Electronic Journal of Statistics
Vol. 4 (2010) 914–931
ISSN: 1935-7524
DOI: 10.1214/10-EJS572

Sparse supervised dimension reduction in high dimensional classification

Junhui Wang*

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607
e-mail: junhui@uic.edu

and

Lifeng Wang

Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824
e-mail: wang@stt.msu.edu

Abstract: Supervised dimension reduction has proven effective in analyzing data with complex structure. The primary goal is to seek the reduced subspace of minimal dimension which is sufficient for summarizing the data structure of interest. This paper investigates the supervised dimension reduction in high dimensional classification context, and proposes a novel method for estimating the dimension reduction subspace while retaining the ideal classification boundary based on the original dataset. The proposed method combines the techniques of margin based classification and shrinkage estimation, and can estimate the dimension and the directions of the reduced subspace simultaneously. Both theoretical and numerical results indicate that the proposed method is highly competitive against its competitors, especially when the dimension of the covariates exceeds the sample size.

AMS 2000 subject classifications: Primary 62H30.
Keywords and phrases: Dimension reduction, SAVE, SIR, large-p-small-n, support vector machine, tuning.

Received January 2010.

Contents

*The authors would like to thank the editor, the associate editor and two referees for their insightful comments, and Dr. Lexin Li from North Carolina State University for his constructive suggestions. The research of the second author was supported by a NSF grant DMS-1007634.

1. Introduction

Recent advances in biomedical sciences have provided statisticians with a large spectrum of new research projects, such as gene microarray analysis, protein structure analysis, and so on. These projects often involve data sets with small number of observations but huge number of covariates. For instance, a gene expression dataset concerning diagnosis of leukaemia [11] consists of only 72 patients with 7,129 genes expressed for each patient. Such a "large-p-small-n" scenario imposes great challenge to conventional statistical techniques due to the curse of dimensionality. A natural way of remedy is to reduce the data dimension, by eliminating some irrelevant covariates or creating some informative combinations of covariates.

In the literature, various dimension reduction techniques have been developed, especially in the route of supervised dimension reduction, where both scalar response Y and p-dimensional covariate $\mathbf{X}$ are utilized. Among other supervised dimension reduction techniques, slice inverse regression (SIR) [18], sliced average variance estimate (SAVE) [7], and principal Hessian directions (pHd) [19] are most popularly used. These methods focus on the conditional distribution of $\mathbf{X}|Y$ and rely on certain distributional assumptions such as the linearity assumption [18]. Recently, the minimum average variance estimation method (MAVE) [29] is proposed for estimating the conditional mean function $E(Y|\mathbf{X})$, which requires no distributional assumption and incorporates estimation of dimension reduction subspace in the framework of local linear smoother [10].

Note that the aforementioned dimension reduction methods are mainly developed in the regression context, and very little effort has been devoted to dimension reduction in the classification context. Most existing methods [2, 4, 6] formulate dimension reduction for classification in a generalized regression framework by treating $P(Y|\mathbf{X})$ as a continuous response, so that successful techniques for regression such as SIR and SAVE can still be applicable. However, as pointed out in [6], dimension reduction for classification may require more careful treatment since the classification decision functions can be substantially affected by some minor change in $P(Y|\mathbf{X})$.

In addition, the effectiveness of conventional dimension reduction methods can be deteriorated when dimension of $\mathbf{X}$ is high, or even greater than the sample size. First, the dimension reduction directions are estimated as a combination of all covariates, and thus can be difficult to interpret. Second, the hypothesis testing procedures for determining the proper dimension of the dimension reduction subspace might become unreliable due to the low power of the hypothesis tests, as demonstrated by the numerical examples in Section 4. To circumvent this difficulty in the regression context, sparse dimension reduction [20] proposes to incorporate variable selection techniques, such as LASSO, into the framework of supervised dimension reduction.

This article introduces a novel sparse supervised dimension reduction technique for high dimensional multicategory classification, which directly estimates the reduced subspace and automatically identifies the low-rank structure of the classification decision functions, while retaining the classification boundary based on the original dataset. The contribution of the proposed method is threefold. First, a margin based loss function is adopted, which directly targets on the classification decision functions rather than the conditional probabilities as in logistic regression. As a result, the proposed method is capable of identifying the central discriminant space (CDS) [6] that is most relevant to the classification, which is in contrast to most existing methods that seek to estimate the larger central space (CS) [3]. Second, a bi-level LASSO type penalty is incorporated that encourages sparsity at both the direction level that forms the basis of the CDS, and the covariate level within each direction. Consequently, the proposed method not only automatically recovers the low-rank structure of the CDS consisting of the most relevant directions, but also produces sparse representation of each direction by removing the redundant covariates. This is advantageous compared to variable selection with a single-level penalty, which does not estimate the CDS at all. Lastly, the sparsity of the proposed method is especially attractive in sparse high dimensional classification, where most covariates are expected to be irrelevant. Furthermore, theoretical study reveals that the proposed method indeed overcomes the difficulty of large-p-small-n, and achieves an consistent estimation.

The rest of this paper is organized as follows. Section 2 briefly reviews the CS and central mean subspace (CMS) [5] for regression, as well as the CDS for classification. Section 3 presents our proposed sparse supervised dimension reduction method for estimating the central discriminant subspace, together with a tuning parameter selection criterion through data perturbation technique. Section 4 compares the proposed method against other top performers through a variety of simulated and real examples, followed by a brief summary in Section 5.

2. Preliminaries

In this section, we briefly review dimension reduction for regression and classification. Particularly, we will focus on the CMS for regression and CDS for classification.

2.1. Dimension reduction for regression

In a regression setting with response $Y \in \mathbb{R}$ and covariate $\mathbf{X} \in \mathbb{R}^p$, assume that $\mathbf{X}$ is standardized such that $E(\mathbf{X}) = \mathbf{0}_p$ and $\mathrm{Var}(\mathbf{X}) = \mathbf{I}_p$, where $\mathbf{0}_p$ is a vector of 0's and $\mathbf{I}_p$ is a $p \times p$ identity matrix.

The goal of supervised dimension reduction is to seek the CS $\mathcal{S}(\mathbf{B}_{cs})$ of minimal dimension such that

$$Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}(\mathbf{B}_{cs})}\mathbf{X}, \tag{2.1}$$

where $\perp\!\!\!\perp$ denotes independence, $\mathbf{B}_{cs}$ is a base of $\mathcal{S}(\mathbf{B}_{cs})$, and $P_{\mathcal{S}(\mathbf{B}_{cs})}$ is an orthogonal projection onto $\mathcal{S}(\mathbf{B}_{cs})$. When the conditional mean function $E(Y|\mathbf{X})$ is of primary interest, the CMS $\mathcal{S}(\mathbf{B}_{cms})$ is defined as the reduced subspace of minimal dimension such that

$$Y \perp\!\!\!\perp E(Y|\mathbf{X}) | P_{\mathcal{S}(\mathbf{B}_{cms})}\mathbf{X}. \tag{2.2}$$

Equivalently, projecting the original data onto $\mathcal{S}(\mathbf{B}_{cms})$ will not lose any information in regression of the conditional mean function.

