 (0.70)	24.87 (0.72)	24.52 (0.72)	24.65 (0.72)	24.32 (0.70)	24.08 (0.71)	25.18 (0.69)	24.94 (0.73)	26.12 (0.61)	24.62 (0.71)	25.46 (0.58)	24.47 (0.76)	31.86 (1.57)
20 (TC)	11.14 (0.23)	11.56 (0.28)	13.09 (0.31)	13.22 (0.33)	12.80 (0.31)	14.54 (0.34)	11.17 (0.23)	10.56 (0.24)	10.59 (0.24)	8.18 (0.14)	11.74 (0.29)	8.49 (0.15)	15.17 (0.61)
100 (TC)	25.30 (0.62)	25.26 (0.60)	25.01 (0.65)	24.86 (0.66)	24.26 (0.63)	24.07 (0.61)	25.98 (0.52)	24.89 (0.55)	25.05 (0.58)	24.89 (0.68)	24.80 (0.62)	25.03 (0.71)	28.39 (1.46)
20 (RPC)	1.94 (0.09)	0.89 (0.03)	2.41 (0.15)	2.19 (0.14)	3.43 (0.27)	2.53 (0.24)	1.87 (0.08)	0.79 (0.02)	0.79 (0.02)	0.91 (0.02)	1.08 (0.06)	1.03 (0.03)	3.71 (0.50)
20 (CPC)	24.00 (0.64)	23.97 (0.71)	23.28 (0.83)	23.72 (0.93)	23.40 (1.21)	24.63 (1.28)	24.31 (0.64)	24.07 (0.65)	24.10 (0.65)	24.29 (0.69)	24.15 (0.72)	24.25 (0.73)	26.56 (1.83)
1000 (All)	26.83 (0.52)	25.69 (0.59)	27.42 (0.56)	27.03 (0.57)	28.10 (0.58)	27.19 (0.62)	26.63 (0.56)	25.03 (0.63)	27.35 (0.48)	24.30 (0.69)	30.32 (0.47)	24.81 (0.79)	30.58 (1.98)

4.4. Simulation study

Several aspects of the simulation study designed were inspired in the previous example. We have considered data sets with the following characteristics: 2 groups, with 40 observations per group, $n_1 = n_2 = 40$, and 1000 variables. The simulations were run $B = 50$ times independently for an “uncontaminated case” (C0) and a “contaminated case” (C1).

The C0 data sets were generated from two multivariate normal distributions with parameters $\mu_1 = \mathbf{0}_{1000}$, $\mu_2 = (0.8 \mathbf{1}_{100}, \mathbf{0}_{900})$ and common covariance Σ , with diagonal equal to $(1.4 \mathbf{1}_{100}, \mathbf{1}_{900})$ and all off-diagonal elements equal to zero, except those related to the covariance matrix of the first 100 variables. This matrix has 5 diagonals with non-zero elements: 1.4 on all elements of the main diagonal; values between -0.8 and 0.8 (randomly generated from a uniform distribution) on the (symmetric) elements of the two diagonals adjacent to the main diagonal; similarly with values between -0.6 and 0.6 on the next diagonals.

The C1 data sets were generated as the C0 data sets but, with probability 0.15, each row was replaced by an observation generated from multivariate normal distributions with parameters $\mu_1^c = \mu_1$, $\mu_2^c = -10\mu_2$ and $\Sigma^c = 25\Sigma$.

For each data set, the combinations of discriminant methods \times selection of variables, with $p = 20$, $p = 100$ and $q = 20$, described in the previous subsection, were applied.

Since the distribution of the data is known, there is no need to use cross-validation for error rate estimation. The actual error rates can be estimated either using Eq. (16), with $C(2|1) = C(1|2) = 1$ and $\pi_1 = \pi_2$, in the cases of linear discrimination, or using the classification results from a large independent test set generated from the model distribution. Note that the model distribution is the uncontaminated one (C0). The first approach was used, except for RSIMCA, which does not produce linear decision surfaces. The comparability between the two methods of estimating the error rate was established by a preliminary application of both to results from LDA.

Tables 5 and 6 report the mean and standard deviation, over the 50 simulation runs, of the total rates observed, for the uncontaminated and contaminated cases, respectively. The additional notation used is the one described in the previous subsection (box: overall minimum; bold: row minimum; underlined: non-significant difference to overall minimum).

Table 6

Simulation results (contaminated case). Actual total error rates: mean (standard deviation) over 50 simulation runs.

Var. (Meth.)	Estimators												
	PPH ₁	PPH ₂	PPM ₁	PPM ₂	PPS ₁	PPS ₂	PPC ₁	PPC ₂	LDA	DDA	OGK ₁	OGK ₂	RSIMCA
20	22.12	22.04	20.73	20.95	20.87	23.39	42.09	42.24	41.87	45.73	21.87	17.55	26.98
(TR)	(1.21)	(1.12)	(0.91)	(0.89)	(0.91)	(0.86)	(1.06)	(0.67)	(0.65)	(1.41)	(0.93)	(0.93)	(1.52)
100	25.12	23.52	23.90	23.62	22.87	22.56	39.06	40.74	41.08	37.02	24.64	20.63	35.75
(TR)	(1.65)	(1.63)	(1.48)	(1.48)	(1.45)	(1.47)	(0.96)	(0.84)	(0.49)	(0.52)	(0.99)	(1.61)	(2.24)
20	46.53	45.88	46.57	46.43	45.23	45.18	47.08	46.07	45.96	50.44	45.11	44.03	47.61
(TC)	(0.86)	(0.83)	(1.01)	(0.99)	(1.13)	(0.98)	(0.88)	(0.65)	(0.65)	(0.89)	(0.96)	(1.11)	(0.80)
100	44.35	42.96	43.25	42.97	41.65	41.08	45.58	44.99	43.32	42.50	39.07	38.85	51.10
(TC)	(1.04)	(0.99)	(1.26)	(1.27)	(1.17)	(1.18)	(0.70)	(0.70)	(0.56)	(0.53)	(0.93)	(1.17)	(1.64)
20	1.29	0.92	1.47	1.44	1.55	1.66	40.76	37.83	37.83	42.52	1.19	1.14	23.84
(RPC)	(0.07)	(0.03)	(0.07)	(0.09)	(0.12)	(0.14)	(1.64)	(1.48)	(1.48)	(2.12)	(0.17)	(0.06)	(2.79)
20	38.01	38.27	36.12	36.34	35.69	36.29	36.81	27.21	26.85	28.01	37.17	38.19	45.81
(CPC)	(1.89)	(2.10)	(1.90)	(2.01)	(1.90)	(1.99)	(2.19)	(1.34)	(1.26)	(1.11)	(2.15)	(2.11)	(3.21)
1000	39.71	38.37	38.74	38.50	37.80	37.06	43.48	45.01	46.34	25.47	29.95	37.55	45.09
(All)	(1.49)	(1.69)	(1.57)	(1.59)	(1.72)	(1.79)	(1.42)	(1.54)	(0.83)	(0.78)	(0.76)	(2.17)	(2.29)

