Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 68(2) 2014 273-283

Applications of Sufficient Dimension Reduction Algorithms on Non-elliptical Data

Andreas Artemiou
School of Mathematics, Cardiff University, Senghennydd Road, Cardiff, Wales CF24 4AG, UK

Received 25 February 2014; Revised 21 May 2014; Accepted 26 May 2014

SUMMARY

Sufficient dimension reduction (SDR) is a class of supervised dimension reduction techniques which generally perform much better than unsupervised dimension reduction techniques like Principal Component Analysis (PCA). In this paper we present classic methodology in the SDR framework that is based on inverse moments and we discuss the theoretical assumptions. At the end we demonstrate the advantage of a recently introduced method known as Principal Support Vector Machine (PSVM) in the presence of predictors which violate the theoretical assumption of ellipticity of the marginal distribution.

Keywords: Suffcient dimension reduction, Categorical predictors, Sliced inverse regression, Principal support vector machine, Principal component analysis.

1. INTRODUCTION

The increase in computing storage capabilities and the low cost of data storage have made high dimensional data a daily phenomenon in a number of sciences, like Biosciences, Geosciences, Medical Sciences, Engineering and Agriculture. These high dimensional datasets create huge challenges for scientists to find effective ways to analyze them. Traditional statistical techniques were derived during an era when datasets were small and most of them lack the ability to work as effectively in the massive datasets we collect today. Therefore, there is a need for techniques that work effectively with large datasets and methodology that can effectively identify the most important features in a large datasets. In this work we will focus on the latter and we will present methods for dimension reduction in a regression setting.

Dimension reduction in regression, goes back to the early 20th century and the introduction by Pearson (1901) of Principal Component Analysis (PCA) which was later formalized by Hotelling (1933). PCA extracted the axis that had the most variation in the cloud of data points. PCA is a simple technique which works well, but at the same time has some shortcomings. The most important shortcoming of PCA is the fact that it is an unsupervised dimension reduction technique and nothing ensures that the extracted features in a regression setting are actually features correlated with the response. This led to a long debate among researchers (see Cook 2007) as some of them (see Joliffe 1982 and Hadi and Ling 1998) showed examples where PCA failed to capture the features most correlated with the response. Artemiou and Li (2009, 2013) and Ni (2011) demonstrated that PCA most of the times captures features that have the highest correlation with the response under different models but at the same time emphasized the fact that there was an unmeasurable risk that PCA will not capture the desired features, but rather capture features uncorrelated with the response.

E-mail address: artemioua@cardiff.ac.uk

This shortcoming of PCA led to a number of solutions in the literature such as projection pursuit and sufficient dimension reduction (SDR). In this work we will focus on the most well known methods in SDR as well as a recent method which has garnered a lot of interest due to its robustness to violation of theoretical assumptions that are commonly used in this framework.

In SDR we assume that we have a univariate (for simplicity) response variable Y and a p dimensional predictor vector $\boldsymbol{X}$. The objective is to estimate a set of d features (where $d \leq p$) without losing information about the conditional distribution of the $Y|\boldsymbol{X}$. In other words we are trying to estimate a $p \times d$ matrix β which satisfies

$$Y \perp\!\!\!\perp \boldsymbol{X} \,|\, \beta^T \boldsymbol{X}. \tag{1}$$

This is known as linear sufficient dimension reduction since the extracted features are linear functions of the original predictors. The space spanned by the columns of β is called a Dimension Reduction Subspace (DRS). There are many β's that satisfy model (1) - for example the identity matrix which of course does not achieve any reduction to the dimension of the regression problem. The intersection of all possible DRSs if it is itself a DRS, it is called the Central Dimension Reduction Subspace (CDRS) and it is denoted with $S_{Y|X}$. CDRS is the space that has the smallest dimension (d) among all DRSs. Although the CDRS doesn't always exist the assumptions required for existence are mild so for the rest of the paper we assume existence of the CDRS (see Cook 1998b and Chiaromonte and Cook 2002). Classic methods are described in Li (1991), Cook and Weisberg (1991), Li (1992), Cook (1998a), Cook (2000), Li (2005) and Li and Wang (2007).

More recently, there has been interest in sufficient dimension reduction under the model:

$$Y \perp\!\!\!\perp \boldsymbol{X} / \phi(\boldsymbol{X}), \tag{2}$$

where $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^d$. Model (2) is more general than (1) since function $\phi(\boldsymbol{X})$ can be nonlinear function of the original predictors. This model has been discussed in Wu (2008), Fukumizu *et al.* (2009), Yeh *et al.* (2009) and Li *et al.* (2011).

Most SDR methodology was introduced under the following assumption which is known as the linear conditional mean (LCM) assumption:

Assumption 1. The $E(\boldsymbol{X}|\beta^T\boldsymbol{X})$ is a linear function of $\beta^T\boldsymbol{X}$ for all possible β's.

It is well known that this assumption needs to be valid only for the β that spans the CDRS (which makes the assumption a lot weaker) but since this β is not known then we have to generally use the stronger assumption. Furthermore, this assumption is equivalent to assuming the ellipticity of the marginal distribution of the predictors.

The second main assumption that appears in some SDR methodology is the constant conditional variance (CCV) assumption:

Assumption 2. var$(\boldsymbol{X}|\beta^T\boldsymbol{X})$ is non random.

Methods which require the second assumption, also require the first assumption. The existence of both LCM and CCV assumptions is equivalent to assuming that the predictors have a Normal marginal distribution. This marginal distribution assumption on the predictors is very strong and it is believed to be the main reason algorithms that require both assumptions are sensitive to the choice of slices (a tuning parameter) and as we demonstrate in our real dataset examples later, these methods are also very sensitive to violations of this assumption.

