Biometrika (2017), **104**, 1, *pp.* 67–81
Printed in Great Britain
doi: 10.1093/biomet/asw057
Advance Access publication 19 January 2017

Principal weighted support vector machines for sufficient dimension reduction in binary classification

By SEUNG JUN SHIN

Department of Statistics, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, South Korea

sjshin@korea.ac.kr

YICHAO WU

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, North Carolina 27695, U.S.A.

wu@stat.ncsu.edu

HAO HELEN ZHANG

Department of Mathematics, University of Arizona, 617 North Santa Rita Ave., P. O. Box 210089, Tucson, Arizona 85721, U.S.A.

hzhang@math.arizona.edu

and YUFENG LIU

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, 354 Hanes Hall, Chapel Hill, North Carolina 27599, U.S.A.

yfliu@email.unc.edu

Summary

Sufficient dimension reduction is popular for reducing data dimensionality without stringent model assumptions. However, most existing methods may work poorly for binary classification. For example, sliced inverse regression (Li, 1991) can estimate at most one direction if the response is binary. In this paper we propose principal weighted support vector machines, a unified framework for linear and nonlinear sufficient dimension reduction in binary classification. Its asymptotic properties are studied, and an efficient computing algorithm is proposed. Numerical examples demonstrate its performance in binary classification.

Some key words: Fisher consistency; Hyperplane alignment; Reproducing kernel Hilbert space; Weighted support vector machine.

1. Introduction

Increasing data dimension can pose challenges at various stages of a statistical analysis. Given a p-dimensional predictor $X \in \mathbb{R}^p$ and a univariate response $Y \in \mathbb{R}$, sufficient dimension reduction assumes that

$$Y \perp\!\!\!\perp X \mid B^{\mathrm{T}}X, \tag{1}$$

where $\perp\!\!\!\perp$ denotes statistical independence. Dimension reduction is achieved by finding a matrix $B \in \mathbb{R}^{p\times k}$ for some $k < p$ while preserving all the information about Y contained in X. The model (1) does not assume any specific relationship between X and Y. Because B in relation (1) is not unique, we define the central subspace, $\mathcal{S}_{Y|X}$, as the intersection of span(B) for all B satisfying (1), where span(B) denotes the space spanned by the columns of B. It is known that $\mathcal{S}_{Y|X}$ exists uniquely under mild conditions (Cook, 1996, 1998b; Yin et al., 2008). One often assumes that span(B) $= \mathcal{S}_{Y|X}$ to facilitate estimation. The dimension k of $\mathcal{S}_{Y|X}$ is called the structural dimension, and its estimation is also crucial.

There are many methods for sufficient dimension reduction. Sliced inverse regression (Li, 1991) and sliced average variance estimation (Cook & Weisberg, 1991) are among the early proposals and continue to be widely used in practice. Others include principal Hessian direction estimation (Li, 1992; Cook, 1998a), iterative Hessian transformation estimation (Cook & Li, 2002), the Fourier method (Zhu & Zeng, 2006), partial least squares estimation (Li et al., 2007), and directional regression (Li & Wang, 2007).

By construction, (1) is linear sufficient dimension reduction. Cook (2007) introduced nonlinear sufficient dimension reduction as a generalization of (1):

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

where $\phi : \mathbb{R}^p \mapsto \mathbb{R}^k$ is an unknown vector-valued function of X. Under (2), dimension reduction is achieved by identifying a possibly nonlinear function ϕ. We assume that ϕ is unique modulo injective transformation to guarantee its identifiability. Several methods have been proposed for nonlinear sufficient dimension reduction (Wu, 2008; Yeh et al., 2009; Wu et al., 2013). Lee et al. (2013) introduced a general theory for nonlinear sufficient dimension reduction.

In this article, we focus on sufficient dimension reduction in binary classification. In the regression context with a continuous response Y, predicting whether Y is greater than a specific value, i.e., $I(Y \geqslant c)$ for a given constant c, can be easier than predicting Y itself, since the resolution of the dichotomized response $I(Y \geqslant c)$ is much lower than that of Y. Here $I(A) = 1$ if event A is true and 0 otherwise. In the sufficient dimension reduction literature, however, many estimators rely on inverse regression, whose target is functionals of X given Y. Thus, they may suffer in binary classification due to the insufficient information provided by the binary response. For example, sliced inverse regression can estimate at most one direction of $\mathcal{S}_{Y|X}$ and sliced average variance estimation is known for its inefficiency in binary classification; see Cook & Lee (1999) and Li & Wang (2007).

In binary classification, Shin et al. (2014b) showed that $\mathcal{S}_{Y|X} = \mathcal{S}_{p(X)|X}$, where $p(X) = \mathrm{pr}(Y = 1 \mid X)$ and $\mathcal{S}_{p(X)|X}$ is analogously defined as $\mathcal{S}_{Y|X}$ by replacing Y with $p(X)$ in (1). The equivalence between $\mathcal{S}_{Y|X}$ and $\mathcal{S}_{p(X)|X}$ provides a natural way to improve the poor resolution of binary Y by replacing it with continuous $p(X)$. Shin et al. (2014b) proposed to apply sliced inverse regression to X and $p(X)$, instead of Y. Although slices based on $p(X)$ are unavailable, Shin et al. (2014b) estimated the slices by exploiting Fisher consistency of the weighted support vector machine (Lin et al., 2002) without knowing $p(X)$.

Li et al. (2011) proposed principal support vector machines, a unified learning framework for sufficient dimension reduction in regression. Li et al. (2011) showed that the normal of an optimal hyperplane separating the sets $S_c^+ = \{X : Y \geqslant c\}$ and $S_c^- = \{X : Y < c\}$ for an arbitrary given $c \in \mathbb{R}$ lies in $\mathcal{S}_{Y|X}$ if $E(X) = 0_p$ and $\mathrm{cov}(X) = I_p$, where 0_p and I_p are the p-dimensional zero vector and identity matrix. The principal support vector machine applies linear support vector machines to $(\tilde{Y}_c, X)$ for different values of c, where $\tilde{Y}_c = 1$ if $Y \geqslant c$ and -1 otherwise. The normals of the optimal hyperplanes obtained from these support vector machines can estimate