The results reported in Tables 5 and 6 are quite surprising. Applying discrimination on the scores of the 20 first robust principal components (RPC20) dramatically improves the classification results. The differences are so large that it does not make sense to analyze the other lines. Fig. 4 shows boxplots of the error rates produced when the discriminant methods are applied to RPC20, in the uncontaminated (a) and contaminated (b) cases.

In the uncontaminated case the best result is obtained, as expected, by LDA (almost the same as PPC₂), followed closely by PPH₂ and DDA, but all the other methods produce good results. In the contaminated case the scenario is completely distinct. Only the robust methods (with the exception of RSIMCA) have error rates similar to the ones in the uncontaminated case. The best result comes from PPH₂, followed by OGK₂.

5. Concluding remarks

In this paper the projection-pursuit method for linear discriminant analysis is presented together with a study of relevant theoretical properties of the pertaining estimators. The ability of the robust estimators produced by the projection-pursuit approach to deal with the presence of outliers, is investigated by comparing their performance with the performance of equivalent estimators produced by other robust methods as well as the classical method of estimation for linear discriminant analysis. Both real and simulated data, with a large number of variables, were used in this exercise.

Data sets with large number of variables are a source of much disturbances that hopefully robust methods may be able to cope with. In this exercise the methods to be compared apply to a reduced number of variables that have been obtained after an operation of variable selection on the original set of variables. Several schemes, resulting from the combination of variable selection procedures and methods of various degrees of robustness, were produced. As a general conclusion one can say that, under realistic conditions for contaminated data, the projection-pursuit method performed well and is a strong competitor of the others methods in the study.

It can be argued that the present analysis includes only two groups with common variance matrices, leading simply to linear boundaries. This situation, commonly encountered in statistical methods (e.g., logistic regression), is just an apparent limitation and does not diminish the illustrated benefits of using the projection-pursuit method of estimation in discriminant analysis. The application of the method to more than two groups and different dispersions is a matter for future research.

Acknowledgments

The authors would like to thank the two referees whose critical but constructive remarks and useful suggestions have greatly improved the contents of this paper. This research was partially supported by the project grant ERA-PG/0002/2006 and by the Center for Mathematics and its Applications, Lisbon, Portugal, through Programa Operacional “Ciência, Tecnologia, Inovação” (POCTI) of the Fundação para a Ciência e a Tecnologia (FCT), cofinanced by the European Community fund FEDER.

Appendix. Proofs

Proof of Proposition 1. If $\hat{\alpha}_N$ is the PP estimate of α_N based on the untransformed samples, it means that for all $\mathbf{a} \in R^m$, such that $\|\mathbf{a}\| = 1$

$$\frac{[T(\hat{\alpha}_N^T \mathbf{x}_1) - T(\hat{\alpha}_N^T \mathbf{x}_2)]^2}{a_1 S^2(\hat{\alpha}_N^T \mathbf{x}_1) + a_2 S^2(\hat{\alpha}_N^T \mathbf{x}_2)} \geq \frac{[T(\mathbf{a}^T \mathbf{x}_1) - T(\mathbf{a}^T \mathbf{x}_2)]^2}{a_1 S^2(\mathbf{a}^T \mathbf{x}_1) + a_2 S^2(\mathbf{a}^T \mathbf{x}_2)}. \quad (39)$$

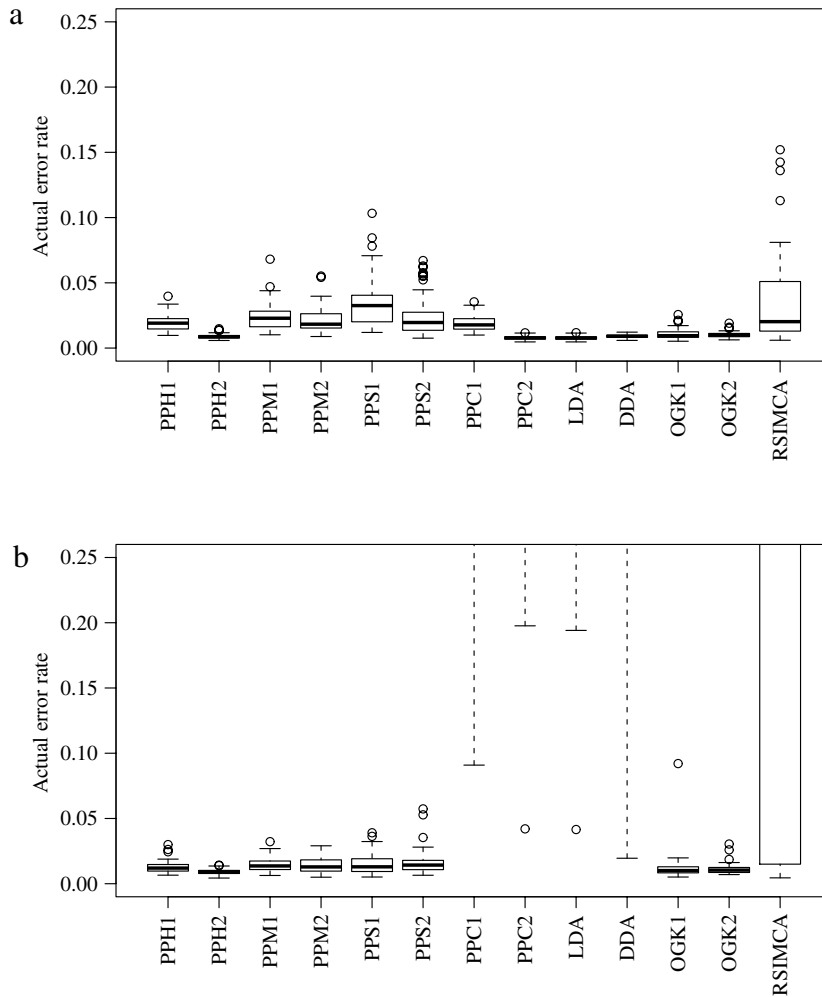


Fig. 4. Boxplots of the simulation results. Error rates obtained when using the first 20 robust principal components: (a) uncontaminated case, (b) contaminated case.

