Robust Linear Discriminant Analysis and the Projection Pursuit Approach

Practical aspects

A. M. Pires[1]

Department of Mathematics and Applied Mathematics Centre, Technical University of Lisbon (IST), Lisboa, Portugal

Received: date / Revised version: date

Abstract This paper starts with a short review of previous work on robust discriminant analysis with emphasis on the projection pursuit approach. Some theoretical aspects are briefly discussed. The core of the paper deals with practical issues related to the projection pursuit approach which generalizes Fisher's linear discriminant analysis. The choices of univariate estimators, starting points and maximization procedure are discussed and exemplified. The results of a simulation study are presented.

1 Introduction

Discriminant analysis is a widely used multivariate statistical method. Interesting applications are found in a broad range of traditional scientific areas. More recent areas of research, such as pattern recognition and data mining, show great intersection with discriminant analysis. The need for robust methods is obvious if one thinks that many problems have a large number of variables and observations, and data seldom coming from nice theoretical models.

Robust methods for discriminant analysis have been proposed, mostly for linear discrimination between two groups. The great majority of this work is concentrated on replacing the classical mean vectors and covariance matrices by robust counterparts. The analogy with regression and the use of robust regression has also been explored (e.g. [21]). The earlier studies (e.g. [1],[18],[2],[20]) used estimators with some drawbacks such as lack of equivariance or low breakdown point and were therefore not very successful. More recent work with high breakdown equivariant estimators led to more promising results: [4] with MVE estimators, [8] with MCD estimators adapted to the two sample situation, [9] and [5] with S-estimators ([9] also

proposed an adaptation to the two sample situation which worked very well on their simulation study). However, when the covariance structure is close to singularity the later methods may fail. Singularity problems may occur if, for instance, the number of variables is large relatively to the number of observations or when there are categorical variables. It is therefore important to explore other possibilities for robustifying discriminant analysis.

Quite surprisingly the projection pursuit approach, briefly suggested in [11] and [12], has not received much attention. The projection index suggested by Huber was considered by [3] and [14] (and referred in [22]). Other projection strategies were considered by: [17], [18] (in two of the proposed methods), [3] and [22]. A reason why these methods have not received much attention may be because they are not so easy to understand and to apply as the "replacement" method by robust location vectors and scatter matrices. This paper intends to be a contribution for reverting this situation by pointing practical difficulties, possible solutions and showing corresponding results.

The paper is organized as follows: Sect. 2 introduces the method. Sect. 3 discusses the implementation problems. Sect. 4 presents results obtained on a simulation study (which replicates the conditions of the simulation study of [9]) and on a single simulated data set where the "replacement" method fails. Finally Sect. 5 is devoted to a concluding remark.

2 A projection pursuit approach

Some of the advantages of projection pursuit (PP) methods are their ability to overcome the so called "curse of dimensionality" and the possibility to extend one-dimensional techniques to higher dimensions. The method is thus suitable for the robustification of multivariate procedures that can already be regarded as projection pursuit methods, such as principal components and linear discriminant analysis (LDA). This was pointed out in [11] and the case of principal components was explored by [13] and revisited recently by [6].

The projection pursuit approach here discussed uses as projection index the squared standardized distance between two groups (Fisher's original idea, [7]). That is, given a training sample, $\mathbf{x}_{ij}$, $i = 1, 2, j = 1, \ldots, n_i$, with n_i m-variate observations from group G_i, the projection pursuit estimate, $\hat{\boldsymbol{\alpha}}$, of the m-dimensional vector characterizing the discriminant direction between the two groups is the value of $\boldsymbol{\alpha}$ which maximizes

$$I(\boldsymbol{\alpha}) = \frac{\left[T(\boldsymbol{\alpha}^T\mathbf{x}_1) - T(\boldsymbol{\alpha}^T\mathbf{x}_2)\right]^2}{a_1 S^2(\boldsymbol{\alpha}^T\mathbf{x}_1) + a_2 S^2(\boldsymbol{\alpha}^T\mathbf{x}_2)} \tag{1}$$