To estimate the CMS, the MAVE method proposes to find a reduced subspace such that the regression mean function can be properly estimated based on the projected data on the reduced subspace. Specifically, the MAVE method estimates the CMS as the solution of

$$\min_{B, a_j, \mathbf{b}_j} \sum_{i,j=1}^{n} w_{ij} \left[Y_i - (a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{X}_i - \mathbf{X}_j)) \right]^2, \tag{2.3}$$

subject to $\mathbf{B}^T\mathbf{B} = I_d$, where $a_j + \mathbf{b}_j^T\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_j)$ is a local linear approximation of $E(Y|\mathbf{X}_i)$ based on point $\mathbf{X}_j$, and $\sum_{i=1}^n w_{ij} = 1$ with w_{ij} being the kernel weights centered at $\mathbf{B}^T\mathbf{X}_j$. The optimization in (2.3) can be solved through an iterative scheme by fixing $\mathbf{B}$ or a_j and $\mathbf{b}_j$ respectively.

2.2. Dimension reduction for classification

In classification, a decision function $\phi : \mathbb{R}^p \to \{1, \ldots, K\}$ is estimated from a training sample $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, independent and identically distributed according to some unknown distribution $P(\mathbf{x}, y)$, where $\mathbf{X}_i \in \mathbb{R}^p$ and $Y_i \in \{1, \ldots, K\}$. The ideal classification decision function is known as the Bayes rule, $\phi_B(\mathbf{X}) = \arg\max_{k \in \{1,\ldots,K\}} p_k(\mathbf{X})$ with $p_k(\mathbf{X}) = P(Y = k|\mathbf{X})$. The primary goal of dimension reduction for classification is to seek the CDS, defined as the subspace $\mathcal{S}(\mathbf{B}_{cds})$ of minimal dimension such that

$$\phi_B(\mathbf{X}) = \phi_B(P_{\mathcal{S}(\mathbf{B}_{cds})}\mathbf{X}), \tag{2.4}$$

for all values of $\mathbf{X}$. That is, the Bayes rule ϕ_B obtained from the projected data on $\mathcal{S}(\mathbf{B}_{cds})$ should be the same as that obtained from the original data.

Note that the CDS can be a proper subspace of the CS, and that the estimation of CDS requires careful treatment as the classification decision function

is highly sensitive to the distribution of $Y|\mathbf{X}$ in many situations. For illustration, consider a simple binary classification problem with $Y \in \{1, 2\}$ and two independent standard normal covariates X_1 and X_2. Suppose the conditional probability is defined as $p(Y = 1|\mathbf{X}) = 0.9$, if $X_1 \geq 0$, and γ otherwise, where $0 < \gamma < 1$. In this example, the CS is spanned by $(1, 0)'$ regarless of the value of γ. However, the Bayes rule $\phi_B(\mathbf{x}) = \text{sign}(x_1)$ if $\gamma < 0.5$, and $\phi_B(\mathbf{x}) \equiv 1$ if $\gamma \geq 0.5$, implying that the CDS is spanned by $(1, 0)'$ if $\gamma < 0.5$, and a empty space if $\gamma \geq 0.5$. Clearly, the CDS is a proper subspace of the CS if $\gamma \geq 0.5$, and it changes substantially when γ changes from above 0.5 to below 0.5.

3. Sparse dimension reduction for classification

This section presents a sparse supervised dimension reduction method for high dimensional classification with multiple classes. The proposed method combines the techniques of margin based classification and shrinkage estimation, and is capable of estimating the dimension and the directions of the CDS simultaneously.

3.1. Margin based multicategory classification

In margin based multicategory classification, a classification function vector $\mathbf{f} = (f_1, \ldots, f_K)$ is constructed by minimizing a cost function of $\mathbf{f}$ over a linear function class $\mathcal{F}^K$ with $\mathcal{F} = \{f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b : \mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}\}$:

$$\min_{\mathbf{f} \in \mathcal{F}^K} \; n^{-1} \sum_{i=1}^{n} L(y_i, \mathbf{f}(\mathbf{x}_i)) + \lambda J(\mathbf{f}), \tag{3.1}$$

subject to the sum-to-zero constraint $\sum_{k=1}^{K} f_k(\mathbf{x}) = 0$; $\forall \mathbf{x} \in \mathbb{R}^p$. Here $L(y_i, \mathbf{f}(\mathbf{x}_i))$ is a loss function, $J(\mathbf{f})$ is a regularization term for penalizing model complexity, $\lambda > 0$ is the degree of penalization, and the zero-sum constraint $\sum_{k=1}^{K} f_k(\mathbf{x}) = 0$ is enforced to avoid redundancy in $\mathbf{f}$. The decision function $\phi(\mathbf{x}) = \arg\max_k f_k(\mathbf{x})$ estimates the Bayes rule $\phi_B(\mathbf{x})$. Note that we restrict the candidate function to be linear because linear classifiers are in general sufficient to separate different classes and often yield more stable performance than their nonlinear competitors in high dimensional settings [12].

The loss $L(y, \mathbf{f}(\mathbf{x}))$ is margin based if it can be written as $L(\mathbf{u}(\mathbf{f}(\mathbf{x}), y))$, where

$$\mathbf{u}(\mathbf{f}(\mathbf{x}), y) = (f_y(\mathbf{x}) - f_1(\mathbf{x}), \ldots, f_y(\mathbf{x}) - f_K(\mathbf{x})) \tag{3.2}$$

is the generalized functional margin [22]. Clearly, larger components of $\mathbf{u}(\mathbf{f}(\mathbf{x}), y)$ imply better separation and more accurate classification. Based on the generalized functional margin, some popularly used loss functions are proposed:

$$L(\mathbf{u}) = \sum_{k \neq y} (1 - u_k)_+ \text{ [24]}, \tag{3.3}$$

$$L(\mathbf{u}) = \sum_{k \neq y} \left(\frac{1}{K-1} - u_k + \frac{\sum_{k=1}^{K} u_k}{K} \right)_+ \text{ [17], and} \tag{3.4}$$

$$L(\mathbf{u}) = (1 - \min_{k \neq y} u_k)_+ \text{ [22]}, \tag{3.5}$$

where $\mathbf{u}(\mathbf{f}(\mathbf{x}), y)$ is written as $\mathbf{u} = (u_1, \ldots, u_K)$ to simplify notations. Although the loss functions (3.3), (3.4) and (3.5) decrease with respect to $\mathbf{u}$ encouraging better classification, they are constructed based on different principles. In specific, (3.3) corresponds to the "one-versus-rest" approach, whereas (3.4) and (3.5) correspond to the simultaneous formulation that treats all classes at one time. Mathematical analysis reveals that only (3.4) is Fisher consistent when no dominating class is present [17, 30]. Particularly, when $L(\mathbf{u})$ is set as in (3.4), the minimizer of $E(L(\mathbf{u}(\mathbf{f}(\mathbf{X}), Y)))$ subject to the sum-to-zero constraint is

$$f_k(\mathbf{x}) = \begin{cases} 1, & \text{if } k = \arg\max_j p_j(\mathbf{x}); \\ -\frac{1}{K-1}, & \text{otherwise,} \end{cases}$$

and hence that $\phi(\mathbf{x}) = \arg\max_k f_k(\mathbf{x}) = \arg\max_k p_k(\mathbf{x}) = \phi_B(\mathbf{x})$, assuming that $\mathcal{F}^K$ is sufficiently rich [17]. Therefore, we focus our attention on (3.4) in this article, although the proposed method can be adapted to other loss functions.