Because of the non-singularity of \mathbf{A} , each \mathbf{x}_i can be premultiplied by $\mathbf{A}^{-1}\mathbf{A}$. Because of the location equivariance of T and the location invariance of S , $\hat{\boldsymbol{\alpha}}_N^T \mathbf{A}^{-1} \mathbf{b}$ can be added to all the arguments of T and S on the left side of the inequality and $\mathbf{a}^T \mathbf{A}^{-1} \mathbf{b}$ to all the arguments on the right side. And because of the scale equivariance of T and S all the arguments of T and S can be divided by $\|(\mathbf{A}^{-1})^T \hat{\boldsymbol{\alpha}}_N\|$ on the left side and by $\|(\mathbf{A}^{-1})^T \mathbf{a}\|$ on the right side. Then (39) is equivalent to

$$\frac{\left[T\left(\frac{\hat{\boldsymbol{\alpha}}_N^T \mathbf{A}^{-1}}{\|(\mathbf{A}^{-1})^T \hat{\boldsymbol{\alpha}}_N\|} \mathbf{x}_1^* \right) - T\left(\frac{\hat{\boldsymbol{\alpha}}_N^T \mathbf{A}^{-1}}{\|(\mathbf{A}^{-1})^T \hat{\boldsymbol{\alpha}}_N\|} \mathbf{x}_2^* \right) \right]^2}{a_1 S^2\left(\frac{\hat{\boldsymbol{\alpha}}_N^T \mathbf{A}^{-1}}{\|(\mathbf{A}^{-1})^T \hat{\boldsymbol{\alpha}}_N\|} \mathbf{x}_1^* \right) + a_2 S^2\left(\frac{\hat{\boldsymbol{\alpha}}_N^T \mathbf{A}^{-1}}{\|(\mathbf{A}^{-1})^T \hat{\boldsymbol{\alpha}}_N\|} \mathbf{x}_2^* \right)} \geq \frac{[T(\mathbf{a}_*^T \mathbf{x}_1^*) - T(\mathbf{a}_*^T \mathbf{x}_2^*)]^2}{a_1 S^2(\mathbf{a}_*^T \mathbf{x}_1^*) + a_2 S^2(\mathbf{a}_*^T \mathbf{x}_2^*)}, \quad (40)$$

for all $\mathbf{a} \in R^m$, such that $\|\mathbf{a}\| = 1$, and $\mathbf{a}_* = (\mathbf{A}^{-1})^T \mathbf{a} / \|(\mathbf{A}^{-1})^T \mathbf{a}\|$, but as there is a one to one linear relation between \mathbf{a} and \mathbf{a}_* , Eq. (40) is valid for all $\mathbf{a}_* \in R^m$, such that $\|\mathbf{a}_*\| = 1$, which shows that (23) is true. Finally, as a consequence of (23) and of the scale equivariance of T and S it is easy to verify that (24), (25) and (26) are also valid, thus completing the proof. \square

Proof of Proposition 2. By (23) $\hat{\boldsymbol{\alpha}}_N = \mathbf{A}^T \hat{\boldsymbol{\alpha}}_N^* / \|\mathbf{A}^T \hat{\boldsymbol{\alpha}}_N^*\|$, therefore, applying the usual differentiation rules,

$$\begin{aligned} PIF_i(\mathbf{x}; \hat{\boldsymbol{\alpha}}_N, \mathbf{F}) &= PIF_i(\mathbf{x}^*; \mathbf{A}^T \hat{\boldsymbol{\alpha}}_N / \|\mathbf{A}^T \hat{\boldsymbol{\alpha}}_N\|, \mathbf{F}^*) \\ &= \frac{\mathbf{A}^T PIF_i(\mathbf{x}^*; \hat{\boldsymbol{\alpha}}_N^*, \mathbf{F}^*)}{\|\mathbf{A}^T \hat{\boldsymbol{\alpha}}_N^*\|} - \frac{\mathbf{A}^T \hat{\boldsymbol{\alpha}}_N^* PIF_i(\mathbf{x}^*; \|\mathbf{A}^T \hat{\boldsymbol{\alpha}}_N\|, \mathbf{F}^*)}{\|\mathbf{A}^T \hat{\boldsymbol{\alpha}}_N^*\|^2}, \end{aligned}$$

but

$$PIF_i(\mathbf{x}^*; \|\mathbf{A}^T \hat{\boldsymbol{\alpha}}_N\|, \mathbf{F}^*) = PIF_i(\mathbf{x}^*; \sqrt{\hat{\boldsymbol{\alpha}}_N^T \mathbf{A} \mathbf{A}^T \hat{\boldsymbol{\alpha}}_N}, \mathbf{F}^*)$$

$$= \frac{PIF_i(\mathbf{x}^*; \hat{\alpha}_N^T \mathbf{A} \mathbf{A}^T \hat{\alpha}_N, \mathbf{F}^*)}{2 \|\mathbf{A}^T \hat{\alpha}_N^*\|} = \alpha_N^T \mathbf{A}^T PIF_i(\mathbf{x}^*; \hat{\alpha}_N, \mathbf{F}^*),$$

which then yields (27) (after noting that $\mathbf{A}^T \alpha_N^* = \alpha_N \|\mathbf{A}^T \alpha_N^*\|$ and that $\|\mathbf{A}^T \alpha_N^*\| = \|\alpha\|/\Delta$, because $\mathbf{U}^T \alpha_N^* = \Sigma^{-1/2}(\mu_1 - \mu_2)/\Delta$).