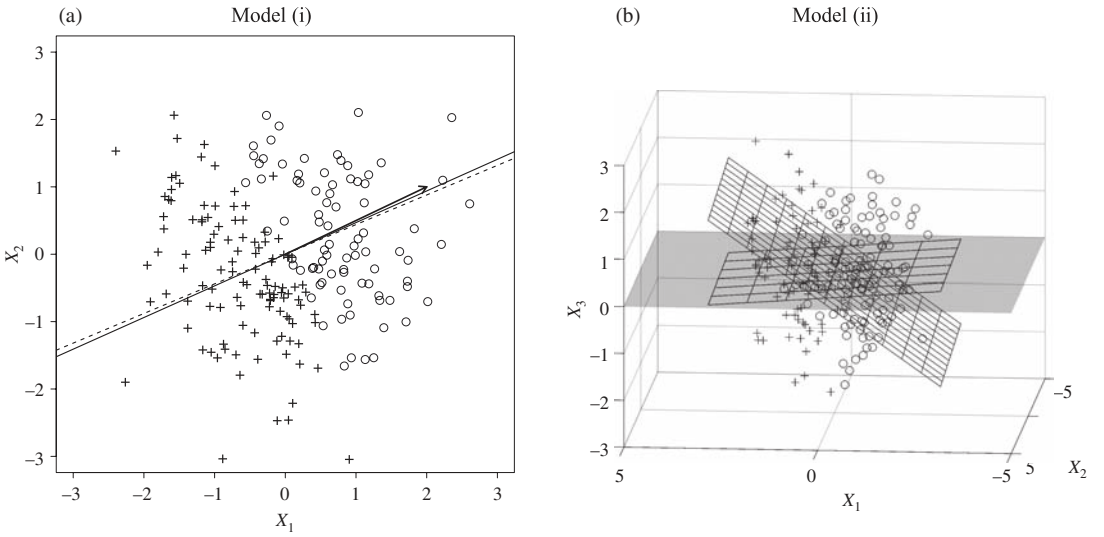

Fig. 1. Motivating example: Panel (a) depicts projections of the estimated $\mathcal{S}_{Y|X}$ onto the (X_1, X_2) plane for model (i) with $k = 1$. The arrow represents the basis of the true $\mathcal{S}_{Y|X}$. Dotted and solid lines are the estimated $\mathcal{S}_{Y|X}$ by sliced inverse regression and the principal weighted support vector machine, respectively. Panel (b) depicts the estimated $\mathcal{S}_{Y|X}$ on the three-dimensional predictor space for model (ii) with $k = 2$. The shaded plane is the true $\mathcal{S}_{Y|X}$, and the fine and coarse meshes represent the estimated $\mathcal{S}_{Y|X}$ by sliced average variance estimation and the principal weighted support vector machine, respectively. Two different symbols, pluses and circles, represent observations from different classes.

$\mathcal{S}_{Y|X}$. The principal support vector machine has promising performance. In binary classification, however, it suffers from estimating at most one direction for $\mathcal{S}_{Y|X}$ as well, since there is only one way of dichotomizing Y, which is already binary.

In this article, motivated by the properties of principal support vector machines for sufficient dimension reduction in regression, we propose a new dimension reduction method for binary classification, called the principal weighted support vector machine. The basic idea is to consider optimal hyperplanes that separate $S_\pi^+ = \{X : p(X) \geqslant \pi\}$ and $S_\pi^- = \{X : p(X) < \pi\}$ for different values of $\pi \in (0, 1)$. If $p(X)$ were known, then it would play the role of a continuous response, and $\mathcal{S}_{Y|X}$ could be estimated from the normals of these hyperplanes since $\mathcal{S}_{Y|X} = \mathcal{S}_{p(X)|X}$, thus avoiding the difficulty with a binary response. However, $p(X)$ is unknown and its estimation is more challenging than classification itself, so it is not desirable to estimate $p(X)$ first in order to perform dimension reduction. Moreover, we cannot impose a strong model assumption on $p(X)$, since sufficient dimension reduction is a model-free approach that only requires conditional independence between Y and X in (1). The proposed method tackles these issues by embedding $p(X)$ in a weighted support vector machine.

As an illustration, we consider two simple models: (i) $Y = \text{sign}(2X_1 + X_2 + 0{\cdot}2\epsilon)$ and (ii) $Y = \text{sign}(2X_1 + \log|X_2| + 0{\cdot}2\epsilon)$, where $\epsilon \sim N(0, 1)$ and $X = (X_1, \ldots, X_p)^{\mathrm{T}} \sim N_p(0_p, I_p)$ with $p = 3$. Notice that $B = (2, 1, 0)^{\mathrm{T}}$ for model (i) and $B = (e_1, e_2)$ for model (ii), where e_1 and e_2 are p-dimensional vectors whose first and second elements, respectively, are whose other elements are all 1 and 0. We randomly generate 200 observations under each model and estimate $\mathcal{S}_{Y|X}$ by different methods. The results are shown in Fig. 1. For model (i), $\mathcal{S}_{Y|X}$ forms a line since $k = 1$, and the angles between the estimated and true $\mathcal{S}_{Y|X}$ are reported as 0·12 and 0·08 radians for sliced inverse regression and the principal weighted support vector machine, respectively. Both approaches perform well in this case, with the principal weighted support vector machine being slightly better. For model (ii), $\mathcal{S}_{Y|X}$ forms a plane since $k = 2$, where

sliced inverse regression can recover only one basis component of the true $\mathcal{S}_{Y|X}$. Sliced averaged variance estimation does not estimate $\mathcal{S}_{Y|X}$ satisfactorily, whereas the principal weighted support vector machine shows substantial improvement in recovering the true $\mathcal{S}_{Y|X}$. Frobenius norm distances $d(\hat{B}, B) = \|P_{\hat{B}} - P_B\|_{\mathrm{F}}$ are 0·709 and 0·094 for sliced average variance estimation and the principal weighted support vector machine, respectively. Here $P_B = B(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}$, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, and B and $\hat{B}$ are the true and estimated bases of $\mathcal{S}_{Y|X}$, respectively. We provide the corresponding 360° rotation animation in the Supplementary Material.

2. LINEAR PRINCIPAL WEIGHTED SUPPORT VECTOR MACHINE

2·1. *Population level*

We start by briefly introducing the weighted support vector machine, which will serve as the building block for the proposed method. For a given set of data $\{(X_i, Y_i) \in \mathbb{R}^p \times \{-1, +1\} : i = 1, \ldots, n\}$, the linear weighted support vector machine (Lin et al., 2002) solves

$$(\hat{a}_n, \hat{b}_n^{\mathrm{T}})^{\mathrm{T}} = \arg\min_{a,b} b^{\mathrm{T}}b + \frac{\lambda}{n}\sum_{i=1}^{n} w_\pi(Y_i)|1 - Y_i(a + b^{\mathrm{T}}X_i)|_+, \tag{3}$$

where $|u|_+ = \max(u, 0)$ and $w_\pi(Y) = 1 - \pi$ if $Y = 1$ and π if $Y = -1$ with a weight $\pi \in (0, 1)$ that controls the relative importance of the two classes. The weighted support vector machine (3) can be viewed as a loss-plus-penalty formulation. The tuning parameter λ balances the goodness of fit and complexity of the model measured by the loss and penalty functions, respectively. In the weighted support vector machine, the loss function changes as π varies, while the observed data input remains unchanged. The weighted support vector machine is Fisher-consistent; the minimizer of the expected loss given X is the Bayes rule (Lin et al., 2002).

For a pair of random variables $(X, Y) \in \mathbb{R}^p \times \{-1, +1\}$, the linear principal weighted support vector machine minimizes

$$\Lambda_\pi(\theta) = \beta^{\mathrm{T}}\Sigma\beta + \lambda E\{w_\pi(Y)|1 - Yf(X;\theta)|_+\}, \tag{4}$$

where $\theta = (\alpha, \beta^{\mathrm{T}})^{\mathrm{T}}$, $\Sigma = \mathrm{cov}(X)$, and $f(X;\theta) = \alpha + \beta^{\mathrm{T}}\{X - E(X)\}$. The differences between (4) and the population counterpart of the objective function in (3) are the matrix Σ in the penalty term and the centring of X in $f(X;\theta)$. For standardized X, with $E(X) = 0_p$ and $\mathrm{cov}(X) = I_p$, (4) becomes the population version of (3):

$$\beta^{\mathrm{T}}\beta + \lambda E\left\{w_\pi(Y)|1 - Y(\alpha + \beta^{\mathrm{T}}X)|_+\right\}.$$

Fisher consistency of the weighted support vector machine ensures that a hyperplane $\{X : f(X;\theta_{0,\pi}) = 0\}$ optimally separates $S_\pi^+ = \{X : p(X) \geqslant \pi\}$ and $S_\pi^- = \{X : p(X) < \pi\}$, where $\theta_{0,\pi} = (\alpha_{0,\pi}, \beta_{0,\pi}^{\mathrm{T}})^{\mathrm{T}} = \arg\min_\theta \Lambda_\pi(\theta)$. Equivalent to $\mathcal{S}_{p(X)|X}$, $\mathcal{S}_{Y|X}$ can be estimated from the normal of this optimal hyperplane using the idea of Li et al. (2011).

