A pairwise subspace projection method for multi-class linear dimension reduction

Diego Tomassi

[1] Instituto de Matemática Aplicada del Litoral, UNL - CONICET
[2] Centro de Investigación en Señales, Sistemas e Inteligencia Computacional FICH, Universidad Nacional del Litoral - CONICET
[3] Departamento de Matemática, FIQ, Universidad Nacional del Litoral
diegot@santafe-conicet.gov.ar

Abstract. Linear feature extraction is commonly applied in an all-at-once way, meaning that a single trasformation is used for all the data regardless of the classes. Very good results can be achieved with this approach when the classification problem involves just a few classes. Nevertheless, when the number of classes grows is often difficult to find a low dimensional subspace while preserving the error rates, due to overlapping between the different populations. In this paper we propose an alternative method based on a collection of transformations, each involving two of the classes in the problem. Each transformation in the collection is estimated using an approximation to the information discriminant analysis, which is found to be equivalent to sufficient dimension reduction for heteroscedastic Gaussian data. A regularized version of the objective function is also introduced, allowing for simultaneous variable selection. In this way, each reduction implies only a subset of the original variables. A probabilistic model is build by means of a simple latent variable, so that classification is carried out using standard Bayes decision rule. Several real data sets are used to compare the performance of the proposed method against similar approaches based on ensembles of binary classifiers.

1 Introduction

Linear feature extraction/dimension reduction is often included in statistical pattern recognition to lower the size of the models to estimate [1, 2]. In many cases with finite sample data this allows for estimates with smaller variance, which translates into better generalization capability of the classifier. In other cases, the recognition rates are not improved but computations take place in a much smaller feature space, a fact that can be of technological significance.

Well known methods for linear dimension reduction are principal component analysis (PCA) [3] and Fisher's linear discriminant analysis (LDA) [4]. PCA actually does not take into account class information and thus can be highly suboptimal for classification tasks. LDA is supervised by class information, but from a theoretical point of view it is optimal only when class-conditional data is normally distributed with constant covariance matrices across the classes [5].

A bunch of methods have been proposed to deal with the heteroscedastic case. The most theoretically grounded alternatives are those that care about the amount of discriminant information available in the data before and after the transformation is applied. Information discriminant analysis (IDA) [6] aims at finding a projetion that peserves mutual information between the features and the labels. Despite it is stated in general terms, the proposed working method actually assumes that observations are normally distributed given the class. Aproximate information discriminant analysis (AIDA) is aimed to approximate IDA but using matrix eigenanalysis instead of the numerical optimization required in IDA [7]. Other methods measure the separability of the classes in terms of the Chernoff distance, whether in the original space [8] or in the reduced feature subspace [9].

A parallel line of research developed mostly from the statistics community is sufficient dimension reduction (SDR) [10, 11]. The goal here is to find a transformation of the explanatory variables $\mathbf{X}$ that retains all the available information about a response Y for a particular objective. For classification tasks, Y is given just by the labels and the transformation is shown to preserve Bayes error [12]. When the observations from each class are normally distributed, an optimal estimator of the reduction can be found using likelihood theory and it is known as likelihood acquired directions (LAD) [13]. This objective function, nevertheless, comes down to be equivalent to IDA [14].

The standard practice with linear feature extraction methods is to apply an all-at-once transformation, meaning that the data is projected into a lower dimensional subspace using the same transformation matrix regardless of the class. When there are many classes involved in the problem, low dimensional projections often reduce the accuracy of the classifier significantly [15, 16]. The reason is that the classes overlap in the reduced subspace. In such cases, in order to preserve the error rates is often necessary to retain a number of directions similar to the original dimensionality of the data. The effect is more evident in small sample problems, when the estimates of covariance matrices are poor.

To overcome this limitation we propose a pairwise linear dimension reduction method. Instead of a single transformation, a collection of transformations is estimated, each one relating two of the classes in the problem. The method is stated under the framework of sufficient reductions, so that a probabilistic model and a likelihood are available. We also introduce a penalized version of AIDA which adds variable selection to the feature extraction process, so that linear combinations in the reduced subspace include only a subset of active variables from the original predictors. Keeping only the relevant variables helps to reduce computations in the classification stage and to identify the most important information to discriminate between pairs of classes.