By (24), $\|\widehat{\alpha}\| = \|\widehat{\alpha}\|^* \|\mathbf{A}^T \hat{\alpha}_N^*\|$, therefore

$$PIF_i(\mathbf{x}; \|\widehat{\alpha}\|, \mathbf{F}) = \|\mathbf{A}^T \alpha_N^*\| PIF_i(\mathbf{x}^*; \|\widehat{\alpha}\|, \mathbf{F}^*) + \|\alpha\|^* PIF_i(\mathbf{x}^*; \|\mathbf{A}^T \hat{\alpha}_N\|, \mathbf{F}^*),$$

which is immediately seen to be equivalent to (28). By (25) $\hat{\alpha}_{0N} = (\hat{\alpha}_{0N}^* + \mathbf{b}^T \hat{\alpha}_N^*)/\|\mathbf{A}^T \hat{\alpha}_N^*\|$, therefore

$$PIF_i(\mathbf{x}; \hat{\alpha}_{0N}, \mathbf{F}) = \frac{PIF_i(\mathbf{x}^*; \hat{\alpha}_{0N}, \mathbf{F}^*) + \mathbf{b}^T PIF_i(\mathbf{x}^*; \hat{\alpha}_N, \mathbf{F}^*)}{\|\mathbf{A}^T \alpha_N^*\|} - \frac{(\alpha_{0N}^* + \mathbf{b}^T \alpha_N^*) PIF_i(\mathbf{x}^*; \|\mathbf{A}^T \hat{\alpha}_N\|, \mathbf{F}^*)}{\|\mathbf{A}^T \alpha_N\|^2},$$

which is equivalent to (29), by noting that $\alpha_{0N}^* = -\log k/\Delta$ and that $\mathbf{b}^T \alpha_N^* = -(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)/\Delta$. Finally the relations in (30) are straightforward. \square

Proof of Theorem 1. Let \mathbf{F}_ε^* , with $\varepsilon = (\varepsilon_1, \varepsilon_2)$ denote the ε -contaminated distribution $(F_{1,\varepsilon_1}^*, F_{2,\varepsilon_2}^*) = [(1 - \varepsilon_1)F_1^* + \varepsilon_1 \Delta_{\mathbf{x}}, (1 - \varepsilon_2)F_2^* + \varepsilon_2 \Delta_{\mathbf{x}}]$, then

$$PIF_i(\mathbf{x}; \hat{\alpha}_N, \mathbf{F}^*) = \left. \frac{\partial}{\partial \varepsilon_i} \alpha_\varepsilon \right|_{\varepsilon=0},$$

where $\alpha_\varepsilon \equiv \alpha_N(\mathbf{F}_\varepsilon^*)$ must verify Eq. (18). A simple way to incorporate the restriction of unit norm of the solution is to work in a system of spherical coordinates. In this system a vector $\alpha_{(m)} \in R^m$ with $\|\alpha_{(m)}\| = 1$ is represented by an $(m - 1)$ -dimensional vector $\theta_{(m)} = (\theta_1, \dots, \theta_{m-1}) \in R^{m-1}$, such that

$$\alpha_{(2)}(\theta_{(2)}) = \begin{bmatrix} \cos \theta_1 \\ \sin \theta_1 \end{bmatrix} \quad \alpha_{(m)}(\theta_{(m)}) = \begin{bmatrix} \sin \theta_{m-1} \sin \theta_{m-2} \cdots \cos \theta_1 \\ \sin \theta_{m-1} \sin \theta_{m-2} \cdots \sin \theta_1 \\ \sin \theta_{m-1} \sin \theta_{m-2} \cdots \cos \theta_2 \\ \sin \theta_{m-1} \sin \theta_{m-2} \cdots \cos \theta_3 \\ \vdots \\ \sin \theta_{m-1} \cos \theta_{m-2} \\ \cos \theta_{m-1} \end{bmatrix} \quad (m \geq 3). \quad (41)$$

The matrix of partial derivatives, $\partial \alpha_{(m)}/\partial \theta_{(m)}$ with dimensions $m \times (m - 1)$ is given by the following recurrence formulae

$$\begin{aligned} \frac{\partial \alpha_{(2)}}{\partial \theta_{(2)}} &= \begin{bmatrix} -\sin \theta_1 \\ \cos \theta_1 \end{bmatrix}, \\ \frac{\partial \alpha_{(m)}}{\partial \theta_{(m)}} &= \begin{bmatrix} \left(\frac{\partial \alpha_{(m-1)}}{\partial \theta_{(m-1)}} \right) \sin \theta_{m-1} & \alpha_{(m-1)} \cos \theta_{m-1} \\ \mathbf{0} & -\sin \theta_{m-1} \end{bmatrix}. \end{aligned} \quad (42)$$

A solution for $\theta_\varepsilon \in R^{m-1}$ can be obtained without imposing restrictions and after that converted to $\alpha_\varepsilon \in R^m$ having automatically unit norm. Eq. (18) can be rewritten as

$$\begin{aligned} \theta_\varepsilon &= \underset{\theta \in R^{m-1}}{\operatorname{argmax}} \frac{[T(F_{1,\varepsilon_1}^{\alpha(\theta)}) - T(F_{2,\varepsilon_2}^{\alpha(\theta)})]^2}{a_1 S^2(F_{1,\varepsilon_1}^{\alpha(\theta)}) + a_2 S^2(F_{2,\varepsilon_2}^{\alpha(\theta)})} \\ &= \underset{\theta \in R^{m-1}}{\operatorname{argmax}} \frac{[T_{1,\varepsilon_1}(\alpha(\theta)) - T_{2,\varepsilon_2}(\alpha(\theta))]^2}{a_1 S_{1,\varepsilon_1}^2(\alpha(\theta)) + a_2 S_{2,\varepsilon_2}^2(\alpha(\theta))} = \underset{\theta \in R^{m-1}}{\operatorname{argmax}} I(\alpha(\theta)). \end{aligned}$$