Theorem 1 provides a theoretical foundation for the linear principal weighted support vector machine for linear sufficient dimension reduction.

THEOREM 1. *Assume that $E(X \mid B^{\mathrm{T}}X)$ is a linear function of $B^{\mathrm{T}}X$. Then for any given weight $\pi \in (0, 1)$, $\beta_{0,\pi} \in \mathcal{S}_{Y|X}$ under (1).*

The assumption in Theorem 1 is known as the linearity condition. It implies that $E(\beta^{\mathrm{T}}X \mid B^{\mathrm{T}}X) = \beta^{\mathrm{T}}P_B(\Sigma)X$ where $P_B(\Sigma) = B(B^{\mathrm{T}}\Sigma B)^{-1}B^{\mathrm{T}}\Sigma$, and it plays an essential role in sufficient dimension reduction. The condition holds when X is elliptically symmetric (Li & Duan, 1989; Li, 1991; Li & Dong, 2009) and approximately holds when p is large (Hall & Li, 1993).

Considering a sequence of weights $0 < \pi_1 < \cdots < \pi_H < 1$, we have $\theta_{0,h} = (\alpha_{0,h}, \beta_{0,h}^{\mathrm{T}})^{\mathrm{T}} = \arg\min_\theta \Lambda_{\pi_h}(\theta)$ $(h = 1, \ldots, H)$. By Theorem 1, $\mathrm{span}(\beta_{0,1}, \ldots, \beta_{0,H}) \subseteq \mathcal{S}_{Y|X}$. Following the usual protocol in sufficient dimension reduction, we assume that $\mathrm{span}(\beta_{0,1}, \ldots, \beta_{0,H}) = \mathcal{S}_{Y|X}$ whenever $\mathrm{span}(\beta_1, \ldots, \beta_H) \subseteq \mathcal{S}_{Y|X}$ and H is large enough. This often holds in practice (Cook & Ni, 2006).

For the principal weighted support vector machine, Y is binary and remains unchanged while the loss function in (4) is changed as the weight π varies. For different π, the target of (4) is changed, enabling us to estimate more than one direction of $\mathcal{S}_{Y|X}$ in binary classification. In this regard, the principal weighted support vector machine is more than a weighted version of the principal support vector machine. Furthermore, π in the principal weighted support vector machine plays a different role from the weight used in weighted least squares regression, where only estimation efficiency is improved by adopting weights. The effect of the weight in the principal weighted support vector machine is analogous to that of the quantile level parameter in the check loss of quantile regression.

2·2. *Finite-sample estimation and solution paths*

Denote the observed data by $\{Z_i = (X_i, Y_i) : X_i \in \mathbb{R}^p, Y_i \in \{-1, +1\}, i = 1, \ldots, n\}$. The sample version of Λ_π in (4) is

$$\hat{\Lambda}_{n,\pi}(\theta) = \beta^{\mathrm{T}}\hat{\Sigma}_n\beta + \frac{\lambda}{n}\sum_{i=1}^{n} w_\pi(Y_i)|1 - Y_i\hat{f}_n(X_i;\theta)|_+, \tag{5}$$

where $\hat{f}_n(X_i;\theta) = \alpha + \beta^{\mathrm{T}}(X_i - \bar{X}_n)$, $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is the sample mean, and $\hat{\Sigma}_n$ denotes the sample covariance matrix. Given a grid $0 < \pi_1 < \cdots < \pi_H < 1$, let $\hat{\theta}_{n,h} = (\hat{\alpha}_{n,h}, \hat{\beta}_{n,h}^{\mathrm{T}})^{\mathrm{T}} = \arg\min_\theta \hat{\Lambda}_{n,\pi_h}(\theta)$ $(h = 1, \ldots, H)$. The candidate matrix of the linear principal weighted support vector machine is

$$\hat{M}_n = \sum_{h=1}^{H} \hat{\beta}_{n,h}\hat{\beta}_{n,h}^{\mathrm{T}}. \tag{6}$$

The first k eigenvectors of $\hat{M}_n$, denoted by $\hat{V}_n = (\hat{v}_1, \ldots, \hat{v}_k)$, estimate a basis of $\mathcal{S}_{Y|X}$. Due to the way it was constructed, $\hat{M}_n$ may have more than one eigenvector with a nonzero eigenvalue.

The above procedure requires minimizing (5) repeatedly for $\pi_1, \ldots, \pi_H$, which can be computationally intensive when the sample size n and/or H are large. With transformations $\eta = \hat{\Sigma}_n^{1/2}\beta$ and $U_i = \hat{\Sigma}_n^{-1/2}(X_i - \bar{X}_n)$, (5) becomes $\eta^{\mathrm{T}}\eta + n^{-1}\lambda\sum_{i=1}^n w_\pi(Y_i)|1 - Y_i(\alpha + \eta^{\mathrm{T}}U_i)|_+$, which is equivalent to training the weighted support vector machine with respect to α and η. Denote the optimizer by $\hat{\alpha}_{n,\pi}$ and $\hat{\eta}_{n,\pi}$. Then the optimizer of (5) is $\hat{\alpha}_{n,\pi}$ and $\hat{\beta}_{n,\pi} = \hat{\Sigma}_n^{-1/2}\hat{\eta}_{n,\pi}$.

To facilitate the computation of the principal weighted support vector machine, we employ a solution path algorithm that computes entire trajectories as a function of π (Wang et al., 2008; Shin et al., 2014a), which we call the π-path. The algorithm has the same computational complexity as solving a single quadratic programming problem (Hastie et al., 2004). As an

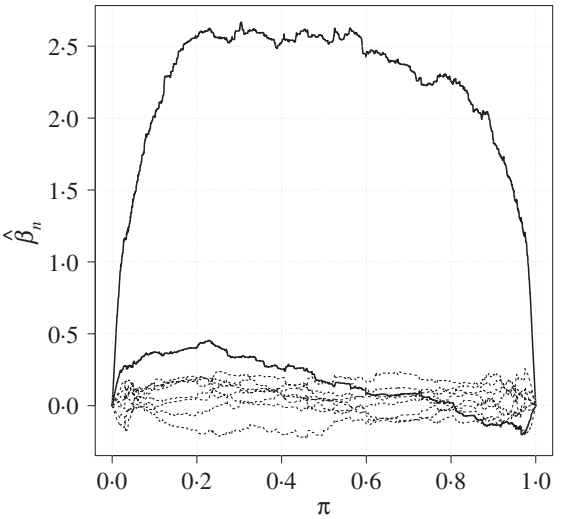

Fig. 2. π-path of $\hat{\beta}_n(\pi) = \{\hat{\beta}_{n,1}(\pi), \ldots, \hat{\beta}_{n,p}(\pi)\}^{\mathrm{T}}$: solid lines represent the π-paths of $\hat{\beta}_{n,1}(\pi)$ and $\hat{\beta}_{n,2}(\pi)$ corresponding to X_1 and X_2, while dotted lines are those of $\hat{\beta}_{n,j}(\pi)(j = 3, \ldots, 10)$ corresponding to other predictors.

illustration we use a simulated dataset with $n = 500$ and $p = 10$ generated from the first model f_1 in § 4, where only X_1 and X_2 are used to define $\mathcal{S}_{Y|X}$. Figure 2 depicts the π-path of $\hat{\beta}_{n,\pi} = \hat{\beta}_n(\pi) = \{\hat{\beta}_{n,1}(\pi), \ldots, \hat{\beta}_{n,p}(\pi)\}^{\mathrm{T}}$ as a function of π. In Fig. 2, the π-paths of $\hat{\beta}_{n,1}(\pi)$ and $\hat{\beta}_{n,2}(\pi)$ associated with relevant predictors X_1 and X_2 show larger variation than the paths of $\hat{\beta}_{n,j}(\pi)$ $(j = 3, \ldots, 10)$ corresponding to irrelevant predictors.