The work closest to this contribution is presented in [16]. Authors also explore pairwise dimension reduction, but their method does not allow for a probabilistic model of the data. Beacuse of this, likelihood computation is not available, and voting strategies are build to obtain the final label assignment after testing the data with each of the binary classifiers. We compare our results with theirs along

the paper. About adding variable selection in the context of sufficient dimension reduction, we are not aware of previous work targetted to heteroscedasticity. For homoscedastic conditions, available procedures are available from [17, 18]. Between them, only [18] preserve invariance properties of the nonregularized methods.

The paper is organized as follows. In Section 2 we start by reviewing the basics of sufficient dimension reduction and presenting the general idea of the proposed pairwise transformation method. We then introduce the penalized form of AIDA that we finally use in estimation. In Section 3 we use real datasets from the UCI repository to illustrate the performance of the method compared to all-at-once projections and other pairwise strategies from the literature. Finally, as the penalized estimator of the reduction is itself a contribution, we also carry out a simulation study to assess the accuracy of the variable selection procedure.

2 Pairwise subspace projections under sufficiency

We start this section by briefly reviewing the basics of sufficient dimension reduction. Then we describe the general idea behind the proposed pairwise scheme for linear feature extraction and the estimation procedure. Special emphasis is given to the introduction of a regularized version of approximate information discriminant analysis that allows for simultaneous variable selection.

2.1 Basics of sufficient dimension reduction

Let $\mathbf{X} \in \mathbf{R}^p$ be a random vector of features and let Y indicate the class labels. The reduction $\boldsymbol{\beta}^T\mathbf{X}$ is said to be *sufficient* if and only if [11]

$$F(\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}, Y) \sim F(\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}); \tag{1}$$

where $F(\cdot)$ and $F(\cdot|\cdot)$ denote distribution and conditional distribution function, respectively, and $\sim$ stands for asymptotic equivalence. In simple words, the above definition tells that $\boldsymbol{\beta}^T\mathbf{X}$ has all of the information about Y that is available in $\mathbf{X}$.

Sufficient reductions preserve Bayes error [12] and allow for a deeper theoretical understanding of the estimators. Though the methodoly is quite general and does not require a model for the data, throughtout this paper we will assume that $\mathbf{X}|(Y = y)$ is normally distributed with density $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$, where $\boldsymbol{\mu}_y = \mathrm{E}(\mathbf{X}|Y = y)$ and $\boldsymbol{\Delta}_y = \mathrm{Var}(\mathbf{X}|Y = y)$. Under this model, a likelihood solution for the sufficient reduction is given in [13]. The corresponding objective function is

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}^T\boldsymbol{\beta}=\mathbf{I}}{\arg\max} \left\{ \log|\boldsymbol{\beta}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}| - \sum_y f_y \log|\boldsymbol{\beta}^T\widehat{\boldsymbol{\Delta}}_y\boldsymbol{\beta}| \right\}, \tag{2}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the sample marginal covariance matrix and f_y is the sample estimate of the prior probability of class y. Solution of (2) is known as the LAD estimator

and requires numerical optimization on the Grassmann manifold [19]. Under the heteroscedastic Gaussian model, LAD is shown to be the minimal sufficient reduction [13]. Because it preserves Bayes error, we can build a classifier that uses $\mathrm{p}(\widehat{\boldsymbol{\beta}}^T\mathbf{X}|Y=y)\mathrm{p}(y)$ instead of the full model $\mathrm{p}(\mathbf{X}|Y=y)\mathrm{p}(y)$ for the decision rule.

