Published in final edited form as:

Biometrics. 2014 September ; 70(3): 546–555. doi:10.1111/biom.12174.

Probability-Enhanced Sufficient Dimension Reduction for Binary Classification

Seung Jun Shin, Yichao Wu, Hao Helen Zhang, and Yufeng Liu

Department of Mathematics, University of Arizona, P.O. Box 210089, Tucson, Arizona 85721-0089, U.S.A.

Hao Helen Zhang: hzhang.work@gmail.com

SUMMARY

In high-dimensional data analysis, it is of primary interest to reduce the data dimensionality without loss of information. Sufficient dimension reduction (SDR) arises in this context, and many successful SDR methods have been developed since the introduction of sliced inverse regression (SIR) [Li (1991) *Journal of the American Statistical Association* 86, 316–327]. Despite their fast progress, though, most existing methods target on regression problems with a continuous response. For binary classification problems, SIR suffers the limitation of estimating at most one direction since only two slices are available. In this article, we develop a new and flexible probability-enhanced SDR method for binary classification problems by using the weighted support vector machine (WSVM). The key idea is to slice the data based on conditional class probabilities of observations rather than their binary responses. We first show that the central subspace based on the conditional class probability is the same as that based on the binary response. This important result justifies the proposed slicing scheme from a theoretical perspective and assures no information loss. In practice, the true conditional class probability is generally not available, and the problem of probability estimation can be challenging for data with large-dimensional inputs. We observe that, in order to implement the new slicing scheme, one does not need exact probability values and the only required information is the relative order of probability values. Motivated by this fact, our new SDR procedure bypasses the probability estimation step and employs the WSVM to directly estimate the order of probability values, based on which the slicing is performed. The performance of the proposed probability-enhanced SDR scheme is evaluated by both simulated and real data examples.

Keywords

Binary classification; Conditional class probability; Fisher consistency; Sufficient dimension reduction; Weighted support vector machines (WSVMs)

1. Introduction

Binary classification is commonly encountered in a variety of biomedical applications such as image classification and cancer type classification. Due to rapid advances of technologies,

high dimensional data are frequently collected in science and industry. When dealing with high dimensional data, it is essential to reduce the dimensionality of the original predictor space prior to conducting statistical analysis, because even simple analysis tools such as histograms and scatter plots can be seriously challenged or even break down when the number of variables is enormously large.

In binary classification, we are given data with a binary response $Y \in \{-1, +1\}$ and a p -dimensional predictor $\mathbf{X} = (X_1, \dots, X_p)^T \in \mathbb{R}^p$. The goal is to train a decision rule which assigns labels to future data based on the predictor $\mathbf{x}$. Classification problems are typically solved by estimating the conditional class probability function $p(\mathbf{x}) = P(Y = 1 | \mathbf{X} = \mathbf{x})$ first, and then making the decision based on the magnitude of $p(\mathbf{x})$. Parametric methods assume that the dependency of $p(\mathbf{x})$ on $\mathbf{x}$ can be modeled by a finite number of parameters. Regularization approaches such as the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and adaptive LASSO (Zou, 2006; Zhang and Lu, 2007) are often employed to reduce the dimensionality of the predictor space under the sparsity assumption.

Another way to handle high dimensional data is sufficient dimension reduction (SDR), which is developed in the recent literature to reduce the dimensionality of the predictor space under weak assumptions. The SDR assumes that

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T \mathbf{X}, \quad (1)$$

where $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_d) \in \mathbb{R}^{p \times d}$ and $\perp\!\!\!\perp$ denotes statistical independence. Essentially, the SDR assumes that the response Y is independent of the predictors $\mathbf{X}$ given some linear combinations $\mathbf{B}^T \mathbf{X}$ of $\mathbf{X}$. Compared to sparse regression, the SDR model assumption is less stringent since (1) does not impose any assumption on the relationship between the response and predictor except the conditional independence. Notice that $\mathbf{B}$ is not unique since any linear combination of $\mathbf{B}$ satisfies (1) as long as $\mathbf{B}$ itself does. Consequently, the target of interest in SDR is not on $\mathbf{B}$ itself but the space spanned by the columns of $\mathbf{B}$. The *central subspace*, denoted by $\mathcal{S}_{Y|\mathbf{X}}$, is the intersection of the spaces spanned by all $\mathbf{B}$ satisfying (1) and hence has the minimal dimension among all such $\mathbf{B}$ s (Cook, 1998b). Under mild conditions (Cook, 1996), $\mathcal{S}_{Y|\mathbf{X}}$ exists uniquely. From now on, we assume that $\mathbf{B}$ spans the central subspace, that is, $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\mathbf{B})$. In the literature, the dimension of $\mathcal{S}_{Y|\mathbf{X}}$, denoted by d , is called the structure dimensionality.

There are a variety of methods to implement SDR in the literature. Two early works are the sliced inverse regression (SIR; Li, 1991) and the sliced average variance estimation (SAVE; Cook and Weisberg, 1991), which continue to be used widely. Other methods include, but are not limited to, principal Hessian direction (pHd, Li, 1992; Cook, 1998a), partial least square (PLS, Boulesteix, 2004; Li, Cook, and Tsai, 2007), Fourier method (Zhu and Zeng, 2006), directional regression (DR, Li and Wang, 2007), kernel dimension reduction (KDR, Fukumizu, Bach, and Jordan, 2009), and sufficient component analysis (SCA, Yamada et al., 2011). However, most existing SDR methods are primarily designed to reduce dimension for data with a continuous response, even though some of them work fairly well for data with a categorical response. The performance of some popular SDR methods may suffer severely when the response is binary. For example, SIR can identify at most one

direction in binary classification since there are only two slices available, and SAVE is known for its estimation inefficiency (Li and Wang, 2007) when the response is binary. The book by Cook (1998b) elaborated the difficulty of SDR for binary classification in Chapter 5. See also Cook and Lee (1999) for more discussions on SDR for data with a binary response.

In the context of classification, there are other dimension reduction methods than SDR in the literature based on the classical Fisher's between-within variance idea. These include, but are not limited to, quadratic discrimination (QD, Young, Marco, and Odell, 1987), computer-intensive dimension reduction (CIDR, Röhl and Weihs, 1999), and discriminant coordinates (DC, Hennig, 2004). These approaches are useful in practice but also have some limitations. For example, both QD and CIDR methods require normality of the covariates, which might be too strong in some applications. The DC method can estimate at most one direction for binary classification problems, just like SIR.

