LINEAR PROJECTION METHODS
An Experimental Study for Regression Problems[*]

Carlos Pardo-Aguilar[1], José F. Diez-Pastor[1], Nicolás García-Pedrajas[2], Juan J. Rodríguez[1]
and César García-Osorio[1]
[1]*Departament of Civil Engineering, University of Burgos, Avd. Cantabría s/n, Burgos, Spain*
[2]*Department of Computing and Numerical Analysis, University of Córdoba, Campus de Rabanales, Córdoba, Spain*

Keywords: Localized sliced inverse regression, Linear discriminant analysis for regression, Weighted principal components analysis, Nonparametric discriminant regression analysis, Localized principal Hessian directions, Hybrid discriminant analysis for regression.

Abstract: Two contexts may be considered, in which it is of interest to reduce the dimension of a data set. One of these arises when the intention is to mitigate the curse of dimensionality, when the data set will be used for training a data mining algorithm with a heavy computational load. The other is when one wishes to identify the data set attributes that have a stronger relation with either the class, if dealing with a classification problem, or the value to be predicted, if dealing with a regression problem. Recently, various linear regression projection models have been proposed that attempt to conserve those directions that show the highest correlation with the value to be predicted: *Localized Slices Inverse Regression*, *Weighted Principal Component Analysis* and *Linear Discriminant Analysis for regression*. However, the papers that have presented these methods use only a small number of data sets to validate their smooth functioning. In this research, a more exhaustive study is conducted using 30 data sets. Moreover, by applying the ideas behind these methods, a further three new methods are also presented and included in the comparative study; one of which is competitive with the methods recently proposed.

1 INTRODUCTION

Very frequently, the intrinsic dimension of a data set —the number of variables or characteristics needed to represent it— is lower or even much lower than the real dimension shown by the data set. One perfect illustration of this is the example provided by (Tenenbaum et al., 2000), in which a data set consisting of photographs of hands may be characterized by two variables (intrinsic dimension 2) —wrist rotation and the angle of finger extension— despite its dimension being 4096 (given that there are 64×64 pixel images). In other words, maintaining a constant distance and similar lighting conditions for the photograph, all the images of the hands taken with the same rotation and finger extension will be approximately equal, such that the value of 4096 pixels may be determined fairly easily, knowing only those two values.

In the field of data mining, there is great interest in the study of methods that will identify the intrinsic dimension of data sets. This has given rise to the area of *manifold learning* (Tenenbaum et al., 2000; Roweis and Saul, 2000; Lee and Verleysen, 2007), which is usually centred on the determination of non-linear relations, and methods for feature selection and extraction (Guyon and Elisseeff, 2003; Liu and Yu, 2005), in which the linear relations are usually more interesting, as they are easier to interpret.

Interest in discovering the intrinsic dimension is twofold. On the one hand, reducing the dimension of the data set mitigates the effects of the *curse of dimensionality*, a term coined by Richard Bellman to describe the fact that some problems become intractable as the number of variables increase. As regards data mining problems, this is related to the fact that the number of necessary instances to solve a learning problem grows exponentially with the number of variables. On the other hand, to possess knowledge of the variables, on which the values to be predicted are more directly dependent, is in itself very valuable for

[*]This work was supported by the projects MAGNO2008-1028-CENIT, TIN2008-03151 and TIN2011-24046 of the Spanish Ministry of Science and Innovation.

Pardo-Aguilar C., F. Diez-Pastor J., García-Pedrajas N., J. Rodríguez J. and García-Osorio C. (2012).
LINEAR PROJECTION METHODS - An Experimental Study for Regression Problems.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 198-204
DOI: 10.5220/0003763301980204

the data analyst.