The regularization term $J(\mathbf{f})$ in high dimensional classification is usually set to be the componentwise L_1-norm of the candidate function, that is, $J(\mathbf{f}) = \sum_{k=1}^{K} \|\mathbf{w}_k\|_1$. The L_1-norm regularization term allows the parameter estimation and variable selection at the same time, which is desirable when the data dimension is high. In addition, λ is a tuning parameter controlling the tradeoff between the loss function and the regularization term, and thus needs to be determined in order to optimize the classification performance.

3.2. Sparse dimension reduction

The proposed sparse supervised dimension reduction for classification is motivated by the MAVE method and the asymptotic consistency of the margin based classification method. In specific, the dimension reduction matrix $\mathbf{B}$ and the corresponding $\mathbf{f}$ can be estimated by solving

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{f} \in \mathcal{F}^K} \quad & n^{-1} \sum_{i=1}^{n} L(\mathbf{u}(\mathbf{f}(\mathbf{B}^T \mathbf{x}_i), y_i)) + \lambda_1 \sum_{k=1}^{K} \|\mathbf{w}_k\|_1 + \lambda_2 \|\mathbf{B}\|_1, \\ \text{subject to} \quad & \sum_{k=1}^{K} \mathbf{B}\mathbf{w}_k = \mathbf{0}_p \text{ and } \sum_{k=1}^{K} b_k = 0. \end{aligned} \tag{3.6}$$

Here $L(\mathbf{u})$ is defined as in (3.4), and $\|\mathbf{B}\|_1 = \sum_{r,s=1}^{p} |\mathbf{B}_{rs}|$ with $\mathbf{B} = (\mathbf{B}_{rs})_{r,s=1}^{p}$, whose columns correspond to the potential directions of the CDS. The original sum-to-zero constraint $\sum_{k=1}^{K} f_k(\mathbf{x}) = 0;\ \forall \mathbf{x} \in \mathbb{R}^p$ is replaced by its equivalent form $\sum_{k=1}^{K} \mathbf{B}\mathbf{w}_k = \mathbf{0}_p$ and $\sum_{k=1}^{K} b_k = 0$ for the ease of implementation [26]. The

resulting classification decision function is $\hat{\phi}(\mathbf{x}) = \arg\max_k \hat{f}_k(\widehat{\mathbf{B}}^T\mathbf{x})$. Note that the key difference between (3.6) and the ordinary margin based classification in (3.1) is the dimension reduction matrix $\mathbf{B}$, which leads to automatic estimation of the CDS without sacrificing the classification accuracy. The estimated CDS spanned by $\widehat{\mathbf{B}}$ allows visualization and exploration of the original dataset with high dimension, which is one of the primary purposes of dimension reduction [3].

To illustrate the validity of (3.6) in estimating the CDS, note that the data fitting component $n^{-1}\sum_{i=1}^n L(\mathbf{u}(\mathbf{f}(\mathbf{B}^T\mathbf{x}_i), y_i))$ in (3.6) approaches the limiting functional $E(L(\mathbf{u}(\mathbf{f}(\mathbf{B}^T\mathbf{X}),\ Y)))$ as n diverges, and hence that the minimizer of (3.6) approaches that of $E(L(\mathbf{u}(\mathbf{f}(\mathbf{B}^T\mathbf{X}), Y)))$ under certain regularity conditions [21]. Analogous to [17], it can be shown that the minimizer of $E(L(\mathbf{u}(\mathbf{f}(\mathbf{B}^T\mathbf{X}), Y)))$ subject to the sum-to-zero constraint is

$$f_k(\mathbf{B}^T\mathbf{x}) = \begin{cases} 1, & \text{if } k = \arg\max_j p_j(\mathbf{x}); \\ -\frac{1}{K-1}, & \text{otherwise,} \end{cases} \tag{3.7}$$

assuming that $\mathcal{F}^K$ is sufficiently rich. Therefore, $\phi(\mathbf{B}^T\mathbf{x}) = \arg\max_k p_k(\mathbf{x}) = \phi_B(\mathbf{x})$, which immediately justifies (3.6) in estimating the CDS. In addition, (3.7) ensures that (3.6) is able to handle the example in Section 2.2 as (3.6) directly targets on $\mathbf{f}(\mathbf{B}^T\mathbf{x})$ as opposed to $p_k(\mathbf{x})$, and hence that it is able to identify the correct CDS regardless of how $p_k(\mathbf{x})$ may change $\mathbf{f}(\mathbf{B}^T\mathbf{x})$.

Next, Lemma 3.1 suggests that (λ_1, λ_2) in (3.6) can be suppressed into one tuning parameter.

Lemma 3.1. *Let* $\hat{\mathbf{f}}$ *be the solution of (3.6), then* $\hat{\mathbf{f}}$ *depends on* (λ_1, λ_2) *only through* $\lambda_1\lambda_2$.

Proof of Lemma 3.1. The desired result immediately follows from the fact that $f_k(\mathbf{B}^T\mathbf{x}_i) = \mathbf{w}_k^T\mathbf{B}\mathbf{x} + b_k = (\mathbf{w}_k/a)^T(a\mathbf{B})\mathbf{x} + b_k;\ k = 1, \ldots, K$, $\lambda_1\sum_{k=1}^K \|\mathbf{w}_k\|_1 = a\lambda_1\sum_{k=1}^K \|\mathbf{w}_k/a\|_1$ and $\lambda_2\|\mathbf{B}\|_1 = (\lambda_2/a)\|a\mathbf{B}\|_1$, for any constant $a \neq 0$. □

Without loss of generality, we then replace the tuning parameters (λ_1, λ_2) by $\lambda = \lambda_1 = \lambda_2$, and the sparse dimension reduction formulation in (3.6) becomes

$$\begin{aligned} \min_{\mathbf{B},\mathbf{f}\in\mathcal{F}^K} \quad & n^{-1}\sum_{i=1}^n L(\mathbf{u}(\mathbf{f}(\mathbf{B}^T\mathbf{x}_i), y_i)) + \lambda\left(\sum_{k=1}^K \|\mathbf{w}_k\|_1 + \|\mathbf{B}\|_1\right), \\ \text{subject to} \quad & \sum_{k=1}^K \mathbf{B}\mathbf{w}_k = \mathbf{0}_p \text{ and } \sum_{k=1}^K b_k = 0. \end{aligned} \tag{3.8}$$

Furthermore, the LASSO type of regularization terms $\|\mathbf{w}_k\|_1$ and $\|\mathbf{B}\|_1$ in (3.8) perform a bi-level variable selection, encouraging sparsity at both the direction level and the covariate level. The zero entries in $\|\mathbf{w}_k\|_1$ automatically remove the redundant columns in $\widehat{\mathbf{B}}$ and recover informative directions of the CDS. In addition, the sparseness of the remaining columns in $\widehat{\mathbf{B}}$ leads to simpler representation and easier interpretation for the informative directions. This is in contrast to the single-level penalty in L_1-norm multicategory support vector

machine (L1MSVM) [26] that performs only variable selection at the covariate level. As a result, our method is capable of identifying a low-rank structure of the CDS that usually has a rank much less than $K-1$, whereas L1MSVM in general yields the decision function vector of dimension $K-1$ without further dimension reduction. This idea of bi-level penalty is closely related to the hierarchical penalization in [27], which performs both group and within-group variable selection for grouped predictors. The novelty of our proposed method is that each dimension reduction direction is a sparse combination adaptively identified out of all the covariates, whereas each direction merely consists of the covariates within a given group in [27].

