Electronic Journal of Statistics
Vol. 0 (0000) 1–19
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

Sufficient Dimension Reduction via Principal Lq Support Vector Machine*

Andreas Artemiou

School of Mathematics, Cardiff University
e-mail: ArtemiouA@cardiff.ac.uk

and

Yuexiao Dong

Department of Statistics, Temple University
e-mail: ydong@temple.edu

Abstract: Principal support vector machine was proposed recently by Li, Artemiou and Li (2011) to combine L1 support vector machine and sufficient dimension reduction. We introduce the principal Lq support vector machine as a unified framework for linear and nonlinear sufficient dimension reduction. By noticing that the solution of L1 support vector machine may not be unique, we set $q > 1$ to ensure the uniqueness of the solution. The asymptotic distribution of the proposed estimators are derived for $q > 1$. We demonstrate through numerical studies that the proposed L2 support vector machine estimators improve existing methods in accuracy, and are less sensitive to the tuning parameter selection.

Keywords and phrases: Inverse regression, L2 support vector machine, Reproducing kernel Hilbert space.

1. Introduction

The emergence of computer power and the increase in storage capabilities have provided scientists the necessary tools to collect and store high dimensional data. In an effort to reduce the dimensionality of the data before applying classical techniques for inference, sufficient dimension reduction has seen great development among recent statistics literature. The main objective of sufficient dimension reduction is to estimate a $p \times d$ matrix $\boldsymbol{\beta}$ with $d < p$, such that

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

where Y is the response and $\boldsymbol{X}$ is a p-dimensional predictor. The column space of $\boldsymbol{\beta}$ in (1) is called the dimension reduction space. Under mild assumptions (Cook, 1998a; Yin, Li and Cook, 2008), the intersection of all dimension reduction spaces is a dimension reduction space itself. This unique minimum dimension reduction space is called the central space, and is denoted by $\mathcal{S}_{Y|\boldsymbol{X}}$. The dimensionality of $\mathcal{S}_{Y|\boldsymbol{X}}$ is called the structural dimension. We assume the existence of the central space throughout this article.

Since the introduction of the seminal sliced inverse regression method in Li (1991), many sufficient dimension reduction procedures have been proposed in the literature, such as Cook and Weisberg (1991), Cook (1998b), Xia et al. (2002), Li, Zha and Chiaromonte (2005), Li and Wang (2007),

*This is an original paper

etc. More recently, Li, Artemiou and Li (2011) proposed the principal support vector machine, which combines the ideas of sliced inverse regression (Li, 1991), contour regression (Li, Zha and Chiaromonte, 2005) and support vector machine (Cortes and Vapnik, 1995; Vapnik 1998). By employing L1 support vector machine and focusing on separating hyperplanes rather than slice means, the principal support vector machine improves the accuracy of popular inverse regression estimators.

As demonstrated elegantly in Li, Artemiou and Li (2011), when we apply a modified L1 support vector machine for binary response Y, the normal vector $\boldsymbol{\psi}$ from the optimal hyperplane $\boldsymbol{\psi}^{\mathsf{T}}\boldsymbol{X}-t=0$ naturally belongs to the central space $\mathcal{S}_{Y|\boldsymbol{X}}$. For continuous response, the predictors are separated into several slices according to the values of the responses, and multiple support vector machines are implemented to find the optimal hyperplanes that separate these slices. The principal eigenvectors of the normal vectors from these hyperplanes, known as the principal L1 support vector machines estimators, thus recover the central space. In spite of the popularity of L1 support vector machine among practitioners and researchers, the corresponding objective function is not strictly convex and may have multiple optimal solutions (Burges and Crisp, 1999). More specifically, if the optimal hyperplane is described by an equation $\boldsymbol{\psi}^{\mathsf{T}}\boldsymbol{X}-t=0$ for some $\boldsymbol{\psi}\in\mathbb{R}^p$ and $t\in\mathbb{R}$, then L1 support vector machine may have multiple optimal solutions where all of them share the same $\boldsymbol{\psi}$ but have different t. On the other hand, Lq support vector machine with $q>1$ enjoys unique solution due to the strict convexity of its objective function. See, for example, Burges and Crisp (1999) and Abe (2002). This motivates us to consider Lq support vector machine for sufficient dimension reduction with $q>1$.

We extend Li, Artemiou and Li (2011) and propose the principal Lq support vector machine with $q>1$ in this article. The principal Lq support vector machine inherits the benefits of the principal L1 support vector machine, and combines both linear and nonlinear sufficient dimension reduction in a general framework. By focusing on the theoretical development of the principal Lq support vector machine with $q>1$, we clearly demonstrate the connections and differences between our proposal and the existing principal L1 support vector machine estimator. As we will see later, both estimators depend on the tuning parameter known as the misclassification penalty. When the misclassification penalty goes to infinity, these estimators become equivalent. Our proposal improves the accuracy of the existing estimators at the sample level, and it enjoys the additional benefit of being less sensitive to the choice of the misclassification penalty. Along with the theoretical development of the principal Lq support vector machine estimator, we develop a more complete asymptotic theory for the existing support vector machine literature.

2. Principal Lq support vector machine

Let $\{(\boldsymbol{X}_i,Y_i), i=1,\ldots,n\}$ be an i.i.d. sample of $(\boldsymbol{X},Y)$. Denote $\boldsymbol{\Sigma}=\mathrm{var}(\boldsymbol{X})$ and $\bar{\boldsymbol{X}}=n^{-1}\sum_{i=1}^n\boldsymbol{X}_i$. Suppose Y is binary random variable with values ±1. The Lq support vector machine (Abe, 2010) is defined through the following optimization problem,

$$\begin{aligned}&\text{minimize}\quad \boldsymbol{\psi}^{\mathsf{T}}\boldsymbol{\psi}+\lambda q^{-1}n^{-1}\sum_{i=1}^n\xi_i^q\ \ \text{among } (\boldsymbol{\psi},t,\boldsymbol{\xi})\in\mathbb{R}^p\times\mathbb{R}\times\mathbb{R}^n\\ &\text{subject to}\quad \xi_i\geq0,\ Y_i\{\boldsymbol{\psi}^{\mathsf{T}}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})-t\}\geq1-\xi_i,\quad i=1,\ldots,n.\end{aligned}\tag{2}$$

Here $\lambda>0$ is a tuning parameter often referred to as the cost or misclassification penalty. The vector $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)^{\mathsf{T}}$, where ξ_i's are the misclassification distances with $\xi_i=0$ for correctly classified

points and $\xi_i > 0$ for incorrectly classified points. The separating hyperplane $\boldsymbol{\psi}^\intercal \boldsymbol{X} - t = 0$ is described by $\boldsymbol{\psi} \in \mathbb{R}^p$ and $t \in \mathbb{R}$. The solution $(\boldsymbol{\psi}^*, t^*)$ to this minimization problem gives the optimal hyperplane. For fixed $\boldsymbol{\psi}$ and t, minimizing (2) over $\boldsymbol{\xi}$ leads to solution $\xi_i^* = [1 - Y_i\{\boldsymbol{\psi}^\intercal(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) - t\}]^+$, where $a^+ = \max(a, 0)$. Plug ξ_i^* into (2) leads to the following unconstrained minimization problem,

$$\boldsymbol{\psi}^\intercal \boldsymbol{\psi} + \lambda q^{-1} n^{-1} \sum_{i=1}^{n} \left([1 - Y_i\{\boldsymbol{\psi}^\intercal(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) - t\}]^+\right)^q. \tag{3}$$

At the population level, (3) corresponds to

$$\boldsymbol{\psi}^\intercal \boldsymbol{\psi} + \lambda q^{-1} E\left([1 - Y\{\boldsymbol{\psi}^\intercal(\boldsymbol{X} - E\boldsymbol{X}) - t\}]^+\right)^q. \tag{4}$$

In a regression setting the response Y is a continuous variable. Let A_1 and A_2 be two disjoint sets of the range of Y and define $\tilde{Y} = I(Y \in A_2) - I(Y \in A_1)$ to be the discretized response variable. We modify (4) and define the following objective function,

$$\Lambda(\boldsymbol{\psi}, t) = \boldsymbol{\psi}^\intercal \boldsymbol{\Sigma} \boldsymbol{\psi} + \lambda q^{-1} E\left([1 - \tilde{Y}\{\boldsymbol{\psi}^\intercal(\boldsymbol{X} - E\boldsymbol{X}) - t\}]^+\right)^q, \tag{5}$$

where $\boldsymbol{\psi}^\intercal \boldsymbol{\Sigma} \boldsymbol{\psi}$ and $\tilde{Y}$ replaces $\boldsymbol{\psi}^\intercal \boldsymbol{\psi}$ and Y in (4) respectively. Replacing Y with $\tilde{Y}$ allows us to handle continuous as well as discrete response Y in (5). As we will see in the next theorem, adding $\boldsymbol{\Sigma}$ in the first term of (5) is essential to the unbiasedness of the resulting principal Lq support vector machine estimator.