A very simple way of reducing the dimension of a set is to find a projection matrix that projects the data set onto a lower dimensional space. In other words, there are lower numbers of variables in the new data set that represent a linear combination of those in the initial data set. The difficulty resides in finding a projection that retains some interesting characteristics of the initial data set. In this work, our interest lies in these types of linear projection methods. Among the non-supervised methods in this category, the most widely used is without a doubt *Principal Component Analysis* (PCA) (Jolliffe, 1986), which attempts to preserve the variance of the data set. The most well known among the supervised classification-based methods are *Linear Discriminant Analysis* (LDA) (Fisher et al., 1936) and *Nonparametric Discriminant Analysis* (NDA) (Fukunaga and Mantock, 1983), both of which try to achieve a projection that maximizes the separation between classes and minimizes the dispersion of the instances within their own class. A third supervised method with the same objective is *Hybrid Discriminant Analysis* (HDA) (Tian et al., 2005), which is proposed as a mixed method that combines PCA and LDA. Finally, supervised methods also exist, oriented towards regression, that attempt to find the linear relation that has the strongest correlation with the dependent variable. Among these, it is worth noting *Sliced Inverse Regression* (SIR) (Li, 1991) and *Principal Hessian Directions* (PHD) (Li, 1992), and the most recent, *Localized SIR* (LSIR) (Wu et al., 2008), *LDA for regression* (LDAr) and *Weighted PCA* (WPCA) (Kwak and Lee, 2010).

Our work here is centred on linear projection methods for regression. An experimental study of LSIR, LDAr and WPCA is completed, given that the articles in which these methods were presented only used two real data sets in the case of LSIR, and three data sets for the two final methods. Furthermore, using the ideas in these methods, new methods are also presented that are included in the comparative study.

The rest of the article is structured as follows. Section 2 presents the details of the methods, as well as a unifying conceptual framework. Section 3 presents the new methods. Section 4 explains the details of how the study was made and presents the results. Finally, section 5 summarises the conclusions.

2 REVIEW OF BACKGROUND

Consider a set of n data and values $\{\mathbf{x}_i, y_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathcal{R}^{d\times 1}$ and $y_i \in \mathcal{R}$ (in a more general context, they would be considered pairs $\{\mathbf{x}_i, \mathbf{y}_i\}$ with $\mathbf{y}_i \in \mathcal{R}^{t\times 1}$, but only the data sets for which $t = 1$ are considered in this article). The question is how to find a linear combination of attributes $f_j = \mathbf{w}_j^T\mathbf{x}$ that will give rise to the characteristics f_j that best explain the value, y, that is to be predicted.

All of the following methods that are presented may be proposed as an optimization problem, in which the function to maximize is of the form:

$$J(W) = \frac{|W^TAW|}{|W^TBW|} \quad (1)$$

in which, the columns of the optimum solution, W, may be obtained by solving the following generalized eigenvalue problems:

$$A\mathbf{w}_k = \lambda_k B\mathbf{w}_k, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \quad (2)$$

It may be solved as $B^{-1}Aw_k = \lambda_k \mathbf{w}_k$, a classic eigenvalue problem that can be sensitive to poor conditioning of B (when the determinant is close to zero).

What changes from one method to the other is the way in which matrices A and B, which appear as numerator and denominator, are calculated.

2.1 Unsupervised Linear Projection (PCA)

In the case of PCA (Jolliffe, 1986), matrix B in equation 2 is nothing other than the identity matrix. Matrix A is the covariance matrix:

$$A = S_x = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$B = I \quad \text{(identity matrix)}$$

where, $\bar{\mathbf{x}} = (1/n)\sum_{i=1}^n \mathbf{x}_i$ is the average of the $\mathbf{x_i}$. Equation 2 is therefore reduced to a classic eigenvalue problem:

$$S_x\mathbf{w}_k = \lambda_k \mathbf{w}_k, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