(subject to $\|\boldsymbol{\alpha}\| = 1$ and $T(\boldsymbol{\alpha}^T\mathbf{x}_1) > T(\boldsymbol{\alpha}^T\mathbf{x}_2)$). T and S are general univariate estimators of location and dispersion, respectively, and $\boldsymbol{\alpha}^T\mathbf{x}_i$ denotes the univariate sample of the observations in group i projected onto $\boldsymbol{\alpha}$, that is, $(\boldsymbol{\alpha}^T\mathbf{x}_{i1}, \ldots, \boldsymbol{\alpha}^T\mathbf{x}_{in_i})$. If T and S are the sample mean and the sample

standard deviation, then Fisher's solution, $\boldsymbol{\alpha} \propto \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, is obtained. If T and S have good properties of efficiency and robustness it is expected that $\hat{\boldsymbol{\alpha}}$ inherits them. The pair (T =median; $S = MAD$) and several pairs of M-estimators will be used as examples in the remainder of the paper but there are many other possibilities. Natural choices for the constants a_1 and a_2 are $a_i = n_i/(n_1 + n_2 - 2)$ or $a_i = (n_i - 1)/(n_1 + n_2)$.

For classification purposes a cut-off or separation point has to be estimated. Fisher ([7]) proposed to use the midpoint between the projected means, that is, $\hat{\alpha}_0 = [T(\hat{\boldsymbol{\alpha}}^T\mathbf{x}_1) + T(\hat{\boldsymbol{\alpha}}^T\mathbf{x}_2)]/2$ (assuming implicitly equal a priori probabilities of group membership, if this is not the case then a suitable constant must be added, see [16]).

The classification rule for a new observation of unknown origin, with feature vector $\mathbf{x}$ is: classify in G_1 if $\hat{\boldsymbol{\alpha}}^T\mathbf{x} > \hat{\alpha}_0$ and in G_2 otherwise.

Affine equivariance is a desirable property for multivariate estimators, meaning that if the original data is affinely transformed, $\mathbf{x}_i^* = \mathbf{A}\mathbf{x_i} + \mathbf{b}$, where $\mathbf{A}$ is a real non-singular $m \times m$ matrix and $\mathbf{b} \in \Re^n$, then there is a rule for transforming the estimates that depends only on $\mathbf{A}$, $\mathbf{b}$ and the solution before the transformation. For the "replacement" method referred in Sect. 1 it is obvious that the linear discriminant estimates are affine equivariant if the multivariate location and scatter estimates are. For PP estimates it is shown in [16] that is also the case if the univariate estimators T and S are location scale equivariant. This property is important for practical reasons, 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. It is also important for theoretical studies since it means that it is possible to fix simple parametric situations and then go back to a general setup. This was used in [16] to obtain the influence functions of $\hat{\boldsymbol{\alpha}}$ and $\hat{\alpha}_0$ (more specifically the partial influence functions, as defined in [15]) for sufficiently regular T and S. Like the influence functions obtained in [6] for the PP estimators of principal components these influence functions depend on the derivatives of the influence function of T and S and are unbounded, meaning that the PP estimator may be very sensitive to small amounts of contamination at special points (for instance small values of a certain variable combined with large values of others).

Using the partial influence functions it is possible to obtain asymptotic variances of the PP estimators (cf. [16]) and make asymptotic comparisons with others methods. Other properties are studied in [16] such as maximum asymptotic biases due to point contaminations and the related asymptotic breakdown point. One aspect which deserves attention is the notion of breakdown point in the discriminant analysis setup. In [22] important considerations are made about this issue.

3 Implementation of the projection pursuit approach

3.1 Choice of the unidimensional estimators

The literature on robust univariate estimators of location and scale is so vast that it can not be reviewed here. As usual one as to made a compromise between efficiency and robustness but because of the special use of the estimators this is by no means so easy as a univariate estimation. Some of the theoretical considerations made previously already give this indication. Location scale equivariance is clearly a natural condition. Also special attention has to be given to the maximum biases of location and scale. For the scale estimator and because it appears in the denominator of (1) the bias due do inliers is of great concern. Another important aspect with consequences on the maximization procedure is the smoothness of the index, which is related to differentiability properties of T and S. At one end, very well known and considered very robust because of the maximal breakdown point, is the pair $T = median$ and $S = MAD$ (these were the only ones considered in [3]). Although this choice is not discarded, it has some drawbacks, namely, the lack of smoothness, low efficiency, large bias due to inliers (in the MAD) and large bias due to "far" outliers (in the median). With the same breakdown point, better efficiency and bias but still non differentiable there is the scale estimator Q_n ([19]) which was used in [6] for PP principal components. This estimator could be associated with any reasonable robust location estimator and has the advantage of having an explicit definition.