The purpose of this paper is mainly to present some of the most common SDR techniques and at the same time demonstrate the advantage one of them has in cases where the above assumptions are violated, due to the presence of categorical predictors. The rest of the paper is organized as follows. In section 2 we will give an overview of some classic and easy to use methods that are probably the most well known methods in SDR, along with a technique called Principal Support Vector Machine (PSVM) that was introduced recently and has many advantages over previous methodology. In the third section we will demonstrate the advantages of the PSVM algorithm with the application of two real dataset applications and a small discussion will close the paper.

2. REVIEW OF SUFFICIENT DIMENSION REDUCTION METHODOLOGY

In this section we discuss some of the most well known Sufficient Dimension Reduction (SDR) methods. These methods (except the first one) have the common theme of slicing the response variable, Y, in

a sense discretizing the response and using information from the discretized response variable. Also, these are fast algorithms that do not need a lot of computation time thus making them more appropriate for use with larger datasets.

2.1 Ordinary Least Squares

The common regression estimator derived by the method of Ordinary Least Squares (OLS) is

$$\beta_{OLS} = \Sigma^{-1}\Sigma_{YX}$$

where Σ^{-1} is the variance matrix of $\boldsymbol{X}$ and Σ_{YX} is the covariance matrix between $\boldsymbol{Y}$ and $\boldsymbol{X}$. It can be shown that under Assumption 1 this vector is actually in the CDRS, that is $\beta_{OLS} \in S_{Y|\boldsymbol{X}}$ (see Li and Duan 1989). One of the main disadvantages of OLS is that it can find at most one direction, so if the $S_{Y|\boldsymbol{X}}$ has dimension greater than 1, OLS will fail to recover the whole space.

2.2 Sliced Inverse Regression

Sliced Inverse Regression (SIR) was introduced in the breakthrough work of Li (1991) which sparked an interesting discussion and initiated interest in the area of sufficient dimension reduction. SIR like OLS requires Assumption 1 to theoretically work, but it can capture more than one direction, therefore if the CDRS, $S_{Y|X}$, has dimension greater than 1, SIR can still perform well.

The algorithm of SIR, requires the slicing of the response variables into H slices containing approximately the same number of points, and the standardization of the predictors using the formula:

$$\boldsymbol{Z} = (\text{diag}(\sigma_i))^{-1}(\boldsymbol{X} - \mu_{\boldsymbol{X}})$$

where $\mu_{\boldsymbol{X}}$ is the p dimensional mean vector of the predictors, σ_i, $i = 1, ..., p$ is the variance of the i^{th} predictor and diag(·) denotes a matrix for which all off main diagonal elements are 0. The idea of SIR is to calculate the mean of the standardized predictors in each slice $\hat{m}_{Z,i}, i = 1, ..., \boldsymbol{H}$ and then use those means to form the candidate matrix:

$$\hat{M}_{SIR} = \frac{1}{H}\sum_{i=1}^{H} \hat{m}_{\boldsymbol{Z},i}\hat{m}_{\boldsymbol{Z},i}^{T}$$

(Here we note that if the slices have different number of observations one can multiply with the proportion of observations in each slice instead of the ratio 1/H in the above formula). Then an eigenvalue decomposition of this matrix will give us p eigenvalues and eigenvectors. Theoretically, we expect $d = \dim(S_{Y|\boldsymbol{Z}}) < p$ to be non-zero and therefore using the eigenvectors corresponding to the largest d eigenvalues will reveal the vectors which span $S_{Y|\boldsymbol{Z}}$. One may notice that these vectors span the CDRS of the regression of $\boldsymbol{Y}$ on $\boldsymbol{Z}$. To find the vectors that span $S_{Y|\boldsymbol{X}}$ one needs to use the following proposition which is very common in the SDR literature and is known as the invariance property of the SDR methodology.

Proposition 1. Let $S_{Y|\boldsymbol{X}}$ be the central dimension reduction subspace for the regression of Y on $\boldsymbol{X}$ spanned by β, and $S_{Y|\boldsymbol{Z}}$ the central dimension reduction subspace for the regression of Y on $\boldsymbol{Z}$, spanned by η, where $\boldsymbol{Z} = \boldsymbol{A}^T\boldsymbol{X}$. Then $S_{Y|\boldsymbol{Z}} = \boldsymbol{A}^{-1}S_{Y|\boldsymbol{X}}$ and $\eta = \boldsymbol{A}^{-1}\beta$.

One example to demonstrate how SIR works is illustrated in Fig. 1. We have simulated 100 observations $\boldsymbol{X}$ from a bivariate standard Normal distribution and 100 errors $\in \sim N(0, 0.2)$. We set the regression function $Y = X_1 + \in$ which is a regression function whose corresponding CDRS is spanned by $\beta = [1\ 0]^T$, that is, it depends only on X_1. In Fig. 1, we have a scatterplot of the points projected on the predictors 2-dimensional plane. We use 4 slices (circles, triangles, crosses, X's) with 25 points each. Each slice has a point denoted with a * which is the mean of each slice. If you connect the mean of each slice to the overall mean of the predictors (which is [0 0] since they