2·3. *Large-sample properties*

Asymptotic results for the linear principal weighted support vector machine are closely connected with those for the linear support vector machine (Jiang et al., 2008; Koo et al., 2008) and the principal support vector machine (Li et al., 2011). Without loss of generality, we assume that $E(X) = 0$ and let $\tilde{X} = (1, X^{\mathrm{T}})^{\mathrm{T}}$. Then $f(X; \theta) = \theta^{\mathrm{T}}\tilde{X}$. Let $\hat{\theta}_{n,\pi} = (\hat{\alpha}_{n,\pi}, \hat{\beta}_{n,\pi}^{\mathrm{T}})^{\mathrm{T}}$ denote the minimizer of (5). We sometimes omit the subscript π for simplicity when the results hold for an arbitrary value of π.

Throughout this subsection, we make the following regularity assumptions:

Assumption 1. X has an open and convex support and $E(\|X\|^2) < \infty$.

Assumption 2. The conditional distribution $X \mid Y$ is dominated by the Lebesgue measure.

Assumption 3. For an arbitrary $\theta \neq \theta_0$,

$$\sum_{y \in \{-1,+1\}} \mathrm{pr}\{Y = y, X \in \Delta(y, \theta)\} > 0,$$

where $\Delta(y, \theta) = \{X : (1 - y\theta^{\mathrm{T}}\tilde{X})(1 - y\theta_0^{\mathrm{T}}\tilde{X}) < 0\}$.

Assumption 4. For arbitrary given vectors $\beta, \delta \in \mathbb{R}^p$, let U and V denote $\beta^{\mathrm{T}}X$ and $\delta^{\mathrm{T}}X$, respectively. Then the map $u \mapsto E(X \mid U = u, V, Y) f_{U|V,Y}(U = u \mid V, Y)$ is continuous for any

$V \in \mathbb{R}$ and $Y \in \{-1, +1\}$, where $f_{U|V,Y}(U \mid V, Y)$ denotes the conditional density of U given V and Y.

Assumption 5. Given $U = u$, there exists a nonnegative $\mathbb{R}^{p+1}$-valued function $c(V, Y)$ such that $E\{c(V, Y)\} < \infty$ and $E(\tilde{X} \mid U = u, V, Y) f_{U|V,Y}(U = u \mid V, Y) < c(V, Y)$, where the inequality holds componentwise.

Assumptions 1 and 2 are standard for sufficient dimension reduction. Assumption 3 is similar to that in Theorem 1 of Jiang et al. (2008), which guarantees the existence of a unique minimum of $\Lambda_\pi(\theta)$ which is not strictly convex with respect to α. Assumptions 4 and 5 essentially ensure the continuity and boundedness of the gradient vector of $\Lambda_\pi(\theta)$, which turns out to be non-Lipschitz with respect to θ. These regularity conditions are required to avoid technical difficulties brought by the use of a non-strictly convex and nondifferentiable function $|\cdot|_+$ in $\Lambda_\pi(\theta)$. In order to relax these conditions, it is possible to consider a strictly convex differentiable loss.

Now, we establish the consistency of $\hat{\theta}_n$.

THEOREM 2. *Suppose that Σ is positive definite and Assumption 2 holds. Then $\hat{\theta}_n \to \theta_0$ in probability as $n \to \infty$.*

The Bahadur representation of $\hat{\theta}_n$ is provided in Theorem 3.

THEOREM 3. *Assume that Σ is positive definite and Assumptions 1– 5 hold. Then*

$$n^{1/2}(\hat{\theta}_n - \theta_0) = -n^{-1/2} H_{\theta_0}^{-1} \sum_{i=1}^{n} D_{\theta_0}(Z_i) + o_{\mathrm{p}}(1),$$

where

$$D_\theta(Z) = (0, 2\beta^{\mathrm{T}}\Sigma)^{\mathrm{T}} - \lambda w_\pi(Y)\tilde{X} Y I(\theta^{\mathrm{T}}\tilde{X}Y < 1),$$

$$H_\theta = 2\,\mathrm{diag}(0, \Sigma) + \lambda \sum_{y=-1,1} \mathrm{pr}(Y = y) w_\pi(y) f_{\beta^{\mathrm{T}}X|Y}(y - \alpha \mid y) E(\tilde{X}\tilde{X}^{\mathrm{T}} \mid \theta^{\mathrm{T}}\tilde{X} = y).$$

Here $\mathrm{diag}(0, \Sigma)$ *denotes the* $(p + 1) \times (p + 1)$ *block-diagonal matrix whose block-diagonal elements are 0 and* Σ.

For any given π_h, let $S(\theta_{0,h}, Z) = F_{\theta_{0,h}} D_{\theta_{0,h}}(Z)$ with $F_{\theta_{0,h}}$ being the last p rows of $H_{\theta_{0,h}}^{-1}$ $(h = 1, \ldots, H)$. A Bahadur representation of $\hat{\beta}_{n,h}$ is then

$$n^{1/2}(\hat{\beta}_{n,h} - \beta_{0,h}) = -n^{-1/2} \sum_{i=1}^{n} S(\theta_{0,h}, Z_i) + o_{\mathrm{p}}(1) \tag{7}$$

by Theorem 3. From (7), asymptotic normality of $\hat{M}_n$ in (6) is established.

THEOREM 4. *Let* $M_0 = \sum_{h=1}^{H} \beta_{0,h}\beta_{0,h}^{\mathrm{T}}$. *Under the conditions of Theorem 3,* $n^{1/2}\mathrm{vec}(\hat{M}_n - M_0) \to N(0_{p^2}, \Sigma_M)$ *in distribution as* $n \to \infty$. *The form of the covariance matrix* Σ_M *is given in the Supplementary Material.*

Finally, asymptotic normality of $\hat{V}_n$ follows from Theorem 4 and Corollary 1 of Bura & Pfeiffer (2008).

COROLLARY 1. *Assume that* $\text{rank}(M_0) = k$ *and let* $V_0 = (v_1, \ldots, v_k)$ *denote the first* k *eigenvectors of* M_0. *Under the conditions of Theorem 3, we have* $n^{1/2}\text{vec}(\hat{V}_n - V_0) \to N(0_{pk}, \Sigma_V)$ *in distribution as* $n \to \infty$, *with* $\Sigma_V = (D^{-1}U^{\mathrm{T}} \otimes I_p)\Sigma_M(UD^{-1} \otimes I_p)$, *where* U *is a* $p \times k$ *matrix with columns being the eigenvectors of* M_0 *corresponding to its nonzero eigenvalues, and* D *is a* $k \times k$ *diagonal matrix with the nonzero eigenvalues as diagonal elements.*

2·4. *Determination of structural dimension*

Under (1), the structural dimension k is a crucial quantity, and must be estimated from the data. In this regard, Li et al. (2011) proposed a procedure that determines k in a data-adaptive manner for the linear principal support vector machine. We consider a criterion similar to theirs,

$$G_n(m; \rho, \hat{M}_n) = \sum_{j=1}^{m} \ell_j - \rho \frac{m \log n}{\sqrt{n}} \ell_1,$$

where ℓ_j is the jth leading eigenvalue of $\hat{M}_n$ in (6) and ρ is a tuning parameter. Following the lines of Theorem 8 of Li et al. (2011) and Theorem 4 above, we can prove that $\hat{k} = \arg\max_{m \in \{1,\ldots,p\}} G_n(m; \rho, \hat{M}_n)$ is a consistent estimator of k, i.e., $\lim_{n\to\infty} \text{pr}(\hat{k} = k) = 1$. We propose an algorithm to select ρ by directly extending the idea of Li et al. (2011); see the Supplementary Material.