On the other hand, there are differentiable estimators (or almost) with controllable trade-off between efficiency and robustness but with non-explicit definitions. Attention will be restricted to simultaneous M-estimators of location and scale, [10] (Chapter 6), The following cases were considered:

(1) Huber's "Proposal 2" with, $\psi_1(u) = \max\{\min\{u, c\}, -c\}$ and $\chi_1(u) = \min\{u^2, c^2\}$.
(2) M-like-S-estimators with Tukey's biweight function: $\psi_2(u) = \chi_2'(u)$, $\chi_2(u) = \min\left\{u^2/2 - u^4/(2c^2) + u^6/(6c^2), c^2/6\right\}$. Denoted by S(c).
(3) Andrew's sine function for location combined with Huber's function for scale but with different tuning constants: $\psi_3(u) = \sin(\pi u/(2c))$, $|u| < 2c$, $\chi_3(u) \equiv \chi_1(u)$. Denoted by A(c).

The last two offer the advantage of finite rejection point for location.

The behavior of the PP indexes can be viewed for bidimensional situations since all directions can be identified by a single coordinate (the angle θ in polar coordinates). Several data sets were generated according to bidimensional situations equivalent to the three dimensional situations A, B, C, D, E considered by [9]. These are:
A. $G_1 : 50N_2(\mathbf{0}, \mathbf{I})$ $G_2 : 50N_2(\boldsymbol{\mu}, \mathbf{I})$
B. $G_1 : 40N_2(\mathbf{0}, \mathbf{I}) + 10N_2(5\boldsymbol{\mu}, .25^2\mathbf{I})$ $G_2 : 40N_2(\boldsymbol{\mu}, \mathbf{I}) + 10N_2(-4\boldsymbol{\mu}, .25^2\mathbf{I})$
C. $G_1 : 80N_2(\mathbf{0}, \mathbf{I}) + 20N_2(5\boldsymbol{\mu}, .25^2\mathbf{I})$ $G_2 : 8N_2(\boldsymbol{\mu}, \mathbf{I}) + 2N_2(-4\boldsymbol{\mu}, .25^2\mathbf{I})$

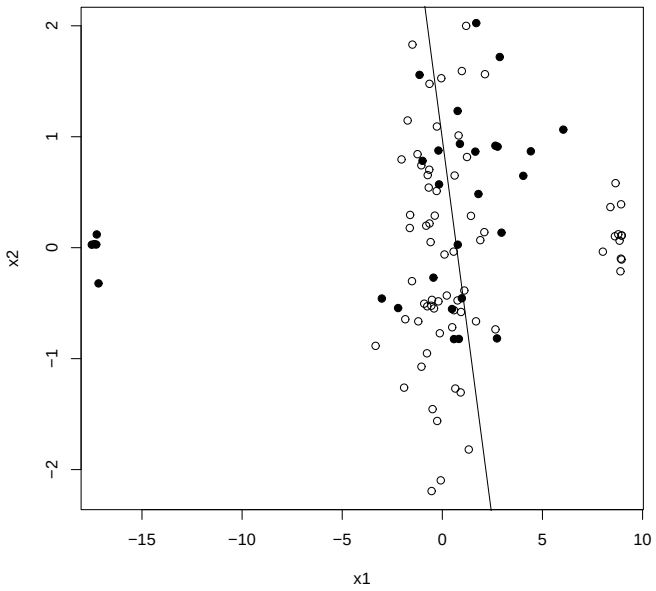

Fig. 1 A simulated data set from case E. Group 1: open circle; Group 2: solid circle. Separation line corresponding to a local maximum of the PP index with A(1.0).