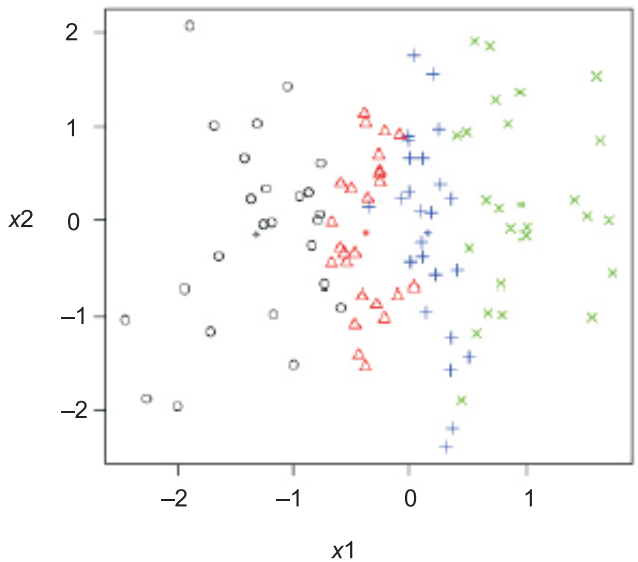

Fig. 1. This figure shows how SIR works when $Y = X_1 + \in$

are simulated from a bivariate standard normal) we get 4 vectors that are almost parallel to the direction of X_1. These 4 vectors are the ones used to construct the candidate matrix.

2.3 Sliced Average Variance Estimation

Sliced Average Variance Estimation (SAVE) was introduced by Cook and Weisberg (1991) as part of the discussion of the SIR work and discussed further by Cook (1998). In their discussion, Cook and Weiberg (1991) pointed out that SIR is based on the first inverse moments, and will have problems identifying the correct directions in cases where the response variable Y depends symmetrically on one or more predictors. Using the fact that not only predictors but also variances vary across the slices of the response variable Y, they proposed SAVE, a method that also exploits the second inverse moments. Although SAVE addresses one issue of the SIR algorithm it does that at the expense of an additional assumption. Therefore in theory SAVE depends on both Assumptions 1 and 2 to capture the vectors that span the CDRS.

The algorithm for SAVE is very similar to the SIR algorithm. The only difference is that within each slice one needs to calculate the covariance matrix of the standardized predictors denoted by $\hat{V}_{\mathbf{Z},i}$ and constructs the candidate matrix:

$$\hat{\boldsymbol{M}}_{\text{SAVE}} = \frac{1}{H}\sum_{i=1}^{H}(\boldsymbol{I} - \hat{V}_{\mathbf{Z},i})(\boldsymbol{I} - \hat{V}_{\mathbf{Z},i})^T$$

where $\boldsymbol{I}$ denotes the $p \times p$ identity matrix. Theoretically SAVE is more powerful than SIR in estimating $S_{Y|\boldsymbol{X}}$ (see Cook and Critchley 2000) but the fact that it requires both Assumptions 1 and 2 restricts it to a smaller set of problems. Since it depends on second inverse moments while SIR only on first moments it needs a larger sample size to work appropriately.

In Fig. 2, we demonstrate an example where SIR fails and SAVE works. We have simulated 100 observations $\boldsymbol{X}$ from a bivariate standard Normal distribution and 100 errors $\in \sim N(0, 0.2)$. We set the regression function $Y = X_1^2 + X_2^2 + \in$ which is a regression function whose corresponding CDRS is spanned by two column vectors $\beta_1 = [1\ 0]^{\mathrm{T}}$ and $\beta_2 = [0\ 1]^{\mathrm{T}}$. In Fig. 2, we have a scatterplot of the points projected on the predictors plane. We use 4 slices (circles, triangles, crosses, X's) with 25 points each and the * denotes the mean of each slice. We can see that

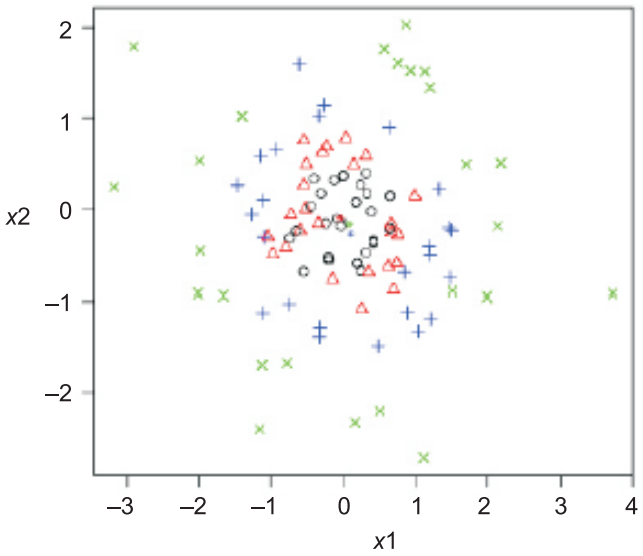

Fig. 2. This figure shows why SIR will fail when $Y = X_1^2 + X_2^2 + \varepsilon$

the means are very close to the overall mean of [0 0]. Therefore SIR will return a degenerate direction as the means do not differ between the slices. Now it is obvious though that the variances of the points differs between slices and therefore SAVE performs much better in capturing the vectors that span the CDRS.

2.4 Directional Regression

Directional Regression (DR) was introduced by Li and Wang (2007) and the main objective was to have a hybrid method that combines the benefits of both SIR and SAVE. It requires both Assumptions 1 and 2 to be true for the theoretical work to hold. The candidate matrix is based on both the mean and the variance of the points in each slice and it works better than both SIR and SAVE in several occasions.

2.5 Principal Support Vector Machine

Principal Support Vector Machine (PSVM) is a recent algorithm proposed by Li *et al.* (2011) which unlike previous methodology we discussed above does not depend on moments. PSVM uses instead a modified version of the Support Vector Machine (SVM) algorithm (Cortes and Vapnik 1995) that has been used effectively for classification problems the last 20 years. PSVM still uses the idea of slicing the response in order to discretize it. Discretization of the response is exactly the main reason one can employ classification techniques to achieve dimension reduction in the SDR framework.