3. KERNEL PRINCIPAL WEIGHTED SUPPORT VECTOR MACHINE

3·1. *Population level*

Under (2), we consider the following objective function as a nonlinear generalization of (4):

$$\Lambda_\pi(\alpha, \psi) = \text{var}\{\psi(X)\} + \lambda E\{w_\pi(Y)|1 - Yf(X; \alpha, \psi)|_+\}, \tag{8}$$

where $f(X; \alpha, \psi) = \alpha + \psi(X) - E\{\psi(X)\}$ with ψ being a function in a Hilbert space $\mathcal{H}$ of functions of X. Notice that (8) is equivalent to (4) if $\psi(X)$ is a linear function of X, i.e., $\psi(X) = \beta^{\mathrm{T}}X$.

Theorem 5 provides a theoretical foundation for the nonlinear principal weighted support vector machine.

THEOREM 5. *Consider the identity mapping from a function* $f \in \mathcal{H}$ *to* $f \in L_2(P_X)$, *where* $L_2(P_X) = \{f : (\int |f|^2 \, \mathrm{d}P_X)^{1/2} < \infty\}$, *with* P_X *the probability measure induced by* X. *Assume that the mapping is continuous and* $\mathcal{H}$ *is a dense subset of* $L_2(P_X)$. *Let* $(\alpha_{0,\pi}, \psi_{0,\pi}) = \arg\min_{\alpha,\psi} \Lambda_\pi(\alpha, \psi)$. *Then for any given weight* $\pi \in (0, 1)$, $\psi_{0,\pi} \in \mathcal{H}$ *has a one-to-one transformation that is measurable with respect to* $\sigma\{\phi(X)\}$, *where* $\sigma\{\phi(X)\}$ *denotes the* σ*-field generated by* $\phi(X)$ *in* (2).

3·2. *Finite representation via kernel trick*

Due to the infinite dimension of $\mathcal{H}$, it is not trivial to estimate $(\alpha_{0,\pi}, \psi_{0,\pi})$. We employ the reproducing kernel Hilbert space $\mathcal{H}_K$ generated by a positive definite kernel $K(\cdot, \cdot)$. By the representer theorem (Kimeldorf & Wahba, 1971), the minimizer of the empirical version of (8) has a n-dimensional representation of $\psi_\pi(\cdot) = \alpha_\pi^{\mathrm{T}} k_n(\cdot)$, where $\alpha_\pi = (\alpha_{1,\pi}, \ldots, \alpha_{n,\pi})^{\mathrm{T}}$ and

$k_n(\cdot) = \{K(\cdot, X_i) : i = 1, \ldots, n\}^{\mathrm{T}}$. However, as pointed out by Li et al. (2011), the solution space spanned by $k_n(\cdot)$ is too rich, so that the solution often overfits the data.

As an alternative, Li et al. (2011) proposed

$$\psi_\pi(X) = \gamma_\pi^{\mathrm{T}} \omega(X), \tag{9}$$

with d-dimensional parameter $\gamma_\pi = (\gamma_{1,\pi}, \ldots, \gamma_{d,\pi})^{\mathrm{T}}$ and the associated basis functions $\omega(X) = \{\omega_j(X) : j = 1, \ldots, d\}^{\mathrm{T}}$. Here $\omega_j(\cdot)$ is the jth leading eigenfunction of the sample covariance operator Σ_n such that $\langle \psi_1, \Sigma_n \psi_2 \rangle_{\mathcal{H}_K}$ computes the sample covariance of $\psi_1(X_i)$ and $\psi_2(X_i)$ $(i = 1, \ldots, n)$. Under the assumptions in Theorem 5, there exists a bounded and self-adjoint operator $\Sigma : \mathcal{H} \to \mathcal{H}$ such that $\langle \psi_1, \Sigma \psi_2 \rangle_{\mathcal{H}} = \mathrm{cov}\{\psi_1(X), \psi_2(X)\}$ (Conway, 1990). This provides a justification of (9), since (8) can be equivalently rewritten as

$$\Lambda_\pi(\alpha, \psi) = \langle \psi, \Sigma \psi \rangle_{\mathcal{H}} + \lambda E\{w_\pi(Y)|1 - Yf(X; \alpha, \psi)|_+\}.$$

By Proposition 2 of Li et al. (2011),

$$\omega_j(X) = \tilde{k}_n(X)^{\mathrm{T}} q_j / \lambda_j \quad (j = 1, \ldots, d),$$

where $\tilde{k}_n(\cdot) = k_n(\cdot) - n^{-1} \sum_{i=1}^n k_n(X_i)$ and q_j and λ_j are the jth leading eigenvector and eigenvalue of the matrix $Q_n = (I_n - J_n/n) K_n (I_n - J_n/n)$. Here K_n is the kernel matrix whose (i,j)th element is $K(X_i, X_j)$ $(i,j = 1, \ldots, n)$ and J_n denotes the n-dimensional square matrix whose elements are all one. Notice that Q_n is the candidate matrix of kernel principal component analysis (Schölkopf et al., 1997) on $\mathcal{H}_K$. In fact, (9) restricts the full solution space to the subspace spanned by the first d principal directions to avoid overfitting.

For the choice of d, we suggest an integer around $n/4$ or even smaller. Li et al. (2011) propose using any integer between $n/3$ and $2n/3$ for the principal support vector machine. As mentioned in § 1, binary responses have weaker signals than continuous responses and therefore a smaller value of d is better to avoid overfitting.

3·3. *Estimation*

Inserting $\alpha + \gamma^{\mathrm{T}} \omega(X)$ into $f(X; \alpha, \psi)$, a sample version of (8) is

$$\hat{\Lambda}_{n,\pi}(\alpha, \gamma) = \gamma^{\mathrm{T}} \Omega^{\mathrm{T}} \Omega \gamma + \frac{\lambda}{n} \sum_{i=1}^n w_\pi(Y_i)|1 - Y_i(\alpha + \gamma^{\mathrm{T}} \Omega_i)|_+, \tag{10}$$

which we call the kernel principal weighted support vector machine. The matrix Ω, whose (i,j)th element is $\omega_j(X_i)$ $(i = 1, \ldots, n; j = 1, \ldots, d)$, is identical to $(q_1, \ldots, q_d)$ with q_j defined above as the eigenvector of Q_n (Li et al., 2011).

Let $(\hat{\alpha}_{n,\pi}, \hat{\gamma}_{n,\pi}^{\mathrm{T}})^{\mathrm{T}} = \arg\min_{\alpha,\gamma} \hat{\Lambda}_{n,\pi}(\alpha, \gamma)$, the minimizer of (10). It is shown that $\hat{\gamma}_{n,\pi} = \lambda \sum_{i=1}^n \hat{v}_{i,\pi} Y_i \{(\Omega^{\mathrm{T}} \Omega)^{-1} \Omega_i\}/2$, where $\hat{v}_\pi = (\hat{v}_{1,\pi}, \ldots, \hat{v}_{n,\pi})^{\mathrm{T}}$ solves

$$\max_{v_1, \ldots, v_n} \sum_{i=1}^n v_i - \frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n v_i v_j Y_i Y_j P_\Omega^{(i,j)} \tag{11}$$

$$\text{subject to } 0 \leqslant v_i \leqslant \lambda w_\pi(Y_i) \quad (i = 1, \ldots, n), \quad \sum_{i=1}^n v_i Y_i = 0,$$

with $P_\Omega^{(i,j)}$ the (i,j)th element of $P_\Omega = \Omega(\Omega^{\mathrm{T}} \Omega)^{-1} \Omega^{\mathrm{T}}$.