D. $G_1 : 16N_2\left(\mathbf{0}, \mathbf{I}\right) + 4N_2(\mathbf{0}, 25\mathbf{I})$ $G_2 : 16N_2\left(\boldsymbol{\mu}, \mathbf{I}\right) + 4N_2(\boldsymbol{\mu}, 25\mathbf{I})$
E. $G_1 : 58N_2\left(\mathbf{0}, \mathbf{I}\right) + 12N_2(5\boldsymbol{\mu}, .25^2\mathbf{I})$ $G_2 : 25N_2\left(\boldsymbol{\mu}, 4\mathbf{I}\right) + 5N_2(-10\boldsymbol{\mu}, .25^2\mathbf{I})$ with $\boldsymbol{\mu} = (\sqrt{3}, 0)^T$. The optimal solution is $\theta = 0$ or $\theta = \pi$. [1] Situation A is the normal control (equal dispersions, no contamination), situations B and C correspond to location outliers while situation D corresponds to dispersion outliers. Situation E has location outliers and different dispersion as well as different number of observations (this is expected to be the most difficult situation). For case E a particularly hard example (of the many tried) was selected. The corresponding data set is represented in Fig. 1. Due to lack of space only some plots for this example are shown (Fig. 2).

In general it can be seen that the non-robust PP based on the mean and standard deviation fails completely for cases other than A. The median-MAD is very irregular and appears to be very sensitive to the location contamination. The Q_n (not shown) combined with M-location from A(c) appears to do much better but is still not smooth enough. For case A everything works nicely. For case B the global maximum points to a good solution but local maxima appear. In case E the PP is in trouble. However, for A(c) and S(c) estimators, there are local maxima which still provide information on the correct direction. To see this, plots of the discriminant scores along the directions corresponding to the two highest local maxima were prepared and are shown in Fig. 3 (only for A(1.0), similar results for the others). In

[1] Or only $\theta = \pi$ if the second restriction mentioned after equation (1) is applied.

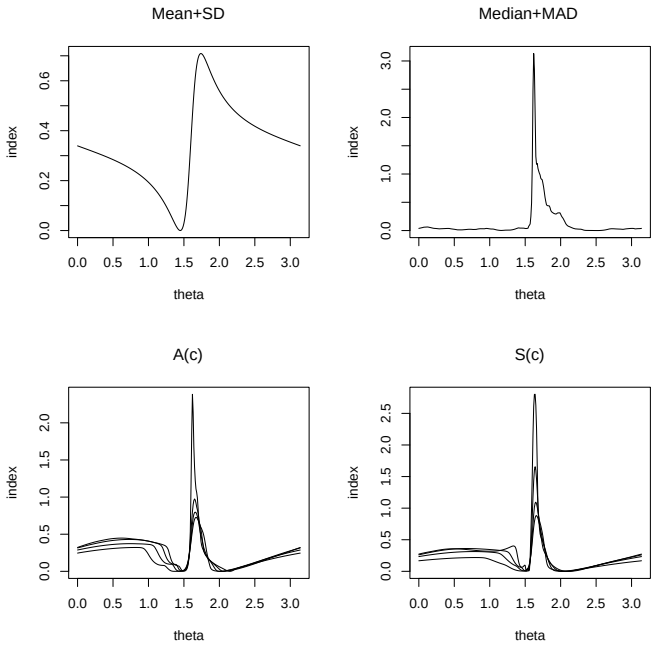

Fig. 2 Projection index, $I(\theta)$, for the simulated data set from case E, based on different univariate estimators. c: 1., 1.25, 1.5, 1.75 (for A) and 1.5, 2., 2.5, 3. (for S). (Higher c always correspond to lower $I(\theta)$.)