Because of the regularity conditions assumed the solution verifies

$$\left. \frac{\partial I(\alpha(\theta))}{\partial \theta} \right|_{\theta=\theta_\varepsilon} = 0 \quad (43)$$

which is equivalent ($\alpha(\theta_\varepsilon) \equiv \alpha_\varepsilon$) to

$$2 [T_{1,\varepsilon_1}(\alpha_\varepsilon) - T_{2,\varepsilon_2}(\alpha_\varepsilon)] \begin{bmatrix} A_1(\alpha_\varepsilon) & A_2(\alpha_\varepsilon) \end{bmatrix} \frac{\partial \alpha(\theta)}{\partial \theta} \bigg|_{\theta=\theta_\varepsilon} = A_3(\alpha_\varepsilon) \begin{bmatrix} A_4(\alpha_\varepsilon) & [T_{1,\varepsilon_1}(\alpha_\varepsilon) - T_{2,\varepsilon_2}(\alpha_\varepsilon)] \end{bmatrix} \frac{\partial \alpha(\theta)}{\partial \theta} \bigg|_{\theta=\theta_\varepsilon}, \quad (44)$$

with

$$\begin{aligned} A_1(\alpha_\varepsilon) &= 2 \left[\frac{\partial T_{1,\varepsilon_1}(\alpha)}{\partial \alpha} - \frac{\partial T_{2,\varepsilon_2}(\alpha)}{\partial \alpha} \right] \Big|_{\alpha=\alpha_\varepsilon}, \\ A_2(\alpha_\varepsilon) &= a_1 S_{1,\varepsilon_1}^2(\alpha_\varepsilon) + a_2 S_{2,\varepsilon_2}^2(\alpha_\varepsilon), \\ A_3(\alpha_\varepsilon) &= \left[a_1 \frac{\partial S_{1,\varepsilon_1}^2(\alpha)}{\partial \alpha} + a_2 \frac{\partial S_{2,\varepsilon_2}^2(\alpha)}{\partial \alpha} \right] \Big|_{\alpha=\alpha_\varepsilon}, \end{aligned}$$

and

$$A_4(\alpha_\varepsilon) = T_{1,\varepsilon_1}(\alpha_\varepsilon) - T_{2,\varepsilon_2}(\alpha_\varepsilon).$$

The trivial solution of Eq. (44), $\alpha_\varepsilon : T_{1,\varepsilon_1}(\alpha_\varepsilon) = T_{2,\varepsilon_2}(\alpha_\varepsilon)$ is not interesting, since it minimizes $I(\theta)$ (instead of maximizing). Because $\mu_1^* \neq \mu_2^*$, there is at least one direction α_ε for which $T_{1,\varepsilon_1}(\alpha_\varepsilon) \neq T_{2,\varepsilon_2}(\alpha_\varepsilon)$, therefore both members of Eq. (44) can be divided by $[T_{1,\varepsilon_1}(\alpha_\varepsilon) - T_{2,\varepsilon_2}(\alpha_\varepsilon)]$. The next step is to differentiate the resulting equation with respect to ε_i and to set $\varepsilon = \mathbf{0}$, which gives

$$\begin{aligned} & \frac{\partial A_1(\alpha_\varepsilon)}{\partial \varepsilon_i} \Big|_{\varepsilon=\mathbf{0}} A_2(\alpha_0) \frac{\partial \alpha(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} + A_1(\alpha_0) \frac{\partial A_2(\alpha_\varepsilon)}{\partial \varepsilon_i} \Big|_{\varepsilon=\mathbf{0}} \frac{\partial \alpha(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \\ & + A_1(\alpha_0) A_2(\alpha_0) \frac{\partial}{\partial \varepsilon_i} \left(\frac{\partial \alpha(\theta)}{\partial \theta} \Big|_{\theta=\theta_\varepsilon} \right) \Big|_{\varepsilon=\mathbf{0}} = \frac{\partial A_3(\alpha_\varepsilon)}{\partial \varepsilon_i} \Big|_{\varepsilon=\mathbf{0}} A_4(\alpha_0) + \frac{\partial \alpha(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \\ & + A_3(\alpha_0) \frac{\partial A_4(\alpha_\varepsilon)}{\partial \varepsilon_i} \Big|_{\varepsilon=\mathbf{0}} \frac{\partial \alpha(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} + A_3(\alpha_0) A_4(\alpha_0) \frac{\partial}{\partial \varepsilon_i} \left(\frac{\partial \alpha(\theta)}{\partial \theta} \Big|_{\theta=\theta_\varepsilon} \right) \Big|_{\varepsilon=\mathbf{0}}. \end{aligned} \quad (45)$$

At the “central” distribution under consideration, \mathbf{F}^* ,

$$\begin{aligned} \alpha_0 &= (1, 0, \dots, 0)^T \quad \theta_0 = \left(0, \frac{\pi}{2}, \dots, \frac{\pi}{2}\right)^T, \\ \frac{\partial \alpha(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} &= \begin{bmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -1 \end{bmatrix} \\ \text{PIF}_i(\mathbf{x}; \hat{\alpha}, \mathbf{F}^*) &= \frac{\partial \alpha_\varepsilon}{\partial \varepsilon_i} \Big|_{\varepsilon=\mathbf{0}} = \frac{\partial \alpha_\varepsilon}{\partial \theta_\varepsilon} \Big|_{\theta=\theta_0} \frac{\partial \theta_\varepsilon}{\partial \varepsilon_i} \Big|_{\varepsilon=\mathbf{0}} = \frac{\partial \alpha_\varepsilon}{\partial \theta_\varepsilon} \Big|_{\theta=\theta_0} \text{PIF}_i(\mathbf{x}; \hat{\theta}, \mathbf{F}^*), \end{aligned}$$

or

$$\text{PIF}_i(\mathbf{x}; \hat{\alpha}, \mathbf{F}^*) = \begin{bmatrix} 0 \\ \text{PIF}_i(\mathbf{x}; \hat{\theta}_1, \mathbf{F}^*) \\ -\text{PIF}_i(\mathbf{x}; \hat{\theta}_2, \mathbf{F}^*) \\ \dots \\ -\text{PIF}_i(\mathbf{x}; \hat{\theta}_{m-1}, \mathbf{F}^*) \end{bmatrix}^T \quad (46)$$