Theorem 1 *Suppose $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal \boldsymbol{X})$ is a linear function of $\boldsymbol{\beta}^\intercal \boldsymbol{X}$, where $\boldsymbol{\beta}$ is as defined in (1). If $(\boldsymbol{\psi}_0, t_0)$ minimizes $\Lambda(\boldsymbol{\psi}, t)$ in (5) among all $(\boldsymbol{\psi}, t) \in \mathbb{R}^p \times \mathbb{R}$, then $\boldsymbol{\psi}_0 \in \mathcal{S}_{Y|\boldsymbol{X}}$.*

Theorem 1 suggests that we can estimate the central space $\mathcal{S}_{Y|\boldsymbol{X}}$ via minimization of objective function (5). Note that for population level objective function such as $\Lambda(\boldsymbol{\psi}, t)$ in (5), the minimizer is denoted by $(\boldsymbol{\psi}_0, t_0)$. For sample level objective function such as (2), we denote the minimizer by $(\boldsymbol{\psi}^*, t^*, \boldsymbol{\xi}^*)$.

With $q = 1$ in (5), $\Lambda(\boldsymbol{\psi}, t)$ reduces to the objective function proposed in Li, Artemiou and Li (2011). Although there is a unique value $\boldsymbol{\psi}_0$ that minimizes $\Lambda(\boldsymbol{\psi}, t)$ in this case, the value of t_0 that minimizes $\Lambda(\boldsymbol{\psi}, t)$ is not unique. This is because the second term of the objective function $\Lambda(\boldsymbol{\psi}, t)$ is not a strictly convex function of t when $q = 1$. On the other hand, the second term becomes a strictly convex function of t when $q > 1$, which guarantes the uniqueness of the solution $(\boldsymbol{\psi}_0, t_0)$. Without interrupting the flow of the main article, we provide in Appendix B the sufficient conditions for the existence of non-unique minimizer t_0 for $\Lambda(\boldsymbol{\psi}, t)$ when $q = 1$.

3. Sample estimation algorithm

Given i.i.d. sample $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, we study the sample algorithm of principal Lq support vector machine to estimate the central space $\mathcal{S}_{Y|\boldsymbol{X}}$. We first develop a general result for $q > 1$ and then focus on $q = 2$ for our implementation. Let $\hat{\boldsymbol{\Sigma}}$ be the sample covariance estimator. The sample version objective function of the principal Lq support vector machine can be modified from (2) as follows.

$$\begin{aligned} &\text{minimize} \quad \boldsymbol{\psi}^\intercal \hat{\boldsymbol{\Sigma}} \boldsymbol{\psi} + \lambda q^{-1} n^{-1} \sum_{i=1}^{n} \xi_i^q \ \text{ among } (\boldsymbol{\psi}, t, \boldsymbol{\xi}) \in \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^n \\ &\text{subject to} \quad \tilde{Y}_i\{\boldsymbol{\psi}^\intercal(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) - t\} \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n, \end{aligned} \tag{6}$$

where $\boldsymbol{\psi}^\intercal\boldsymbol{\psi}$ and Y in (2) are substituted by $\boldsymbol{\psi}^\intercal\hat{\boldsymbol{\Sigma}}\boldsymbol{\psi}$ and $\tilde{Y}$ respectively. Let $\boldsymbol{Z}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$ and $\boldsymbol{\zeta} = \hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\psi}$. Then (6) becomes

$$\text{minimize } \boldsymbol{\zeta}^\intercal\boldsymbol{\zeta} + \lambda q^{-1}n^{-1}\sum_{i=1}^{n}\xi_i^q, \text{ subject to } \xi_i - 1 + \tilde{Y}_i(\boldsymbol{Z}_i^\intercal\boldsymbol{\zeta} - t) \geq 0, \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$$

Denote $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\intercal$, $\tilde{\boldsymbol{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^\intercal$, $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^\intercal$, $\mathbf{1}_n = (1, \ldots, 1)^\intercal$, and $\mathbf{0}_n = (0, \ldots, 0)^\intercal$. Let $\odot$ be the Hadamard product, and suppose the symbol $\geq$ denotes componentwise inequality. Then we get the equivalent matrix form optimization problem

$$\text{minimize } \boldsymbol{\zeta}^\intercal\boldsymbol{\zeta} + \lambda q^{-1}n^{-1}\mathbf{1}_n^\intercal\boldsymbol{\xi}^q, \text{ subject to } \boldsymbol{\xi} - \mathbf{1}_n + \tilde{\boldsymbol{Y}} \odot (\boldsymbol{Z}^\intercal\boldsymbol{\zeta} - t) \geq \mathbf{0}_n, \quad \boldsymbol{\xi} \geq \mathbf{0}_n \tag{7}$$

The next proposition will facilitate finding the solution of (7).

Proposition 1 *For $q > 1$, the solution $\boldsymbol{\zeta}^*$ of (7) is given by $\boldsymbol{\zeta}^* = \frac{1}{2}\boldsymbol{Z}^\intercal(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})$, where $\boldsymbol{\alpha}$ is the solution to the following optimization problem:*

$$\begin{aligned} &\textit{maximize} \quad \boldsymbol{\alpha}^\intercal\mathbf{1}_n - \frac{1}{4}(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal\boldsymbol{Z}\boldsymbol{Z}^\intercal(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}) + \frac{1-q}{q}(\lambda n^{-1})^{\frac{1}{1-q}}(\boldsymbol{\alpha}^\intercal)^{\frac{q}{q-1}}\mathbf{1}_n \\ &\textit{subject to} \quad \boldsymbol{\alpha} \geq \mathbf{0}_n, \quad (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal\mathbf{1}_n = 0. \end{aligned} \tag{8}$$

We relegate the proof of Proposition 1 in Appendix A.

Note that $\tilde{\boldsymbol{Y}}$ has entries ± 1 and $\boldsymbol{\alpha}^\intercal\boldsymbol{\alpha} = (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})$. Thus for the special case of $q = 2$, (8) reduces to the following quadratic programming problem,

$$\begin{aligned} &\text{maximize} \quad \boldsymbol{\alpha}^\intercal\mathbf{1}_n - \frac{1}{4}(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal\left(\boldsymbol{Z}\boldsymbol{Z}^\intercal + \frac{2n}{\lambda}\boldsymbol{I}_n\right)(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}) \\ &\text{subject to} \quad \boldsymbol{\alpha} \geq \mathbf{0}_n, \quad (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal\mathbf{1}_n = 0. \end{aligned} \tag{9}$$

For the corresponding problem with $q = 1$, one can show that solving the sample version of (5) leads to $\boldsymbol{\zeta}^* = \frac{1}{2}\boldsymbol{Z}^\intercal(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})$, with $\boldsymbol{\alpha}$ being the solution to

$$\begin{aligned} &\text{maximize} \quad \boldsymbol{\alpha}^\intercal\mathbf{1}_n - \frac{1}{4}(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal\boldsymbol{Z}\boldsymbol{Z}^\intercal(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}) \\ &\text{subject to} \quad \mathbf{0}_n \leq \boldsymbol{\alpha} \leq \lambda\mathbf{1}_n, \quad (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal\mathbf{1}_n = 0. \end{aligned} \tag{10}$$

One can follow the proof of Theorem 3 in Li, Artemiou and Li (2011) for the derivation of (10). We notice an interesting fact by comparing (9) and (10). Namely, the two problems become equivalent as $\lambda \to \infty$. We will discuss this property further in our numerical studies section. It is easy to see that using $q > 2$ in Proposition 1 will not give a quadratic programming problem. While the asymptotic theory will be developed for any $q > 1$, our numerical studies focus on $q = 2$.

We present the principal L2 support vector machine algorithm to conclude this section. Suppose for now the structural dimension d of the central space $\mathcal{S}_{Y|\boldsymbol{X}}$ is known.

1. Calculate the sample mean $\bar{\boldsymbol{X}}$, sample variance matrix $\hat{\boldsymbol{\Sigma}}$, and the standardized predictor $\boldsymbol{Z}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$.
2. Let $q_r, r = 1, \ldots, H-1$, be equally spaced sample percentiles of $\{Y_1, \ldots, Y_n\}$.

3. For each q_r, construct $\tilde{Y}_i^r = I(Y_i > q_r) - I(Y_i \leq q_r)$. Let $\hat{\boldsymbol{\zeta}}_r$ be the solution of

$$\text{minimize } \boldsymbol{\zeta}^\intercal \boldsymbol{\zeta} + 2^{-1}\lambda n^{-1}\mathbf{1}_n^\intercal \boldsymbol{\xi}^2, \text{ subject to } \boldsymbol{\xi} - \mathbf{1}_n + \tilde{\boldsymbol{Y}}^r \odot (\boldsymbol{Z}^\intercal \boldsymbol{\zeta} - t) \geq \mathbf{0}_n,$$

where $\tilde{\boldsymbol{Y}}^r = (\tilde{Y}_1^r, \ldots, \tilde{Y}_n^r)^\intercal$. Calculate $\hat{\boldsymbol{\psi}}_r = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\zeta}}_r$.
4. Calculate the d leading eigenvectors $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_d$ of

$$\hat{\boldsymbol{V}} = \sum_{r=1}^{H-1} \hat{\boldsymbol{\psi}}_r \hat{\boldsymbol{\psi}}_r^\intercal. \tag{11}$$

5. Estimate $\mathcal{S}_{Y|\boldsymbol{X}}$ by the subspace spanned by $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_d$.