As aforementioned, many SDR methods were originally developed for data with a continuous response. Even if some methods are designed to work regardless of the response type, such as the PLS, Fourier method, and SCA, they are not very efficient for binary classification since the information contained in binary responses is generally not so much as that in continuous responses. This motivates us to develop a SDR method specially for dimension reduction in binary classification. In order to illustrate the unsatisfactory performance of commonly-used SDR methods in binary classification, we now give an illustration using the Wisconsin Diagnostic Breast Cancer data (WDBC; Mangasarian, Street, and Wolberg, 1995). The data set is available at the UCI machine learning repository website (<http://archive.ics.uci.edu/ml/index.html>). The WDBC data record diagnosis results of breast cancer for 569 subjects. For each subject, ten features of cell nuclei are measured from a digital image of a fine needle aspirate (FNA) of a breast mass. The mean, standard error, and the largest values are computed for each feature, leading to 30 predictors in total. Figure 1 shows the scatter plot of the projected data into the low-dimensional central subspace spanned by the estimated directions for three popular SDR estimation methods, SIR, SAVE and pHd. Note that SIR can estimate only one direction for binary classification problems, while SAVE and pHd can estimate more than one direction. In panel (a) we plot the binary response versus the SDR predictor $\hat{\mathbf{b}}_1^T \mathbf{x}$ estimated by SIR. Panels (b) and (c) show the scatter plots of the estimated first two SDR predictors ($\hat{\mathbf{b}}_1^T \mathbf{x}$ vs. $\hat{\mathbf{b}}_2^T \mathbf{x}$) for SAVE and pHd, respectively. In Figure 1, the data points from two classes are respectively denoted by circles and plus marks. In order to achieve better classification accuracy, it is desirable to make the data points from two classes more separable after projecting them into the estimated central subspace, where all the discrimination information in predictors is concentrated. Yet there is no clear separation shown in Figure 1. The SDR predictors estimated by SAVE and pHd suggest that some complex and nonlinear classification rules would be needed to separate two classes better.

In this article, we develop and study a new class of SDR approaches to effectively estimating the central subspace for binary classification. The key idea is to slice data based on the conditional class probability $p(\mathbf{x})$ instead of the binary response $\mathbf{Y}$. We first show that

the conditional class probability $p(\mathbf{x})$ contains the same amount of information as the response $\mathbf{Y}$ in terms of SDR. This is achieved by proving that $\mathcal{S}_{Y|X}$ is identical to $\mathcal{S}_{p(X)|X}$, the central subspace for the regression of $p(\mathbf{X})$ on $\mathbf{X}$. We then propose the new “PRobability-Enhanced” SDR (PRE-SDR) schemes which estimate the central subspace by slicing observations based on $p(\mathbf{x})$. Since the true conditional class probability $p(\mathbf{x})$ is generally unknown, intuitively one would need to first estimate $p(\mathbf{x})$ before slicing. However, we observe that probability estimation is not needed to implement the PRE-SDR schemes, as the only required information is the relative orders of probabilities rather than their exact values. In order to learn the orders of probabilities, we propose to employ the weighted support vector machines (WSVM, Lin, Lee, and Wahba, 2004), which is known to be Fisher consistent and hence provides a consistent estimator of $\text{sign}[p(\mathbf{x}) - \pi]$ for any $\pi \in (0, 1)$ and any $\mathbf{x}$. Then we slice the observations based on a sequence of WSVM decision rules associated with different weights. In order to maintain the model-free property for the SDR (1), kernel WSVMs are used at the slicing step. In this paper, we develop three PRE-SDR schemes for dimension reduction, each motivated from a different perspective, and study their empirical performance. The WDBC data set is revisited in Section 7 to illustrate performance of the proposed schemes in real applications.

The rest of this article is organized as follows. In Section 2 we rigorously prove the equivalence between $\mathcal{S}_{Y|X}$ and $\mathcal{S}_{p(X)|X}$. Fisher consistency of the WSVM and WSVM-based slicing techniques are introduced in Section 3. In Section 4, we propose three types of probability-enhanced SDR estimation procedures for binary classification. The issue of the structure dimensionality is addressed in Section 5. Simulated examples and the WDBC data analysis results are presented in Sections 6 and 7. Final remarks are in Section 8.

2. Equivalence of $\mathcal{S}_{Y|X}$ and $\mathcal{S}_{p(X)|X}$

We first show an important theoretical result regarding the central subspace for binary classification. Specifically, we prove that the central subspace based on the binary response and that based on the conditional class probability are the same subspace, namely $\mathcal{S}_{Y|X} = \mathcal{S}_{p(X)|X}$. Intuitively, conditional on $\mathbf{X} = \mathbf{x}$ we can think of the binary response $\mathbf{Y}$ as a function $\mathbf{Y}(p(\mathbf{x}), \epsilon)$ of $p(\mathbf{x})$ and a random variable ϵ , where $\epsilon \sim \text{Uniform}(0, 1)$ and is independent of predictors $\mathbf{x}$, defined as $\mathbf{Y}(p(\mathbf{x}), \epsilon) = 1$ if $\epsilon \leq p(\mathbf{x})$ and -1 otherwise. Consequently $\mathbf{Y}$ and $p(\mathbf{x})$ contain exactly the same amount of information about $\mathbf{x}$ since ϵ is independent of $\mathbf{x}$. This property is rigorously shown in Lemma 1.

Lemma 1. $\mathcal{S}_{Y|X} = \mathcal{S}_{p(X)|X}$

Proof. We first prove that $\mathcal{S}_{Y|X} \supseteq \mathcal{S}_{p(X)|X}$. It is enough to show that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{X}^T \mathbf{B}$ for any $\mathbf{B}$ satisfying $\mathbf{X} \perp\!\!\!\perp p(\mathbf{X}) | \mathbf{X}^T \mathbf{B}$. Notice that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | p(\mathbf{X})$. Consequently $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | (\mathbf{X}^T \mathbf{B}, p(\mathbf{X}))$ which, together with $\mathbf{X} \perp\!\!\!\perp p(\mathbf{X}) | \mathbf{X}^T \mathbf{B}$, implies $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{X}^T \mathbf{B}$ due to Proposition 4.6 of Cook (1998b).

We next prove that $\mathcal{S}_{Y|X} \subseteq \mathcal{S}_{p(X)|X}$, which requires to show that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{X}^T \mathbf{B}$ implies $\mathbf{X} \perp\!\!\!\perp p(\mathbf{X}) | \mathbf{X}^T \mathbf{B}$ for any $\mathbf{B}$. Note that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{X}^T \mathbf{B}$ implies that $p(\mathbf{X}) = E((\mathbf{Y} + 1)/2 | \mathbf{X}) = E(\mathbf{Y} | \mathbf{X}^T \mathbf{B})/2 + 1/2$, and consequently $\mathbf{X} \perp\!\!\!\perp p(\mathbf{X}) | \mathbf{X}^T \mathbf{B}$. This completes the proof.

3. Probability-Based Slicing Via WSVM

As shown in Section 2, the conditional class probability $p(\mathbf{x})$ contains predictors' complete information in binary classification. In practice, however, conditional class probabilities of data are generally not available. This seems to suggest that probability estimation is necessary prior to dimension reduction. Yet it turns out that, similar to classical slicing methods, the proposed PRE-SDR methods use only the relative order of responses rather than their exact values to slice observations. Therefore, one can implement the new SDR methods without estimating the conditional probabilities. To learn the relative order of the conditional probabilities, we propose to train a sequence of WSVMs, each associated with a different weight parameter, and then summarize the WSVM solutions. In the following, we first provide a brief review on the WSVM.