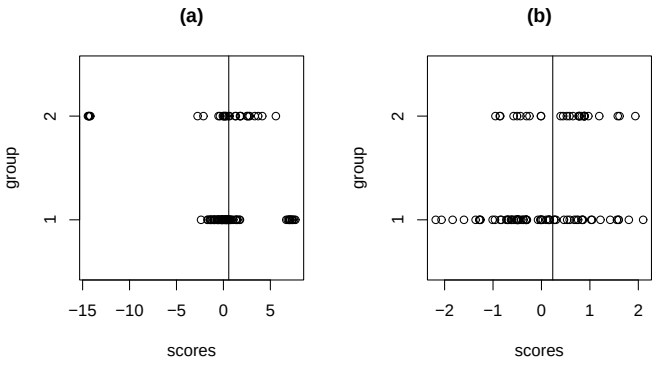

Fig. 3 Plot of the discriminant scores versus group for the two maxima of $I(\theta)$ with A(1.0). (a) local maximum, (b) global maximum.

face of the two plots it is an easy decision to choose the solution given by the local maximum. The separation line obtained from this local solution was added to the data plot (Fig. 1), confirming that it is indeed a reasonable solution, taking into account the difficulty of the case.

Two main conclusions are drawn from this empirical study: (1) a good strategy can not look blindly for the global maximum but it has also to detect and examine any existing local maxima. To do so it is imperative that

there are not so many of those, which stresses the importance of smoothness; (2) the $A(c)$ and $S(c)$ estimators emerge as good candidates for further studies.

3.2 Search procedure

The objective of the procedure is to find as many possible local maxima of $I(\boldsymbol{\alpha})$ as possible, subject to the two restrictions. The first restriction can be easily accommodated by working in spherical coordinates. The second restriction is easy to handle, by simply reverting the direction.

In view of the objective it is important to start from a large number of initial points. These can be chosen on a grid on the half-unit sphere, randomly generated or obtained by a procedure similar to the one used in [6] for PP principal components. For each starting point a numerical optimization procedure must then be applied. It is important to note that for the simultaneous M-estimators explicit derivatives can be computed ([14], [16]), therefore it is possible to use for instance a quasi-Newton conjugate gradient method. After that step the solution is tested to find out if it is different from all the previous solutions. The final output consists of the different solutions found. These final solutions must then be evaluated under another criterion, for instance the apparent error rate (after removing obvious outliers) either in the training sample or, if possible, in a separate test sample.

This procedure was implemented in Fortran and S-plus for the A(c) and S(c) univariate estimators. The Fortran version was used in the simulations described in the next Section.

4 Results

4.1 Simulation study

The simulation study performed replicates the one in [9] for two and three dimensions. The five distributional situations considered are the same described in Subsect. 3.1 (for dimension 3 use N_3 and $\boldsymbol{\mu} = (\sqrt{3}, 0, 0)^T$). The results are in terms of mean and standard deviation (over 500 replicates) of misclassification probabilities, evaluated on separate test samples of total size 2000, and are organized as in [9]. Several values of the tuning constant c were chosen for A(c) and S(c) but results are given for only two. For control Fisher's LDA was computed with all the observations (MLE-A) and with the clean (uncontaminated) observations (MLE-C).

The choice of the final solution, which was left open in the description of the search procedure, had to be incorporated in the simulation program. The option used was to decide on the basis of the apparent error rate on the training sample (it would be non realistic and misleading to use the test sample in these cases, since it would naturally yield better results). For

Table 1 Results of the simulation (3 variables, PP estimators based on $A(c)$ and $S(c)$ each with two values of c): mean and standard deviation of misclassification probability estimates evaluated under the contaminated (M_c, S_c) and uncontaminated (M_u, S_u) distributions.

	A(1.50)	A(1.75)	S(2.5)	S(3.0)	MLE-C	MLE-A
Case A						
M_u	.204	.202	.206	.204	.200	.200
S_u	.013	.012	.015	.013	.010	.010
Case B						
M_c	.363	.362	.363	.361	.362	.524
S_c	.010	.009	.011	.010	.010	.048
M_u	.203	.203	.203	.201	.202	.655
S_u	.012	.012	.013	.012	.013	.060
Case C						
M_c	.371	.370	.370	.364	.369	.492
S_c	.018	.018	.014	.011	.014	.047
M_u	.215	.217	.212	.205	.211	.615
S_u	.037	.049	.017	.013	.017	.059
Case D						
M_c	.273	.273	.271	.271	.259	.293
S_c	.033	.032	.029	.029	.019	.047
M_u	.231	.231	.229	.229	.214	.256
S_u	.038	.037	.033	.033	.021	.056
Case E						
M_c	.402	.400	.402	.400	.400	.461
S_c	.019	.017	.019	.017	.017	.033
M_u	.281	.279	.282	.279	.278	.555
S_u	.030	.028	.029	.025	.021	.040

distributional situations A and D all the observations were used, while for the other situations only the clean observations were used (assuming that, as happened in the very difficult example discussed before, the location outliers will always be clearly identified, but the dispersion outliers may not always be identified, case D).