In its simpler form SVM is applied when we have two classes of data and the objective is to construct a hyperplane that separates the data. The hyperplane that achieves the maximum separation and minimizes the misclassification distance of the incorrectly classified points is called the optimal hyperplane. The optimal hyperplane has equation $\psi^T x - t = 0$ where $\psi \in \mathbb{R}^p$ and $t \in \mathbb{R}$. In the SDR framework Li *et al.* (2011) showed that under Assumption 1, $\psi \in S_{Y|X}$. In Fig. 1 one can see that the separating hyperplane between any two slices (*i.e*. triangles and X's) will be a line that is parallel to X_2 and therefore the normal vector of the hyperplane (which is vertical to the hyperplane) will be parallel to X_1 which is the correct direction in the CDRS.

The algorithm of PSVM can be described as follows. For all possible pairs of slices, we find the optimal separating hyperplane by minimizing the following objective function:

$$\psi^T \hat{\Sigma} \psi + \lambda E_n\{1 - \hat{Y}^{rs}[(\boldsymbol{X} - \bar{\boldsymbol{X}})^T \psi - t]\}^+$$

over $(\psi, t) \in \mathbb{R}^p \times \mathbb{R}$ and $\boldsymbol{X}$ is the predictor matrix, $\hat{\Sigma}$ is the estimator of $\Sigma = \text{cov}(\boldsymbol{X})$ and $\tilde{Y}^{rs} \in \mathbb{R}^n$ is a vector with entries

$$\tilde{Y}_i^{rs} = I(q_{s-1} < Y_i \le q_s) - I(q_{r-1} < Y_i \le q_r),$$

where Y_i is the response for the i^{th} observation and q_j, $j = 1, ..., H - 1$ are the cutoff points between the slices (with $q_0 < \min Y_i$ and $q_H \ge \max Y_i$). Using the vector minimizer ψ one can construct the candidate matrix:

$$\hat{\boldsymbol{M}} = \sum_{i=1}^{k} \psi_i \psi_i^T$$

where k is the number of total comparisons $\left($in the case that we use all possible pairs it is $\binom{H}{2}\right)$. It was shown that the eigenvectors corresponding to the d non-zero eigenvalues of this matrix span $S_{Y|X}$. Since using all possible pairs might be time consuming especially when we have a large number of slices, Li *et al.* (2011) proposed the LVR ("left vs right") approach which uses the dividing points between slices to apply SVM and therefore does only $H - 1$ comparisons.

PSVM is a method that combines machine learning and sufficient dimension reduction, which both have been used independently to handle high dimensional problems.

It has several advantages over previous methodology. First of all, it is the first method that can do linear and nonlinear dimension reduction in a unified framework. Since SVM uses kernel functions in Hilbert spaces, one can use the linear kernel to do linear dimension reduction under model (1) or use one of many nonlinear kernels to do nonlinear dimension reduction under model (2). It was also shown that in the linear case one can do dimension reduction without the need for matrix inversion which is a numerically unstable computation. Finally, for the nonlinear dimension reduction it was shown that PSVM does not depend on the two assumptions that classic SDR methodology depends on - that is to show that the normal of the separating hyperplane is in the CDRS in the nonlinear case does not need Assumption 1 which is required for linear dimension reduction. The great performance and applicability of PSVM has already led to several extensions as Artemiou and Shu (to appear) and Shin (2013) demonstrate.

Based on this final observation of Li *et al.* (2011) we investigate how robust the linear version of PSVM is in cases where Assumption 1 is violated and more specifically when the predictors are discrete. As it is shown in the real applications section later in this work, although the theoretical framework requires this assumption, the algorithm is not as sensitive as other methodology to this assumption.

2.6 Other Methodology

Our choice for this work is to focus on methodology that is based on slicing the response variable, but the SDR framework is rich with other methodology which would need a long review article to describe in full detail. Each method has its own advantages and shortcomings. Some examples are principal Hessian directions (pHd) by Li (1992) and Cook (1998), Minimum Average Variance Estimation (MAVE) by Xia *et al.* (2002) and Contour Regression (CR) by Li *et al.* (2005). Methods like MAVE and CR are themselves powerful, but they are at the same time very expensive computationally so it is not efficient to use them in large datasets.

Related to this work is the work by Chiaromonte *et al.* (2002), Li *et al.* (2003) and Wen and Cook (2007)

who proposed algorithms to perform dimension reduction in the presence of a categorical predictor. Their method proposed to perform dimension reduction only on the continuous predictors for the different levels of the categorical predictor. Therefore, they applied SDR techniques only on predictors satisfying Assumption 1.

2.7 Other Issues of Sufficient Dimension Reduction

The use of the aforementioned methodology in SDR creates a number of questions as to how the tuning parameters should be treated. One such question for the methods that require the slicing of the response variable is how many slices to use. Although there is not a clear answer to the question, it has been shown through extensive simulation that methods that depend only on Assumption 1 are more robust to the number of slices while methods that depend on both 1 and 2 are very sensitive to the number of slices. Generally having about 10-20 points in each slice is considered a good choice. PSVM gets better as the number of slices increases but it is obvious that having less than 10 points per slice gives very little additional accuracy, that is the benefit is almost non-existent beyond that point.

An even more important question is how to determine the dimension d of the CDRS. In most cases d is unknown and it has to be inferred from the data. In the literature two different ways have been proposed, one is using sequential tests based on the asymptotic distribution of the candidate matrix and the other is using a BIC type criterion.