4. Asymptotic results for LqSVM

In this section we discuss the asymptotic results for the PLqSVM with $q > 1$. Assume $E(\boldsymbol{X}) = \mathbf{0}$ without loss of generality. First we introduce the following notations: $\boldsymbol{\theta} = (\boldsymbol{\psi}^\intercal, t)^\intercal, \boldsymbol{W} = (\boldsymbol{X}^\intercal, \tilde{Y})^\intercal, \boldsymbol{X}^\dagger = (\boldsymbol{X}^\intercal, -1)^\intercal$, $\lambda^\dagger = \lambda q^{-1}$, and $\boldsymbol{\Sigma}^\dagger = \text{diag}(\boldsymbol{\Sigma}, 0)$, where $\text{diag}(\mathbf{A}, \mathbf{B})$ denotes a block diagonal matrix with $\mathbf{A}$ and $\mathbf{B}$ on the block diagonals. $\Lambda(\boldsymbol{\psi}, t)$ in (5) can be rewritten as $E\{m(\boldsymbol{\theta}, \boldsymbol{W})\}$, where

$$m(\boldsymbol{\theta}, \boldsymbol{W}) = \boldsymbol{\theta}^\intercal \boldsymbol{\Sigma}^\dagger \boldsymbol{\theta} + \lambda^\dagger \{(1 - \boldsymbol{\theta}^\intercal \boldsymbol{X}^\dagger \tilde{Y})^+\}^q. \tag{12}$$

Denote the corresponding sample version objective function as $E_n\{m(\boldsymbol{\theta}, \boldsymbol{W})\}$. Let $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$ be the minimizer of $E\{m(\boldsymbol{\theta}, \boldsymbol{W})\}$ and $E_n\{m(\boldsymbol{\theta}, \boldsymbol{W})\}$ respectively. Before we state the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ in Theorem 2, the gradient and the Hessian matrix of the Lq objective function $E\{m(\boldsymbol{\theta}, \boldsymbol{W})\}$ are provided in the next two propositions.

Proposition 2 *Suppose for each $\tilde{y} \in \{-1, 1\}$, the distribution of $\boldsymbol{X}|\tilde{Y} = \tilde{y}$ is dominated by the Lebesgue measure. In addition, suppose $E(\|\boldsymbol{X}\|^2) < \infty$ and $E(\|\boldsymbol{X}\|^{q-1}) < \infty$. Let $D_{\boldsymbol{\theta}}$ be the $(p+1)$-dimensional column vector of differential operators $(\partial/\partial\theta_1, \ldots, \partial/\partial\theta_{p+1})^\intercal$. Then*

$$D_{\boldsymbol{\theta}}[E\{m(\boldsymbol{\theta}, \boldsymbol{W})\}] = (2\boldsymbol{\psi}^\intercal \boldsymbol{\Sigma}, 0)^\intercal - q\lambda^\dagger E\{\boldsymbol{X}^\dagger \tilde{Y}(1 - \boldsymbol{\theta}^\intercal \boldsymbol{X}^\dagger \tilde{Y})^{q-1} I(1 - \boldsymbol{\theta}^\intercal \boldsymbol{X}^\dagger \tilde{Y} > 0)\}. \tag{13}$$

Proposition 3 *Suppose $\boldsymbol{X}$ has a convex and open support, and for each $\tilde{y} \in \{-1, 1\}$, the distribution of $\boldsymbol{X}|\tilde{Y} = \tilde{y}$ is dominated by the Lebesgue measure. Let $f_{\cdot|\cdot}$ denote the conditional probability density function. Suppose, moreover:*

1. *for any linearly independent $\boldsymbol{\psi}, \boldsymbol{\delta} \in \mathbb{R}^p$, $\tilde{y} = -1, 1$, and $v, \epsilon \in \mathbb{R}$, the function*

$$u \mapsto \tilde{y}(1 - \tilde{y}(u - t) - \epsilon v)^{q-1} E\{\boldsymbol{X}^\dagger | \boldsymbol{\psi}^\intercal \boldsymbol{X} = u, \boldsymbol{\delta}^\intercal \boldsymbol{X} = v, \tilde{Y} = \tilde{y}\} f_{\boldsymbol{\psi}^\intercal \boldsymbol{X}|\boldsymbol{\delta}^\intercal \boldsymbol{X}, \tilde{Y}}(u|v, \tilde{y})$$

is continuous;
2. *for any $i = 1, \ldots, p$, and $\tilde{y} = -1, 1$, there is a nonnegative function $c_i(v, \tilde{y})$ with $E\{c_i(V, \tilde{Y})|\tilde{Y}\} < \infty$ such that*

$$\tilde{y}(1 - \tilde{y}(u - t) - \epsilon v)^{q-1} E\{X_i | \boldsymbol{\psi}^\intercal \boldsymbol{X} = u, \boldsymbol{\delta}^\intercal \boldsymbol{X} = v, \tilde{Y} = \tilde{y}\} f_{\boldsymbol{\psi}^\intercal \boldsymbol{X}|\boldsymbol{\delta}^\intercal \boldsymbol{X}, \tilde{Y}}(u|v, \tilde{y}) \leq c_i(v, \tilde{y});$$

3. *for any $\tilde{y} = -1, 1$, there is a nonnegative function $c_0(v, \tilde{y})$ with $E\{c_0(V, \tilde{Y})|\tilde{Y}\} < \infty$ such that $f_{\boldsymbol{\psi}^\intercal \boldsymbol{X}|\boldsymbol{\delta}^\intercal \boldsymbol{X}, \tilde{Y}}(u|v, \tilde{y}) \leq c_0(v, \tilde{y})$.*

Then the function $\boldsymbol{\theta} \mapsto D_{\boldsymbol{\theta}}[E\{m(\boldsymbol{\theta}, \boldsymbol{W})\}]$ *is differentiable in all directions with derivative matrix*

$$\boldsymbol{H} = 2\text{diag}(\boldsymbol{\Sigma}, 0) + q(q-1)\lambda^{\dagger} \sum_{\tilde{y}=-1,1} P(\tilde{Y} = \tilde{y}) E\{(1 - \boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\dagger}\tilde{y})^{q-2}\boldsymbol{X}^{\dagger}(\boldsymbol{X}^{\dagger})^{\intercal} I(1 - \boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\dagger}\tilde{y} > 0) | \tilde{Y} = \tilde{y}\} \tag{14}$$

Let $\hat{\boldsymbol{\theta}}$ is the solution to the sample version objective function $E_n\{m(\boldsymbol{\theta}, \boldsymbol{W})\}$, with $m(\boldsymbol{\theta}, \boldsymbol{W})$ defined in (12). We provide the influence function for the principal Lq support vector machine in the next theorem.

Theorem 2 *Suppose the conditions in Propositions 2 and 3 are satisfied. Then*

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \boldsymbol{H}^{-1}\left[(2\boldsymbol{\psi}_0^{\intercal}\boldsymbol{\Sigma}, 0)^{\intercal} - q\lambda^{\dagger} E_n\{\boldsymbol{X}^{\dagger}\tilde{Y}(1 - \boldsymbol{\theta}_0^{\intercal}\boldsymbol{X}^{\dagger}\tilde{Y})^{q-1} I(1 - \boldsymbol{\theta}_0^{\intercal}\boldsymbol{X}^{\dagger}\tilde{Y} > 0)\}\right] + o_P(n^{-1/2}),$$