2.2 Methods of Supervised Linear Projection for Classification

2.2.1 Linear Discriminant Analysis

In LDA (Fisher et al., 1936), the numerator matrix is known as the *between-covariance matrix* and that of the denominator the *within-covariance matrix* defined as:

$$A = S_b = \frac{1}{n}\sum_{c=1}^{N_c} n_c(\bar{\mathbf{x}}_c - \bar{\mathbf{x}})(\bar{\mathbf{x}}_c - \bar{\mathbf{x}})^T$$

$$B = S_w = \frac{1}{n}\sum_{c=1}^{N_c}\sum_{i\in\text{class } c}(\mathbf{x}_i - \bar{\mathbf{x}}_c)(\mathbf{x}_i - \bar{\mathbf{x}}_c)^T$$

where, N_c is the number of classes, n_c the number of instances in class c, and $\bar{\mathbf{x}}_c = (1/n_c)\sum_{i\in \text{class } c} \mathbf{x}_i$ is the mean of the instances of class c. Matrix S_w may be considered as the weighted sum of the covariance matrices for each class.

2.2.2 Non-parametric Discriminant Analysis

In NDA (Fukunaga and Mantock, 1983), the LDA matrices S_b and S_w are replaced by the following ones:

$$A = S_b^{\text{NDA}} = \sum_{c=1}^{N_c} P_c \sum_{\substack{d=1\\ d\neq c}}^{N_c} \sum_{i\in \text{class } c} \frac{w_i^{(c,d)}}{n_c} D_d(\mathbf{x}_i^{(c)}) \cdot D_d(\mathbf{x}_i^{(c)})^T$$

$$B = S_w^{\text{NDA}} = \sum_{c=1}^{N_c} P_c \sum_{i\in \text{class } c} \frac{w_i^{(c,c)}}{n_c} D_c(\mathbf{x}_i^{(c)}) \cdot D_c(\mathbf{x}_i^{(c)})^T$$

where, N_c is the number of classes, n_c is the number of instances in class c, P_c is the *a priori* probability of class c, $D_d(\mathbf{x}_i^{(c)}) = \mathbf{x}_i^{(c)} - M_d^k(\mathbf{x}_i^{(c)})$ the difference between instance $\mathbf{x}_i^{(c)}$ and $M_d^k(\mathbf{x}_i^{(c)}) = (1/k)\sum_{t=1}^{k} \mathbf{x}_{\text{tNN}}^{(d)}$, the mean of the nearest neighbours k in class d to the instance $\mathbf{x}_i^{(c)}$ in class c, its "k-NN local mean", and the weighting factor $w_i^{(c,d)}$, which depends on a control parameter ρ (with a value of between 0 and infinite), is defined as:

$$w_i^{(c,d)} = \frac{\min\left\{\text{dist}(\mathbf{x}_i^{(c)}, \mathbf{x}_{\text{kNN}}^{(c)})^\rho, \text{dist}(\mathbf{x}_i^{(c)}, \mathbf{x}_{\text{kNN}}^{(d)})^\rho\right\}}{\text{dist}(\mathbf{x}_i^{(c)}, \mathbf{x}_{\text{kNN}}^{(c)})^\rho + \text{dist}(\mathbf{x}_i^{(c)}, \mathbf{x}_{\text{kNN}}^{(d)})^\rho}$$

where, $\text{dist}(\mathbf{x}_i^{(c)}, \mathbf{x}_{\text{kNN}}^{(d)})$ is the distance of $\mathbf{x}_i^{(c)}$ in class c to its k-nth nearest neighbour in class d.

2.2.3 Hybrid Discriminant Analysis

This method is presented in (Tian et al., 2005) as a combination of PCA and LDA. The numerator and denominator matrices of equation 1 are obtained by a linear combination of the corresponding PCA and LDA matrices:

$$A = (1-\lambda)S_b + \lambda S_x$$
$$B = (1-\eta)S_w + \eta I$$

where, I is the identity matrix. For $\lambda = 1$ and $\eta = 1$, HDA is reduced to PCA, for $\lambda = 0$ and $\eta = 0$, HDA corresponds entirely to LDA, we can obtain projections for other values with intermediary characteristics between both methods. In addition, we obtain a simple regularization of B from $\eta > 0$.