Let $IF(z; T, F_0)$ and $IF(z; S, F_0)$ denote the influence functions of T and S at F_0 . When F_i^* is ε -contaminated at \mathbf{x} then $F_i^{*\alpha_\varepsilon}$ is ε -contaminated at $\alpha_\varepsilon^T \mathbf{x}$, and using

$$T(F_\varepsilon) = T(F) + \varepsilon IF(\mathbf{x}; T, F) + R(\varepsilon), \quad \text{with } R(0) = 0 \quad \text{and} \quad R'(0) = 0,$$

then

$$\begin{aligned} \frac{A_1(\alpha_\varepsilon)}{2} &= \left[\frac{\partial T_{1,\varepsilon_1}(\alpha)}{\partial \alpha} - \frac{\partial T_{2,\varepsilon_2}(\alpha)}{\partial \alpha} \right] \Big|_{\alpha=\alpha_\varepsilon} \\ &= \frac{\partial}{\partial \alpha} \left[T_1(\alpha) + \varepsilon_1 \sqrt{\alpha^T \alpha} IF \left(\frac{\alpha^T (\mathbf{x} - \mu_1^*)}{\sqrt{\alpha^T \alpha}}; T, F_0 \right) + R(\varepsilon_1) \right. \\ &\quad \left. - T_2(\alpha) - \varepsilon_2 \sqrt{\alpha^T \alpha} IF \left(\frac{\alpha^T (\mathbf{x} - \mu_2^*)}{\sqrt{\alpha^T \alpha}}; T, F_0 \right) - R(\varepsilon_2) \right] \Big|_{\alpha=\alpha_\varepsilon}. \end{aligned}$$

Then, at $\varepsilon = \mathbf{0}$, and because $T_i(\alpha) = \alpha^T \mu_i^*$, we have $A_1(\alpha_0) = 2(\Delta, 0, \dots, 0)$ and, after some algebra

$$\left. \frac{\partial A_1(\alpha_\varepsilon)}{\partial \varepsilon_i} \right|_{\varepsilon=0} = (-1)^{i+1} [\alpha_0 IF(x_1 - \mu_{i1}^*; T, F_0) + (0, x_2, \dots, x_m) IF'(x_1 - \mu_{i1}^*; T, F_0)],$$

with $IF'(z; T, F_0) = \frac{d}{dz} IF(z; T, F_0)$. Proceeding in a similar way (details omitted), the following results are obtained: $A_2(\alpha_0) = 1, A_3(\alpha_0) = [2, 0, \dots, 0], A_4(\alpha_0) = \Delta$ and

$$\left. \frac{\partial A_2(\alpha_\varepsilon)}{\partial \varepsilon_i} \right|_{\varepsilon=0} = 2a_i IF(x_1 - \mu_{i1}^*; S, F_0),$$

$$\left. \frac{\partial}{\partial \varepsilon_i} \left(\frac{\partial \alpha}{\partial \theta} \right) \right|_{\theta=\theta_\varepsilon, \varepsilon=0} = \sum_{j=1}^{m-1} \left. \frac{\partial}{\partial \theta_j} \left(\frac{\partial \alpha}{\partial \theta} \right) \right|_{\varepsilon=0} \left. \frac{\partial \theta_j}{\partial \varepsilon_i} \right|_{\varepsilon=0} = \begin{bmatrix} -1 & 0 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -1 \\ 0 & 0 & \dots & 0 \end{bmatrix} PIF_i(\mathbf{x}; \hat{\theta}, \mathbf{F}^*),$$

$$\left. \frac{\partial A_3(\alpha_\varepsilon)}{\partial \varepsilon_i} \right|_{\varepsilon=0} = 2 \left[0, PIF_i(\mathbf{x}; \hat{\theta}_1, \mathbf{F}^*), PIF_i(\mathbf{x}; \hat{\theta}_2, \mathbf{F}^*), \dots, PIF_i(\mathbf{x}; \hat{\theta}_{m-1}, \mathbf{F}^*) \right] \\ + 2\alpha_0 IF(x_1 - \mu_{i1}^*; S, F_0) + (0, x_2, \dots, x_m) IF'(x_1 - \mu_{i1}^*; S, F_0),$$

with $IF'(z; S, F_0) = \frac{d}{dz} IF(z; S, F_0)$,

$$\left. \frac{\partial A_4(\alpha_\varepsilon)}{\partial \varepsilon_i} \right|_{\varepsilon=0} = (-1)^{i+1} IF(x_1 - \mu_{i1}^*; T, F_0).$$

Finally, putting everything back into Eq. (45), solving for $PIF_i(\mathbf{x}; \hat{\theta}, \mathbf{F}^*)$, and using (46), the result in (31) is obtained. The other results in Theorem 1 follow easily. The case $\hat{\Delta}$ is exemplified below.

$$PIF_i(\mathbf{x}; \hat{\Delta}, \mathbf{F}^*) = \frac{1}{2\Delta} PIF_i(\mathbf{x}; \hat{\Delta}^2, \mathbf{F}^*) = \frac{1}{2\Delta} \left. \frac{\partial A_4^2(\alpha_\varepsilon)}{\partial \varepsilon_i A_2(\alpha_\varepsilon)} \right|_{\varepsilon=0} \\ = \left. \frac{2A_4(\alpha_\varepsilon)A_2(\alpha_\varepsilon) \frac{\partial A_4(\alpha_\varepsilon)}{\partial \varepsilon_i} - A_4^2(\alpha_\varepsilon) \frac{\partial A_2(\alpha_\varepsilon)}{\partial \varepsilon_i}}{2\Delta A_2^2(\alpha_\varepsilon)} \right|_{\varepsilon=0}. \quad \square$$