where $\boldsymbol{H}$ *is given in Proposition 3.*

To apply Theorem 2 for the proposed estimator of $\mathcal{S}_{Y|\boldsymbol{X}}$, recall from the algorithm in Section 3 that for a fixed dividing point q_r, we have a corresponding $\tilde{Y}^r$, $r = 1, \ldots, H-1$. Let $\boldsymbol{W}^r = (\boldsymbol{X}^{\intercal}, \tilde{Y}^r)$ and $m(\boldsymbol{\theta}, \boldsymbol{W}^r) = \boldsymbol{\theta}^{\intercal}\boldsymbol{\Sigma}^{\dagger}\boldsymbol{\theta} - \lambda^{\dagger}\{(1 - \boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\dagger}\tilde{Y}^r)^{+}\}^q$. Let the minimizer of $E\{m(\boldsymbol{\theta}, \boldsymbol{W}^r)\}$ over $\boldsymbol{\theta}$ be $\boldsymbol{\theta}_{0r} = (\boldsymbol{\psi}_{0r}^{\intercal}, t_{0r})^{\intercal}$. The population correspondence of $\hat{\boldsymbol{V}}$ in (11) is thus $\boldsymbol{V} = \sum_{r=1}^{H-1} \boldsymbol{\psi}_{0r}\boldsymbol{\psi}_{0r}^{\intercal}$. Furthermore, let $\boldsymbol{K}_{p,p}$ be the unique matrix satisfying $\boldsymbol{K}_{p,p}\text{vec}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A}^{\intercal})$ for any $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, let $\mathbf{F}_r$ be the first r rows of $\boldsymbol{H}_r^{-1}$ with $\boldsymbol{H}_r$ the Hessian of $E\{m(\boldsymbol{\theta}, \boldsymbol{W}^r)\}$ and denote $\boldsymbol{s}_r(\boldsymbol{\theta}, \boldsymbol{W}^r) = \mathbf{F}_r[(2\boldsymbol{\psi}^{\intercal}\boldsymbol{\Sigma}, 0)^{\intercal} - q\lambda^{\dagger}E\{\boldsymbol{X}^{\dagger}\tilde{Y}^r(1 - \boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\dagger}\tilde{Y}^r)^{q-1} I(1 - \boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\dagger}\tilde{Y}^r > 0)\}]$. Now we present the asymptotic distribution of $\hat{\boldsymbol{V}}$.

Theorem 3 *Suppose the conditions in Propositions 2 and 3 are satisfied. Then* $\sqrt{n}\text{vec}(\hat{\boldsymbol{V}} - \boldsymbol{V})$ *converges to multivariate normal with mean* $\mathbf{0}$ *and variance* $\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_1$*, were* $\boldsymbol{\Lambda}_1 = \boldsymbol{I}_{p^2} + \boldsymbol{K}_{p,p}$ *and* $\boldsymbol{\Lambda}_2 = \sum_{r=1}^{H-1}\sum_{i=1}^{H-1}[\boldsymbol{\psi}_{0r}\boldsymbol{\psi}_{0i}^{\intercal} \otimes E\{\boldsymbol{s}_r(\boldsymbol{\theta}_{0r}, \boldsymbol{W}^r)\boldsymbol{s}_i^{\intercal}(\boldsymbol{\theta}_{0i}, \boldsymbol{W}^i)\}]$.

Let $\boldsymbol{D}$ be a diagonal matrix with diagonal elements being the d leading eigenvalues of $\boldsymbol{V}$. Let $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d)$, where $\boldsymbol{u}$'s are the d leading eigenvectors of $\boldsymbol{V}$. Denote $\hat{\boldsymbol{U}} = (\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_d)$ correspondingly. We get the asymptotic distribution of as a result of Corollary 1 in Bura and Pfeiffer (2008).

Corollary 1 *Suppose the conditions in Propositions 2 and 3 are satisfied, and* $\boldsymbol{V}$ *has rank* d*. Then* $\sqrt{n}\,\text{vec}(\hat{\boldsymbol{U}} - \boldsymbol{U}) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, (\boldsymbol{D}^{-1}\boldsymbol{U}^{\intercal} \otimes \boldsymbol{I}_p)\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_1(\boldsymbol{D}^{-1}\boldsymbol{U}^{\intercal} \otimes \boldsymbol{I}_p)\right)$.

In the last step of the principal L2 support vector machine algorithm in Section 3, we extract d leading eigenvectors of $\hat{\boldsymbol{V}}$. Since d is unknown in practice, one needs to estimate the dimensionality of $\mathcal{S}_{Y|\boldsymbol{X}}$ before successful implementation of the proposed algorithm. We propose a modified BIC criterion for this purpose. Define

$$G_n(k) = \sum_{i=1}^{k} \rho_i(\hat{\boldsymbol{V}}) - \rho_1(\hat{\boldsymbol{V}})n^{-\frac{3}{8}}\log(\lambda + 2)\left(\frac{p}{H}\right)^{\frac{1}{4}} k, \tag{15}$$

where $\rho_i(\hat{\boldsymbol{V}})$ denotes the ith largest eigenvalue of $\hat{\boldsymbol{V}}$. Then we estimate d by $\hat{d}$, the maximizer of $G_n(k)$ over $k = 0, 1, \ldots, p$. Similar criteria have been used in Zhu, Miao and Peng (2006), Wang and Yin (2008). Our criterion is different from the one used in Li, Artemiou and Li (2011) as we include number of slices H, the predictor dimensionality p, and the misclassification penalty λ in (15). The consistency of $\hat{d}$ is provided next.

Theorem 4 *Suppose the conditions in Propositions 2 and 3 are satisfied, and* $\boldsymbol{V}$ *has rank* d*. Then* $\lim_{n\to\infty} P(\hat{d} = d) = 1$.

5. Nonlinear sufficient dimension reduction

The dimension reduction in (1) aims at finding d features that are linear combinations of the original predictors, and will be referred to as the *linear sufficient dimension reduction*. Let $\boldsymbol{\phi} : \mathbb{R}^p \mapsto \mathbb{R}^d$ be nonlinear functions satisfying

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\phi}(\boldsymbol{X}). \tag{16}$$

Identifying $\boldsymbol{\phi}(\boldsymbol{X})$ is known as *nonlinear sufficient dimension reduction*. Model (16) was first formulated in Cook (2007), and has been studied in Wu (2008), Yeh, Huang and Lee (2009) and Fukumizu, Bach and Jordan (2009). Following Li, Artemiou and Li (2011), we discuss nonlinear sufficient dimension via the principal L2 support vector machine in this section.

Let $\mathcal{H}$ be a reproducing kernel Hilbert space of the functions of $\boldsymbol{X}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. We assume $\mathcal{H}$ to have finite dimensionality for technical convenience, although this is not required in general. Let $\Sigma : \mathcal{H} \mapsto \mathcal{H}$ be the covariance operator such that $\langle f_1, \Sigma f_2 \rangle_{\mathcal{H}} = \text{cov}\{f_1(\boldsymbol{X}), f_2(\boldsymbol{X})\}$ for any $f_1, f_2 \in \mathcal{H}$. Consider the population level objective function

$$\Lambda(\psi, t) = \langle \psi, \Sigma\psi \rangle_{\mathcal{H}} + \frac{\lambda}{2} E\left(\{1 - \tilde{Y}[\psi(\boldsymbol{X}) - E\psi(\boldsymbol{X}) - t]\}^+\right)^2. \tag{17}$$

Note that (17) is parallel to (5) with $q = 2$. Let (ψ_0, t_0) be the minimizer of $\Lambda(\psi, t)$ over all $(\psi, t) \in \mathcal{H} \times \mathbb{R}$. Suppose $\sigma\{\boldsymbol{\phi}(\boldsymbol{X})\}$ is the σ-field generated by $\boldsymbol{\phi}(\boldsymbol{X})$. Under proper conditions, one can show that $\psi_0(\boldsymbol{X})$ is measurable $\sigma\{\boldsymbol{\phi}(\boldsymbol{X})\}$, which means ψ_0 is a function of the sufficient predictor $\boldsymbol{\phi}(\boldsymbol{X})$. The derivation follows Theorem 2 in Li, Artemiou and Li (2011) and is thus omitted.

Based on i.i.d. sample $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, we now describe the principle for the sample level estimation. Suppose $\mathcal{H}$ can be spanned by $\{h_1, \ldots, h_G\}$, where we choose $h_j \in \mathcal{H}$ to satisfy $E_n(h_j(\boldsymbol{X})) = 0$, $j = 1, \ldots, G$. Define $\boldsymbol{\Psi} \in \mathbb{R}^{n \times G}$ with the element in the ith row and jth column being $h_j(\boldsymbol{X}_i)$. The sample version of (17) becomes

$$\frac{1}{n}\boldsymbol{c}^\intercal \boldsymbol{\Psi}^\intercal \boldsymbol{\Psi} \boldsymbol{c} + \frac{\lambda}{2n} \sum_{i=1}^{n} [\{1 - \tilde{Y}_i(\boldsymbol{\Psi}_i^\intercal \boldsymbol{c} - t)\}^+]^2, \tag{18}$$

where $\boldsymbol{\Psi}_i = \{h_1(\boldsymbol{X}_i), \ldots, h_G(\boldsymbol{X}_i)\}^\intercal$. Let $(\boldsymbol{c}^*, t^*)$ be the minimizer of (18) over $(\boldsymbol{c}, t) \in \mathbb{R}^G \times \mathbb{R}$. Denote $\tilde{\boldsymbol{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^\intercal$ and $\boldsymbol{P}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}(\boldsymbol{\Psi}^\intercal \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^\intercal$. Parallel to Proposition 1, we have the following result