The results for dimension 3 are given in Table 1 and are similar to the results reported in [9] using replacement by multivariate S-estimators. There are minor differences between methods A(c) and S(c) and some differences between the tuning constants deserving further investigation. To choose c based on asymptotic efficiency at the normal model is a possibility.

Additionally a similar situation in dimension 8 but with a reduced number of observations was run in order to show some results where the replacment method by multivariate high breakdown estimators (S or MCD, for instance) would fail. The distributional situations were

A'. $G_1 : 7N_8(\mathbf{0}, \mathbf{I})$ $G_2 : 7N_8(\mathbf{1}, \mathbf{I})$

B'. $G_1 : 6N_8(\mathbf{0}, \mathbf{I}) + 1N_8(\mathbf{5}, 0.25^2\mathbf{I})$ $G_2 : 6N_8(\mathbf{1}, \mathbf{I}) + 1N_8(-\mathbf{4}, 0.25^2\mathbf{I})$

D'. $G_1 : 6N_8(\mathbf{0}, \mathbf{I}) + 1N_8(\mathbf{0}, 25\mathbf{I})$ $G_2 : 6N_8(\mathbf{1}, \mathbf{I}) + 1N_8(\mathbf{1}, 25\mathbf{I})$.

Table 2 Results of the simulation study (8 variables, PP estimators based on M-Andrews type estimators with different tuning constants): mean and standard deviation of misclassification probability under the contaminated (M_c, S_c) and uncontaminated (M_u, S_u) distributions.

	A(1.00)	A(1.25)	A(1.50)	A(1.75)	MLE-C	MLE-A
Case A						
M_u	.222	.222	.213	.212	.212	.212
S_u	.077	.079	.077	.080	.072	.076
Case B						
M_c	.338	.342	.343	.347	.358	.463
S_c	.067	.069	.072	.071	.075	.015
M_u	.229	.235	.236	.241	.253	.523
S_u	.083	.089	.089	.091	.092	.016
Case D						
M_c	.260	.256	.256	.256	.284	.270
S_c	.072	.076	.076	.075	.084	.081
M_u	.231	.230	.226	.227	.258	.241
S_u	.079	.085	.084	.084	.094	.090

The results for method A(c) (500 replicates) are given in Table 2. Although much higher than the optimum error rate ($e_{opt} = \Phi(-\sqrt{2}) \simeq .079$) the obtained misclassification probabilities are far from the "toss coin decision rule" (0.5 error rate), showing that there is valuable information taken out of such sparse data, either with or without outliers. For cases B and D, PP is even better than MLE-C. These results are encouraging but there is certainly room for improvement.

4.2 Simulated example with a categorical variable

As mentioned in the Introduction another situation causing problems to robust multivariate estimators is the existence of categorical variables. The use of Fisher's LDA in such situations may be controversial but there are many examples where it has given better results than other theoretically more adequate methods, and it continues being widely used in practice.

A small example with one continuous variable (x_1) and one binary variable (x_2) is used here to show the potentiality and also the difficulties of the PP method. The original uncontaminated data set was generated as follows:

G_1: 80 obs. with $x_2 = 0$ and $X_1|x_2 = 0 \sim N(0, 0.3^2)$ and 20 obs. with $x_2 = 1$ and $X_1|x_2 = 1 \sim N(-1, 0.3^2)$;

G_2: 20 obs. with $x_2 = 0$ and $X_1|x_2 = 0 \sim N(2, 0.3^2)$ and 80 obs. with $x_2 = 1$ and $X_1|x_2 = 1 \sim N(1, 0.3^2)$.

It is possible to separate the two groups completely by a straight line. A contaminated data set was then obtained by adding to G_1 10 observations with $x_2 = 1$ and $X_1|x_2 = 1 \sim N(10, 0.1^2)$ and to G_2 10 observations with $x_2 = 0$ and $X_1|x_2 = 0 \sim N(-9, 0.1^2)$.