In a binary classification problem, we are given a data set $\{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$ of *i.i.d.* observations from some distribution, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$ denote the predictor vector and the binary response, respectively. Then, for a nonnegative definite kernel function $K(\cdot, \cdot)$ and any weight parameter $\pi \in (0, 1)$, the kernel WSVM estimates a decision function f_π by solving

$$\min_{f_\pi \in \mathcal{F}_K} (1 - \pi) \sum_{i: y_i = 1} H_1 \left(y_i f_\pi(\mathbf{x}_i) \right) + \pi \sum_{i: y_i = -1} H_1 \left(y_i f_\pi(\mathbf{x}_i) \right) + \frac{\lambda}{2} \|f_\pi\|_{\mathcal{F}_K}^2, \quad (2)$$

where $H_1(u) = (1 - u)_+ = \max(1 - u, 0)$ is the hinge loss, $\mathcal{F}_K$ is the reproducing kernel

Hilbert space (RKHS, Wahba, 1990) associated with the kernel K , and $\|f\|_{\mathcal{F}_K}^2$ represents the squared RKHS norm of f in $\mathcal{F}_K$. Here $\lambda > 0$ is a regularization parameter controlling the balance between data fit and model complexity. According to the Representer Theorem (Kimeldorf and Wahba, 1971), the solution to (2) has a finite representation form

$f_\pi(\mathbf{x}) = b_\pi + \frac{1}{\lambda} \sum_{i=1}^n \alpha_{i,\pi} y_i K(\mathbf{x}, \mathbf{x}_i)$, where $(b_\pi, \alpha_{1,\pi}, \dots, \alpha_{n,\pi})$ solves the optimization problem

$$\min_{b_\pi, \alpha_{1,\pi}, \dots, \alpha_{n,\pi}} (1 - \pi) \sum_{i: y_i = 1} H_1(y_i f_\pi(\mathbf{x}_i)) + \pi \sum_{i: y_i = -1} H_1(y_i f_\pi(\mathbf{x}_i)) + \frac{1}{2\lambda} \sum_{i=1}^n \sum_{j=1}^n \alpha_{i,\pi} \alpha_{j,\pi} y_i y_j K(\mathbf{x}_i, \mathbf{x}_j). \quad (3)$$

We denote the optimizer of (3) by $(\hat{b}_\pi, \hat{\alpha}_{1,\pi}, \dots, \hat{\alpha}_{n,\pi})$ and the estimated decision function by

$\hat{f}_\pi(\mathbf{x}) = \hat{b}_\pi + \frac{1}{\lambda} \sum_{i=1}^n \hat{\alpha}_{i,\pi} y_i K(\mathbf{x}, \mathbf{x}_i)$. More information about the WSVM can be found in Lin et al. (2004).

One important property of the WSVM solution is its Fisher consistency, shown by Lin et al. (2004). In other words, for any $\pi \in (0, 1)$, the WSVM classification rule $\text{sign}\{\hat{f}_\pi(\mathbf{x})\}$ is a consistent estimator of the Bayes rule $\text{sign}[p(\mathbf{x}) - \pi]$. To implement the SDR, it is enough to know whether $p(\mathbf{x}_i)$ lies in a pre-specified slice, say (π_1, π_2) for some $0 < \pi_1 < \pi_2 < 1$, for $i = 1, \dots, n$. It is easily seen that, by training a sequence of WSVMs for different weights, the set $\{i : \pi_1 < p(\mathbf{x}_i) < \pi_2\}$ can be estimated consistently. In particular, let $\hat{f}_{\pi_1}(\cdot)$ and $\hat{f}_{\pi_2}(\cdot)$

denote the decision functions from (3), respectively corresponding to π_1 and π_2 . Then the index set $\{i: \hat{f}_{\pi_1}(x_i) > 0 \text{ and } \hat{f}_{\pi_2}(x_i) < 0\}$ provides a consistent estimator of $\{i: \pi_1 < p(x_i) < \pi_2\}$.

4. Probability-Enhanced (PRE) Dimension Reduction Methods

For simplicity, we assume for now that the structure dimensionality d is known. In order to conduct SDR, we need to estimate a candidate matrix, denoted by $\mathbf{M}$, whose first d leading eigenvectors span the central subspace. In this section we propose three estimators of $\mathbf{M}$ for binary classification problems based on the aforementioned probability-enhanced slicing scheme via the WSVM (3). We refer them as PRObability-Enhanced SDR (PRE-SDR) procedures for binary classification.

4.1. PRE-SIR₁

We first propose a straightforward extension of SIR using $p(x)$. Given a fixed grid, $\pi_0 = 0 < \pi_1 < \dots < \pi_h < \dots < \pi_{H-1} < \pi_H = 1$, we repeatedly train the WSVM by solving (3). For π_h , $h = 1, \dots, H-1$, denote the solution as $\hat{f}_{\pi_h}(\cdot)$. then the h th slice is defined as

$$I_{\text{SIR1}}^{(h)} := S_{\text{SIR1}}^{(h)} \setminus S_{\text{SIR1}}^{(h-1)}, h=1, \dots, H,$$

where $S_{\text{SIR1}}^{(h)} = \{i: \hat{f}_{\pi_h}(x_i) < 0\}$ for $h = 1, 2, \dots, H-1$, $S_{\text{SIR1}}^{(0)} = \emptyset$ and $S_{\text{SIR1}}^{(H)} = \{1, \dots, n\}$.

Theoretically $S_{\text{SIR1}}^{(h)}$ is a non-decreasing sequence of sets satisfying

$$S_{\text{SIR1}}^{(1)} \subseteq \dots \subseteq S_{\text{SIR1}}^{(H)},$$

due to the Fisher consistency of the WSVM. In practice, however, since these sets are estimated based on a finite sample, the non-decreasingness might be violated (Wang, Shen, and Liu, 2008). To handle this potential issue, we instead use

$$\tilde{I}_{\text{SIR1}}^{(h)} = \tilde{S}_{\text{SIR1}}^{(h)} \setminus \tilde{S}_{\text{SIR1}}^{(h-1)},$$

where $\tilde{S}_{\text{SIR1}}^{(h)} = \bigcup_{k=1}^h S_{\text{SIR1}}^{(k)}$ for $h = 1, \dots, H$. An estimate of the candidate matrix is constructed as

$$\hat{\mathbf{M}}_{\text{SIR1}} = \sum_{h=1}^H \frac{n_h}{n} \hat{\mathbf{z}}_h \hat{\mathbf{z}}_h^T, \quad (4)$$

where $\hat{\mathbf{z}}_h = n_h^{-1} \sum_{i=1}^n \hat{\Sigma}_n^{-1/2} (x_i - \bar{x}_n) \cdot \mathbb{1}\{i \in \tilde{I}_{\text{SIR1}}^{(h)}\}$, $n_h = \sum_{i=1}^n \mathbb{1}\{i \in \tilde{I}_{\text{SIR1}}^{(h)}\}$, and $\bar{x}_n$ and $\hat{\Sigma}_n$ denote the usual sample mean and covariance of $\{x_1, \dots, x_n\}$, respectively. We denote the first d leading eigenvectors of $\hat{\mathbf{M}}_{\text{SIR1}}$ as $\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_d$. Then $(\hat{\mathbf{b}}_1, \dots, \hat{\mathbf{b}}_d) = \hat{\Sigma}_n^{-1/2} (\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_d)$ gives an estimate of the basis of $\mathcal{S}_{p(X)|X}$, or equivalently $\mathcal{S}_{Y|X}$.