Proposition 4 *The solution* $\boldsymbol{c}^*$ *of (18) is given by* $\boldsymbol{c}^* = \frac{1}{2}(\boldsymbol{\Psi}^\intercal \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^\intercal (\tilde{\boldsymbol{Y}} \odot \boldsymbol{\alpha}^*)$*, where* $\boldsymbol{\alpha}^*$ *is the solution to the quadratic programming problem:*

$$\begin{aligned} &\textit{maximize} \quad \boldsymbol{\alpha}^\intercal \mathbf{1}_n - \frac{1}{4}(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal \left(\boldsymbol{P}_{\boldsymbol{\Psi}} + \frac{2n}{\lambda} \boldsymbol{I}_n\right)(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}) \\ &\textit{subject to} \quad \boldsymbol{\alpha} \geq \mathbf{0}_n, \quad \boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{Y}} = 0. \end{aligned} \tag{19}$$

Following similar procedures as in Li, Artemiou and Li (2011), we describe the details of carrying out nonlinear sufficient dimension reduction through reproducing kernel Hilbert space as follows. For

the function class $\mathcal{H}$, we use the reproducing kernel Hilbert space based on mapping $\kappa : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$. Common choices of κ include the Gaussian radial kernel and the polynomial kernel. Define kernel matrix $\boldsymbol{K}_n \in \mathbb{R}^{n\times n}$, with $\kappa(\boldsymbol{X}_i, \boldsymbol{X}_j)$ being the element in the ith row and jth column of $\boldsymbol{K}_n$. Define $\boldsymbol{Q}_n = \boldsymbol{I}_n - \boldsymbol{J}_n/n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{J}_n$ is the $n \times n$ matrix whose entries are 1. Let $\boldsymbol{w}_g$ be the eigenvector corresponding to λ_g, the gth largest eigenvalue of $\boldsymbol{Q}_n\boldsymbol{K}_n\boldsymbol{Q}_n$ for $g = 1, \ldots, n$. From Proposition 2 in Li, Artemiou and Li (2011), we know $\boldsymbol{\Psi}$ becomes $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_G)$. After plugging $\boldsymbol{\Psi} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_G)$ into (19) and applying Proposition 4, we get $\boldsymbol{c}^* \in \mathbb{R}^G$. Recall from the sample level algorithm in Section 3 that $\tilde{\boldsymbol{Y}}^r = (\tilde{Y}_1^r, \ldots, \tilde{Y}_n^r)^\intercal$, where $\tilde{Y}_i^r = I(Y_i > q_r) - I(Y_i \leq q_r)$ and q_r denotes the equally spaced sample percentiles of $\{Y_1, \ldots, Y_n\}$ for $r = 1, \ldots, H-1$. When we replace $\tilde{\boldsymbol{Y}}$ in (19) with $\tilde{\boldsymbol{Y}}^r$, the corresponding solution $\boldsymbol{c}^*$ becomes $\boldsymbol{c}^{r*}$. Let $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_d$ be the d leading eigenvectors of $\sum_{r=1}^{H-1} \boldsymbol{c}^{r*}(\boldsymbol{c}^{r*})^\intercal$. For $t = 1, \ldots, d$ and $g = 1, \ldots, G$, denote the gth component of $\hat{\boldsymbol{u}}_t$ as $\hat{u}_{tg}$. For $i = 1, \ldots, n$ and $g = 1, \ldots, G$, denote the ith component of $\boldsymbol{w}_g$ as w_{gi}. From (16), we have $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^d$ as a nonlinear reduction of $\boldsymbol{x} \in \mathbb{R}^p$. At the sample level, the tth component of $\boldsymbol{\phi}(\boldsymbol{x})$ is then estimated by $\sum_{g=1}^{G} \hat{u}_{tg} h_g(\boldsymbol{x})$, where $h_g(\boldsymbol{x}) = \lambda_g^{-1} \sum_{i=1}^{n} w_{gi}[\kappa(\boldsymbol{x}, \boldsymbol{X}_i) - E_n\kappa(\boldsymbol{x}, \boldsymbol{X})]$.

6. Numerical studies

We use synthetic examples as well as real data analysis to demonstrate the finite-sample performance of the proposed methods in this section.

Example 1: This example is designed to compare the principal Lq support vector machine estimators for linear sufficient dimension reduction. As it has been demonstrated in Li, Artemiou and Li (2011) that the principal L1 support vector machine can consistently outperform popular methods such as sliced inverse regression (Li, 1991), sliced average variance estimation (Cook and Weisberg, 1991), and directional regression (Li and Wang, 2007), we focus on comparing the principal L1 support vector machine with the newly proposed principal L2 support vector machine estimator. Consider

$$\begin{aligned} \text{Model I}: &\quad Y = X_1 + X_2 + \sigma\varepsilon, \\ \text{Model II}: &\quad Y = X_1/\{0.5 + (X_2 + 1)^2\} + \sigma\varepsilon, \\ \text{Model III}: &\quad Y = X_1(X_1 + X_2 + 1) + \sigma\varepsilon, \end{aligned}$$

where $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$, $\sigma = .2$, and $\varepsilon \sim N(0, 1)$ independent of $\boldsymbol{X}$. We set q_r as equally spaced sample percentiles of $\{Y_1, \ldots, Y_n\}$ for $r = 1, \ldots, H-1$, and define $\tilde{Y}_i^r = I(Y_i > q_r) - I(Y_i \leq q_r)$. Let sample size $n = 100$, number of slices $H = 10, 20, 50$, and $p = 10, 20, 30$. Suppose $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is the basis of the central space. Denote its sample estimator as $\hat{\boldsymbol{\beta}}$. We measure the accuracy of $\hat{\boldsymbol{\beta}}$ by $\Delta = \|\boldsymbol{P}_{\boldsymbol{\beta}} - \boldsymbol{P}_{\hat{\boldsymbol{\beta}}}\|$, where $\boldsymbol{P}_{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}^\intercal\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal$, $\boldsymbol{P}_{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}^\intercal\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}}^\intercal$, and $\|\cdot\|$ is the Frobenius norm. The results are summarized in Table 1. The entries are of the form $a(b)$, which are the means and the standard errors of Δ based on 200 repetitions. Smaller values in Table 1 mean better estimation. In all models across different combinations of p and H, we see that the principal L2 support vector machine can consistently improve over its L1 counterpart for $\lambda = 1$. When λ increases to 10 and 100, the estimation improves for both L1 and L2 support vector machine, and the difference between the two methods become smaller. This verifies the theoretical finding from Section 3, where we showed that the two algorithms become equivalent as $\lambda \to \infty$.

Example 2: This example is to examine the validity of estimating the structural dimension d via the modified BIC criterion (15). We include Model I and Model III from the previous example, and compare principal Lq support vector machine with $q = 1$ or $q = 2$. The misclassification penalty is

TABLE 1
Estimating $\mathcal{S}_{Y|\boldsymbol{X}}$ via the principal Lq support vector machine. The means and standard errors of Δ are reported based on 200 repetitions in Example 1.

λ	Model	p	$H=10$		$H=20$		$H=50$	
			$q=1$	$q=2$	$q=1$	$q=2$	$q=1$	$q=2$
$\lambda=1$	I	10	.22 (.058)	.15 (.045)	.22 (.054)	.15 (.042)	.22 (.053)	.15 (.044)
		20	.33 (.070)	.25 (.062)	.33 (.060)	.24 (.054)	.33 (.067)	.24 (.056)
		30	.43 (.074)	.32 (.069)	.44 (.074)	.33 (.065)	.43 (.084)	.32 (.070)
	II	10	.94 (.217)	.74 (.193)	.93 (.218)	.71 (.170)	.94 (.212)	.73 (.155)
		20	1.23 (.159)	1.06 (.156)	1.17 (.160)	1.01 (.153)	1.21 (.132)	1.03 (.145)
		30	1.35 (.123)	1.23 (.138)	1.34 (.120)	1.21 (.122)	1.35 (.114)	1.23 (.125)
	III	10	1.19 (.252)	1.10 (.256)	1.20 (.216)	1.13 (.197)	1.16 (.252)	1.08 (.245)
		20	1.48 (.164)	1.43 (.177)	1.45 (.186)	1.43 (.190)	1.47 (.172)	1.43 (.182)
		30	1.62 (.144)	1.57 (.146)	1.64 (.146)	1.61 (.149)	1.60 (.162)	1.57 (.173)
$\lambda=10$	I	10	.14 (.038)	.11 (.031)	.13 (.036)	.10 (.025)	.13 (.038)	.09 (.027)
		20	.21 (.034)	.17 (.031)	.20 (.043)	.17 (.034)	.19 (.042)	.16 (.035)
		30	.28 (.052)	.25 (.045)	.27 (.062)	.24 (.054)	.25 (.052)	.22 (.042)
	II	10	.79 (.196)	.67 (.158)	.80 (.193)	.69 (.166)	.76 (.182)	.65 (.148)
		20	1.09 (.149)	.98 (.146)	1.03 (.158)	.93 (.149)	1.06 (.154)	.96 (.151)
		30	1.24 (.133)	1.18 (.132)	1.25 (.117)	1.19 (.120)	1.21 (.128)	1.15 (.138)
	III	10	1.06 (.270)	1.05 (.240)	1.02 (.244)	1.01 (.208)	1.01 (.243)	1.00 (.215)
		20	1.35 (.188)	1.41 (.166)	1.38 (.175)	1.41 (.160)	1.35 (.188)	1.39 (.195)
		30	1.54 (.161)	1.58 (.150)	1.52 (.160)	1.55 (.164)	1.52 (.171)	1.57 (.162)
$\lambda=100$	I	10	.10 (.026)	.09 (.024)	.09 (.022)	.08 (.023)	.09 (.023)	.08 (.019)
		20	.19 (.035)	.17 (.034)	.17 (.034)	.15 (.033)	.15 (.034)	.15 (.029)
		30	.27 (.051)	.25 (.044)	.26 (.048)	.24 (.045)	.24 (.042)	.22 (.036)
	II	10	.67 (.180)	.64 (.158)	.63 (.167)	.62 (.166)	.62 (.156)	.60 (.151)
		20	.98 (.177)	.96 (.176)	.94 (.157)	.93 (.168)	.95 (.169)	.94 (.167)
		30	1.24 (.148)	1.22 (.145)	1.17 (.140)	1.17 (.142)	1.16 (.134)	1.14 (.133)
	III	10	0.96 (.243)	1.09 (.226)	0.91 (.240)	1.04 (.276)	0.90 (.262)	1.03 (.280)
		20	1.29 (.212)	1.42 (.207)	1.22 (.220)	1.35 (.220)	1.27 (.202)	1.41 (.192)
		30	1.58 (.180)	1.62 (.163)	1.56 (.171)	1.61 (.150)	1.53 (.196)	1.59 (.162)