If $\Omega = K_n$, then $P_\Omega = I_p$ and (11) becomes independent of X, which explains the overfitting problem on the full solution space spanned by $k_n(\cdot)$. Notice that (11) is the same as the dual problem of the weighted support vector machine with a kernel matrix P_Ω. The π-path algorithm can be exploited to obtain the entire solution profile of $\hat{v}_\pi$ for all $\pi \in (0, 1)$.

Finally, we have a candidate matrix $\sum_{h=1}^{H} \hat{\gamma}_{n,h}\hat{\gamma}_{n,h}^{\mathrm{T}}$ with $(\hat{\alpha}_{n,h}, \hat{\gamma}_{n,h}^{\mathrm{T}})^{\mathrm{T}}$ the minimizer of $\hat{\Lambda}_{n,\pi_h}(\alpha, \gamma)$ $(h = 1, \ldots, H)$. The jth component of $\phi(X) = \{\phi_1(X), \ldots, \phi_k(X)\}^{\mathrm{T}}$ in (2) is then estimated by $\hat{\phi}_j(X) = \hat{v}_j^{\mathrm{T}}\omega(X)$ $(j = 1, \ldots, k)$, where $\hat{V}_n = (\hat{v}_1, \ldots, \hat{v}_k)$ contains the first k eigenvectors of $\sum_{h=1}^{H} \hat{\gamma}_{n,h}\hat{\gamma}_{n,h}^{\mathrm{T}}$.

4. SIMULATION STUDIES

4·1. *Linear sufficient dimension reduction*

The following model is assumed for the simulation: $Y_i = \mathrm{sign}\{f(X_i) + \epsilon_i\}$ $(i = 1, \ldots, n)$, where $X_i \sim N_p(0_p, I_p)$ and the random error $\epsilon_i \sim N(0, 0{\cdot}2^2)$. Nine different combinations of $(n, p) \in \{100, 500, 1000\} \times \{10, 20, 30\}$ are considered.

Five different decision functions are considered: $f_1(X) = X_1/\{0{\cdot}5 + (X_2 + 1)^2\}$, $f_2(X) = (X_1 + 0{\cdot}5)(X_2 - 0{\cdot}5)^2$, $f_3(X) = \sin(X_1/\mathrm{e}^{X_2})$, $f_4(X) = X_1(X_1 + X_2 + 1)$, and $f_5(X) = (X_1^2 + X_2^2)^{1/2}\log(X_1^2 + X_2^2)^{1/2}$. They all share a common central subspace, $\mathrm{span}(B)$ with $B = (e_1, e_2)$. Notice that f_4 is approximately and f_5 exactly symmetric about the origin. Under the exactly symmetric scenario, the $\beta_{0,h}$ $(h = 1, \ldots, H)$ are identical, so the linear principal support vector machine fails. This is similar to the failure of sliced inverse regression for regression with a symmetric regression function (Cook & Weisberg, 1991).

For the principal weighted support vector machine, we use twenty π values equally spaced between 0 and 1 and set $\lambda = 1$. Our method is not overly sensitive to the choice of H and λ provided that the weights π_h are well-spread over the interval $(0, 1)$ and both H and λ are not too small. The rationale is to avoid the situation where all the solutions are similar and it is hard to extract the variation efficiently.

Sliced inverse regression, sliced average variance estimation, partial least squares estimation, the Fourier method, iterative Hessian transformation, directional regression and probability-enhanced sliced inverse regression (Shin et al., 2014b) are considered as competing methods. We use the Frobenius norm distance $d(\hat{B}, B) = \|P_{\hat{B}} - P_B\|_{\mathrm{F}}$ defined in § 1 to evaluate the performance of an estimator $\hat{B}$. We assume that the true $k = 2$ is known. Sliced inverse regression and partial least squares estimation can estimate at most one direction. Table 1 reports $d(\hat{B}, B)$ averaged over 100 repetitions for $n = 500$ and $p = 10$. See the Supplementary Material for results for all combinations of n and p.

The linear principal weighted support vector machine outperforms the other methods under f_1, f_2 and f_3, which are not symmetric about the origin. The improvement of the linear principal weighted support vector machine for f_4 is not significant compared to the others, partly because f_4 is approximately symmetric about the origin. For f_5, both the probability-enhanced sliced inverse regression and linear principal weighted support vector machine fail, since f_5 is exactly symmetric about the origin, while sliced average variance estimation performs best. In § 4·2, we will see that this limitation can be resolved by employing the kernel principal weighted support vector machine.

In order to provide a more complete picture, we also consider models with $k = 1$ by simply replacing X_2 with X_1 in f_1, f_2 and f_3: $f_1'(X) = X_1/\{0{\cdot}5 + (X_1 + 1)^2\}$, $f_2'(X) = (X_1 + 0{\cdot}5) \times$

Table 1. *Performance of the linear sufficient dimension reduction methods for* $n = 500$ *and* $p = 10$. *Reported values are the averaged Frobenius norm distances between the projection matrices of the true and estimated* $\mathcal{S}_{Y|X}$ *over* 100 *independent repetitions*

	SIR	SAVE	PLS	FCN	IHT	DR	PRE	PWSVM
f_1	1·30	1·29	1·02	1·29	1·32	1·28	1·13	0·75
f_2	1·24	1·27	1·03	1·21	1·14	1·26	1·06	1·02
f_3	1·30	1·26	1·02	1·28	1·30	1·25	1·11	0·80
f_4	0·64	0·77	1·07	0·47	0·47	0·54	0·59	0·53
f_5	1·38	0·27	1·64	0·49	1·42	0·27	1·70	1·60
f_1'	0·20	0·21	0·21	0·20	1·28	0·21	0·19	0·17
f_2'	0·29	0·42	0·29	0·30	0·66	0·30	0·32	0·26
f_3'	0·19	0·20	0·20	0·19	1·21	0·19	0·17	0·15

SIR, sliced inverse regression; SAVE, sliced average variance estimation; PLS, partial least squares estimation; FCN, Fourier method; IHT, iterative Hessian direction; DR, directional regression; PRE, probability-enhanced sliced inverse regression; PWSVM, principal weighted support vector machine.

$(X_1 - 0{\cdot}5)^2$ and $f_3'(X) = \sin(X_1/\mathrm{e}^{X_1})$. The linear principal weighted support vector machine outperforms all others even when $k = 1$.

In practice, the structural dimension k is unknown. The numerical performance of the approach proposed in § 2·4 to determine k is also investigated under various scenarios. The results are relegated to the Supplementary Material.

4·2. *Nonlinear sufficient dimension reduction*

The linear principal weighted support vector machine fails when the true decision curve f is symmetric about the origin. Next we apply the kernel principal weighted support vector machine using the Gaussian kernel $K(X, X') = \exp\{-\|X - X'\|^2/(2\sigma^2)\}$, with σ the median of pairwise Euclidean distances between the two classes (Jaakkola et al., 1999). Both λ and π grids are set to be the same as the linear principal weighted support vector machine in § 4·1. Figure 3 compares the linear and kernel principal weighted support vector machines for a dataset with $n = 200$ and $p = 10$ simulated from f_5. The linear principal weighted support vector machine fails as the radial distinction becomes much less clear than that in the true $\mathcal{S}_{Y|X}$, while the kernel principal weighted support vector machine recovers the true f_5 very well.