4.2. PRE-SIR₂

For slicing-based dimension reduction procedures, it is sometimes recommended to use similar numbers of observations in each slice. Note that PRE-SIR1 directly controls the number of slices but not the number of observations in each slice, since it is difficult to decide how many observations in each slice before training the WSVMs. In order to tackle this issue, we propose to use a solution path of the WSVM as a function of π , which is called the π -path.

The piecewise linearity of the π -path for the solution to (3) with varying π is shown by Wang et al. (2008). We have recently developed the path algorithm in R by Shin, Wu, and Zhang (in press). The piecewise linearity property suggests that the WSVM solutions $\hat{\alpha}_{i,\pi}, i=1, \dots, n$ as well as $\hat{b}_\pi$ move piecewise linearly in π for any fixed regularization λ . As a consequence, the estimated decision function $\hat{f}_\pi(\mathbf{x})$ is also piece-wise linear in π for any given $\mathbf{x}$. To better illustrate this, we consider a simulated dataset with $n = 30$ and $p = 10$ and plot the solution paths for $\hat{f}_\pi(\mathbf{x}_i)$. The data are generated from Model I as described in Section 6. Employing the Gaussian kernel, the corresponding π -paths of the decision functions $\hat{f}_\pi(\mathbf{x}_i), i=1, 2, \dots, n$, for a fixed λ , are depicted in Figure 2.

In Figure 2, each dashed line corresponds to one particular observation. With these π -paths, one can readily compute a π such that $\hat{f}_\pi(\mathbf{x}_i)=0$ and denote such π by $\hat{\pi}_i$ for $i = 1, \dots, n$. Similarly, such a π value exists uniquely at the theoretical level due to Fisher consistency of the WSVM. However this might not be the case in practice, as these decision functions are estimated from a finite sample. To deal with this, we have considered three different definitions of $\hat{\pi}_i$: (i) $\hat{\pi}_i := \min\{\pi: \hat{f}_\pi(\mathbf{x}_i)=0\}$ (minimum); (ii) $\hat{\pi}_i := \max\{\pi: \hat{f}_\pi(\mathbf{x}_i)=0\}$ (maximum); and (iii) $\hat{\pi}_i := \text{average}\{\pi: \hat{f}_\pi(\mathbf{x}_i)=0\}$ (average). Our limited empirical results suggest that the average type performs the best. Therefore the definition (iii) is used in our numerical examples.

Notice that $\hat{\pi}_i$ can be regarded as a reasonable surrogate of $p(\mathbf{x})$ by the consistency of the WSVM (Wang et al., 2008). Therefore we can readily define slices based on the order of $\hat{\pi}_i, i=1, \dots, n$. Finally, we estimate the basis of $\mathcal{S}_{Y|X}$ from $\hat{M}_{\text{SIR}2}$, which is defined in a similar manner to $\hat{M}_{\text{SIR}1}$ in (4) by replacing $I_{\text{SIR}1}^{(h)}$ with accordingly defined slices based on $\hat{\pi}_i, i=1, \dots, n$.

4.3. PRE-CUME

Zhu, Zhu, and Feng (2010) proposed a cumulative slicing estimation (CUME) method for SDR, also motivated from SIR. We first briefly review the CUME for the case with a continuous response y and then develop a probability-enhanced CUME, PRE-CUME, for binary classification problems.

Suppose the response variable is continuous. If $H = 2$ and the two intervals of the response are sliced by a fixed single point $\tilde{y}$, then it is not difficult to show that the SIR estimate can be written as a function of $\tilde{y}$ as follows

$$\hat{\mathbf{m}}(\tilde{y}) = n^{-1} \sum_{i=1}^n z_i z_i^T \mathbb{1}(y_i \leq \tilde{y}) \cdot \omega(\tilde{y}),$$

where $z_i = \hat{\Sigma}_n^{-1/2}(x_i - \bar{x}_n)$, $i=1, \dots, n$, and $\omega(\cdot)$ is a weight function with range $[0, 1]$. Since Zhu et al. (2010) have empirically showed that $\omega(\cdot) = 1$ works reasonably well, we omit it thereafter for simplicity. The CUME estimate of the candidate matrix is given by

$n^{-1} \sum_{i=1}^n \hat{\mathbf{m}}(y_i)$. The CUME method is different from the SIR in the sense that it aggregates all information for the unconditional mean $E(\mathbf{x} \mathbb{1}(Y \leq \tilde{y}))$ for all $\tilde{y}$ in order to estimate $\mathcal{S}_{Y|X}$, while SIR uses that for the conditional mean $E(\mathbf{x}|Y)$. We notice further that the CUME exploits the observed responses through their order only, making it natural to develop its probability-enhanced version for binary classification problems.

In the context of binary responses, we propose to use the WSVM to obtain the π -paths and hence $\hat{\pi}_i$, $i=1, \dots, n$. Then the PRE-CUME estimate for binary classification is defined as

$$\hat{\mathbf{M}}_{\text{CUME}} := n^{-1} \sum_{i=1}^n \hat{\mathbf{m}}(\hat{\pi}_i),$$

where $\hat{\mathbf{m}}(\hat{\pi}) = n^{-1} \sum_{i=1}^n z_i z_i^T \mathbb{1}(\hat{\pi}_i \leq \hat{\pi})$. Denote the first d leading eigenvectors of $\hat{\mathbf{M}}_{\text{CUME}}$ by $\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_d$. Then we use $(\hat{\mathbf{b}}_1, \dots, \hat{\mathbf{b}}_d) = \hat{\Sigma}_n^{-1/2}(\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_d)$ as an estimate of the basis.

5. Estimation of Structure Dimensionality

In the previous section, the structure dimensionality d is assumed to be known. However, this information is not available in real problems. Consequently, another important issue is to estimate the structure dimensionality d in the SDR. In the literature, each SDR method has its own procedure for determining d . For example, in the context of SIR procedures, various chi-squared tests have been developed (Li, 1991; Schott, 1994; Bura and Cook, 2001). However, their extension to other methods beyond SIR, even for SAVE, is not straightforward. The BIC-type approach is a popular alternative in the literature (Zhu et al., 2006, 2010). Both of these approaches are developed based on the asymptotic distribution (or normality) of the candidate matrix estimator.