To solve (3.8), we implement an iterative algorithm.

Algorithm 3.1.

Step 1. Initialize $\mathbf{B}^{(0)} = \mathbf{I}_p$, and set precision $\epsilon = 10^{-6}$.

Step 2. Given $\mathbf{B}^{(m)}$, find $\mathbf{w}_k^{(m+1)}$ and $b_k^{(m+1)}$; $k = 1, \ldots, K$ by solving

$$\min_{\mathbf{w}_k, b_k} \; n^{-1} \sum_{i=1}^{n} L(\mathbf{u}(\mathbf{f}((\mathbf{B}^{(m)})^T \mathbf{x}_i), y_i)) + \lambda \sum_{k=1}^{K} \|\mathbf{w}_k\|_1,$$

subject to $\sum_{k=1}^{K} \mathbf{B}^{(m)} \mathbf{w}_k = 0$ and $\sum_{k=1}^{K} b_k = 0$.

Step 3. Given $\mathbf{w}^{(m+1)}$ and $b_k^{(m+1)}$, find $\mathbf{B}^{(m+1)}$ by solving

$$\min_{\mathbf{B}} \; n^{-1} \sum_{i=1}^{n} L(\mathbf{u}(\mathbf{f}(\mathbf{B}^T \mathbf{x}_i), y_i)) + \lambda \|\mathbf{B}\|_1.$$

Step 4. Repeat *Steps 2* and *3* until $\|\mathbf{B}^{(m)} - \mathbf{B}^{(m+1)}\|_1 \leq \epsilon$.

Step 5 (optional). Orthonormalize $\widehat{\mathbf{B}}$ by Gram-Schmidt process, and update $\hat{\mathbf{w}}_k$ accordingly.

As computational remarks, the sub-optimization problems in *Steps 2* and *3* can be solved by any linear programming algorithm, which is available in most standard statistical softwares. When p is large, to expedite *Steps 2* and *3*, one can restrict the dimension of $\mathbf{B}$ to be $p \times (K-1)$, or solve for $\mathbf{w}^{(m+1)}$ or $\mathbf{B}^{(m+1)}$ column by column. In addition, *Step 5* is mainly for orthonormalizing $\widehat{\mathbf{B}}$, which is optional as the primary goal of dimension reduction in classification is to recover $\mathcal{S}(\mathbf{B}_{\mathbf{cds}})$ rather than the base of $\mathcal{S}(\mathbf{B}_{\mathbf{cds}})$.

3.3. Tuning parameter selection

The estimation accuracy of the proposed dimension reduction method largely depends on the tuning parameters λ, and hence that they need to be properly determined. In this section, we rewrite $\hat{\phi}$ as $\hat{\phi}_\lambda$ to indicate its dependency on λ, and present a tuning parameter selection criterion based on the prediction accuracy of $\hat{\phi}_\lambda$, obtained from the projected data on the reduced subspace.

To assess the prediction accuracy of $\hat{\phi}_\lambda$, the generalization error (GE) of $\hat{\phi}_\lambda$ is used, defined as $GE(\hat{\phi}_\lambda) = E(I(Y \neq \hat{\phi}_\lambda(\mathbf{X})))$, where $I(\cdot)$ is an indicator

function, and the expectation is taken with respect to the unknown $P(\mathbf{x}, y)$ and thus needs to be estimated from data.

In the literature, estimation of the GE given fixed $\mathbf{X}$'s has been extensively investigated; c.f., [8, 9, 23] for more details. Wang and Shen [25] proposes a data adaptive GE estimation technique for random $\mathbf{X}$ in the context of classification, where $GE(\hat{\phi}_\lambda)$ is decomposed as a sum of GE's of binary classifiers,

$$GE(\hat{\phi}_\lambda) = \frac{1}{2}\sum_{k=1}^{K} E(I(t(Y)_k \neq t(\hat{\phi}_\lambda(\mathbf{X}))_k)), \tag{3.9}$$

where $t : (1, \ldots, K) \to \{0,1\}^K$ maps j to a vector of length K which has all entries equal to 0 except the jth one equal to 1. Furthermore, $E(I(t(Y)_k \neq t(\hat{\phi}_\lambda(\mathbf{X}))_k))$ is estimated by

$$EGE_k(\hat{\phi}_\lambda) + 2n^{-1}\sum_{i=1}^{n} \mathrm{Cov}(t(Y_i)_k, t(\hat{\phi}_\lambda(\mathbf{X}_i))_k|\mathbf{X}^n) + D_{1k}(\mathbf{X}^n, \hat{\phi}_\lambda) + D_{2k}(\mathbf{X}^n).$$

Here $EGE_k(\hat{\phi}_\lambda) = \frac{1}{n}\sum_{i=1}^{n} I(t(Y_i)_k \neq t(\hat{\phi}_\lambda(\mathbf{X}_i))_k)$ is the empirical GE for the kth component of (3.9), $\mathrm{Cov}(t(Y_i)_k, t(\hat{\phi}_\lambda(\mathbf{X}_i))_k|\mathbf{X}^n)$ with $\mathbf{X}^n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is known as covariance penalty [9] measuring the prediction accuracy of ϕ on $\mathbf{X}^n$,

$$D_{1k}(\mathbf{X}^n, \hat{\phi}_\lambda) = E\big(E(E(t(Y)_k|\mathbf{X}) - t(\hat{\phi}_\lambda(\mathbf{X}))_k)^2 - \frac{1}{n}\sum_{i=1}^{n}(E(t(Y_i)_k|\mathbf{X}_i) - t(\hat{\phi}_\lambda(\mathbf{X}_i))_k)^2|\mathbf{X}^n\big)$$

accounts for the randomness of $\mathbf{X}$, and

$$D_{2k}(\mathbf{X}^n) = E(\mathrm{Var}(t(Y)_k|\mathbf{X})) - \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(t(Y_i)_k|\mathbf{X}_i)$$

is independent of $\hat{\phi}_\lambda$ and thus can be omitted in the estimation of $GE(\hat{\phi}_\lambda)$.

To estimate Cov and D_{1k} terms in (3.9), a data perturbation technique [25] can be employed. The idea is to generate local perturbations of $\mathbf{X}$ and Y to evaluate sensitivity of $\hat{\phi}_\lambda$ by estimating its classification accuracy via its derivative estimation. The formulas are given in (11) and (12) of [25]. The proposed tuning parameter selection technique yields higher estimation accuracy than cross validation in a wide variety of numerical examples while achieving asymptotic optimality for both fixed and random $\mathbf{X}$'s [25]. The numerical experiments in Section 4 also demonstrate that this selection criterion yields satisfactory performance in estimating the dimension and the directions of the CDS.

3.4. Asymptotic properties

In this section, we present some asymptotic results showing that the proposed method is able to estimate the CDS, while yielding comparable asymptotical

classification performance to standard classification methods, such as L1MSVM, in terms of the generalized hinge loss for the large-p-small-n classification. Similar results have been established in [13, 29] in the regression context.