2.2 Proposed method for pairwise linear dimension reduction

To derive our method, we assume a Gaussian mixture model for the class-conditional data and set a simple latent variable model. Assume there are h classes in the discrimination problem and let G be the latent variable. Each state $G=g$ indicates one out of $\binom{h}{2}$ groups involving just two classes from where the observaton comes. As we do not have certainty about the specific group for a given data point $\mathbf{X}=\mathbf{x}$, the likelihood reads

$$\mathrm{p}(\mathbf{x}|Y=y)=\sum_g \mathrm{p}(\mathbf{x}|Y=y,G=g)\mathrm{p}(G=g|Y=y).$$

Provided $\boldsymbol{\beta}_g^T\mathbf{X}$ is a sufficient dimension reduction for $\mathbf{X}|(Y,G=g)$, we can rewrite

$$\mathrm{p}(\mathbf{x}|Y=y)=\sum_g \mathrm{p}(\boldsymbol{\beta}_g^T\mathbf{x}|Y=y,G=g)\mathrm{p}(G=g|Y=y).$$

Given a new observation $\mathbf{x}$ to be classified, we then assign the class label according to Bayes rule

$$\begin{aligned}\hat{y} &= \arg\max_j \mathrm{p}(Y=j|\mathbf{x}),\\ &= \arg\max_j \mathrm{p}(Y=j)\mathrm{p}(\mathbf{x}|Y=j),\\ &= \arg\max_j \left\{\mathrm{p}(Y=j)\sum_g \mathrm{p}(\boldsymbol{\beta}_g^T\mathbf{x}|Y=j,G=g)\mathrm{p}(G=g|Y=j)\right\}.\end{aligned}$$

Thus, to build the classifier we form $\binom{h}{2}$ groups from the training set, each one containing the observations from two classes only. Using these datasets, we estimate the collection of reductions $\{\boldsymbol{\beta}_g\}$ using LAD. When all the transformations have been obtained, we can estimate the model parameters for each of the $h(h-1)$ normal densities in the model. Then, given $\mathbf{x}$ we can compute $\mathrm{p}(\beta_g^T\mathbf{x}|Y=j,G=g)$ for $j=1,2,\ldots,h$ and $g=1,2,\ldots,\binom{h}{2}$. To estimate $\mathrm{p}(G=g|Y=y)$ we first use Bayes rule

$$\mathrm{p}(G=g|Y=j)=\frac{\mathrm{p}(Y=j|G=g)\mathrm{p}(G=g)}{\sum_k \mathrm{p}(Y=j|G=k)\mathrm{p}(G=k)}.$$

The priors $\mathrm{p}(G=g)$ are estimated from the whole training sample, whereas $\mathrm{p}(Y=y|G=g)$ are estimated from the training subset corresponding to each

group. With the estimates obtained in the training phase, classification of a new observation $\mathbf{x}$ is carried out using:

$$\hat{y} = \arg\max_j \sum_g \mathrm{p}(\boldsymbol{\beta}_g^T x | Y = y, G = g)\mathrm{p}(Y = y | G = g)\mathrm{p}(G = g).$$

One can argue that computing all the projections $\{\boldsymbol{\beta}_i^T \mathbf{X}\}$ can be computationally too expensive. In the next subsections we propose alternatives to reduce the computations and to include variable selection into the reduction process.

2.3 Dimension reduction as an eigenvalue problem

The LAD estimator, albeit its optimal properties under conditional normality, requieres numerical optimization to compute the projection. In addition, as the likelihood function is not convex, it requires the computation of a suitable initial estimate to start the iterative process. Despite the final estimate is found to be stable under different initializations provided they are consistent estimators [13], it is clear that the total amount of time needed to get the projection is considerably larger than when using methods based in eigendecomposition of a symmetric matrix. Thus, one way to reduce the computational cost at the training stage is to use a simple estimation for the reduction that does not require numerical optimization. In this sense, many dimension reduction methods have been proposed where the projection is obtained from eigenanalysis of a suitable matrix. Among them, AIDA can can be thought of as a quadratic approximation to LAD when it is stated in terms of $\mathrm{E}_Y(\boldsymbol{\Delta}_y)^{-1/2}\mathbf{X}$ instead of the original random vectors $\mathbf{X}$. The objective function is

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}^T\boldsymbol{\beta}=\mathbf{I}}{\arg\max}\ \mathrm{tr}(\boldsymbol{\beta}^T S_{AIDA}\boldsymbol{\beta}), \tag{3}$$

with

$$S_{AIDA} = \log(\widehat{\boldsymbol{\Sigma}}) - \sum_y f_y \log(\boldsymbol{\Delta}_y).$$

Though it is not the optimal estimator, it is found to perform remarkably well in most situations with real data [14]. Notice that replacing LAD with AIDA reduces the computation time at the training stage, but there are no gains at the classification stage.