We note that it is very challenging to fully explore asymptotic properties of the estimators proposed in this paper. The main reason is that Fisher consistency of the WSVM solution holds pointwisely and hence it is not straightforward to rigorously study the random behavior of slices induced from the consistency of the WSVM solution. One may use resampling techniques such as bootstrap (Ye and Weiss, 2003; Zhu and Zeng, 2006). However, they can be computationally intensive because the WSVM itself employs a numerical procedure to obtain the solution. A simple alternative is to use the cumulative ratio of the eigenvalues to choose d . It is in common use among many other ad hoc criteria used by classical dimension reduction methods such as principal component analysis

(Jolliffe, 2002). In the following real data analysis in Section 7, for simplicity, we choose d such that its cumulative ratio firstly exceeds 70%.

6. Simulated Examples

We use several simulation examples to illustrate finite sample performance of three probability-enhanced SDR methods proposed in Section 4. All of these methods have a regularization parameter λ to be tuned for training the WSVM. In the context of SDR, the selection of λ is not straightforward, since the target of interest is a space rather than a numerical statistic or function. Nevertheless, it is reasonable to indirectly tune λ in the context of the (weighted) SVM, since the proposed PRE-SDR methods work well as long as the associated WSVMs perform reasonably. We use fivefold cross validation to choose λ by minimizing the cross validation error rate of the unweighted SVM in the numerical examples. There are various choices of kernel functions for the WSVM. Our limited numerical investigations show that the performance of the PRE-SDR methods is not very sensitive to the choice of the kernel. We employ Gaussian kernel $\mathbf{K}(\mathbf{x}, \mathbf{x}') = \exp\{-\|\mathbf{x} - \mathbf{x}'\|^2/(2\sigma^2)\}$ due to its flexibility. The linear kernel $\mathbf{K}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$ may not be adequate since the linear classification boundary is a strong assumption, while the SDR only assumes conditional independence. The choice of the kernel parameter is important to guarantee good performance of the PRE-SDR schemes. The hyper parameter σ in Gaussian kernel is chosen to be the median of pairwise distances between all pairs of points from two classes (Jaakkola, Diekhans, and Haussler, 1999). The number of slices H is another quantity that may affect the performance of the first two proposed methods: PRE-SIR₁ and PRE-SIR₂. However, we observe that the proposed methods are not sensitive to the choice of H and thus we set $H = 10$ in our numerical examples.

For comparison, we consider five popular SDR methods: SAVE, pHd, PLS, Fourier method (Zhu and Zeng, 2006), and SCA (Yamada et al., 2011). The standard SIR is not considered in the simulation study since the true underlying structure dimensionality of the examples is large than one, while SIR can estimate at most one direction for binary classification problems.

6.1. Examples for the Central Subspace with $d = 2$

We generate the data from the following three models,

- Model I: $Y = \text{sign} \{X_1 / [0.5 + (X_2 + 1)^2] + 0.2\varepsilon\}$
- Model II: $Y = \text{sign} \{(X_1 + 0.5)(X_2 - 0.5)^2 + 0.2\varepsilon\}$
- Model III: $Y = \text{sign} \{\sin X_1 / e^{X_2} + 0.2\varepsilon\},$

where $\mathbf{X} \sim N_p(\mathbf{0}_p, \mathbf{I}_p)$ with a p -dimensional zero vector $\mathbf{0}_p$ and the identity matrix $\mathbf{I}_p$, and $\varepsilon \sim N(0, 1)$. The true $\mathbf{B} = (\mathbf{e}_1, \mathbf{e}_2)$, where $\mathbf{e}_i$ is a column vector with the i th element being 1 and 0 for the rest. The true structure dimensionality $d = 2$ is assumed to be known. For

performance evaluation, we use $\|\mathbf{P}_B - \mathbf{P}_{\hat{B}}\|_F$, where $\mathbf{P}_A = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ is the projection matrix and $\|\mathbf{A}\|_F$ denotes the Frobenius norm of a matrix $\mathbf{A}$, defined as the square root of the sum of the squares of its elements. Therefore smaller values of the distance indicate better

performance. We consider nine different combinations of sample sizes and input dimensions, with $(n \times p) = \{100, 200, 400\} \times \{10, 20, 30\}$.

Table 1 reports the average Frobenius distance between $\mathbf{P}_B$ and $\mathbf{P}_{\hat{B}}$ over 100 independent repetitions under three model settings for various SDR methods under comparison. The proposed probability-enhanced methods appear to outperform existing ones in terms of the Frobenius distance measure under all the scenarios. As expected, all the methods get better as n increases and worse as p increases. However, it is noteworthy to point out that the performance of the PRE-SDR methods improves faster than the others when the sample size increases. This is because the PRE-SDR methods rely on Fisher consistency of the WSVM. Three PRE-SDR methods perform quite similarly, with the PRE-SIR₁ and the PRE-CUME being slightly better than the PRE-SIR₂. In summary, the proposed PRE-SDR methods outperform existing SDR methods in binary classification, and their advantages become more substantial as the sample size increases.

6.2. Examples for the Central Subspace with $d > 2$

To evaluate performance of the PRE-SDR methods for data with more complex structures, we further consider examples with the underlying structure dimensionality higher than 2 (following the AE's suggestions). We generate the binary response $Y = \text{sign}\{f(\mathbf{x}) + 0.2\varepsilon\}$

with four decision functions: and $f_{31}(\mathbf{x}) = \frac{X_1}{2 + (X_2 + 0.5)(X_3 + 1.5)}$,

$f_{32}(\mathbf{x}) = \frac{\sin(X_1)}{2|X_3|\sqrt{\exp(X_2)}}$, $f_{41}(\mathbf{x}) = \frac{X_1}{2 + (X_2 + 0.5)(X_3X_4 + 1.5)}$ and

$f_{42}(\mathbf{x}) = \frac{\sin(X_1)}{2|X_3X_4|\sqrt{\exp(X_2)}}$, the predictors and ε are generated in the same way as above.

We set $n = 400$ and $p = 10$. The true central subspaces have $\mathbf{B} = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and $d = 3$ for f_{31} and f_{32} , and $\mathbf{B} = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4)$ and $d = 4$ for f_{41} and f_{42} . Results over 100 independent repetitions are reported in Table 2. We again observe that, in these examples with higher dimensional central spaces, the proposed PRE-SDR methods give better performance than other methods in binary classification.