To handle the large-p-small-n problem, we first introduce some notations. Write $\mathbf{X}(p) = (X^{(1)}, \ldots, X^{(p)})^T$ as a truncated infinite-dimensional random vector $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots)^T$, and p is allowed to grow with n at a much faster rate. To relate to the existing learning theory, we work on the following equivalent formulation of (3.6),

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{f}\in\mathcal{F}^K} \quad s(\mathbf{B},\mathbf{f}) = n^{-1}\sum_{i=1}^{n} L(\mathbf{u}(\mathbf{f}(\mathbf{B}^T\mathbf{x}_i), y_i)) \\ &\text{subject to} \quad \sum_{k=1}^{K}\|\mathbf{w}_k\|_1 + \|\mathbf{B}\|_1 \le s, \ \sum_{k=1}^{K}\mathbf{w}_k = \mathbf{0}_p \text{ and } \sum_{k=1}^{K} b_k = 0, \end{aligned} \tag{3.10}$$

and define $\hat{\mathbf{f}}^{(p)}(\widehat{\mathbf{B}}^T\mathbf{x})$ as the solution of (3.10). The optimal performance is defined by $\mathbf{f}^{(p)} = \arg\min_{\mathbf{f}\in\mathcal{F}(p)} EL(\mathbf{u}(\mathbf{f}(\mathbf{X}), Y))$, where $\mathcal{F}(p) = \{\mathbf{f} : f_k = \mathbf{w}_k^T\mathbf{x}(p) + b, \mathbf{w}_k \in R^p\}$ is the class of all linear decision functions. For any $\mathbf{f} \in \mathcal{F}(p)$, its performance is measured by the excess hinge risk $e_L(\mathbf{f}, \mathbf{f}^{(p)}) = EL(\mathbf{u}(\mathbf{f}(\mathbf{X}), Y)) - EL(\mathbf{u}(\mathbf{f}^{(p)}(\mathbf{X}), Y)) \ge 0$.

We now quantify the magnitude of $e_L(\hat{\mathbf{f}}^{(p)}(\widehat{\mathbf{B}}^T\mathbf{x}), \mathbf{f}^{(p)})$ under the following assumptions.

Assumption A: Assume that $\sup_{0<j<\infty} X^{(j)} < \infty$.

Assumption B: There exists a finite s^* such that $\mathbf{f}^{(p)} \in \mathcal{F}(p, s^*)$ for all p, where $\mathcal{F}(p, s^*) = \{\mathbf{f} : f_k = \mathbf{w}_k^T\mathbf{B}^T\mathbf{x}(p) + b, \ \sum_{k=1}^{K}\|\mathbf{w}_k\|_1 + \|\mathbf{B}\|_1 \le s^*\}$.

Theorem 3.1. *Suppose Assumptions A and B hold, and* $(n^{-1}\log p_n)^{1/2} \to 0$. *Then for* $s = s^*$,

$$e_L(\hat{\mathbf{f}}^{(p_n)}(\widehat{\mathbf{B}}^T\mathbf{x}), \mathbf{f}^{(p_n)}) = O\left((n^{-1}\log p_n)^{1/2}\log(n(\log p_n)^{-1})\right), \ a.s..$$

Proof of Theorem 3.1. Note that $\hat{\mathbf{f}}^{(p)}(\widehat{\mathbf{B}}^T\mathbf{x}) = \arg\min_{\mathbf{f}\in\mathcal{F}(p,s^*)} s(\mathbf{B}, \mathbf{f})$, and for all $\mathbf{f} \in \mathcal{F}(p, s^*)$, its L_1-norm, $\|\mathbf{f}\|_1 = \sum_{k=1}^{K}\|\mathbf{w}_k^T\mathbf{B}^T\|_1 \le \sum_{k=1}^{K}\|\mathbf{w}_k^T\|_1 s^* \le s^{*2}$, is bounded. As a result, $\mathcal{F}(p, s^*) \subset \mathcal{F}_{L_1}(p, s^{*2})$ with $\mathcal{F}_{L_1}(p, s^{*2}) = \{\mathbf{f} : f_k = \mathbf{w}_k^T\mathbf{x}(p) + b, \ \sum_{k=1}^{K}\|\mathbf{w}_k\|_1 \le s^{*2}\}$, and

$$\begin{aligned} &P(e_L(\hat{\mathbf{f}}^{(p)}(\widehat{\mathbf{B}}^T\mathbf{x}), \mathbf{f}^{(p)}) > \delta) \\ &\le P^*\left(\sup_{\{\mathbf{f}\in\mathcal{F}(p,s^*):\ e_L(\mathbf{f},\mathbf{f}^{(p)})>\delta\}} n^{-1}\sum_{i=1}^{n}(L(\mathbf{u}(\mathbf{f}^{(p)}(\mathbf{X}_i), Y_i)) - \right. \\ &\qquad\qquad \left. L(\mathbf{u}(\mathbf{f}(\mathbf{X}_i), Y_i)) > 0\right) \\ &\le P^*\left(\sup_{\{\mathbf{f}\in\mathcal{F}_{L_1}(p,s^{*2}):\ e_L(\mathbf{f},\mathbf{f}^{(p)})>\delta\}} n^{-1}\sum_{i=1}^{n}(L(\mathbf{u}(\mathbf{f}^{(p)}(\mathbf{X}_i), Y_i)) - \right. \\ &\qquad\qquad \left. L(\mathbf{u}(\mathbf{f}(\mathbf{X}_i), Y_i)) > 0\right). \end{aligned} \tag{3.11}$$

The probability in inequality (3.11) can be bounded by Theorem 2 of [26] and the desired result immediately follows from Corollary 1 therein. □

Assumption B describes an L_1-norm "sparse scenario", which is weaker than the commonly used assumption on the L_0-norm sparseness. Specifically, suppose that the true decision functions are sparse in L_0-norm, depending on only a finite number of predictors, that is $f_k^{(p)}(\mathbf{x}) = (\mathbf{w}^*{}_k)^T(\mathbf{B}^*)^T\mathbf{x} + \mathbf{b}^*$, where the number of non-zero entries in $\mathbf{b}^*$ and $\mathbf{B}^*$ is finite, independent of p. Then, it is straightforward to verify that $\mathbf{f}^{(p)} \in \mathcal{F}(p, s^*)$ for some constant s^*, which satisfies Assumption B.

Under the spareness assumption, Theorem 3.1 yields an error rate tending to zero as long as p_n grows no faster than $\exp(n)$, which indicates the optimal performance can be obtained even for $p \gg n$. More importantly, the rate of convergence in Theorem 3.1 is comparable to that of L1MSVM [26], implying that the proposed method achieves the purpose of dimension reduction without sacrificing its classification performance.