2.3 Supervised Linear Projection Methods for Regression

2.3.1 Sliced Inverse Regression

In SIR (Li, 1991), the data set is at first ordered in accordance with the values of y and divided into L slices[2]. The matrices A and B are then defined as:

$$A = S_\eta = \frac{1}{n}\sum_{l=1}^{L} n_l(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})^T$$
$$B = S_x \quad \text{(covariance matrix)}$$

where, n_l is the number of instances in the slice l and $\bar{\mathbf{x}}_l = (1/n_l)\sum_{i\in \text{slice } c} \mathbf{x}_i$ is the mean for each slice. If S_η is calculated on the basis of the data set once it is sphered[3], the solution to equation 2 could be treated as a classic eigenvalue problem:

$$S_\eta \mathbf{w}_k = \lambda_k \mathbf{w}_k, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

2.3.2 Localized SIR

In this variant of SIR (Wu et al., 2008), the means of the slices are replaced by local means.

$$A = S_\eta^{\text{loc}} = \frac{1}{n}\sum_{i=1}^{n} (\bar{\mathbf{x}}_{i,\text{loc}} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{i,\text{loc}} - \bar{\mathbf{x}})^T$$
$$B = S_x \quad \text{(covariance matrix)}$$

where, $\bar{\mathbf{x}}_{i,\text{loc}} = (1/k)\sum_{j\in I_i} \mathbf{x}_j$, in which I_i is the set of indices of the nearest k neighbours of $\mathbf{x}_i$ in its same slice, such that the method now depends on two parameters: the number of slices, L, and the number of nearest neighbours, k.

2.3.3 Principal Hessian Directions

This method (Li, 1992; Li, 2000) is based on resolving a problem of eigenvalues for which it is necessary to calculate the Hessian matrix mean $\overline{H}$, which is related to the weighted covariance matrix $S_{yxx} = E\{(Y-\bar{y})(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T\}$ through equality $\overline{H} = S_x^{-1} S_{yxx} S_x^{-1}$. From the point of view of the unified approach that we propose, this method could be likened to solving the generalized eigenvalue problem of equation 2 in which matrices A and B would be:

$$A = S_{yxx} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

[2] Note that this process may be seen as a discretization of the values of y.

[3] In other words, after projecting it onto the principal components and dividing each variable by the square root of the corresponding eigenvalue.

$$B = S_x \quad \text{(covariance matrix)}$$

As in the case of SIR, if the data set is sphered before calculation of S_{yxx}, the solution could also be obtained by solving the eigenvalue problem:

$$S_{yxx}\mathbf{w}_k = \lambda_k \mathbf{w}_k, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

2.3.4 Weighted PCA

In the weighted PCA (Kwak and Lee, 2010), as in other methods, matrix B is the covariance matrix, S_x. Matrix A is defined as:

$$A = S_{yx} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} g(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$$

where, $g(\cdot)$ is a positive and symmetric function the value of which does not decrease when the absolute value of its argument increases. Two possible examples would be $g(x) = |x|$ and $g(x) = \sqrt{|x|}$, which can be generalized as a function $g(x) = |x|^p$, in which p would be a parameter of the method, and the earlier ones would be special cases for $p = 1$ y $p = 0.5$. Moreover, when $p = 0$, matrix S_{yx} would be equivalent to S_x

2.3.5 Linear Discriminant Analysis for Regression (LDAr)

This method (Kwak and Lee, 2010) based on LDA, uses the following variants of matrices S_b y S_w:

$$A = S_{br} = \frac{1}{n_b} \sum_{(i,j)\in I_{br}} f(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$$

$$B = S_{wr} = \frac{1}{n_w} \sum_{(i,j)\in I_{wr}} f(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$$

where, the sets of pairs of indices I_{br} and I_{wr} are defined as:

$$I_{br} = \{(i,j) : |y_i - y_j| \geq \tau, i < j\}$$
$$I_{wr} = \{(i,j) : |y_i - y_j| < \tau, i < j\}$$

n_b and n_w are the cardinalities of these sets, and the function $f(\cdot)$ could be any of the following $f(x) = ||x| - \tau|$ or $f(x) = \sqrt{||x| - \tau|}$; as in WPCA, it can be generalized as a function $f(x) = ||x| - \tau|^p$, in which p would be a parameter of the method, and the earlier ones would be special cases for $p = 1$ and $p = 0.5$.