2.4 Variable selection within groups

In common linear dimension reduction methods, all the original coordinates in $\mathbf{X}$ are included in the linear combinations. Nevertheless, because we are dealing with transformations involving two classes only, we can expect some of the original variables to be irrelevant for a particular transformation, although important for other ones. Adding variable selection to the dimension reduction

process helps to identify the subset of the predictors that is relevant to discriminate between classes in a specific group. The resulting $\boldsymbol{\beta}_g$s have only a subset of active coordinates, thus computations at the classification stage are also reduced.

For a given estimation method, variable selection is commonly achieved penalizing the original objective function. This regularization is a challenging problem for the LAD estimator, since the likelihood function is not convex. Nevertheless, the trace operator in AIDA is convex and therefore it is a more friendly alternative to penalize terms for variable selection. In addition, it is desirable to preserve the invariance properties of the original estimator [7]. It must be noticed that this goal cannot be achieved using common regularization methods based on the ℓ_1 norm as in LASSO [20]. To account for this invariance, the method we use here is an extension to heteroscedastic data of the method introduced in [18]. As a first step, it is showed in [17] that many SDR methods can be cast in the form of a generalized eigenvalue problem

$$\mathbf{M}_n\boldsymbol{\delta}_{ni} = \lambda_i\mathbf{N}_n\boldsymbol{\delta}_{ni}. \tag{4}$$

In these expressions, the subscript n refers to the sample version of the quantity. Under certain conditions, the set $(\boldsymbol{\delta}_{n1}\boldsymbol{\delta}_{n2}\ldots\boldsymbol{\delta}_{nd})$, corresponding to the d largest eigenvalues λ_i form a basis for the smallest reduction subspace. In [18], authors introduced a coordinate-independent method that finds a sparse sufficient dimension reduction for problems that can be written in this way. Their development, however, extents only to homoscedastic data. It is easy to show that we can obtain a sparse estimator in this framework from AIDA. Let $\boldsymbol{\Delta}_n = \sum_y \frac{n_y}{n}\boldsymbol{\Delta}_y$ with (n_y/n) the fraction of observations from population y in the sample. Set $\mathbf{N}_n = \boldsymbol{\Delta}_n$ and $\mathbf{M}_n = \boldsymbol{\Delta}_n^{1/2}S_{AIDA}\boldsymbol{\Delta}_n^{1/2}$ and let $\mathbf{z}_i = \boldsymbol{\Delta}_n^{1/2}\boldsymbol{\delta}_{ni}$. Then

$$\begin{aligned}\boldsymbol{\Delta}_n^{1/2}S_{AIDA}\boldsymbol{\Delta}_n^{1/2}\boldsymbol{\delta}_{ni} &= \lambda_i\boldsymbol{\Delta}_n\boldsymbol{\delta}_{ni}\\ \boldsymbol{\Delta}_n^{1/2}S_{AIDA}\mathbf{z}_i &= \lambda_i\boldsymbol{\Delta}_n^{1/2}\mathbf{z}_i\end{aligned}$$

Premultiplying both sides with $\boldsymbol{\Delta}_n^{-1/2}$ we get $S_{AIDA}\mathbf{z}_i = \lambda_i\mathbf{z}_i$. Thus, the $\mathbf{z}_i$ are eigenvectors of S_{AIDA} and the basis matrix for the smallest reduction subspace is $\boldsymbol{\Delta}_n^{-1/2}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_d)$. This result allows us to use results from [18] to achive joint variable selection and dimension reduction by optimizing the objective function

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\{-\operatorname{tr}(\boldsymbol{\beta}^T S_{AIDA}\boldsymbol{\beta}) + \rho(\boldsymbol{\Delta}^{-1/2}\boldsymbol{\beta})\},$$

with

$$\rho(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d) = \sum_{i=1}^{p}\theta_i||\mathbf{v}_i||_2.$$

This form of regularization is known as *group lasso* [21] Implementation uses actually a local quadratic approximation for $\rho(\boldsymbol{\Delta}^{-1/2}\boldsymbol{\beta})$. The procedure is iterative. Let $\widehat{\boldsymbol{\beta}}_i^k$ be the i-th row of $\widehat{\boldsymbol{\beta}}$ at iteration k and let $H^k = \operatorname{diag}(\frac{\theta_1}{||\widehat{\boldsymbol{\beta}}_1^k||}, \ldots, \frac{\theta_p}{||\widehat{\boldsymbol{\beta}}_p^k||})$.