6.3. Empirical Computing Time

Computational efficiency is also an important factor to be considered in practice. Table 3 reports the empirical computing times of all the methods under Model I, with various values of n and p . It is observed that all the methods are reasonably fast except the SCA. The proposed PRE-SDR methods are slower than some existing methods such as SAVE, pHd, PLS, and Fourier methods, but their speed are still quite reasonable. For example, when $n = 400$, $p = 30$, it takes fewer than 0.5, 4.3, and 4.8 seconds, respectively, for PRE-SIR₁, PRE-SIR₂, and PRE-CUME methods, to identify the central subspace. This can be seen as the trade-off between computational time and estimated accuracy. One main reason of extra time for the PRE-SDR methods is that they require to solve WSVMs repeatedly.

7. Application to Wisconsin Diagnosis Breast Cancer Data

We now revisit the WDBC data introduced in Section 1 to illustrate the PRE-SDR estimation schemes for real-world problems. We use the Gaussian kernel with σ and tuning parameter λ chosen in the same manner as in Section 6, and $H = 10$. Figure 3 depicts the scatter plots of the first two sufficient predictors estimated by the proposed three probability-enhanced methods. Comparing Figure 3 to Figure 1, we observe that the first two sufficient predictors estimated from the PRE-based approaches give a better separation between the subjects from two classes.

Table 4 contains the first five leading eigenvalues, denoted by $\lambda_1, \dots, \lambda_5$, of the estimated candidate matrix from different methods. These eigenvalues are used to determine d . Since SCA solves an optimization problem to estimate a basis set of the central subspace and hence there is no eigenvalue associated with SCA. The SIR only gives one eigenvalue greater than zero and therefore it estimates one direction only. It seems that pHd breaks down in this example since there are negative eigenvalues observed. We choose d for the proposed PRE-SDR methods based on cumulative ratios of the eigenvalues, for example $\hat{d}=1, 2$, and 4, respectively, for PRE-SIR₁, PRE-SIR₂, and PRE-CUME when 70% is used as a cut-off value.

In order to see how the estimated central subspace help to improve classification accuracy, we then apply the k -nearest neighbor (kNN) classifier to the projected WDBC data on the central subspaces estimated by SDR. Without loss of generality, we suppose all the estimated $\hat{\mathbf{B}}$ are column-wisely normalized for fair comparison based on the kNN classifiers. First, we randomly split the WDBC data into the training set and the test set, as $\{(y_j^{\text{tr}}, \mathbf{x}_j^{\text{tr}}): j=1, \dots, 284\}$ and $\{(y_{j'}^{\text{ts}}, \mathbf{x}_{j'}^{\text{ts}}): j'=1, \dots, 285\}$. The SDR methods considered in Section 6 are carried out on the training data to obtain $\hat{\mathbf{B}}_{\text{tr}}$. Next, we fit the kNN classifier with respect to $(y_j^{\text{tr}}, \hat{\mathbf{B}}_{\text{tr}}^T \mathbf{x}_j^{\text{tr}})$ and predict $\hat{y}_{j'}^{\text{ts}}$ from $\hat{\mathbf{B}}_{\text{tr}}^T \mathbf{x}_{j'}^{\text{ts}}$ for different values of d . Finally, the test error rates are obtained. Each procedure is independently repeated 100 times. To implement the kNN classifier, we consider four different numbers of neighbors to be averaged, $k = 3, 5, 7, 9$. Figure 4 plots the average test error rates of the kNN with $k = 7$ for different SDR methods as a function of the number of sufficient predictors used for classification, $d = 1, \dots, 30$. We report only the case $k = 7$ to avoid redundancy since the patterns for different k s values are similar. We observe that all the PRE-SDR methods give promising performance in the sense that the corresponding test error rates are less than the one computed from original data represented by solid horizontal line, regardless of d . The SAVE, pHd, and SCA perform unsatisfactorily for this data set while the Fourier method and PLS performs quite well. This is consistent with the simulation results in Section 6.

8. Concluding Remarks

We develop a new class of probability-enhanced sufficient dimension reduction methods for binary classification. The proposed methods employ the WSVM and its Fisher consistency property, and they tactically avoid direct probability estimation. Although asymptotic properties of the proposed methods are not yet studied due to the WSVM's inherent

difficulty, the methods show favorable performance numerically in both simulated and real examples. Based on our limited numerical evidence, while both PRE-SIR₁ and PRE-CUME performs comparably, one may prefer to use PRE-SIR₁ in practice due to its simpler implementation.

We would like to point out that the probability-enhanced schemes can be naturally generalized to other SDR methods, in addition to SIR and CUME. For example, PRE-SAVE and PRE-SDR can be developed. As frequently encountered in recent applications, for example disease detection in genomics, p is larger than n . The PRE-SDR schemes proposed in this paper are not directly applicable to such a case, as they are all based on SIR which does not work when $p > n$. In the context of regression, several methods have been developed to handle “large p , small n ” situations via regularization. See Wu and Li (2011) and references therein. Since the proposed probability-enhanced SDR schemes are applicable wherever a slicing technique is employed, the regularized PRE-SIR for $p > n$ problems can be readily developed for binary classification.

9. Supplementary Materials

R-code for PRE-SIR₁, PRE-SIR₂, and PRE-CUME and the Wisconsin breast cancer data analysis are available with this paper at the *Biometrics* website on Wiley Online Library.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the co-editor, the associate editor, and three reviewers for their constructive comments and suggestions which have led to significant improvement of the article. The authors are supported in part by NSF grants DMS-0747575 (Liu), DMS-1055210 (Wu), DMS-1309507 (Zhang) and DMS-1347844 (Zhang), and NIH grants R01 CA-149569 (Liu, Shin, and Wu), R01 CA-085848 (Zhang), and P01 CA-142538 (Wu, Zhang, and Liu).

References

- Boulesteix A-L. PIs dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*. 2004; 3:1–30.
- Bura E, Cook R. Extending sliced inverse regression: The weighted chi-square test. *Journal of the American Statistical Association*. 2001; 96:996–1003.
- Cook R. Graphics for regressions with a binary response. *Journal of the American Statistical Association*. 1996; 91:983–992.
- Cook R. Principal hessian directions revisited. *Journal of the American Statistical Association*. 1998a; 93:84–94.
- Cook, R. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. New York, NY: Wiley; 1998b.
- Cook R, Lee H. Dimension reduction in binary response regression. *Journal of the American Statistical Association*. 1999; 94:1187–1200.
- Cook R, Weisberg S. Discussion of “sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*. 1991; 86:28–33.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.