3 PROPOSALS FOR NEW METHODS

In this section, new supervised projection methods are proposed to approach regression problems, by adapting some of the ideas of the earlier methods.

3.1 Localized Principal Hessian Directions

This method is proposed as an extension of PHD, as in Local SIR, the local information is used at each instance. The new matrix for A would be:

$$A = S_{yxx}^{\text{loc}} = \frac{1}{n} \sum_{i=1}^{n} (\bar{y}_{i,\text{loc}} - \bar{y})(\bar{\mathbf{x}}_{i,\text{loc}} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{i,\text{loc}} - \bar{\mathbf{x}})^T$$

3.2 Hybrid Discriminant Analysis for Regression

This method proposes to use the same idea as in HDA, but using the matrices WPCA and LDAr,

$$A = (1-\lambda)S_{br} + \lambda S_{yx}$$
$$B = (1-\eta)S_{wr} + \eta I$$

in which, I is the identity matrix. Thus, for $\lambda = 1$ and $\eta = 1$, HDAr is reduced to WPCA, for $\lambda = 0$ y $\eta = 0$, HDA corresponds entirely to LDAr, for other values we can obtain projections with intermediate characteristics between both methods. In addition, with $\eta > 0$ a simple regularization of B is obtained.

3.3 Sliced Nonparametric Discriminant

This proposal consists in using NDA, but after completing the discretization of the values of the dependent variable, as was done for SIR. After discretization, the instances that belong to the slices may be considered classes, which allows classic NDA to be applied.

4 COMPARATIVE STUDY

4.1 Validity of the New Methods

In the first place, the validity of the new proposals will be tested by using a pair of artificial data sets, the structures of which are known, for which reason it is easy to validate whether the methods identify the structure. The same artificial data sets of (Kwak and Lee, 2010) were used, both having 1000 instances

Table 1: Absolute value of the cosine of the angle between the optimum direction and that found by the supervised projection methods for regression.

	linear	non-linear
Method	problem	problem
SIR	0.9999332	0.9863278
LSIR	0.9997051	0.9953802
WPCA	0.9999590	0.9520269
LDAr	0.9999995	0.9912704
PHD	0.3764493	0.8009470
LPHD	0.4467953	0.3439180
HDAr55	0.9997744	0.9139320
HDAr83	0.9999450	0.9486312
HDAr38	0.9984839	0.8658341
SNDA	0.9999735	0.9979509

and 5 dimensions that follow a normal distribution of mean 0 and variance 1. In one of them, the output variable is linearly dependent on two of the attributes $y = 2x_1 + 3x_3$, such that the direction of optimal projection would be $\mathbf{w}_1 = (2,0,3,0,0)^T$; in the other, the relation with the ouput is not linear $y = \sin(x_2 + 2x_4)$, and the optimal projection is $\mathbf{w}_1 = (0,1,0,2,0)^T$.

As a reference, the results were also calculated for the other methods. The number of slices was 12 (for SIR, LSIR and SNDA). The number of neighbours for the localized methods (LSIR and LPHD) was 5. A value of 0.5 for parameter p was used in WPCA, LDAr and HDAr. The value of τ was 0.3 in LDAr and HDAr. Three configurations —$(\lambda = 0.5, \eta = 0.5)$, $(\lambda = 0.8, \eta = 0.3)$, and $(\lambda = 0.3, \eta = 0.8)$— were tested for the HDAr method, labelled in the table as HDAr55, HDAr83 and HDAr38, respectively.