Li et al. (2011) argued that the nonlinear principal support vector machine can transform a difference in variability in the original space to a difference in location in the reduced feature space obtained from the kernel principal support vector machine. This is how the kernel support vector machine works for nonlinear classification, and a similar phenomenon is observed in the kernel principal weighted support vector machine.

To evaluate its performance, we consider Hotelling's T^2 statistic between the two classes in the reduced space whose dimension is $k = 2$. If the sufficient dimension reduction method performs well, the two classes should be clearly separated and hence the associated Hotelling's T^2 statistic would be large. Table 1 reports the averaged T^2 over 100 independent repetitions when $n = 500$ and $p = 10$. Hotelling's T^2 statistic may fail to evaluate performance when a classification pattern is nonlinear. In this sense, a large T^2 statistic is a sufficient but not necessary condition for good classification. Table 1 reports its smallest values for sliced average variance estimation, which outperforms other methods according to Table 1, especially for f_5. Nevertheless, the kernel principal weighted support vector machine shows extremely large values of T^2, which is sufficient to guarantee good classification performance. Combining this with Fig. 3, we conclude that the

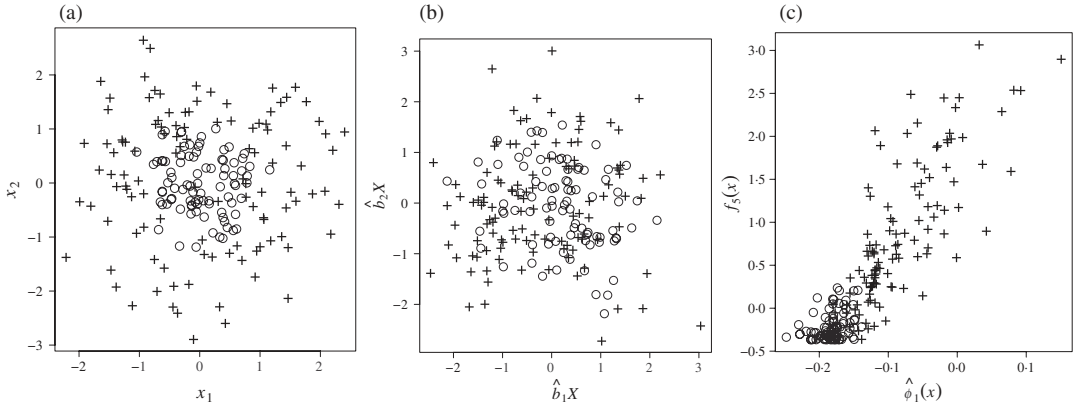

Fig. 3. Sufficient dimension reduction under f_5: panel (a) is a scatterplot of $b_1^{\mathrm{T}}X$ and $b_2^{\mathrm{T}}X$, i.e., X_1 and X_2; panel (b) depicts a scatterplot of $\hat{b}_1^{\mathrm{T}}X$ and $\hat{b}_2^{\mathrm{T}}X$, where $\hat{B} = (\hat{b}_1, \hat{b}_2)$ is estimated by the linear principal weighted support vector machine; panel (c) shows a scatterplot of the true $f_5(X)$ versus the first sufficient predictor $\hat{\phi}_1(X)$ estimated by the kernel principal weighted support vector machine.

kernel principal weighted support vector machine performs very well even when the true decision curve is symmetric about the origin.

To provide a fair comparison, we use $k = 2$ for all sufficient dimension reduction methods, linear or nonlinear, to report performance in the above. Yet it is obvious that Y is conditionally independent of X given $(X_1^2 + X_2^2)$ for f_5. Consequently, the corresponding structural dimension should be 1 for nonlinear sufficient dimension reduction. If using $k = 1$, the corresponding average Hotelling's T^2 is 616·4 with a standard deviation of 71·8 for the kernel principal weighted support vector machine. This large Hotelling's T^2 value echoes the good performance illustrated in Fig. 3(c).

5. WISCONSIN BREAST CANCER DATA

We use the Wisconsin Diagnostic Breast Cancer data available at http://archive.ics.uci.edu/ml/index.html. The dataset contains diagnoses of breast cancer for 569 subjects with 30 predictors. For the linear principal weighted support vector machine, λ and π grids are set in the same way as in § 4·1. We employ the procedure proposed in § 2·4 to determine k. Figure 4(a) depicts $G_n(m; \rho, \hat{M}_n)$ as a function of m at an optimal ρ selected as 0·009, and hence the structural dimension is estimated as $\arg\max_m G_n(m; \rho, \hat{M}_n) = 3$. Figure 4(b) shows a three-dimensional scatterplot of predictors projected onto the estimated $\mathcal{S}_{Y|X}$ by the linear principal weighted support vector machine. It shows that the two classes are well separated in the estimated $\mathcal{S}_{Y|X}$. See the Supplementary Material for a 360° rotation animation of the scatterplot. Finally, we also apply the kernel principal weighted support vector machine to the data, with the Gaussian kernel and σ, λ and π grids chosen in the same way as in § 4·2. The kernel principal weighted support vector machine seems to work well in that the two classes are clearly separated by $\hat{\phi}_1(X)$. See Fig. 4(c).

In practice, the goal is often the improvement of classification accuracy after dimension reduction. In this regard, we carry out validation analysis as follows. First, we randomly split the data into training and test sets of equal size. Different sufficient dimension reduction methods are then applied to the training set, and the five-nearest-neighbour classifiers are applied to train a classifier on the estimated $\mathcal{S}_{Y|X}$. Finally, the test responses are predicted by plugging the test

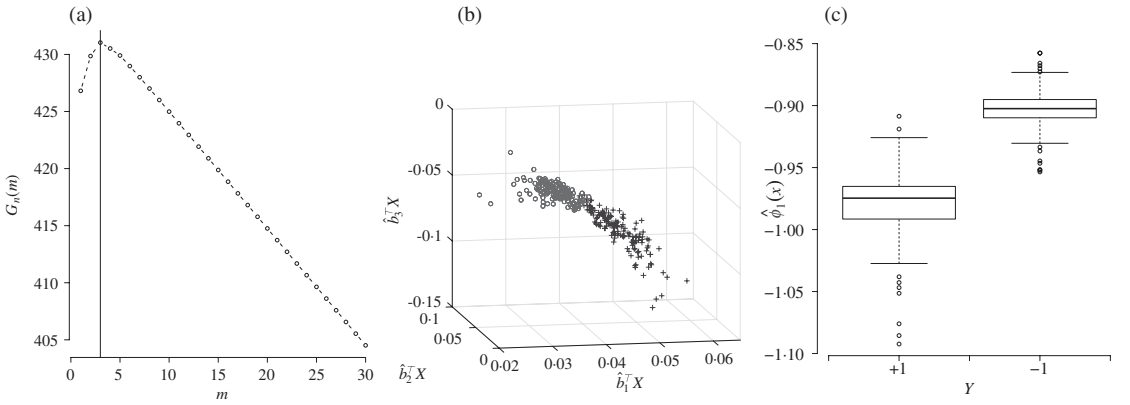

Fig. 4. Wisconsin diagnostic breast cancer data: panel (a) depicts $G_n(m;\rho,\hat{M}_n)$ as a function of m at $\rho = 0{\cdot}009$, which is maximized at 3 (vertical line); panel (b) shows the scatterplot of $(\hat{b}_1^{\mathrm{T}}X \times \hat{b}_2^{\mathrm{T}}X \times \hat{b}_3^{\mathrm{T}}X)$ where $\hat{B} = (\hat{b}_1, \hat{b}_2, \hat{b}_3)$ is estimated by the linear principal weighted support vector machine; panel (c) depicts boxplots of $\hat{\phi}_1(X)$ estimated by the kernel principal weighted support vector machine for the two classes.