- Fukumizu K, Bach F, Jordan M. Kernel dimension reduction in regression. *Annals of Statistics*. 2009; 37:1871–1905.
- Hennig C. Asymmetric linear dimension reduction for classification. *Journal of Computational and Graphical Statistics*. 2004; 13:930–945.
- Jaakkola, T.; Diekhans, M.; Haussler, D. Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology. Heidelberg, Germany: AAAI; 1999. Using the fisher kernel method to detect remote protein homologies; p. 149-158.
- Jolliffe, I. *Principal Component Analysis*. 2nd edition. Berlin, Germany: Springer; 2002.
- Kimeldorf G, Wahba G. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*. 1971; 33:82–95.
- Li B, Wang S. On directional regression for dimension reduction. *Journal of the American Statistical Association*. 2007; 102:997–1008.
- Li K-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*. 1991; 86:316–327.
- Li K-C. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*. 1992; 87:1025–1039.
- Li L, Cook R, Tsai C-L. Partial inverse regression. *Biometrika*. 2007; 94:615–625.
- Lin Y, Lee Y, Wahba G. Support vector machines for classification in nonstandard situations. *Machine Learning*. 2004; 46:191–202.
- Mangasarian O, Street W, Wolberg W. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*. 1995; 43:570–577.
- Röhl, M.; Weihs, C. Optimal vs. classification linear dimension reduction. In: Gaul, W.; Locarek-Junge, H., editors. *Classification in the Information Age*. Berlin, Germany: Springer; 1999. p. 252-259.
- Schott J. Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*. 1994; 89:141–148.
- Shin S, Wu Y, Zhang HH. Two-dimensional solution surface for weighted support vector machines. *Journal of Computational and Graphical Statistics*. 2013 [Epub ahead of print].
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B*. 1996; 58:267–288.
- Wahba, G. *Spline models for observational data*. SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v; 1990. p. 59
- Wang J, Shen X, Liu Y. Probability estimation for large-margin classifier. *Biometrika*. 2008; 95:149–167.
- Wu Y, Li L. Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statistica Sinica*. 2011; 21:707–730. [PubMed: 22140299]
- Yamada, M.; Niu, G.; Takagi, J.; Sugiyama, M. Computationally efficient sufficient dimension reduction via squared-loss mutual information. In: Chun-Nan; Lee, WS., editors. *Proceedings of the Third Asian Conference on Machine Learning (ACML2011), JMLR Workshop and Conference*. Vol. 20. Taoyuan, Taiwan: 2011. p. 247-162.
- Ye Z, Weiss R. Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*. 2003; 98:968–979.
- Young DM, Marco VR, Odell PL. Quadratic discrimination: Some results on optimal low-dimensional representation. *Journal of Statistical Planning and Inference*. 1987; 17:307–319.
- Zhang H, Lu W. Adaptive lasso for cox's proportional hazards model. *Biometrika*. 2007; 94:691–703.
- Zhu Y, Zeng P. Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*. 2006; 101:1638–1651.
- Zhu L, Miao B, Peng H. On sliced inversed regression with high-dimensional covariates. *Journal of the American Statistical Association*. 2006; 101:630–643.
- Zhu L-P, Zhu L-X, Feng Z-H. Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*. 2010; 105:1455–1466.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.

Figure 1.

Results of classical SDR methods for WDBC data: panel (a) plots the response versus the sufficient dimension reduction predictor estimated by SIR; panels (b) and (c) show the first two directions $\hat{b}_1^T x$ vs. $\hat{b}_2^T x$ estimated by SAVE and pHd, respectively. Here, circle – denotes one class while plus the other.

Figure 2.

Solution paths of $\hat{f}(x_i)$ for simulated example: Vertical lines represent $\hat{\pi}_i, i=1, \dots, n$.

Figure 3.

Probability-enhanced SDR for WDBC data: Scatter plots of the first two sufficient predictors estimated by PRE-SIR₁, PRE-SIR₂, and PRE-CUME are depicted, respectively.

7nn: test error rate

Figure 4.

Plot of average test error rates of the kNN classifier with $k = 7$ over 100 times of random partitioning for the WDBC data with respect to the number ($d = 1, 2, \dots, 30$) of leading sufficient predictors that are estimated by different SDR methods. The proposed PRE-SDR methods outperform other existing methods regardless of the value of d . The horizontal solid line corresponds to the average test error rate using the original data.

Table 1

Averaged Frobenius distances between P_B and $P_{\hat{B}}$ over 100 independent repetitions are shown under various scenarios Corresponding standard deviations are given in parentheses. In each scenario, bold case is used to emphasize a winning method in terms of considered measure (and may not be statistically significant)