In Table 1, the absolute value of the cosine of the angle between the optimum directions and the direction found by each different method is shown. It can be seen that both SNDA as well as the various configurations of HDAr achieve good approximations to the optimum, both in the linear as well as the non-linear case. In the linear case, the best approximation is given by LDAr, followed closely by SNDA. In addition, the local version of PHD is able to slightly improve PHD in the linear case, although its results are very bad in the non-linear case. The best approximation in the non-linear problem is given by SNDA.

4.2 Experimental Comparison

The regression data sets shown in Table 2 were used (all are available in the arff Weka format[4]), the majority of which are taken from the UCI machine learning respository (Frank and Asuncion, 2010) and from the collection of Luis Torgo[5].

The results of the prediction were obtained with the same regressor used in (Kwak and Lee, 2010), a weighted nearest neighbour regressor, which normalizes the attributes in the range $[0,1]$ and uses the weighting function $q(\mathbf{x},\mathbf{x}_i) = 1/(1+\sqrt{||\mathbf{x}-\mathbf{x}_i||})$ and 5 neighbours.

The effect of projecting onto dimensions 1, 2, 3, $0.5d$, $0.75d$ and d has been tested, where d is the dimension of the data set and the non integer values were rounded to the nearest integer.

In the experiments, each data set was randomly divided into 90% for training and 10% for test, and this was repeated 10 times, calculating the mean error of each repetition, which was measured as the square root of the mean quadratic error.

For each of the dimensions, the methods were organized in accordance with the regressor results, assigning range 1 to the best, range 2 to the following and so on, successively (Demšar, 2006). The ranges obtained for all the data sets were used to calculate the average ranges, which are those shown in Table 3 (a). One of the proposed methods, SNDA, was the best method when used to project the data set without reducing its dimension and when used to reduce the dimension to 75% of its original size. It also remains among the first three in another three cases. Moreover, it may be seen that the HDAr proposal is not a very good idea, given that in no case was it able to outperform WPCA and LDAr, simultaneously. Neither does the localization of PHD appear to contribute much, even though it outperformed PHD in two cases, results were worse in the others. From among the two best methods, WPCA appears to be the best method for all the dimensions, at all times better than LDAr, which contradicts the conclusions of the designers of this method, who established in (Kwak and Lee, 2010) that LDAr was better than WPCA, although they used only three data sets.

Finally, the ranges of the 66 combinations of methods and possible dimensions were also globally calculated (6 different projection dimensions × 11 methods) together with the result of applying the base regressor directly to the data set (denoted in the tables as ORI). These results are shown in Table 3 (b). A surprising result occurred here, as the majority of the methods were unable to improve on the results of applying the base regressor directly to the initial data set without an associated reduction in dimensionality.

[4] http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html

[5] http://www.liaad.up.pt/∼ltorgo/Regression/DataSets.html

Table 2: Data sets used in the experiments.

#N: Numerical attributes, #D: Discrete attributes, #I: Instances.

Dataset	#N	#D	#I	Dataset	#N	#D	#I
abalone	7	1	4177	housing	12	1	506
auto93	16	6	93	hungarian	6	7	294
auto-horse	17	8	205	lowbwt	2	7	189
auto-mpg	4	3	398	machine-cpu	6	0	209
auto-price	15	0	159	meta	19	2	528
bodyfat	14	0	256	pbc	10	8	418
breast-tumor	1	8	286	pharynx	1	10	195
cholesterol	6	7	303	pw-linear	10	0	200
cleveland	6	7	303	sensory	0	11	576
cloud	4	2	108	servo	0	4	167
cpu	6	1	209	stock	9	0	950
cpu-small	12	0	8192	strike	5	1	625
delta-ailerons	5	0	7129	triazines	60	0	186
echo-months	6	3	130	veteran	3	4	137
fishcatch	5	2	158	wisconsin	32	0	194

Table 3: Ranking of the methods.