Proof of Theorem 2. Asymptotic normality is an immediate consequence of the assumptions. So is the application of (36) and (37). Then, using (31),

$$V_i(\hat{\alpha}_N, \mathbf{F}^*) = \int (\mathbf{I}_m - \alpha_N^* \alpha_N^{*T}) \mathbf{x} \mathbf{x}^T (\mathbf{I}_m - \alpha_N^* \alpha_N^{*T}) K_i(x_1) dF_i^*(\mathbf{x}) \\ = \int \mathbf{A}(\mathbf{x}) K_i(x_1) dF_i^*(\mathbf{x}),$$

where

$$\mathbf{A}(\mathbf{x}) = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & x_2^2 & x_2 x_3 & \dots & x_2 x_m \\ \dots & \dots & \dots & \dots & \dots \\ 0 & x_2 x_m & x_3 x_m & \dots & x_m^2 \end{bmatrix}$$

and

$$K_i(x_1) = \left(\frac{(-1)^{i+1} IF'_T(x_1)}{\Delta} - a_i IF'_S(x_1) \right)^2,$$

where $IF'_T(x) = IF'(x - \mu_{i1}^*; T, F_0)$ and $IF'_S(x) = IF'(x - \mu_{i1}^*; S, F_0)$. Under the “central” model, $x_i, i = 2, \dots, m$, are identically distributed, therefore

$$V_i(\hat{\alpha}_N, \mathbf{F}^*) = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & C_i & D_i & \dots & D_i \\ \dots & \dots & \dots & \dots & \dots \\ 0 & D_i & D_i & \dots & C_i \end{bmatrix}$$

but, by symmetry, $D_i = \int x_2 x_3 K(x_1) dF_i^*(\mathbf{x}) = 0$, and

$$C_i = \int x_2^2 \left\{ \frac{[IF'_T(x_1)]^2}{\Delta^2} - \frac{2a_i(-1)^{i+1}}{\Delta} IF'_T(x_1) IF'_S(x_1) + a_i^2 [IF'_S(x_1)]^2 \right\} dF_i^*(\mathbf{x}).$$

The term involving $IF'_T(x_1) IF'_S(x_1)$ vanishes due to symmetry reasons and so

$$C_i = \frac{1}{\Delta^2} \int x_2^2 [IF'_T(x_1)]^2 dF_i^*(\mathbf{x}) + a_i^2 \int x_2^2 [IF'_S(x_1)]^2 dF_i^*(\mathbf{x}) = \frac{C_T}{\Delta^2} + a_i^2 C_S.$$

Finally, inserting this result into (36) leads to (38). \square

References

- [1] S.W. Ahmed, P.A. Lachenbruch, Discriminant analysis when scale contamination is present in the initial sample, in: J. Van Rysin (Ed.), *Classification and Clustering*, Academic Press, New York, 1977, pp. 331–353.
- [2] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Science* 96 (1999) 6745–6750.
- [3] B. Broffitt, W.R. Clarke, P.A. Lachenbruch, The effect of huberizing and trimming on the quadratic discriminant function, *Communications in Statistics – Theory and Methods* 9 (1980) 13–25.
- [4] A.-L. Boulesteix, C. Strobl, T. Augustin, M. Daumer, Evaluating microarray-based classifiers: an overview, *Cancer Informatics* 6 (2008) 77–97.
- [5] N.A. Campbell, Robust procedures in multivariate analysis II: robust canonical variate analysis, *Applied Statistics* 31 (1982) 1–8.
- [6] Z.Y. Chen, R.J. Muirhead, A comparison of robust linear discriminant procedures using projection-pursuit methods, in: T.W. Anderson, K.T. Fang, I. Olkin (Eds.), *Multivariate Analysis and its Applications*, IMS Monograph Series, Hayward, 1994, pp. 163–176.
- [7] C.Y. Chork, P.J. Rousseeuw, Integrating a high-breakdown option into discriminant analysis in exploration geochemistry, *Journal of Geochemical Exploration* 43 (1992) 191–203.
- [8] W.R. Clarke, P.A. Lachenbruch, B. Broffitt, How non-normality affects the quadratic discriminant function, *Communications in Statistics* 8 (1979) 1285–1301.
- [9] R.A. Cooper, T.J. Weekes, *Data, Models, and Statistical Analysis*, Barnes and Noble, New York, 1983.
- [10] C. Croux, C. Dehon, Robust linear discriminant analysis using S-estimators, *Canadian Journal of Statistics* 29 (2001) 473–492.
- [11] C. Croux, P. Filzmoser, K. Joossens, Classification efficiencies for robust linear discriminant analysis, *Statistica Sinica* 18 (2008) 581–599.
- [12] C. Croux, K. Joossens, Empirical comparison of the classification performance of robust linear and quadratic discriminant analysis, in: M. Hubert, G. Pison, A. Struyf, S. Van Aelst (Eds.), *Theory and Applications of Recent Robust Methods*, Birkhäuser, Basel, 2004, pp. 131–140.
- [13] C. Croux, K. Joossens, Influence of observations on the misclassification probability in quadratic discriminant analysis, *Journal of Multivariate Analysis* 96 (2005) 384–403.
- [14] C. Croux, A. Ruiz-Gazen, A fast algorithm for robust principal components based on projection-pursuit, in: A. Prat (Ed.), *Compstat 1996: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, 1996, pp. 211–216.
- [15] C. Croux, A. Ruiz-Gazen, High breakdown estimators for principal components: The projection-pursuit approach revisited, *Journal of Multivariate Analysis* 95 (2005) 206–226.
- [16] P.L. Davies, Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices, *The Annals of Statistics* 15 (1987) 1269–1292.
- [17] M. Daszykowski, K. Kaczmarek, I. Stanimirova, Y. Vander Heyden, B. Walczak, Robust SIMCA-bounding influence of outliers, *Chemometrics and Intelligent Laboratory Systems* 87 (2007) 95–103.
- [18] P. Filzmoser, K. Joossens, C. Croux, Multiple group linear discriminant analysis: robustness and error rate, in: A. Rizzi, M. Vichi (Eds.), *Compstat 2006: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, 2006, pp. 521–532.
- [19] P. Filzmoser, S. Serneels, R. Maronna, P.J. Van Espen, Robust multivariate methods in chemometrics, in: B. Walczak, R.T. Tauler, S. Brown (Eds.), *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Elsevier, Amsterdam, 2009, pp. 681–722.
- [20] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics* 7 (1936) 179–188.
- [21] J.H. Friedman, J.W. Tukey, A projection-pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers* C-23 (1974) 881–890.
- [22] F.R. Hampel, The influence curve and its role in robust estimation, *Journal of the American Statistical Association* 69 (1974) 383–393.
- [23] P.Y. Han, A.T.B. Jin, Random projection with robust linear discriminant analysis model in face recognition, in: *Proceedings in Computer Graphics, Imaging and Visualization*, IEEE, 2007, pp. 11–15.
- [24] D.M. Hawkins, G.J. McLachlan, High-breakdown linear discriminant analysis, *Journal of the American Statistical Association* 92 (1997) 136–143.
- [25] X. He, W.K. Fung, High breakdown estimation for multiple populations with applications to discriminant analysis, *Journal of Multivariate Analysis* 72 (2000) 151–162.
- [26] P.J. Huber, Robust estimation of a location parameter, *Annals of Mathematical Statistics* 35 (1964) 73–101.
- [27] P.J. Huber, Projection-pursuit (with discussion), *Annals of Statistics* 13 (1985) 435–475.
- [28] P.J. Huber, Projection-pursuit and robustness, in: S. Morgenthaler, E. Ronchetti, W.A. Stahel (Eds.), *New Directions in Statistical Data Analysis and Robustness*, Birkhäuser, Basel, 1993, pp. 139–146.
- [29] M. Hubert, P.J. Rousseeuw, S. Van Aelst, High-breakdown robust multivariate methods, *Statistical Science* 23 (2008) 92–119.
- [30] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, ROBPCA: a new approach to robust principal component analysis, *Technometrics* 47 (2005) 64–79.
- [31] M. Hubert, K. Van Driessen, Fast and robust discriminant analysis, *Computational Statistics and Data Analysis* 45 (2004) 301–320.
- [32] M.C. Jones, R. Sibson, What is projection-pursuit? (with discussion), *Journal of the Royal Statistical Society A* 150 (1987) 1–36.
- [33] P.A. Lachenbruch, C. Sneeringer, L.T. Revo, Robustness of the linear and quadratic discriminant function to certain types of non-normality, *Communications in Statistics* 1 (1973) 39–56.
- [34] J. Lee, J. Lee, M. Park, S. Song, An extensive comparison of recent classification tools applied to microarray data, *Computational Statistics and Data Analysis* 48 (2005) 869–885.
- [35] G. Li, Z. Chen, Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo, *Journal of the American Statistical Association* 80 (1985) 759–766.
- [36] R. Maronna, R. Zamar, Robust estimation of location and dispersion for high-dimensional datasets, *Technometrics* 44 (2002) 307–317.
- [37] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.
- [38] J.A. Nelder, R. Mead, A simplex algorithm for function minimization, *Computer Journal* 7 (1965) 308–313.
- [39] A.M. Pires, J.A. Branco, Partial influence functions, *Journal of Multivariate Analysis* 83 (2002) 451–468.
- [40] A.M. Pires, Robust discriminant analysis and the projection-pursuit approach: Practical aspects, in: R. Dutta, P. Filzmoser, U. Gather, P.J. Rousseeuw (Eds.), *Developments in Robust Statistics*, Springer-Verlag, Heidelberg, 2003, pp. 317–329.
- [41] C. Posse, Projection-pursuit discriminant analysis for two groups, *Communications in Statistics – Theory and Methods* 21 (1992) 1–19.