Table 2. *Averaged test error rate* (%) *of the five-nearest-neighbour classifier for the Wisconsin diagnostic breast cancer data*

k	SAVE	FCN	IHT	DR	PRE	PWSVM	
						Linear	Kernel
1	17·9	12·1	22·0	4·4	4·5	5·2	8·6
2	12·9	8·8	12·2	4·5	4·3	5·1	8·3
3	11·6	7·8	6·9	5·7	4·5	5·3	7·9
4	11·9	7·4	5·9	5·9	4·5	5·3	7·7
5	12·2	7·2	5·9	6·2	4·5	5·4	7·8

The largest standard error for the results is 0·2%. See Table 1 for the abbreviations of the methods.

predictors projected onto the estimated $\mathcal{S}_{Y|X}$ into the five-nearest-neighbour classifier trained in the previous step. Each procedure is independently repeated 100 times and the averaged test error rates are given in Table 2. The averaged test error rate with the original data without applying sufficient dimension reduction is 7·2%.

Most sufficient dimension reduction methods with a carefully selected k, except sliced average variance estimation, perform reasonably well, in the sense that they do not lose information for classification after sufficient dimension reduction. In particular, probability-enhanced sliced inverse regression outperforms the others regardless of the number of sufficient predictors used, but the linear principal weighted support vector machine also shows promising performance. The kernel principal weighted support vector machine performs unsatisfactorily in terms of classification accuracy, and linear sufficient dimension reduction seems to be enough for this example.

6. Discussion

The proposed principal weighted support vector machine is not directly applicable when the number of predictors p is larger than the sample size n, because the estimated sample covariance matrix in (5) is singular and hence the minimizer of (5) may not be unique. One possible remedy is to employ an additional penalty for β. This resolves the problem but requires additional

assumptions to guarantee consistency. Selection of the associated penalty parameter needs to be appropriately addressed. Another way to handle the case where p is larger then n is to use marginal screening, which is popular for dimension reduction. Many screening methods possess the so-called sure screening property. However, it is not desirable to use model-based screening methods since sufficient dimension reduction does not impose any explicit relationship between the response and predictors. Recently, Mai & Zou (2013) developed a model-free screening method, called the Kolmogorov filter for binary responses. A natural idea would be to apply the principal weighted support vector machine after the Kolmogorov filter if p is excessively large.

In multiclass classification, Wu et al. (2010) proposed a model-free probability estimation method based on robust weighted multiclass support vector machines by using their Fisher consistency. It is possible to develop the principal weighted multiclass support vector machine for sufficient dimension reduction in multiclass classification.

Acknowledgement

We thank three reviewers, an associate editor, and the editor for their most helpful comments. Our research is partially supported by the National Institutes of Health, the National Science Foundation and the National Research Foundation of Korea.

Supplementary material

Supplementary Material available at *Biometrika* online includes proofs of the technical results, the details of the modified algorithm to estimate the structural dimension for the linear principal weighted support vector machine, additional simulation results, and real data validation.

References

Bura, E. & Pfeiffer, C. (2008). On the distribution of the left singular vectors of a random matrix and its applications. *Statist. Prob. Lett.* **78**, 2275–80.

Conway, J. (1990). *A Course in Functional Analysis*, vol. 96 of *Graduate Texts in Mathematics*. New York: Springer, 2nd ed.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Am. Statist. Assoc.* **91**, 983–92.

Cook, R. D. (1998a). Principal Hessian directions revisited. *J. Am. Statist. Assoc.* **93**, 84–94.

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

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22**, 1–26.

Cook, R. D. & Lee, H. (1999). Dimension reduction in binary response regression. *J. Am. Statist. Assoc.* **94**, 1187–200.

Cook, R. D. & Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455–74.

Cook, R. D. & Ni, L. (2006). Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika* **93**, 65–74.

Cook, R. D. & Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *J. Am. Statist. Assoc.* **86**, 28–33.

Hall, P. & Li, K.-C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.* **21**, 867–89.

Hastie, T. J., Rosset, S., Tibshirani, R. J. & Zhu, J. (2004). The entire regularization path for the support vector machine. *J. Mach. Learn. Res.* **5**, 1391–415.

Jaakkola, T. S., Diekhans, M. & Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. In *Proc. 7th Int. Conf. Intelligent Systems for Molecular Biology*, T. Lengauer, R. Schneider, P. Bork, D. L. Brutlag, J. I. Glasgow, H.-W. Mewes & R. Zimmer, eds. Menlo Park, California: AAAI Press, pp. 149–58.

Jiang, B., Zhang, X. & Cai, T. (2008). Estimating the confidence interval for prediction errors of support vector machine classifiers. *J. Mach. Learn. Res.* **9**, 521–40.

Kimeldorf, G. & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.* **33**, 82–95.

Koo, J.-Y., Lee, Y., Kim, Y. & Park, C. (2008). A Bahadur representation of the linear support vector machine. *J. Mach. Learn. Res.* **9**, 1343–68.
Lee, K.-Y., Li, B. & Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Ann. Statist.* **41**, 221–49.
Li, B., Artemiou, A. & Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann. Statist.* **39**, 3182–210.
Li, B. & Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Ann. Statist.* **37**, 1272–98.
Li, B. & Wang, S. (2007). On directional regression for dimension reduction. *J. Am. Statist. Assoc.* **102**, 997–1008.
Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Am. Statist. Assoc.* **86**, 316–42.
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**, 1025–39.
Li, K.-C. & Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009–52.
Li, L., Cook, R. D. & Tsai, C.-L. (2007). Partial inverse regression. *Biometrika* **94**, 615–25.
Lin, Y., Lee, Y. & Wahba, G. (2002). Support vector machines for classification in nonstandard situations. *Mach. Learn.* **46**, 191–202.
Mai, Q. & Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100**, 229–34.
Schölkopf, B., Smola, A. & Müller, K.-R. (1997). Kernel principal component analysis. In *Proc. 7th Int. Conf. Artificial Neural Networks*, W. Gerstner, A. Germond, M. Hasler & J.-D. Nicoud, eds. Berlin: Springer, pp. 583–8.
Shin, S. J., Wu, Y. & Zhang, H. H. (2014a). Two-dimensional solution surface of the weighted support vector machines. *J. Comp. Graph. Statist.* **23**, 383–402.
Shin, S. J., Wu, Y., Zhang, H. H. & Liu, Y. (2014b). Probability enhanced sufficient dimension reduction in binary classification. *Biometrics* **70**, 546–55.
Wang, J., Shen, X. & Liu, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika* **95**, 149–67.
Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *J. Comp. Graph. Statist.* **17**, 590–610.
Wu, Q., Liang, F. & Mukherjee, S. (2013). Kernel sliced inverse regression: Regularization and consistency. *Abstract Appl. Anal.* **2013**, 1–11. Article no. 540725.
Wu, Y., Zhang, H. H. & Liu, Y. (2010). Robust model-free multiclass probability estimation. *J. Am. Statist. Assoc.* **105**, 424–46.
Yeh, Y.-R., Huang, S.-Y. & Lee, Y.-Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Trans. Know. Data Eng.* **21**, 1590–603.
Yin, X., Li, B. & Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Mult. Anal.* **99**, 1733–57.
Zhu, Y. & Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Am. Statist. Assoc.* **101**, 1638–51.

[*Received on* 8 *June* 2015. *Editorial decision on* 15 *October* 2016]