4. Numerical experiments

This section presents numerical studies to examine the finite-sample effectiveness of the proposed sparse supervised dimension reduction procedure (SSDR). The purpose of the numerical studies is two-fold. First, we compare the estimation accuracy of $\mathcal{S}(\widehat{\mathbf{B}})$ by SSDR to some popular competitors, including SIR, SAVE and pHd, when $\mathbf{B}_{cds}$ is know. Specifically, the vector correlation coefficient q^2 [16] and the trace correlation r^2 [15] between $\widehat{\mathbf{B}}$ and $\mathbf{B}_{cds}$,

$$q^2(\widehat{\mathbf{B}}, \mathbf{B}_{cds}) = \prod_{s=1}^{d} \rho_s, \quad r^2(\widehat{\mathbf{B}}, \mathbf{B}_{cds}) = \frac{1}{d}\sum_{s=1}^{d} \rho_s,$$

are employed to assess the closeness between $\mathcal{S}(\widehat{\mathbf{B}})$ and $\mathcal{S}(\mathbf{B}_{cds})$, where $\rho_1, \ldots, \rho_d$ are eigenvalues of $\mathbf{B}_{cds}^T\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T\mathbf{B}_{cds}$ and d is the dimension of $\mathbf{B}_{cds}$. If $d = 0$, for simplicity, we define $q^2(\widehat{\mathbf{B}}, \mathbf{B}_{cds}) = r^2(\widehat{\mathbf{B}}, \mathbf{B}_{cds}) = 1$ if the dimension of $\widehat{\mathbf{B}}$ is also 0, and 0 if the dimension of $\widehat{\mathbf{B}}$ is greater than 0. Note that both $q^2(\widehat{\mathbf{B}}, \mathbf{B}_{cds})$ and $r^2(\widehat{\mathbf{B}}, \mathbf{B}_{cds})$ range from 0 to 1, and larger values of $q^2(\widehat{\mathbf{B}}, \mathbf{B}_{cds})$ and $r^2(\widehat{\mathbf{B}}, \mathbf{B}_{cds})$ indicate $\mathcal{S}(\widehat{\mathbf{B}})$ and $\mathcal{S}(\mathbf{B}_{cds})$ are closer.

Second, we compare the classification accuracy of the classification decision functions constructed based on the corresponding $\widehat{\mathbf{B}}$. This can be viewed as a complementary assessment of the estimation accuracy of $\widehat{\mathbf{B}}$, when $\mathbf{B}_{cds}$ is unknown as in the real examples. In specific, we first obtain $\widehat{\mathbf{B}}$ through various dimension reduction methods, then project the original dataset onto $\mathcal{S}(\widehat{\mathbf{B}})$ and construct the classification decision function $\widehat{\phi}(\widehat{\mathbf{B}}^T\mathbf{x})$ by applying multicategory SVM (MSVM) [17] to the projected data on $\mathcal{S}(\widehat{\mathbf{B}})$. A test error, defined as

$$TE(\widehat{\phi}, \widehat{\mathbf{B}}) = \frac{1}{\#\{\text{test set}\}} \sum_{\text{test set}} I(y_i \neq \widehat{\phi}(\widehat{\mathbf{B}}^T\mathbf{x}_i)),$$

is used to measure the classification accuracy of $\widehat{\phi}$ and the estimation accuracy of $\widehat{\mathbf{B}}$. Furthermore, to illustrate that SSDR can retain the classification accuracy while estimating $\mathcal{S}(\mathbf{B}_{cds})$, we also compare its test error with the standard classification methods including L1MSVM and linear discriminant analysis (LDA).

All numerical analyses are conducted in R2.7.2. The "dr" routine in the dr package is employed for SIR, SAVE and pHd, which estimates the dimension reduction spaces and performs marginal tests concerning their dimensions, and the "solve" routine in lpSolveAPI package is employed for solving the linear programming problems in *Algorithm 3.1*.

4.1. Simulation

Two simulated examples are examined.

Example 1. Data $\{(X_{i1},\ldots,X_{i10},Y_i)\}$; $i=1,\ldots,1000$ are generated as follows. First, $\{X_{ij}\}$; $i=1,\ldots,1000$, $j=1,\ldots,10$ are sampled from independent standard normal distribution. Next, $Y_i-1|\mathbf{X}_i\sim Bernoulli(p_1(\mathbf{X}_i))$ with

$$p_1(\mathbf{X}_i)=\begin{cases}0.9 & \text{if } X_{i1}\geq 0;\\ \gamma & \text{otherwise.}\end{cases}$$

Two situations with different γ's are considered: (a) $\gamma=0.1$; (b) $\gamma=0.6$. This yields the first simulated example, in which 100 randomly selected cases are used for training and the remaining 900 for testing. According to the generating probability distribution, $\mathcal{S}(\mathbf{B}_{cds})$ for $\gamma=0.1$ and $\gamma=0.6$ are spanned by $(1,\mathbf{0}_9^T)^T$ and a empty set, respectively.

Example 2. Data $\{(X_{i1},\ldots,X_{i100},Y_i)\}$; $i=1,\ldots,1000$ are generated as follows. First, $X_{i1}^a,\ldots,X_{i100}^a$ are sampled from independent standard normal distribution. Second, denote $Z_{i1}=\sum_{j=1}^{50}X_{ij}^a$ and $Z_{i2}=\sum_{j=51}^{100}X_{ij}^a$, and

$$Y_i=1+I(Z_{i1}<=0)+2*I(Z_{i2}<=0);\ i=1,\ldots,100.$$

Third, $X_{ij}=X_{ij}^a+\theta*\text{sign}(Z_{i1})$; $j=1,\ldots,50$, and $X_{ij}=X_{ij}^a+\theta*\text{sign}(Z_{i2})$; $j=51,\ldots,100$, for $\theta=0.2$ or 1, which specifies the degree of separation among the four classes. This yields the second simulated example, where 80 randomly selected cases are used for training and the remaining 920 for testing. Clearly, $\mathcal{S}(\mathbf{B}_{cds})$ is spanned by $(\mathbf{1}_{50}^T,\mathbf{0}_{50}^T)^T$ and $(\mathbf{0}_{50}^T,\mathbf{1}_{50}^T)^T$.

To eliminate dependency of our proposed method on tuning parameter λ, we apply the tuning parameter selection criterion in Section 3.3 and search for the minimizer of the estimated GE over 61 equally spaced grid points on $\{10^{-3+t/10};\ t=0,\ldots,60\}$. The tuning parameter in MSVM is selected through the same grid search method with the same 61 grid points. Finally, the numerical results, averaged over 100 simulation replications, are summarized in Table 1.

Evidently, SSDR delivers superior performance over its competitors. Specifically, in Examples 1a and 1b, SSDR recovers $\mathcal{S}(\mathbf{B}_{cds})$ in almost all replications, and yields smaller test errors than SIR, SAVE and pHd. In Example 1a, SAVE

TABLE 1
Averaged q^2, r^2 and test errors as well as estimated standard errors (in parenthesis) of SIR, SAVE, pHd, SSDR, L1MSVM and LDA in the simulated examples. Here Example 1a and 1b correspond respectively to the simulated Example 1 with $\gamma = 0.1$ and $\gamma = 0.6$, and Example 2a and 2b correspond respectively to the simulated Example 2 with $\theta = 0.2$ and $\theta = 1$