Model	n	p	SAVE	pHd	PLS	Fourier	SCA	PRE-SIR ₁	PRE-SIR ₂	PRE-CUME
I	100	10	1.431 (0.13)	1.734 (0.15)	1.413 (0.10)	1.373 (0.11)	1.459 (0.14)	1.316 (0.16)	1.318 (0.17)	1.294 (0.18)
		20	1.769 (0.14)	1.879 (0.08)	1.548 (0.07)	1.506 (0.06)	1.710 (0.14)	1.476 (0.10)	1.473 (0.10)	1.471 (0.10)
		30	1.928 (0.06)	1.933 (0.04)	1.635 (0.06)	1.585 (0.07)	1.792 (0.11)	1.581 (0.07)	1.579 (0.07)	1.571 (0.07)
	200	10	1.331 (0.14)	1.682 (0.20)	1.372 (0.10)	1.322 (0.14)	1.361 (0.12)	1.261 (0.17)	1.268 (0.16)	1.248 (0.18)
		20	1.531 (0.10)	1.867 (0.09)	1.467 (0.05)	1.427 (0.07)	1.544 (0.12)	1.370 (0.10)	1.372 (0.11)	1.364 (0.11)
		30	1.756 (0.13)	1.904 (0.06)	1.530 (0.04)	1.499 (0.04)	1.692 (0.12)	1.455 (0.06)	1.462 (0.06)	1.455 (0.06)
	400	10	1.316 (0.15)	1.587 (0.21)	1.348 (0.11)	1.314 (0.14)	1.324 (0.12)	1.142 (0.23)	1.230 (0.21)	1.175 (0.22)
		20	1.407 (0.05)	1.803 (0.13)	1.421 (0.04)	1.403 (0.05)	1.423 (0.07)	1.293 (0.14)	1.318 (0.13)	1.303 (0.14)
		30	1.478 (0.05)	1.881 (0.09)	1.459 (0.03)	1.440 (0.04)	1.543 (0.11)	1.344 (0.10)	1.357 (0.09)	1.345 (0.10)
	100	10	1.612 (0.15)	1.722 (0.20)	1.468 (0.11)	1.427 (0.12)	1.600 (0.15)	1.395 (0.13)	1.382 (0.16)	1.369 (0.14)
		20	1.876 (0.09)	1.887 (0.08)	1.609 (0.07)	1.596 (0.08)	1.805 (0.11)	1.585 (0.09)	1.585 (0.09)	1.563 (0.09)
		30	1.936 (0.04)	1.928 (0.05)	1.695 (0.07)	1.689 (0.07)	1.883 (0.08)	1.689 (0.09)	1.694 (0.08)	1.674 (0.08)
II	200	10	1.450 (0.14)	1.637 (0.18)	1.401 (0.09)	1.353 (0.12)	1.488 (0.12)	1.301 (0.16)	1.298 (0.14)	1.288 (0.16)
		20	1.775 (0.12)	1.852 (0.12)	1.517 (0.06)	1.502 (0.06)	1.727 (0.12)	1.460 (0.09)	1.452 (0.10)	1.436 (0.11)
		30	1.912 (0.06)	1.920 (0.05)	1.588 (0.06)	1.583 (0.06)	1.846 (0.08)	1.553 (0.07)	1.549 (0.08)	1.536 (0.08)
	400	10	1.337 (0.15)	1.452 (0.16)	1.358 (0.11)	1.278 (0.17)	1.411 (0.09)	1.118 (0.23)	1.196 (0.22)	1.173 (0.21)
		20	1.558 (0.13)	1.716 (0.14)	1.453 (0.04)	1.419 (0.07)	1.602 (0.14)	1.349 (0.11)	1.357 (0.11)	1.346 (0.11)
		30	1.781 (0.12)	1.844 (0.10)	1.500 (0.04)	1.487 (0.04)	1.781 (0.13)	1.424 (0.07)	1.427 (0.07)	1.411 (0.08)
	100	10	1.431 (0.12)	1.712 (0.15)	1.408 (0.11)	1.386 (0.10)	1.421 (0.15)	1.315 (0.17)	1.347 (0.17)	1.330 (0.17)
		20	1.748 (0.15)	1.848 (0.10)	1.544 (0.07)	1.497 (0.06)	1.646 (0.12)	1.457 (0.10)	1.465 (0.10)	1.463 (0.09)
		30	1.925 (0.06)	1.911 (0.05)	1.631 (0.06)	1.580 (0.07)	1.746 (0.12)	1.578 (0.06)	1.581 (0.07)	1.571 (0.07)
	200	10	1.341 (0.12)	1.655 (0.19)	1.372 (0.10)	1.338 (0.11)	1.341 (0.12)	1.242 (0.20)	1.265 (0.18)	1.236 (0.19)
		20	1.514 (0.12)	1.812 (0.11)	1.466 (0.04)	1.417 (0.07)	1.493 (0.10)	1.365 (0.11)	1.379 (0.10)	1.368 (0.10)
		30	1.726 (0.13)	1.873 (0.07)	1.528 (0.04)	1.489 (0.05)	1.627 (0.12)	1.450 (0.08)	1.461 (0.07)	1.457 (0.07)
III	400	10	1.307 (0.15)	1.535 (0.21)	1.351 (0.11)	1.306 (0.14)	1.319 (0.13)	1.163 (0.25)	1.202 (0.23)	1.164 (0.24)
		20	1.407 (0.06)	1.740 (0.12)	1.420 (0.04)	1.405 (0.05)	1.405 (0.06)	1.306 (0.14)	1.316 (0.13)	1.289 (0.14)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Model	<i>n</i>	<i>p</i>	SAVE	pHd	PLS	Fourier	SCA	PRE-SIR ₁	PRE-SIR ₂	PRE-CUME
	30		1.471 (0.13)	1.843 (0.15)	1.588 (0.06)	1.438 (0.21)	1.832 (0.12)	1.330 (0.19)	1.351 (0.17)	1.339 (0.17)

Table 2

Averaged Frobenius distances between P_B and $P_{\hat{B}}$ over 100 independent repetitions are shown under various scenarios Corresponding standard deviations are given in parentheses. In each scenario, bold case is used to emphasize a winning method in terms of considered measure (and may not be statistically significant)

f	SAVE	pHd	PLS	Fourier	SCA	PRE-SIR ₁	PRE-SIR ₂	PRE-CUME
f_{31}	1.708 (0.169)	1.885 (0.179)	1.764 (0.168)	1.689 (0.179)	1.727 (0.142)	1.595 (0.174)	1.652 (0.150)	1.641 (0.136)
f_{32}	1.765 (0.135)	1.906 (0.176)	1.766 (0.163)	1.747 (0.150)	1.723 (0.138)	1.558 (0.171)	1.669 (0.170)	1.645 (0.142)
f_{41}	2.152 (0.188)	2.152 (0.188)	2.005 (0.175)	1.994 (0.170)	1.968 (0.141)	1.957 (0.151)	2.014 (0.139)	2.013 (0.136)
f_{42}	2.093 (0.181)	2.093 (0.181)	2.011 (0.167)	2.008 (0.153)	1.954 (0.147)	1.822 (0.160)	1.940 (0.171)	1.938 (0.150)

Table 3

Empirical computing time (in seconds) for different SDR methods averaged over 100 independent repetitions under Model I.

<i>n</i>	<i>p</i>	SAVE	pHd	PLS	Fourier	SCA	PRE		
							SIR ₁	SIR ₂	CUME
100	10	0.0047	0.0040	0.0135	0.0031	1.27	0.130	0.278	0.284
	20	0.0068	0.0062	0.0182	0.0052	1.48	0.149	0.333	0.333
	30	0.0082	0.0070	0.0176	0.0102	1.40	0.161	0.344	0.378
200	10	0.0054	0.0048	0.0150	0.0075	16.86	0.196	0.786	0.882
	20	0.0070	0.0066	0.0187	0.0206	17.44	0.217	0.944	1.087
	30	0.0083	0.0077	0.0192	0.0395	25.81	0.229	0.981	1.242
400	10	0.0060	0.0056	0.0172	0.0297	62.13	0.428	3.049	3.756
	20	0.0089	0.0085	0.0224	0.0773	112.34	0.492	3.610	4.443
	30	0.0123	0.0117	0.0283	0.1396	110.53	0.549	4.342	4.773

Table 4

The first five leading eigenvalues ($\lambda_1, \dots, \lambda_5$) of the candidate matrices estimated by different SDR methods for the WDBC data. Cumulative ratios of the values in percentage are given in parentheses

	SIR	SAVE	pHd	PLS	Fourier	PRE-SIR₁	PRE-SIR₂	PRE-CUME
λ_1	0.774 (100%)	1.666 (11.9%)	0.459 (52.8%)	98.204 (98.2%)	0.061 (26.9%)	0.833 (74.0%)	0.916 (46.9%)	0.092 (58.1%)
λ_2	0.000 (100%)	1.564 (23.0%)	0.453 (104.9%)	1.530 (99.7%)	0.013 (32.7%)	0.112 (84.0%)	0.469 (71.0%)	0.013 (66.5%)
λ_3	0.000 (100%)	1.383 (32.9%)	-0.301 (70.3%)	0.231 (100.0%)	0.011 (37.7%)	0.059 (89.2%)	0.270 (84.9%)	0.004 (68.8%)
λ_4	0.000 (100%)	1.127 (41.0%)	0.295 (104.2%)	0.025 (100.0%)	0.011 (42.4%)	0.044 (93.1%)	0.110 (90.5%)	0.002 (70.3%)
λ_5	0.000 (100%)	0.986 (48.0%)	-0.228 (78.0%)	0.009 (100.0%)	0.010 (46.9%)	0.034 (96.1%)	0.080 (94.6%)	0.002 (71.6%)