The sequential tests, are used to test the hypothesis $H_0 : d = j$ vs $H_A : d > j$ for $j = 0, ..., p$. Starting from 0, if the null hypothesis is rejected we repeat the test, increasing the value of j at each iteration. The smallest value of j for which the test does not reject the null, is the estimated dimension of CDRS. For SIR a number of different sequential tests have been proposed under different assumptions; Li (1991) assumes normality, Bura and Cook (2001) require elliptic distribution of the predictors and finite second moments, while Velilla (1998) impose no assumption on the marginal distribution of the predictors but assumes the number of observations per slice to be fixed and impose some regularity conditions on Y and the regression curve $E(Y|\boldsymbol{X})$. For SAVE sequential tests are developed by Shao *et al.* (2007) and for DR by Li and Wang (2007). The idea of using BIC type criteria to determine the dimensionality of CDRS is more recent and is due to Zhu *et al.* (2006). The estimated dimension of the CDRS in this case will be the value d which maximizes the criterion. For PSVM a cross validated BIC type criterion was proposed that has the form:

$$\sum_{i=1}^{d} \lambda_i - a\lambda_1 d n^{-1/2} \log n$$

where λ_i is the i^{th} eigenvalue of the candidate matrix and a is a quantity that is chosen such that the number of misclassifications is minimized in PSVM. The value of a is estimated using a cross validation procedure and therefore this is known as the CVBIC criterion.

3. REAL DATA APPLICATIONS

As we have explained earlier the main objective of this paper is to demonstrate that the dependence of different methods on Assumption 1 might not be as crucial for some methods as it is for some others. In this section we will demonstrate the clear advantage of the linear version of PSVM in this direction and its robustness in handling data which violate Assumption 1. We have chosen two datasets from the UC Irvine Machine Learning Repository (Bache and Lichman 2013), where the predictors violate Assumption 1 and we compare the performance the four methods SIR, SAVE, DR and the linear version of PSVM on these datasets with non-elliptical predictors.

3.1 Soybean Dataset

The Soybean dataset (Michalski 1980) contains 47 plants with 4 different diseases, and reports 35 characteristics of the plants. All the predictor variables are categorical and 14 of them were excluded from the analysis since they had the same value for all plants involved. The purpose of this is to see if we can extract a direction that distinguishes the plants based on the 4 diseases. There are 10 plants for each disease, 1, 2 and 3 and 17 plants with disease 4 (there is no description available for the type of each disease). To run the analysis we are using the 4 naturally defined slices with appropriate reweighting due to the imbalance of the slices. As we can see from the Fig. 3 (where we emphasize that on y-axis in each plot we put the label of the disease) SIR essentially finds a direction that puts almost all the plants in each disease at the same point and fails to distinguish between diseases 3 and 4; SAVE finds the direction where disease 3 has a lot of variation, while the other 3 diseases are concentrated in one point