At each iteration we solve

$$\widehat{\boldsymbol{\Gamma}} = \underset{\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}=\mathbf{I}}{\arg\min}\, \boldsymbol{\Gamma}^T(-\mathbf{N}_n^{-1/2}\mathbf{M}_n\mathbf{N}_n^{-1/2} + \frac{1}{2}\mathbf{N}_n^{-1/2}H^k\mathbf{N}_n^{-1/2})\boldsymbol{\Gamma} \tag{5}$$

and get $\widehat{\boldsymbol{\beta}}^{k+1} = \mathbf{N}_n^{-1/2}\boldsymbol{\Gamma}^k$. If $||\widehat{\boldsymbol{\beta}}_i^k|| < \epsilon$, the i-th variable of the original dataset is removed and procedure is repeated until convergence. We will refer to this method as *pairwise penalized approximate information discriminant analysis* (PPAIDA).

3 Experiments

3.1 Penalized version vs all predictors

To illustrate the performance of the method proposed in Section 2, we take as example the classification of the letter recognition dataset from the UCI machine learning repository. 10-fold cross validation is used to compare the obtained error rates using the penalized method versus those obtained from multiple projections without variable selection and from single-step projection using LAD. The experiment was carried out using $\epsilon = 10^{-4}$, though it can be further tunned using cross validation. Obtained results are shown in Figure 1. It can be seen that the multiple-projection scheme consistently outperfoms the all-at-once transformation. Furthermore, the penalized version of the algortihm gives results almost identical to the non-penalized version, which uses all the original variables to form the linear combinations (note that curves overlap). This parameter can be further tunned using cross validation. The right panel in the figure shows the fraction of variables that are found active in the selection process, averaged over all the estimated reductions. It is clearly seen that the fraction of retained variables increases with the dimension of the reduced subspace. Despite the figure is for the pendigits dataset, the picture is quite descriptive of the general trend with all the tested data sets.

3.2 Comparison with other pairwise heteroscedastic approaches

We now assess the performance of the proposed method for classification of several real data sets taken from the UCI machine learning database. In particular, we want to compare results with the scores reported in [16] which uses other strategies for multiclass pairwise linear dimension reduction. There, final class assignment is decided upon a voting strategy or through a decision tree. For a direct comparison of results, we chose the same datasets:; Iris, Pendigits, Thyroid, Wine and the Vowel recognition database. We also took a 10-fold cross-validation design for the experiment and the dimension of $\boldsymbol{\beta}_i$ is assumed to be the same across i. Obtained results are shown in Table 1. SV stands for the simple voting strategy in [16], WV for the wighted voting strategy, and DT for the decision rule based on decision trees. The Iris, Thyroid and Wine data sets are standard easy classification tasks comprising only three classes. In such settings, there is

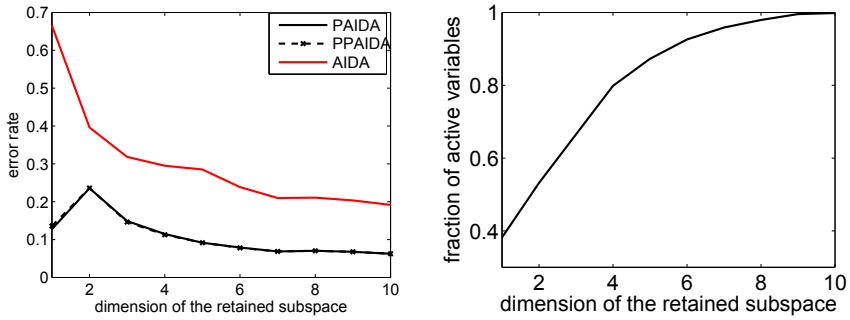

Fig. 1. Comparison of the error rates for classification of the letter recognition data set obtained with multiple subspace projections using penalized estimators against: i) multiple projections without variable selection; ii) single projection using LAD. The right panel shows the average fraction of active variables selected as a function of the dimension d of the retained feature subspace.

not much to be gained from the pairwise scheme. Even with that, it can be seen that the proposed method does not degrade the performance of the all-at-once transformation, which gets excellent results. The proposed method also achieves slightly better error rates that ones reported in [16], though thay are of little significance and can be due to variability in the cross validation procedure.