TABLE 2
Estimating structural dimension d via the principal Lq support vector machine. The proportions that $\hat{d}=d$ are reported based on 200 repetitions in Example 2.

H	p	Method	Model I			Model III		
			$n=200$	$n=300$	$n=400$	$n=200$	$n=300$	$n=400$
10	10	$q=1$	.89	.94	.92	.56	.56	.46
		$q=2$	1	1	1	.78	.85	.93
	20	$q=1$	.95	.93	.91	.47	.62	.68
		$q=2$	1	1	1	.53	.74	.85
	30	$q=1$	1	.97	.98	.25	.52	.68
		$q=2$	1	1	1	.50	.73	.87
20	10	$q=1$	.97	.99	1	.61	.70	.64
		$q=2$	.97	1	1	.85	.92	1
	20	$q=1$	.98	.99	.98	.61	.67	.64
		$q=2$	1	1	1	.77	.94	.98
	30	$q=1$	.98	1	.99	.29	.61	.74
		$q=2$	1	1	1	.74	.91	.97

TABLE 3
Proportion of correct estimation of structural dimension d via the principal Lq support vector machine for $\lambda = 10, 100$ and $p = 10$ for $q = 1$ and $q = 2$.

H	λ	Method	Model I			Model III		
			$n = 200$	$n = 300$	$n = 400$	$n = 200$	$n = 300$	$n = 400$
10	1	$q = 1$	.89	.94	.92	.56	.56	.46
		$q = 2$	1	1	1	.78	.85	.93
	10	$q = 1$	.79	.85	.80	.32	.35	.42
		$q = 2$	1	1	1	.51	.56	.47
	100	$q = 1$	0.62	0.69	0.72	.02	.13	.10
		$q = 2$	1	1	1	.10	.12	.13
20	1	$q = 1$	.97	.99	1	.61	.70	.64
		$q = 2$	.97	1	1	.85	.92	1
	10	$q = 1$	.95	.93	.99	.20	.20	.31
		$q = 2$	1	1	1	.44	.41	.53
	100	$q = 1$	.89	.92	.99	.10	.11	.17
		$q = 2$	.1	1	1	.30	.31	.33

fixed to be $\lambda = 1$. Across $p = 10, 20, 30$, $H = 10, 20$ and $n = 200, 300, 400$, we report in Table 2 the proportions that d is correctly estimated based on 200 repetitions. We see that both principal support vector machine estimators work reasonably well for Model I where true $d = 1$. In the more challenging case of Model III where $d = 2$, the superiority of principal L2 support vector machine becomes more obvious. The estimator $\hat{d}$ based on the principal L1 support vector machine could lead to very bad performances, especially when $n = 200$ or $p = 30$. As n increases and p decreases, both methods improve and get a higher proportion of correctly identified d.

We repeated the experiments for $p = 10$ and $\lambda = 10, 100$ to further investigate the role of λ. We show the results in the Table 3 along with the results for $\lambda = 1$ (which are the same from Table 2). For model 1, the criterion was still perfect when $q = 2$ while it's performance was decreasing when $q = 1$. For model 3, for both $q = 1$ and $q = 2$ the performance was decreasing for larger λ's. When $\lambda = 10$, $q = 2$ was still outperforming and for $\lambda = 100$ the two were mostly equivalent.

Example 3: This real data analysis is to demonstrate the effect of misclassification penalty λ on principal support vector machine estimators. Consider the concrete slump test data studied in Yeh (2007). The response variable is the concrete flow. There are 7 predictors: cement, slag, fly ash, water, superplasticizer (SP), coarse aggregate, and fine aggregate. The sample size is $n = 103$. Fix $H = 20$ and $d = 1$, we compare the L1 and the L2 estimators across $\lambda = 1, 10, 1000$. We report the components of $\hat{\boldsymbol{\beta}}$ in Table 4. Although the two estimator are seemingly different when $\lambda = 1$, they become very close to each other when $\lambda = 1000$. This confirms our findings in Section 3. In the first row of Figure 1, we provide scatterplots of Y versus $\hat{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{X}$ based on the principal L1 support vector machine estimators. We see the patterns change significantly while λ increases. From the scatterplots in the second row of Figure 1, we see that the principal L2 support vector machine estimator are less sensitive to the choice of λ.

Example 4: We study nonlinear sufficient dimension reduction via the principal Lq support vector machine in this example. In addition to Model III: $Y = X_1(X_1 + X_2 + 1) + \sigma\varepsilon$ from Example 1, consider

$$\begin{aligned} \text{Model IV}: &\quad Y = X_1^2 + X_2 + \sigma\epsilon, \\ \text{Model V}: &\quad Y = (X_1^2 + X_2^2)^{1/2}\ \log(X_1^2 + X_2^2)^{1/2} + \sigma\epsilon, \end{aligned}$$

TABLE 4
Comparing the principal Lq support vector machine estimators across different λ. Components of $\hat{\boldsymbol{\beta}}$ are reported based on real data in Example 3.

	Method	Cement	Slag	Fly Ash	Water	SP	Coarse	Fine
$\lambda = 1$	$q = 1$	.157	.215	.170	.464	.794	.164	.164
	$q = 2$	.136	.174	.156	.634	.697	.133	.144
$\lambda = 10$	$q = 1$	.143	.194	.159	.512	.779	.152	.160
	$q = 2$	.127	.148	.146	.710	.629	.135	.149
$\lambda = 1000$	$q = 1$	.113	.122	.132	.610	.727	.162	.165
	$q = 2$	.114	.121	.131	.619	.722	.160	.161

where $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$, $\sigma = .2$, and $\varepsilon \sim N(0, 1)$ independent of $\boldsymbol{X}$. Set $\lambda = 1$, $n = 100$, $p = 10, 20, 30$, and $H = 10, 20, 50$. Based on the description in Section 5, we aim to find a monotone transformation of the sufficient predictor $\boldsymbol{\phi}(\boldsymbol{X})$, which is $X_1(X_1 + X_2 + 1)$, $X_1^2 + X_2$ and $X_1^2 + X_2^2$ for Models III, IV and V respectively. To measure the accuracy of the nonlinear sufficient dimension reduction estimators, we report the absolute value of Spearman correlation between $\boldsymbol{\phi}(\boldsymbol{X})$ and $\hat{\boldsymbol{\phi}}(\boldsymbol{X})$. Note that this measure is invariant under monotone transformation. Table 5 is based on 200 repetitions, where values closer to 1 means better estimation. The Gaussian radial basis kernel $\kappa(\boldsymbol{X}_i, \boldsymbol{X}_j) = e^{-\gamma\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2}$ is used. We set the tuning parameter as $\gamma = 1/(E\|\boldsymbol{X} - \boldsymbol{X}'\|)^2$, where $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent copies of $N(\boldsymbol{0}, \boldsymbol{I}_p)$. Since the principal L2 support vector machine is mainly designed to improve existing linear sufficient dimension reduction estimators, it is comforting to observe in Table 5 that the principal L2 support vector machine is slightly better than its L1 counterpart for nonlinear sufficient dimension reduction.

TABLE 5
Estimating $\phi(\boldsymbol{X})$ for nonlinear sufficient dimension reduction. The means and standard errors of Spearman correlation are reported based on 200 repetitions in Example 4.