		Example 1a	Example 1b	Example 2a	Example 2b
SIR	q^2	0.884(.0059)	0.420(.0496)	–	–
	r^2	0.884(.0059)	0.420(.0496)	–	–
SAVE	q^2	0.009(.0094)	0.990(.0100)	–	–
	r^2	0.009(.0094)	0.990(.0100)	–	–
pHd	q^2	0.000(.0000)	1.000(.0000)	–	–
	r^2	0.000(.0000)	1.000(.0000)	–	–
SSDR	q^2	0.987(.0019)	1.000(.0000)	0.189(.0026)	0.242(.0039)
	r^2	0.987(.0019)	1.000(.0000)	0.441(.0029)	0.492(.0040)
SIR	TE	0.163(.0024)	0.267(.0016)	–	–
SAVE	TE	0.500(.0035)	0.254(.0004)	–	–
pHd	TE	0.504(.0004)	0.254(.0004)	–	–
SSDR	TE	0.124(.0025)	0.254(.0004)	0.217(.0041)	0.000(.0000)
L1MSVM	TE	0.111(.0017)	0.254(.0004)	0.218(.0025)	0.000(.0001)
LDA	TE	0.163(.0024)	0.278(.0012)	0.414(.0074)	0.025(.0036)

and pHd can hardly identify any direction of $\mathcal{S}(\mathbf{B}_{cds})$, whereas both SSDR and SIR deliver satisfactory performance and $\mathcal{S}(\widehat{\mathbf{B}})$ by SSDR is closer to $\mathcal{S}(\mathbf{B}_{cds})$ than that by SIR as it yields larger value of q^2 and r^2. In Example 1b, SSDR, SAVE and pHd are able to recover $\mathcal{S}(\mathbf{B}_{cds})$ in all replications, while SIR tends to overestimate $\mathcal{S}(\mathbf{B}_{cds})$. In Examples 2a and 2b, the number of covariates is larger than that of observations and hence that SIR, SAVE and pHd are not applicable, whereas SSDR is able to handle this large-p-small-n scenario. From Table 1, the small values of q^2 and r^2 suggest that $\mathcal{S}(\widehat{\mathbf{B}})$ by SSDR is somewhat different from $\mathcal{S}(\mathbf{B}_{cds})$. However q^2 and r^2 are no longer appropriate to assess the estimation accuracy of $\mathcal{S}(\widehat{\mathbf{B}})$ in Example 2 as there can be multiple CDS's leading to perfect classification due to the high dimension. Figure 1 displays averaged q^2 and r^2 of SSDR in Example 2 with $\theta = 1$ and $p = 10, 20, 50, 100, 200, 500$ over 100 independent replications. Clearly, both q^2 and r^2 decrease as p increases, while the test error of SSDR remains 0 for all p's. This suggests that although $\mathcal{S}(\widehat{\mathbf{B}})$ by SSDR deviates from $\mathcal{S}(\mathbf{B}_{cds})$ when p is large, the classification decision function based on $\mathcal{S}(\widehat{\mathbf{B}})$ remains the same as that based on $\mathcal{S}(\mathbf{B}_{cds})$.

Finally, we compare the test error of SSDR to that of standard classification methods including L1MSVM and LDA that do not provide estimate of $\mathcal{S}(\mathbf{B}_{cds})$ at all. It is clear that SSDR outperforms LDA in all examples, and compares similarly to L1MSVM. In Example 2b, although $\mathcal{S}(\widehat{\mathbf{B}})$ is seemingly different from $\mathcal{S}(\mathbf{B}_{cds})$, SSDR is able to yield perfect classification in all replications. In Example 2a, due to the high dimension but relatively small separation with $\theta = 0.2$, all methods fail to achieve the perfect classification but SSDR is still able to deliver smaller test error than both L1MSVM and LDA.

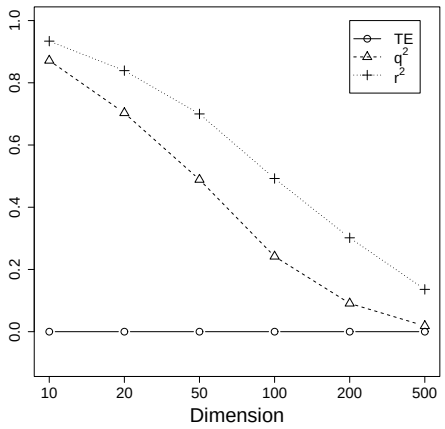

FIG 1. *The averaged q^2, r^2 and test error in Example 2 with $\theta = 1$ and various p's.*

4.2. Real examples

We now examine two real examples, Wisconsin breast cancer and Lung cancer. Both datasets are available from the University of California at Irvine machine learning data repository at http://www.ics.uci.edu/~mlearn/MLRepository.html. The Wisconsin breast cancer dataset, collected at University of Wisconsin Hospitals, consists of 569 cases with two diagnoses, 212 malignant and 357 benign. The main goal is to develop a diagnosis method to determine whether a case is benign or malignant, based on 30 features computed from a digitized image of a fine needle aspiration. More details about the data is given in [28]. The Lung cancer dataset consists of only 32 patients with three different types of pathological lung cancers, 9 from the first, 13 from the second and 11 from the third type, and 56 discrete features extracted from clinical and X-ray data. In Lung cancer, there are four missing values in the 4th feature and one missing value in the 38th feature, which are replaced by the modes of the 4th and 38th features, respectively. The data is described in details by [14].

In the Wisconsin breast cancer example, we randomly choose 200 cases for training and the remainder for testing; in the Lung cancer example, 24 randomly chosen cases are for training and the remainder for testing. For each pair of training and testing sets, tuning is conducted for λ using the same grid search scheme as in the simulated examples. Since $\mathcal{S}(\mathbf{B}_{cds})$ is unavailable in these real examples, the test errors and the dimension of $\widehat{\mathbf{B}}$, averaged over 100 simulation replications, are summarized in Table 2.

In the Wisconsin breast cancer example, SSDR yields the smallest test error than other dimension reduction methods, and compares favorably to L1MSVM and LDA as well. Furthermore, both SSDR and SIR find one significant direction for the CDS in all replications, SAVE identifies too many significant directions for the CDS which deteriorates its classification accuracy, and pHd does not

TABLE 2
Averaged test errors and dimensions as well as estimated standard errors (in parenthesis) of SIR, SAVE, pHd, SSDR, L1MSVM and LDA in the real examples. Here WBC and Lung correspond respectively to the Wisconsin breast cancer and the Lung cancer examples

		WBC	Lung
SIR	Dim	1.00(.000)	–
	TE	0.048(.0010)	–
SAVE	Dim	8.12(.263)	–
	TE	0.269(.0095)	–
pHd	Dim	0.00(.000)	–
	TE	0.373(.0016)	–
SSDR	Dim	1.00(.000)	1.74(.073)
	TE	0.024(.0006)	0.368(.0104)
L1MSVM	Dim	–	–
	TE	0.029(.0009)	0.381(.0174)
LDA	Dim	–	–
	TE	0.052(.0011)	0.529(.0151)

find any significant direction and yields the worst classification performance. In the Lung cancer example, SIR, SAVE and pHd are not applicable because of the large-p-small-n scenario, while SSDR yields averaged test error 0.368, which outperforms L1MSVM, LDA and the classification results (test error 0.375, the best among eight popular classification methods) reported in [1], where they use all 36 cases and 56 features for training and estimate the test error through leave-one-out cross validation.