The vowel recognition and the pendigits datasets are of greater interest. Results for the vowels recognition dataset show that obtained error rates are smaller when using the multiple subspace projections approach. In this case, the proposed method outperforms the pairwise methods proposed in [16], although the best score occurs at a subspace that is larger ($d = 6$ vs $d = 4$). It is fair to say, nevertheless, that even for the all-at-once projection approach, AIDA gets a better score than the all-at-once version of the heteroscedastic methods used in [16] (0.26 vs 0.30). Although this can have some influence on the results from the pairwise methods, it seems clear from the overall experiment that the proposed method performs at least as good as the existing pairwise alternatives which are paired with more complex decision rules.

For the pendigits dataset, the lowest error rate achieved by all the methods is the same (0.02). Nevertheless, this minimum occours at a smaller subspace for PPAIDA. It is important to note that performance may be improved by allowing different dimensions for the subspaces spanned by each $\boldsymbol{\beta}_i$. Such relaxation, however, would require testing the most likely dimension for each pair of classes.

Results commented above show that that the proposed method obtains recognition rates similar to those reported in [16]. Unlike those methods, PPAIDA provides a probabilistic model for the data. This is an advantage when the reduction has to be embedded in likelihood-based approaches; in addition, it allows for simple methods based on information criteria to make inference about the dimensions of the reduced subspaces.

Dataset			**AIDA**		**PPAIDA**		**SV**		**WV**		**DT**	
	h	p	error	d	error	d	error	d	eror	d	error	d
Iris	3	4	0.02	1	0.02	1	0.02	1	0.02	1	0.02	1
Thyroid	3	5	0.04	1	**0.03**	1	0.04	1	0.04	1	0.04	1
Wine	3	13	**0**	5	**0**	5	0.01	7	0.01	7	0.01	7
Vowel	11	10	0.26	6	**0.21**	6	0.29	4	0.28	4	0.28	4
Pendigits	10	16	0.02	15	0.02	**10**	0.02	15	0.02	15	0.02	15

Table 1. Performance of the proposed method compared to all-at-once AIDA and the pairwise transformation methods proposed in [16]. Shown scores are the best average error rate obtained with the corresponding method and the dimension d at which it occurs. h is the number of classes in the problem and p the original dimension of the data.

3.3 Assessment of the variable selection procedure

As the addition of variable selection to the dimension reduction task is a contribution in its own right, it is fair to assess its performance more deeply. With this aim, we carried out a simple simulation experiment to study the accuracy of the procedure in choosing the right active variables. For simplicity we ran a regression experiment with data generated according to

$$Y = 4a(\boldsymbol{\beta}_1^T\mathbf{X})^2 + 0.75\sin(\boldsymbol{\beta}_2^T\mathbf{X}).$$

Notice that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1\boldsymbol{\beta}_2)$ is a basis matrix for the smallest dimension reduction subspace. The dimension of the data was set to $p = 20$ and the columns of $\boldsymbol{\beta}$ were generated so that the first three rows were nonzero in $\boldsymbol{\alpha}_1$, and the first row and the last two were nonzero in $\boldsymbol{\beta}_2$. All the other elements are null. The number of actives variables is then five out of twenty. To assess the performance of the selection procedure, we used the following measures proposed in [18]: r_1, the average fraction of nonzero rows in $\widehat{\boldsymbol{\beta}}$ corresponding to relevant variables; r_2, the average fraction of zero rows in $\widehat{\boldsymbol{\beta}}$ corresponding to nonrelevant variables; r_3, the average fraction of perfect variable selection. Obtained results over 200 replicates of the experiment are shown in Table 2. It can be seen that the method does an excellent job in identifying the irrelevant variables (r_2) and just misses some relevant variables in a few times (r_1). The fraction of times it distinguishes perfectly between the actives and irrelevant variables is always equal or greater than 94%.