Table 3 Simulated example with a categorical variable: Intercept (a) and slope (b) of the separation line between the two groups, and total number of misclassified observations (MIS), for several situations and methods.

	Uncontaminated			Contaminated		
Method	a	b	MIS	a	b	MIS
unregularized data						
LDA	.901	-.846	0	.462	.079	40
Logistic	.867	-.815	0	.462	.079	40
regularized data						
LDA	.928	-.895	0	.462	.079	40
Logistic	.908	-1.038	0	.465	.072	40
MCD(50%)	.570	-.145	40	.562	-.128	60
MCD(25%)	.566	-.119	40	.568	-.137	60
MultS(50%)	.575	-.155	40	.573	-.152	60
MultS(25%)	.885	-.802	0	.909	-.856	20
PPmedian	.812	-.642	0	.884	-.794	20
PPA(1.5)	.884	-.794	0	.904	-.844	20

It is easy to see that a non-robust PP method such as Fisher's LDA can be directly applied to this kind of data but a robust PP can not. For instance, the MAD on the direction of x_2 will be zero unless there are exactly half of the observations on each category. Any other robust scale estimator would give either 0 or very small dispersion on that direction.

However, a very simple trick can make the robust PP work on such a situation. The trick is to add a small white noise to the categorical variable. In this example $N(0, 0.1^2)$ was used. This trick will be referred as "regularization". The same procedure could also put multivariate MCD and S to work, but it happens that they still give nearly singular estimates of the covariance matrix in most cases. In the example S with 25% breakdown point gave reliable estimates for each group, but this is indeed a failure rather than a success, because the same estimator would fail if there were, for instance 15 observations on one category rather than 20 (20 happened to be quite close to the breakdown point and produce a large bias). Logistic discrimination is a natural method for a data set like this and it was also used.

The results are given in Table 3. Plots of some PP indexes are given in Fig. 4. It is again clear that the interesting solution is not at the global maximum but at a local maximum. (An interesting solution is considered one close to the theoretical Fisher's solution, which has $\theta_F \simeq 0.86$. Such a solution will usually misclassify none of the observations.)

5 Conclusions

In this paper it is shown that good practical results can be obtained with the proposed PP method. The method is highly robust, fairly efficient and

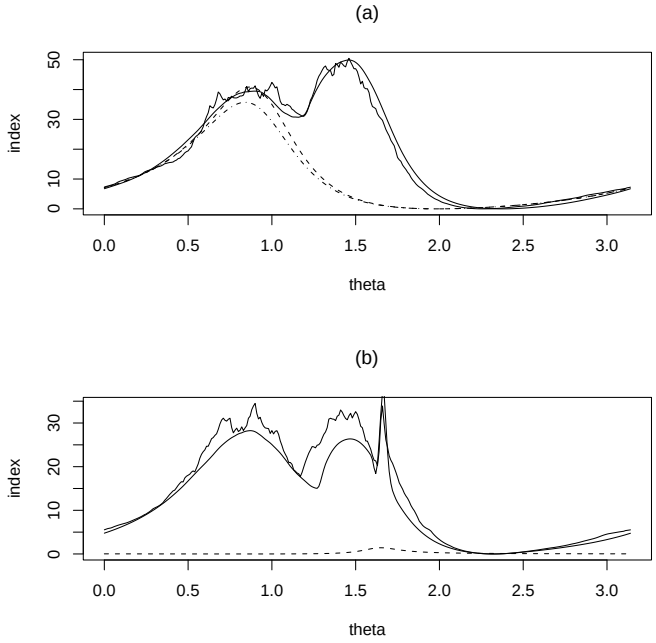

Fig. 4 Projection index, $I(\theta)$, for the simulated example with a categorical variable, based on different univariate estimators. (a) at the uncontaminated regularized data; (b) at the contaminated regularized data. Solid smooth line: A(1.5); solid irregular line: Median and MAD; dashed-dotted line: mean and standard deviation; dashed line: mean and standard deviation for the original data.

gives reliable results where other popular robust procedures fail. It was however necessary to solve several implementation difficulties and to modify the original definition. It is also important to mention that the proposed method is computationally intensive and may be not suitable, yet, for high dimensional problems.