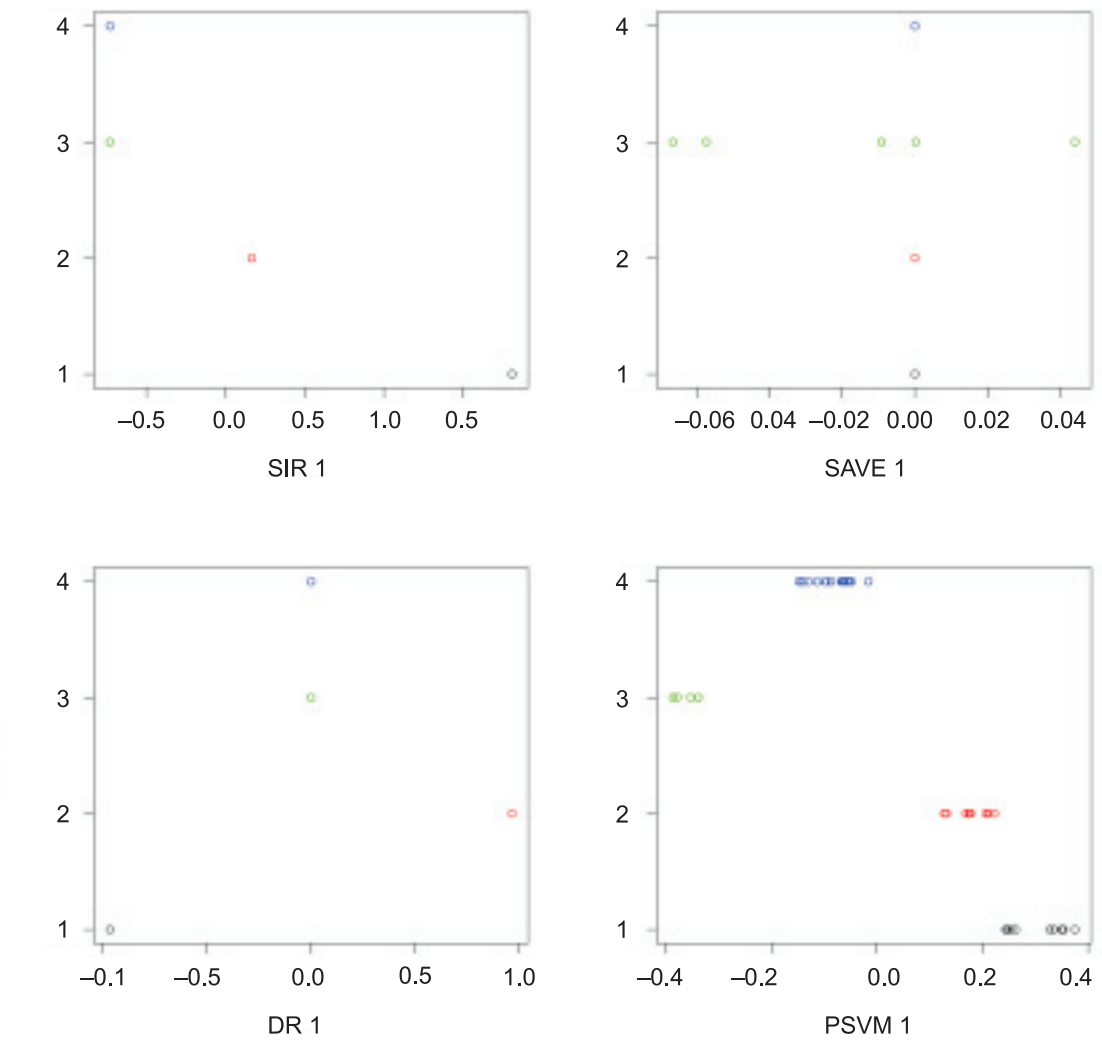

Fig. 3. The first direction of SIR, SAVE, DR and PSVM when the methods are applied on the Soybean dataset. Upper panel: SIR left, SAVE right and lower panel: DR left, PSVM right

and they cannot be distinguished; DR separates the 4 diseases (although this is not really clear in the graph) since disease 3 and 4 are very close and within distance 10^{-7}; and finally PSVM achieves clear separation between all 4 diseases with the closest distance being about 10^{-2} between diseases 1 and 2. In this example although not shown SIR and SAVE actually need more than 2 directions to achieve separation of the diseases.

3.2 *E.coli* Dataset

The *E.coli* dataset (Horton and Nakai 1996) consists of 336 proteins from *E.coli* with 7 different predictors which are biologically related characteristics, and a categorical response variable which indicates from which cell component part the protein is coming from. The predictors are x_1 = McGeoch's method for signal sequence recognition, x_2 = von Heijne's method for signal sequence recognition, x_3 = von Heijne's signal peptide II consensus sequence score (binary), x_4 = presence of charge on *N*-terminus of predicted lipoproteins (binary), x_5 = score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins, x_6 = score of the ALOM membrane spanning region prediction program and x_7 =score of the ALOM program after excluding putative cleavable signal regions from the sequence. The response variable has 8 levels to indicate proteins from the cytoplasm (category 1), inner membrane (categories 2, 3, 4, 5), outer membrane (categories 6 and 7), periplasm (category 8). More details are shown in Table 1. In the analysis we are using the 8 naturally defined datasets with appropriate reweighting due to the unequal size.

Table 1. Categories, number of data and color used in graphs for the E.coli dataset

Class	Number of points	Color (symbol) in graphs
cytoplasm	143	black (rectangle)
inner membrane without signal sequence	77	red (circle)
perisplasm	52	light blue (×ed rectangle)
inner membrane, uncleavable signal sequence	35	yellow (×)
outer membrane	20	purple (diamond)
outer membrane lipoprotein	5	grey (reversed triangle)
inner membrane lipoprotein	2	blue (cross)
inner membrane, cleavable signal sequence	2	green (triangle)