- [42] R Development Core Team, R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>, 2009.
- [43] R.H. Randles, J.D. Broffitt, J.S. Ramberg, R.V. Hogg, Generalized linear and quadratic discriminant functions using robust estimates, *Journal of the American Statistical Association* 73 (1978) 564–568.
- [44] D.M. Rocke, Robustness properties of S-estimators of multivariate location and shape in high dimension, *Annals of Statistics* 24 (1996) 1327–1345.
- [45] P.J. Rousseeuw, Multivariate estimation with high breakdown point, in: W. Grossman, et al. (Eds.), *Mathematical Statistics and Applications*, vol. B, Reidel Publishing, Dordrecht, 1985, pp. 283–297.
- [46] P.J. Rousseeuw, A.M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.
- [47] P.J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, *Technometrics* 41 (1998) 212–223.
- [48] P.J. Rousseeuw, V.J. Yohai, Robust regression by means of S-estimators, in: J. Franke, et al. (Eds.), *Robust and Nonlinear Time Series Analysis*, in: *Lecture Notes in Statistics*, vol. 26, Springer Verlag, New York, 1984, pp. 256–272.
- [49] W.B. Stern, J.-P. Descouedres, X-ray fluorescence analysis of archaic greek pottery, *Archaeometry* 19 (1977) 73–86.
- [50] M.L. Tiku, N. Balakrishnan, Robust multivariate classification procedures based on the MML estimators, *Communications in Statistics – Theory and Methods* 13 (1984) 967–986.
- [51] V. Todorov, N. Neykov, P. Neytchev, Robust two-group discrimination by bounded influence regression, A Monte-Carlo simulation, *Computational Statistics and Data Analysis* 17 (1994) 289–302.
- [52] V. Todorov, A.M. Pires, Comparative performance of several robust linear discriminant analysis methods, *REVSTAT, Statistical Journal* 5 (2007) 63–83.
- [53] J.W. Van Ness, J.J. Yang, Robust discriminant analysis: training data breakdown point, *Journal of Statistical Planning and Inference* 67 (1998) 67–83.
- [54] K. Vanden Branden, M. Hubert, Robust classification in high dimensions based on the SIMCA method, *Chemometrics and Intelligent Laboratory Systems* 79 (2005) 10–21.
- [55] A. Wald, Contributions to the theory of statistical estimation and testing hypothesis, *Annals of Mathematical Statistics* 10 (1939) 299–326.
- [56] B.L. Welch, Note on discriminant functions, *Biometrika* 31 (1939) 218–220.
- [57] S. Wold, Pattern recognition by means of disjoint principal components models, *Pattern Recognition* 8 (1976) 127–139.
- [58] P. Xu, G.N. Brock, R.S. Parrish, Modified linear discriminant analysis approaches for classification of high-dimensional microarray data, *Computational Statistics and Data Analysis* 53 (2009) 1674–1687.