4 Conclusions

We have presented a method for linear feature extraction based on a collection of projections for pairs of classes. The method is motivated for the approach of model-based sufficient dimension reductions, which allows to build a probabilistic model for the data using a latent variable. Unlike previous multiple subspace

measure	a						
	1	3	5	7	9	11	13
r_1	1	0.97	0.77	0.78	0.78	0.81	0.80
r_2	0.98	0.98	1	1	1	1	1
r_3	0.98	0.98	0.94	0.94	0.95	0.95	0.95

Table 2. Accuracy of the proposed method for joint variable selection and dimension reduction.

projection approaches based on pairwise transformations, evaluation of a likelihood allows for class assignment using simple Bayes rule instead of voting strategies or decision trees. As a by-product of this contribution, a regularized version of AIDA have been derived, which allows for simultaneous dimension reduction and variable selection preserving the equivariance property of AIDA. Experiments with real data sets have shown that the proposed pairwise method outperforms the all-at-once approach when the number of classes grows and that its performance is at least as good as that from existing alternatives based on pairwise linear dimension reduction.

Acknowledgments

The author wants to thank Dr. Diego Milone for insightful discussions on the problem of multiple subspace projections and for many suggestions to improve this manuscript. He is also indebted with Prof. R. Dennis Cook and with Dr. Liliana Forzani for allowing him to use partial results from an ongoing colaboration, obtained during his short-term visit at the University of Minnesota. This research was carried out with financial support from UNL, ANPCyT and CONICET.

References

1. Jain, A., Duin, R., Mao, J.: Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 4–37
2. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, New York (1990)
3. Jolliffe, I.: Principal Component Analysis, Second Edition. Springer, New York (2002)
4. Fisher, R.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7** (1936) 179–188
5. Petridis, S., Perantonis, S.: On the relation between discriminant analysis and mutual information for supervised linear feature extraction. Pattern Recognition **37** (2004) 857–874
6. Nenadic, Z.: Information discriminant analysis: Feature extraction with an information-theoretic objective. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(8) (2007) 1394–1407

7. Das, K., Nenadic, Z.: Approximate information discriminant analysis: A computationally simple heteroscedastic feature extraction technique. Pattern Recognition **41**(5) (2008) 1565–1574
8. Loog, M., Duin, R.P.W.: Linear dimensionality reduction via a heteroscedastic extension of lda: The chernoff criterion. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(6) (2004) 732–739
9. Rueda, L., Herrera, M.: Linear dimensionality reduction by maximizing the chernoff distance in the transformed space. Pattern Recognition **41** (2008) 3138–3152
10. Li, K.: Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association **86** (1991) 316–342
11. Cook, R.: Regression Graphics. Wiley, New York (1998)
12. Cook, R., Yin, X.: Dimension reduction and visualization in discriminant analysis (with discussion). Australia New Zeland Journal of Statistics (1994) 18–25
13. Cook, R., Forzani, L.: Likelihood-Based sufficient dimension reduction. Journal of the American Statistical Association **104**(485) (2008) 197–208
14. Tomassi, D., Forzani, L., Milone, D., Cook, R.: Likelihood-based sufficient dimension reduction for statistical pattern recognition. Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (2011)
15. Loog, M., Duin, R.P.W., Haeb-Umbach, R.: Multiclass linear dimension reduction by weighted pairwise fisher criteria. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(7) (2001) 762–766
16. Rueda, L., Oommen, B., Henriquez, C.: Multi-class pairwise linear dimensionality reduction using heteroscedastic schemes. Pattern Recognition **43** (2010) 2456–2465
17. Li, L.: Sparse sufficient dimension reduction. Biometrika **94** (2007) 603–613
18. Chen, B., Zou, C., Cook, R.: Coordinate-independent sparse sufficient dimension reduction and variable selection. Annals of Statistics **38** (2010) 3696–3723
19. Cook, R., Forzani, L., Tomassi, D.: LDR: a package for likelihood-based dimension reduction. Journal of Statistical Software **39** (2011)
20. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B **58**(2) (1996) 267–288
21. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B **68** (2007) 49–67