References

1. S. W. Ahmed, and P. A. Lachenbruch, Discriminant Analysis when Scale Contamination is Present in the Initial Sample, In: J. Van Rysin, editor, *Classification and Clustering*, pp. 331–353, Academic Press: New York, 1977.
2. N. A. Campbell, Robust Procedures in Multivariate Analysis II: Robust Canonical Variate Analysis, Applied Statistics **31,** (1982) 1–8.
3. Z. Y. Chen, and R. J. Muirhead, , A Comparison of Robust Linear Discriminant Procedures Using Projection Pursuit Methods. In: T. W. Anderson, K. T. Fang and I. Olkin, editors, *Multivariate Analysis and its Applications*, pp. 163–176, IMS Monograph Series: Hayward, 1994.
4. C. V. Chork and P. J. Rousseeuw, Integrating a High-Breakdown Option into Discriminant Analysis in Exploration Geochemistry, Journal of Geochemical Exploration **43,** (1992) 191–203.

5. C. Croux and C. Dehon, Robust linear discriminant analysis using S-estimators. To appear in Canadian Journal of Statistics **29,** (2001).
6. C. Croux and A. Ruiz-Gazen, High breakdown estimators for principal components: the projection-pursuit approach revisited. Preprint, University of Brussels (ULB). (2000).
7. R. A. Fisher, The use of multiple measurements in taxonomic problems. Annals of Eugenics **7,** (1936) 179–188.
8. D. M. Hawkins and G. J. McLachlan, High-Breakdown Linear Discriminant Analysis, Journal of the American Statistical Association **92,** (1997) 136–143.
9. X. He and W. K. Fung, High breakdown estimation for multiple populations with applications to discriminant analysis. Journal of Multivariate Analysis **72,** (2000) 151–162.
10. P. J. Huber, *Robust Statistics*, Wiley: New York, 1981.
11. P. J. Huber, Projection pursuit (with discussion). Annals of Statistics **13,** (1985) 435–475.
12. P. J. Huber, Projection Pursuit and Robustness. In: S. Morgenthaler, E. Ronchetti and W. A. Stahel, editors, *New Directions in Statistical Data Analysis and Robustness*, pp. 139–146, Birkäuser: Basel, 1993.
13. G. Li and 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.
14. A. M. Pires and J. A. Branco, A robust linear discriminant by projection pursuit. In: R. Dutter and W. Grossman, editors, *COMPSTAT 1994: Short Communications in Computational Statistics*, pp. 51–52, VIS: Vienna, 1994.
15. A. M. Pires and J. A. Branco, Partial Influence Functions, Preprint 23/2000, Instituto Superior Técnico (Technical University of Lisbon), Dept. of Mathematics. (2000).
16. A. M. Pires and J. A. Branco, Projection-pursuit Approach to Robust Linear Discriminant Analysis. Preprint, Instituto Superior Técnico (Technical University of Lisbon), Dept. of Mathematics. (2001).
17. C. Posse, Projection pursuit discriminant analysis for two groups. Communications in Statistics – Theory and Methods **21,** (1992) 1–19.
18. R. H. Randles, J. D. Broffitt, J. S. Ramberg and R. V. Hogg, Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates, Journal of the American Statistical Association **73,** (1978) 564–568.
19. P. J. Rousseeuw and C. Croux, Alternatives to the Median Absolute Deviation, Journal of the American Statistical Association **88,** (1993) 1273–1283.
20. M. L. Tiku and N. Balakrishnan, Robust Multivariate Classification Procedures Based on the MML Estimators, Communications in Statistics – Theory and Methods, **13,** (1984) 967–986.
21. V. Todorov, N. Neykov, and P. Neytchev, Robust two-group discrimination by bounded influence regression. A Monte-Carlo simulation. Computational Statistics and Data Analysis **17,** (1994) 289–302.
22. J. W. Van Ness and J. J. Yang, Robust discriminant analysis: training data breakdown point. Journal of Statistical Planning and Inference **67,** (1998) 67–83.