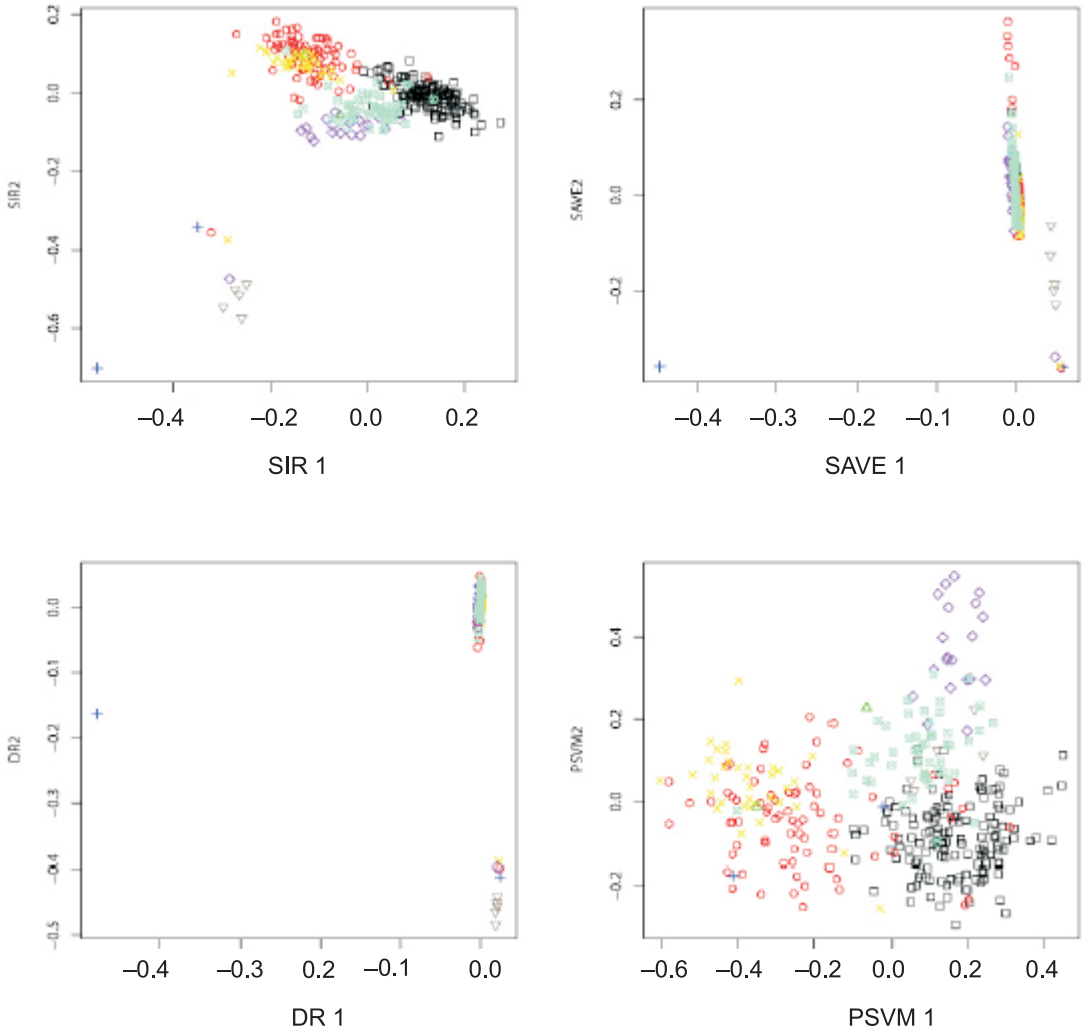

Fig. 4. The first and second directions of SIR, SAVE, DR and PSVM when the methods are applied on the E.coli dataset. Upper panel: SIR left, SAVE right and lower panel: DR left, PSVM right

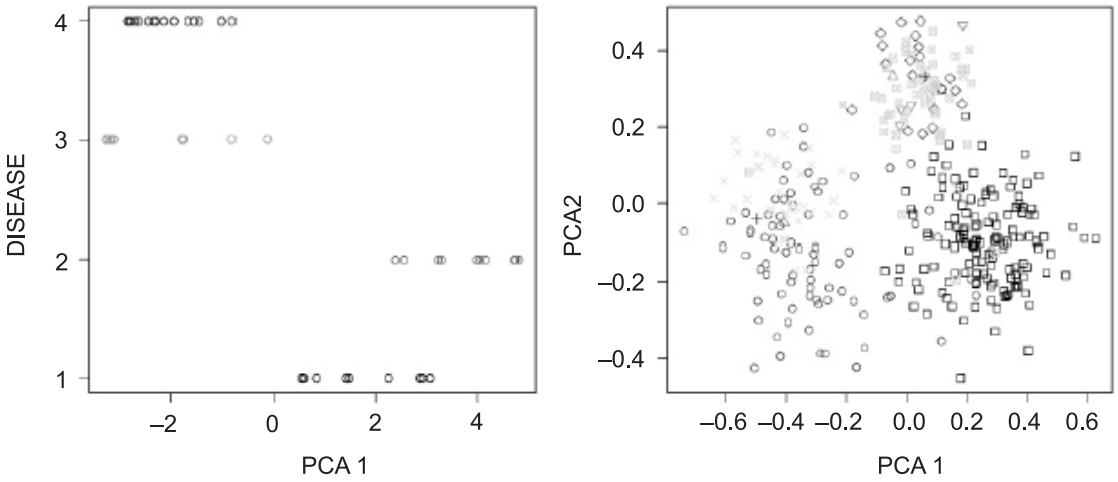

Fig. 5. Using PCA on both the soybean (left) and *E.coli* (right) datasets

As we can see from Fig. 4 SAVE and DR completely fail to capture anything meaningful in the first two directions. SIR performs inefficiently due to 10 points (which are also visible in SAVE and DR) that are completely separated from the rest of the points. Excluding those 10 points and zooming in the area where the other 326 points are concentrated one can see that one gets a nice picture that looks similar to the one by PSVM. PSVM does not achieve perfect separation, but it achieves a meaningful separation of the large slices. The vast majority of the points in the four categories that represent proteins from the inner membrane are grouped together (circle, ×, cross, triangle), the rectangle points representing proteins from the cytoplasm are grouped together, the ×ed rectangle points that are the proteins from the periplasm and are grouped together and the diamond and reversed triangle points that represent proteins from the outer membrane are grouped together.

Finally, a closer look at the 10 points that seem to affect greatly SIR, SAVE and DR, revealed that these are the only proteins that differ from the other proteins on the binary predictors. That is, all other 326 proteins have the same values on the two binary predictors, while those 10 had a different value at one or both of the binary predictors.

Here we feel the need to say that although both datasets have discrete responses and we are discussing how well each method separates the different levels of the response, these methods are not necessarily classification methods. As was demonstrated by Li *et al.* (2011) in the final section of their work, SDR methodology has its own power in analyzing, visualizing and explaining high dimensional datasets. It is though, easier to visualize the advantages and shortcomings of this methodology in cases where the response is discrete.

Before we close the section we want to discuss the advantage of PSVM over unsupervised dimension reduction techniques and more importantly over PCA. As one can see from Fig. 5 when PCA is applied on the soybean dataset the first principal component cannot separate the 4 diseases (left figure) and when PCA is applied on the *E.coli* dataset the first two PCA directions give better separation than SIR, SAVE and DR, (although for SIR if we exclude the 10 strange points and we zoom in at the mass of the rest of the points we get a similar separation to PCA) but this separation is not as good as the separation achieved by PSVM since it cannot discriminate among proteins on the periplasm and outer membrane.

4. DISCUSSION

The SDR framework, consists of a number of methods for reducing the dimension of the predictor vector in high dimensional regression problems through feature extraction. This methodology is supervised, in the sense that information of the response is used in extracting the lower dimensional predictors. In this article we gave a brief overview of well known methodology in the SDR framework. We focus on classic algorithms like SIR, SAVE and DR which are based on the idea of slicing the response and we included also a powerful newly developed algorithm, known as PSVM, which is also based on the idea of slicing the response. Unlike the other methods, PSVM

does not depend on using inverse moments to derive the vectors which span the CDRS, but rather it depends on a modified version of the SVM algorithm.