1		2		3		*.5d*		*.75d*		*d*	
WPCA	(4.27)	SIR	(4.33)	WPCA	(3.80)	WPCA	(4.30)	SNDA	(4.80)	SNDA	(4.70)
SIR	(4.53)	WPCA	(4.47)	SIR	(5.27)	SNDA	(4.73)	WPCA	(4.87)	WPCA	(5.07)
SNDA	(5.37)	SNDA	(4.90)	LSIR	(5.27)	LDAR	(4.97)	SIR	(4.97)	HDAr83	(5.60)
LDAR	(5.47)	LDAR	(5.17)	LDAR	(5.30)	LSIR	(5.07)	LSIR	(5.10)	SIR	(5.70)
HDAr55	(5.80)	HDAr83	(5.57)	SNDA	(5.40)	SIR	(5.10)	LDAR	(5.60)	LDAR	(5.73)
HDAr83	(5.83)	HDAr55	(6.13)	HDAr83	(5.60)	HDAr83	(6.40)	HDAr83	(5.73)	LSIR	(6.00)
LSIR	(6.33)	LSIR	(6.67)	HDAr55	(6.27)	PCA	(6.47)	HDAr55	(6.37)	PCA	(6.07)
HDAr38	(6.40)	HDAr38	(6.67)	HDAr38	(6.57)	HDAr55	(6.60)	PCA	(6.57)	HDAr55	(6.20)
PHD	(7.17)	PCA	(6.67)	PCA	(7.03)	PHD	(7.37)	HDAr38	(7.23)	HDAr38	(6.37)
PCA	(7.37)	PHD	(7.63)	LPHD	(7.73)	LPHD	(7.37)	LPHD	(7.33)	PHD	(7.17)
LPHD	(7.47)	LPHD	(7.80)	PHD	(7.77)	HDAr38	(7.63)	PHD	(7.43)	LPHD	(7.40)

(a) Ranking for each one of the dimensions under consideration.

${}_{3}$WPCA	(20.77)	${}_{.5d}$WPCA	(22.27)	$\underline{{}_{.75d}\text{SNDA}}$	(23.07)	${}_{.75d}$WPCA	(23.43)
$\underline{{}_{d}\text{SNDA}}$	(23.53)	**ORI**	(24.00)	${}_{d}$WPCA	(24.47)	${}_{.5d}$SNDA	(24.90)
${}_{.75d}$LSIR	(25.57)	${}_{.5d}$SIR	(25.73)	${}_{.75d}$SIR	(25.97)	${}_{.5d}$LDAR	(26.13)
${}_{.5d}$LSIR	(26.37)	${}_{d}$SIR	(27.03)	${}_{2}$WPCA	(27.40)	${}_{d}$LDAR	(27.43)
${}_{3}$SIR	(27.73)	${}_{d}$HDAr83	(27.77)	${}_{.75d}$LDAR	(28.00)	${}_{2}$SIR	(28.33)
${}_{3}$SNDA	(28.63)	${}_{d}$LSIR	(28.70)	${}_{d}$PCA	(28.77)	${}_{3}$LDAR	(28.97)
${}_{2}$SNDA	(29.27)	${}_{3}$LSIR	(29.60)	${}_{d}$HDAr55	(30.30)	${}_{.75d}$HDAr83	(30.37)
${}_{d}$HDAr38	(31.07)	${}_{3}$HDAr83	(31.83)	${}_{2}$LDAR	(31.90)	${}_{.5d}$HDAr83	(32.60)
${}_{.75d}$HDAr55	(32.67)	${}_{.75d}$PCA	(33.07)	${}_{.5d}$PCA	(33.63)	${}_{.5d}$HDAr55	(33.73)
${}_{3}$HDAr55	(33.77)	${}_{2}$HDAr83	(34.30)	${}_{d}$PHD	(35.63)	${}_{1}$SIR	(35.67)
${}_{.75d}$HDAr38	(36.00)	${}_{3}$HDAr38	(36.17)	${}_{d}$LPHD	(36.30)	${}_{.75d}$PHD	(36.47)
${}_{2}$HDAr55	(36.67)	${}_{1}$WPCA	(36.87)	${}_{.75d}$LPHD	(37.13)	${}_{1}$SNDA	(37.57)
${}_{3}$PCA	(38.47)	${}_{.5d}$HDAr38	(38.50)	${}_{2}$LSIR	(38.67)	${}_{.5d}$LPHD	(38.83)
${}_{2}$HDAr38	(39.40)	${}_{.5d}$PHD	(39.50)	${}_{2}$PCA	(39.80)	${}_{1}$LDAR	(40.37)
${}_{1}$HDAr55	(43.00)	${}_{3}$PHD	(43.30)	${}_{1}$HDAr83	(44.10)	${}_{3}$LPHD	(44.67)
${}_{1}$LSIR	(46.47)	${}_{1}$HDAr38	(46.63)	${}_{2}$PHD	(47.13)	${}_{2}$LPHD	(50.07)
${}_{1}$PCA	(53.10)	${}_{1}$PHD	(53.60)	${}_{1}$LPHD	(54.87)		