Next, we display both real examples in the obtained reduced subspace via SSDR. As illustrated in Figure 2, the estimated directions by SSDR provide good separation of different classes of samples in both examples. In Wisconsin breast cancer example, patients with smaller values on direction 1 are much more likely to be malignant. In Lung cancer example, patients with negative values on both directions are more likely to have the third type of lung cancer, and patients of the first and second type of lung cancer can be discriminated according the relative largeness in these two directions. Furthermore, in the Wisconsin breast example, the estimated CDS is spanned by

$$\begin{aligned}\text{Direction 1} \quad = \quad & 0.097X_2 + 0.184X_7 + 0.467X_8 + 0.753X_{11} + 0.065X_{15} - \\ & 0.286X_{20} + 0.543X_{22} + 0.582X_{23} + 0.074X_{25} + 0.064X_{27} + \\ & 0.737X_{28};\end{aligned}$$

and in the Lung cancer example, the estimated CDS is spanned by

$$\begin{aligned}\text{Direction 1} \quad = \quad & 0.237X_2 + 0.191X_6 + 0.019X_{13} + 0.033X_{14} - 0.23X_{19} - \\ & 0.651X_{20} - 0.013X_{25} + 0.199X_{28} + 0.161X_{34} - 0.089X_{41}, \text{ and} \\ \text{Direction 2} \quad = \quad & -0.077X_2 - 0.132X_3 - 0.05X_4 + 0.114X_8 + 0.025X_{15} - \\ & 0.105X_{19} - 0.251X_{20} + 0.05X_{23} + 0.127X_{29} + 0.249X_{30} + \\ & 0.031X_{31} - 0.036X_{34} + 0.222X_{35} - 0.035X_{37} - 0.053X_{39} + \\ & 0.117X_{41} - 0.017X_{53} + 0.024X_{55}.\end{aligned}$$

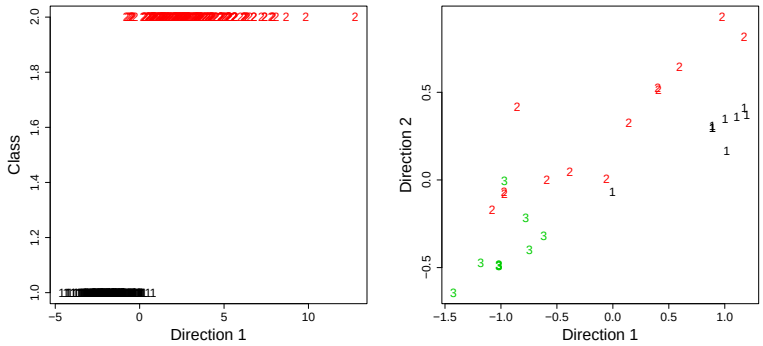

FIG 2. *The estimated CDS based on two randomly chosen replications from the Wisconsin breast cancer and the Lung cancer example. The left panel is the Wisconsin breast cancer example and the right panel is the Lung cancer example, where the digits denote different classes respectively.*

Evidently, SSDR yields sparse representation of the dimension reduction directions in both examples. In Wisconsin breast cancer example, one direction with only 11 out of 30 covriates is identified, and in Lung cancer example, two directions with only 10 and 18 out of 56 covariates are identified.

5. Summary

This article proposes a novel methodology for estimating the dimension reduction subspace for classification. In contrast to existing methods viewing the classification problem as a generalized regression problem, our proposed method directly pursues the minimal sufficient discriminant subspace to retain the classification boundary based on the original dataset. In addition, it enables estimation of the dimension and the sparse directions of the dimension reduction subspace simultaneously. Its asymptotic classification performance is shown to be comparable to the standard classification techniques in large-p-small-n scenario. Numerical analyses demonstrate that the proposed method outperforms several other top competitors in both simulated and real examples.

It is worthy of pointing out that we assume that the true classification boundaries are linear as linear classification boundaries seem more appropriate when the data dimension is high [12]. This assumption may fail when the dimension is relatively small and the true classification boundaries are often nonlinear. The extension of the proposed dimension reduction approach to nonlinear case is under investigation.

References

[1] AEBERHARD, S., COOMANS, D. AND DE VEL, O. (1994). Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recog.*, **27**, 1065-1077.

[2] Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, **19**, 563-570.
[3] Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York. MR1645673
[4] Cook, R.D. and Lee, H. (1999). Dimension reduction in regressions with a binary response. *J. Am. Stat. Assoc.*, **94**, 1187-1200. MR1731482
[5] Cook, R.D. and Li, B. (2002). Dimension reduction for the conditional mean. *Ann. Statist.*, **30**, 455-474. MR1902895
[6] Cook, R.D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Aust. N. Z. J. Statist.*, **43**, 147-199. MR1839361
[7] Cook, R.D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction" by K.C. Li. *J. Am. Statist. Assoc.*, **86**, 328-332. MR1137117
[8] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Statist. Assoc.*, **78**, 316-331. MR0711106
[9] Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *J. Am. Statist. Assoc.*, **99**, 619-632. MR2090899
[10] Fan, J. and Gijbels, I. (1996). *Local Polynomical Modelling and Its Applications*. Chapman and Hall, London. MR1383587
[11] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J. and Caligiuri, M. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-6.
[12] Hastie, Tibshirani and Friedman (2009). *The elements of statistical learning, 2nd Edition*. Springer-Verlag, New York.
[13] Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Am. Statist. Assoc.*, **84**, 986-995. MR1134488
[14] Hong, Z. and Yang, J. (1991). Optimal discriminant plane for a small number of samples and design method of classifier on the plane, *Pattern Recog.*, **24**, 317-324. MR1103954
[15] Hooper, J. (1959). Simultaneous equations and canonical correlation theory. *Econometrica*, **27**, 245-256. MR0105769
[16] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321-377.
[17] Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines, theory and application to the classification of microarray data and satellite radiance data. *J. Am. Statist. Assoc.*, **99**, 67-81. MR2054287
[18] Li, K.C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Am. Statist. Assoc.*, **86**, 316-342. MR1137117
[19] Li, K.C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Assoc.*, **87**, 1025-1039. MR1209564

[20] Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, **94**, 603-613. MR2410011
[21] Lin, Y. (2002). Support Vector Machines and the Bayes Rule in Classification. *Data Mining and Knowledge Discovery*, **6**, 259-275. MR1917926
[22] Liu, Y. and Shen, X. (2006). Multicategory ψ-learning. *J. Am. Statist. Assoc.*, **101**, 500-509. MR2256170
[23] Shen, X. and Huang, H-C. (2006). Optimal model assessment, selection and combination. *J. Am. Statist. Assoc.*, **101**, 554-568. MR2281243
[24] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York. MR1641250
[25] Wang, J. and Shen, X. (2006). Estimation of generalization error: random and fixed inputs. *Statist. Sinica*, **16**, 569-588. MR2267250
[26] Wang, L. and Shen, X. (2007). On L1-Norm Multiclass Support Vector Machines: Methodology and Theory. *J. Am. Statist. Assoc.*, **102**, 583-594. MR2370855
[27] Wang, S., Nan, B., Zhou, N. and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika*, **96**, 307-322. MR2507145
[28] Wolberg, W.H. and Mangasarian, O.L. (1990). Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology. *Proc. Natl. Acad. of Sci.*, **87**, 9193-9196.
[29] Xia, Y., Tong, H., Li, W.K. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc., Ser. B*, **64**, 363-410. MR1924297
[30] Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*, **5**, 1225-1251. MR2248016