Among the methodology presented SIR and the linear version of PSVM depend on Assumption 1 which implies the elliptical distribution of the predictors, while SAVE and DR depend on both Assumptions 1 and 2 which together imply that the predictors are normally distributed. In this work we focus on two datasets violating Assumption 1 because they have all or some categorical predictors. We demonstrate using two datasets that PSVM is the most robust among all four methods to violations of this assumption and is able to capture meaningful results in both cases.

Overall, SDR methodology includes some very powerful supervised dimension reduction techniques that usually perform much better than unsupervised dimension reduction techniques like PCA. In the previous section we have shown this superiority of newer SDR methodology like PSVM through two real data examples where assumptions are violated. Although SDR methodology can be useful in all sciences handling large dimensional datasets including, Biosciences, Medical Sciences, Geosciences, Engineering and Agriculture they are not used as frequently as PCA and other unsupervised dimension reduction techniques as most scientists are not familiar with them. We hope that this article will serve the purpose of introducing the SDR methodology to a wider audience of scientists and help increase the use of this methodology in the future.

ACKNOWLEDGEMENTS

The author would like to thank an Associate Editor and a Reviewer for their valuable comments which improved the presentation of the manuscript.

REFERENCES

Artemiou, A. and Li, B. (2009). On principal components and regression: A statistical explanation of a natural phenomenon. *Statistica Sinica*, **19,** 1557-1565.

Artemiou, A. and Li, B. (2013). Predictive power of principal components for single-index model and sufficient dimension reduction. *J. Multi. Anal.*, **119,** 176-184.

Artemiou, A. and Shu, M. (to appear). A cost based reweighed scheme of Principal Support Vector Machine. In: *Contemporary Developments in Statistical Theory*.

Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Bura, E. and Cook, D.R. (2001). Extending SIR: the weighted chi-square test. *J. Amer. Statist. Assoc.*, **96,** 996-1003.

Chiaromonte, F. and Cook, R.D. (2002). Sufficient dimension reduction and graphics in regression. *Ann. Instt. Statist. Maths.,* **54,** 768-795.

Chiaromonte, F., Cook, R.D. and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.*, **30,** 475-497.

Cook, R.D. (1998b). Principal Hessian directions revisited (with discussion). *J. Amer. Statist. Assoc.*, **93**, 84-100.

Cook, R.D. (1998b). *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York.

Cook, R.D. (2000). SAVE: A method for dimension reduction and regression graphics. *Comm. Statist.-Theory Methods,* **29,** 2109-2121.

Cook, R.D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.*, **22,** 1-40.

Cook, R.D. and Critchley, F. (2000). Identifying regression outliers and mixtures graphically. *J. Amer. Statist. Assoc.,* **95,** 781-794.

Cook, R.D. and Weisberg, S. (1991). Discussion of sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, **86,** 316-342.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20,** 1-25.

Fukumizu, Bach and Jordan (2009). Kernel dimension reduction in regression. *Ann. Statist.*, **4,** 1871-1905.

Hadi, A.S. and Ling, R.F. (1998). Some cautionary notes on the use of principal components in regression. *Amer. Statistician*, **52**, 15-19.

Horton, P. and Nakai, K. (1996). A probablistic classification system for predicting the cellular localization sites of proteins. *Intell. Sys. Mole. Biol.*, 109-115.

Hotelling, H. (1933). Analysis of a complex statistical variable into its principal components. *J. Edu. Psychol.*, **24,** 417-441.

Joliffe, I.T. (1982). A note on the use of principal components in regression. *Appl. Statist.*, **31**, 300-303.

Li, B., Artemiou, A. and Li, L. (2011). Principal Support Vector Machine for linear and nonlinear sufficient dimension reduction. *Ann. Statist.*, **39,** 3182-3210.

Li, B., Cook, R.D. and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *Ann. Statist.*, **31,** 1636-1668.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.*, **102,** 997-1008.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.*, **33,** 1580-1616.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.*, **86,** 316-342.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.*, **86**, 316- 342.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.,* **17**, 1009-1052.

Michalski, R.S. (1980) Learning by being told and learning from examples: an experimental comparison of the two methodes of knowledge acquisition in the context of developing an expert system for soybean desease diagnoiss. *Int. J. Policy Anal. Inform. Sys.*, **4(2)**, 125-161.

Ni, L. (2011). Principal component regression revisited. *Statistica Sinica,* **21**, 741-747.

Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. *Philosophical Magazine* **2(6),** 559-572.

Shao, Y., Cook, R.D. and Weisberg S. (2007). Marginal tests with Sliced Average Variance Estimation. *Biometrika,* **94,** 285-296.

Shin, S.J. (2013). New Techniques for High Dimensional and Complex Data Analysis Based on Weighted Learning. Unpublished manuscript. Ph.D. Thesis, Department of Statistics, North Carolina State University.

Velilla, S. (1998). Assessing the number of linear components in a general regression problem. *J. Amer. Statist. Assoc.*, **93,** 1088-1098.

Wen, X. and Cook, R.D. (2007). Optimal sufficient dimension reduction in regressions with categorical predictors. *J. Statist. Plann. Inf.*, **137**, 1961 { 1978.

Wu, H.M. (2008). Kernel sliced inverse regression with applications on classification. *J. Comput. Graph. Statist.*, **17**, 590-610.

Xia, Y., Tong, H., Li, W.K. and Zhu, L.X. (2002). An adaptive estimation of optimal regression subspace. *J. Roy. Statist. Soc., Series B*, **64,** 363-410.

Yeh, Y.-R., Huang, S.-Y. and Lee, Y.-Y. (2009). Nonlinear Dimension Reduction with Kernel Sliced Inverse Regression. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1590-1603.

Zhu, L.X., Miao, B. and Peng, H. (2006). On sliced inverse regression with large dimensional covariates. *J. Amer. Statist. Assoc.*, **101,** 630-643.