Models	p	$H = 10$		$H = 20$		$H = 50$	
		$q = 1$	$q = 2$	$q = 1$	$q = 2$	$q = 1$	$q = 2$
Ⅲ	10	.92 (.018)	.93 (.017)	.92 (.020)	.93 (.018)	.92 (.018)	.93 (.017)
	20	.85 (.030)	.87 (.029)	.86 (.029)	.88 (.028)	.86 (.032)	.88 (.030)
	30	.83 (.037)	.86 (.036)	.83 (.035)	.85 (.034)	.83 (.036)	.85 (.035)
Ⅳ	10	.95 (.010)	.97 (.008)	.95 (.009)	.97 (.008)	.95 (.008)	.97 (.007)
	20	.91 (.018)	.93 (.017)	.91 (.018)	.93 (.017)	.91 (.020)	.93 (.018)
	30	.89 (.023)	.91 (.022)	.89 (.023)	.91 (.021)	.89 (.024)	.91 (.023)
V	10	.89 (.022)	.91 (.022)	.90 (.026)	.91 (.024)	.90 (.023)	.91 (.023)
	20	.81 (.042)	.82 (.040)	.81 (.040)	.83 (.039)	.81 (.034)	.83 (.033)
	30	.77 (.040)	.79 (.040)	.78 (.046)	.79 (.043)	.77 (.045)	.79 (.044)

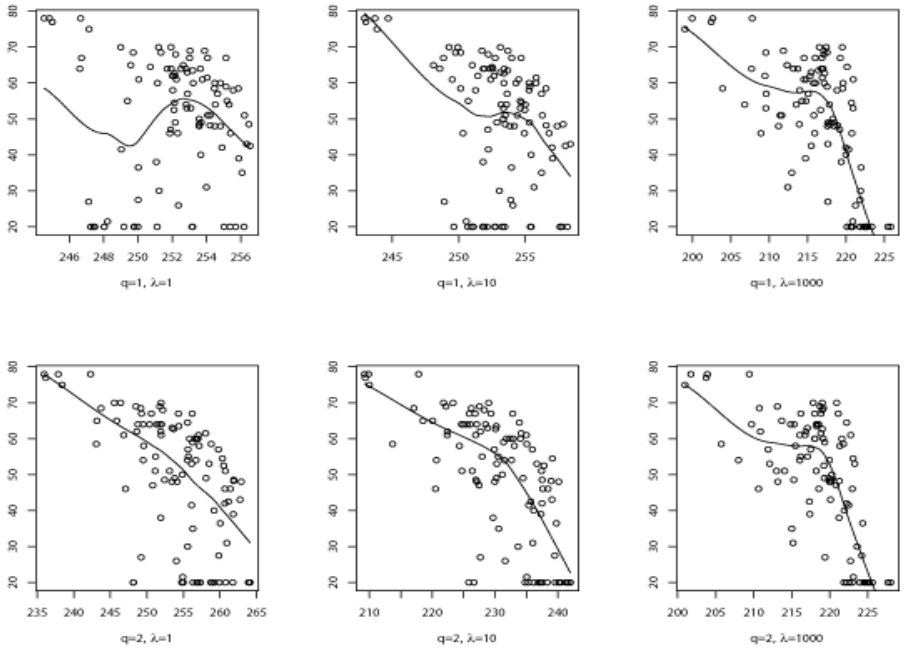

FIG 1. *Scatterplots of Y versus $\hat{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{X}$ across $\lambda = 1$ (first column), $\lambda = 10$ (second column), and $\lambda = 1000$ (third column), $q = 1$ (first row), and $q = 2$ (second row).*

7. Discussion

We propose the principal Lq support vector machine for sufficient dimension reduction. Compared with its L1 counterpart, the principal Lq support vector machine estimator is more robust to the choice of the misclassification parameter, and enjoys more accurate estimation of the central space.

In an effort to combine weighted support vector machine and sufficient dimension reduction, Shin et al. (2014) proposed probability-enhanced sufficient dimension reduction. The misclassification reweighted scheme for the principal L1 support vector machine was studied in Artemiou and Shu (2014). Development of weighted Lq support vector machines is worth exploration. Another open question is about the choice of the tuning parameter λ. The bootstrap method in Ye and Weiss (2003) could potentially be used to facilitate the selection of λ, and the theoretical justification of such procedures needs future investigation. Further to this, a limitation of our study comes with the lack of investigation of the role of λ in the theoretical framework. Another interesting question currently investigated by the authors is the use of equality instead of inequality in the constraint in (2). This leads to the Least Squares SVM (LSSVM) introduced by Suykens et al (2002). In the classification context, LSSVM give an analytic solution compared to the LqSVM which require quadratic programming but suffer in the sense that every point is considered a support vector. Whether similar advantages will hold in the dimension reduction framework is still under investigation.

Acknowledgements

We would like to thank the Editor, an Associate Editor and a reviewer, whose valuable comments help significantly improve the manuscript.

Appendix A

Proof of Theorem 1. The proof follows Theorem 1 of Li, Artemiou and Li (2011), with the key observation that $a \mapsto (a^+)^q$ is convex. We omit the details here. □

Proof of Proposition 1. The Lagrangian form of (7) is

$$L(\boldsymbol{\zeta}, t, \boldsymbol{\alpha}) = \boldsymbol{\zeta}^\tau \boldsymbol{\zeta} + \lambda (nq)^{-1} \mathbf{1}_n^\tau \boldsymbol{\xi}^q - \boldsymbol{\alpha}^\tau \{\boldsymbol{\xi} - \mathbf{1}_n + \tilde{\boldsymbol{Y}} \odot (\boldsymbol{Z}^\tau \boldsymbol{\zeta} - t)\} - \boldsymbol{\beta}^\tau \boldsymbol{\xi}. \tag{20}$$

Apply Kuhn-Tucker Theorem to (20) and we have

$$\begin{cases} \partial L/\partial \boldsymbol{\zeta} = 2\boldsymbol{\zeta} - \boldsymbol{Z}^\tau (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}) = \mathbf{0}_p, \\ \partial L/\partial \boldsymbol{\xi} = \lambda n^{-1} \boldsymbol{\xi}^{(q-1)} - \boldsymbol{\alpha} - \boldsymbol{\beta} = \mathbf{0}_n, \\ \partial L/\partial t = (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\tau \mathbf{1}_n = 0. \end{cases} \tag{21}$$

From the second equation of (21), we have $\boldsymbol{\xi} = (n\lambda^{-1}(\boldsymbol{\alpha} + \boldsymbol{\beta}))^{1/(q-1)}$. Next we use the Karush Kuhn Tucker (KKT) conditions to argue that $\boldsymbol{\beta} = \mathbf{0}$. The KKT conditions state that for the optimal hyperplane, we have $\beta_i \xi_i = 0$ and $\beta_i \geq 0$ for $i = 1, \ldots, n$. This implies that $\xi_i = 0$ whenever $\beta_i > 0$. Suppose we have $\beta_i > 0$, then the fact $\boldsymbol{\xi} = (n\lambda^{-1}(\boldsymbol{\alpha} + \boldsymbol{\beta}))^{1/(q-1)}$ and the KKT conditions together guarantee that $\alpha_i + \beta_i = 0$. It follows that $\alpha_i = -\beta_i < 0$. On the other hand, apply KKT conditions to the Lagrangian multiplier $\boldsymbol{\alpha}$ and we have $\alpha_i \geq 0$. This contradiction guarantees that $\beta_i = 0$, $i = 1, \ldots, n$.

As a result, we have $\boldsymbol{\xi} = (n\lambda^{-1}\boldsymbol{\alpha})^{1/(q-1)}$. From the first equation of (21), we get $\boldsymbol{\zeta} = \frac{1}{2}\boldsymbol{Z}^\tau(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})$.

Plug them into (20) and we have

$$L(\boldsymbol{\zeta}, t, \boldsymbol{\alpha}) = \frac{1}{4}(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal \boldsymbol{Z}\boldsymbol{Z}^\intercal (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}) + \lambda(nq)^{-1} \left(\frac{\boldsymbol{\alpha}^\intercal}{\lambda n^{-1}}\right)^{\frac{q}{q-1}} \mathbf{1}_n - \left(\frac{\boldsymbol{\alpha}^\intercal}{\lambda n^{-1}}\right)^{\frac{1}{q-1}} \boldsymbol{\alpha} + \boldsymbol{\alpha}^\intercal \mathbf{1}_n - \frac{1}{2}(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal \boldsymbol{Z}\boldsymbol{Z}^\intercal (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}) + (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal \mathbf{1}_n t.$$

Using the fact that $(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}})^\intercal \mathbf{1}_n = 0$ and

$$\lambda(nq)^{-1} \left(\frac{\boldsymbol{\alpha}^\intercal}{\lambda n^{-1}}\right)^{\frac{q}{q-1}} \mathbf{1}_n - \left(\frac{\boldsymbol{\alpha}^\intercal}{\lambda n^{-1}}\right)^{\frac{1}{q-1}} \boldsymbol{\alpha} = \frac{1-q}{q} \frac{1}{(\lambda n^{-1})^{\frac{1}{q-1}}} (\boldsymbol{\alpha}^\intercal)^{\frac{q}{q-1}} \mathbf{1}_n,$$

we get the desired result. □

Some preparation is needed before we prove Proposition 2. Recall that $\boldsymbol{W} = (\boldsymbol{X}^\intercal, \tilde{Y})^\intercal$. Denote $\boldsymbol{w} = (\boldsymbol{x}^\intercal, \tilde{y})^\intercal$ and define $N_{\boldsymbol{\theta}}(m) = \{\boldsymbol{w} : m(\cdot, \boldsymbol{w}) \text{ is not differentiable at } \boldsymbol{\theta}\}$. We state the following Lemma which is similar to a result in Li, Artemiou and Li (2011) without proof. Here we use local Lipschitz condition.