(b) Global ranking.

5 CONCLUSIONS

This article has described some of the classic methods of obtaining supervised linear projections for regression problems, together with some new proposals, by using the common conceptual framework of solving generalized eigenvalue problems.

After testing the validity of the new proposals on a pair of artificial data sets with a well known structure, an experimental study was conducted of all the methods using 30 data sets. The most surprising conclusion of this study was that many of the projection methods are unable to improve on the regression results of the regressor used as the basis for the study; a weighted nearest neighbour regressor.

SNDA, one of the new methods proposed in the article, has a performance comparable to WPCA for low dimensions, and it shown to perform better at higher dimensions. It is also worth noting that WPCA performs better than LDAr, which contradicts the results of (Kwak and Lee, 2010), in which LDAr outperformed WPCA.

Possible future work could determine whether the conclusions obtained here might extend to cases in which other regressors are used, as well as considering the effect of the parameters of the methods. Another interesting line of work would be to use these methods as inductors of diversity in the algorithms for building ensemble of regressors. This would be motivated by the results obtained for Rotation Forest using PCA (Rodríguez et al., 2006), or Nonlinear Boosting Projection using NDA (García-Pedrajas and García-Osorio, 2011). It is tempting to think that the substitution of PCA and NDA in regression problems for some of the proposed methods in this article could improve the results.

REFERENCES

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

Fisher, R. et al. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Frank, A. and Asuncion, A. (2010). UCI machine learning repository. Stable URL: http://archive.ics.uci.edu/ml/.

Fukunaga, K. and Mantock, J. (1983). Nonparametric discriminant analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6(5):671–678.

García-Pedrajas, N. and García-Osorio, C. (2011). Constructing ensembles of classifiers using supervised projection methods based on misclassified instances. *Expert Systems with Applications*, 38(1):343–359. DOI: 10.1016/j.eswa.2010.06.072.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Jolliffe, I. (1986). *Principal Component Analysis*. Springer-Verlag.

Kwak, N. and Lee, J.-W. (2010). Feature extraction based on subspace methods for regression problems. *Neurocomputing*, 73(10-12):1740–1751.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Springer.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 84(420):1025–1039. Stable URL: http://www.jstor.org/stable/229064.

Li, K. C. (2000). High dimensional data analysis via the SIR/PHD approach. Available at http://www.stat.ucla.edu/∼kcli/sir-PHD.pdf.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transanction on Knowledge and Data Engineering*, 17:491–502.

Rodríguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323.

Tian, Q., Yu, J., and Huang, T. S. (2005). Boosting multiple classifiers constructed by hybrid discriminant analysis. In Oza, N. C., Polikar, R., Kittler, J., and Roli, F., editors, *Multiple Classifier Systems*, volume 3541 of *Lecture Notes in Computer Science*, pages 42–52, Seaside, CA, USA. Springer.

Wu, Q., Mukherjee, S., and Liang, F. (2008). Localized sliced inverse regression. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *NIPS*, pages 1785–1792. MIT Press.