Lemma 1 *Let $\Theta \subset \mathbb{R}^{p+1}$ an open convex set. Suppose that $m : \Theta \times \Omega_{\boldsymbol{W}} \to \mathbb{R}$ satisfies the following conditions*

1. *(almost surely differentiable) For each $\boldsymbol{\theta} \in \Theta$, $P\{\boldsymbol{W} \in N_{\boldsymbol{\theta}}(m)\} = 0$;*
2. *(local Lipschitz condition) For any $\boldsymbol{\theta}_0 \in \Theta$ there is an integrable function $c(\boldsymbol{w})$, independent of $\boldsymbol{\theta}$, and a spherical neighborhood of $\boldsymbol{\theta}_0$, denoted as $A \subset \Theta$, such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in A$, $|m(\boldsymbol{\theta}_2, \boldsymbol{w}) - m(\boldsymbol{\theta}_1, \boldsymbol{w})| \le c(\boldsymbol{w})\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|$.*

Then $D_{\boldsymbol{\theta}}\{m(\boldsymbol{\theta}, \boldsymbol{W})\}$ is integrable, $E\{m(\boldsymbol{\theta}, \boldsymbol{W})\}$ is differentiable, and $D_{\boldsymbol{\theta}}[E\{m(\boldsymbol{\theta}, \boldsymbol{W})\}] = E[D_{\boldsymbol{\theta}}\{m(\boldsymbol{\theta}, \boldsymbol{W})\}]$.

Proof of Proposition 2. First we denote with $H(\boldsymbol{\psi}, a)$ the hyperplane $\{\boldsymbol{x} : \boldsymbol{\psi}^\intercal \boldsymbol{x} = a\}$ and we verify the conditions of Lemma 1 hold. Note that

$$P\{(\boldsymbol{X}, \tilde{Y}) \in N_{\boldsymbol{\theta}}(m)\} = \sum_{\tilde{y} \in \{-1,1\}} P(\tilde{Y} = \tilde{y}) P\{\boldsymbol{X} \in H(\boldsymbol{\psi}, t + \tilde{y}) | \tilde{Y} = \tilde{y}\}.$$

Since the Lebesgue measure of $H(\boldsymbol{\psi}, t + \tilde{y})$ is 0 for $\tilde{y} \in \{-1, 1\}$, the above probability is 0. Therefore condition 1 of Lemma 1 holds true.

Let $m(\boldsymbol{\theta}, \boldsymbol{w}) = m_1(\boldsymbol{\theta}, \boldsymbol{w}) + \lambda m_2(\boldsymbol{\theta}, \boldsymbol{w})$ where $m_1(\boldsymbol{\theta}, \boldsymbol{w}) = \boldsymbol{\psi}^\intercal \boldsymbol{\Sigma} \boldsymbol{\psi}$ and $m_2(\boldsymbol{\theta}, \boldsymbol{w}) = [\{1 - \tilde{Y}(\boldsymbol{\psi}^\intercal \boldsymbol{X} - t)\}^+]^q$. We need to show that m_1 and m_2 are locally Lipschitz. For m_1 this is obvious. To verify that m_2 is locally Lipschitz, let's take $\boldsymbol{\theta}_0$ any point in an open convex set $\Theta \in \mathbb{R}^{p+1}$ and let $A \subset \Theta$ be a spherical neighborhood around $\boldsymbol{\theta}_0$. Then, let $(\boldsymbol{\psi}_1, t_1), (\boldsymbol{\psi}_2, t_2) \in A$. Then

$$m_2(\boldsymbol{\theta}_2, \boldsymbol{x}, \tilde{y}) - m_2(\boldsymbol{\theta}_1, \boldsymbol{x}, \tilde{y}) = [\{1 - \tilde{y}(\boldsymbol{\psi}_2^\intercal \boldsymbol{x} - t_2)\}^+]^q - [\{1 - \tilde{y}(\boldsymbol{\psi}_1^\intercal \boldsymbol{x} - t_1)\}^+]^q. \tag{22}$$

Note that for numbers a and b, $|(b^+)^q - (a^+)^q| \le q(\max\{b^+, a^+\})^{q-1}|(b^+ - a^+)|$ and $|b^+ - a^+| \le |b - a|$. Since $q\left(\{1 - \tilde{y}(\boldsymbol{\psi}^\intercal \boldsymbol{x} - t)\}^+\right)^{q-1}$ is convex it is also locally bounded in A, let's say by $M(\boldsymbol{x}) < m\|\boldsymbol{x}\|^{q-1}$ where m a constant. Therefore (22) together with the other assumptions of Proposition 2 implies

$$\begin{aligned} |m_2(\boldsymbol{\theta}_2, \boldsymbol{x}, \tilde{y}) - m_2(\boldsymbol{\theta}_1, \boldsymbol{x}, \tilde{y})| &\le M(\boldsymbol{x}) \left|\{1 - \tilde{y}(\boldsymbol{\psi}_2^\intercal \boldsymbol{x} - t_2)\}^+ - \{1 - \tilde{y}(\boldsymbol{\psi}_1^\intercal \boldsymbol{x} - t_1)\}^+\right| \\ &\le M(\boldsymbol{x}) |\boldsymbol{\psi}_1^\intercal \boldsymbol{x} - \boldsymbol{\psi}_2^\intercal \boldsymbol{x} + t_2 - t_1| \\ &\le m\|\boldsymbol{x}\|^{q-1} |\boldsymbol{\psi}_1^\intercal \boldsymbol{x} - \boldsymbol{\psi}_2^\intercal \boldsymbol{x} + t_2 - t_1|. \end{aligned} \tag{23}$$

From the definition that $\boldsymbol{\theta} = (\boldsymbol{\psi}^\intercal, t)^\intercal$, we have $|\boldsymbol{\psi}_1^\intercal \boldsymbol{x} - \boldsymbol{\psi}_2^\intercal \boldsymbol{x} + t_2 - t_1| \leq (1 + \|\boldsymbol{x}\|^2)^{\frac{1}{2}} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|$. Plug it into (23) and we get

$$|m_2(\boldsymbol{\theta}_2, \boldsymbol{x}, \tilde{y}) - m_2(\boldsymbol{\theta}_1, \boldsymbol{x}, \tilde{y})| \leq m\|\boldsymbol{x}\|^{q-1}(1 + \|\boldsymbol{x}\|^2)^{\frac{1}{2}} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|. \tag{24}$$

Because $E(\|\boldsymbol{X}\|^2) < \infty$, $E(1 + \|\boldsymbol{X}\|^2)^{\frac{1}{2}} \leq [1 + E(\|\boldsymbol{X}\|^2)]^{\frac{1}{2}} < \infty$ and from $E(\|\boldsymbol{X}\|^{q-1}) < \infty$ we have that condition 2 of Lemma 1 is implied by (24).

Now that the two conditions of Lemma 1 are verified, for $\boldsymbol{w} \notin N_{\boldsymbol{\theta}}(m)$, take derivatives of $m(\boldsymbol{\theta}, \boldsymbol{w})$ and we get $D_{\boldsymbol{\psi}}\{m(\boldsymbol{\theta}, \boldsymbol{w})\} = 2\boldsymbol{\Sigma}\boldsymbol{\psi} - q\lambda^\dagger \boldsymbol{x}\tilde{y}\{1 - \tilde{y}(\boldsymbol{\psi}^\intercal \boldsymbol{x} - t)\}^{q-1}[I\{1 - \tilde{y}(\boldsymbol{\psi}^\intercal \boldsymbol{x} - t) > 0\}]$, $D_t\{m(\boldsymbol{\theta}, \boldsymbol{w})\} = q\lambda^\dagger \tilde{y}\{1 - \tilde{y}(\boldsymbol{\psi}^\intercal \boldsymbol{x} - t)\}^{q-1}[I\{1 - \tilde{y}(\boldsymbol{\psi}^\intercal \boldsymbol{x} - t) > 0\}]$. Thus we have

$$D_{\boldsymbol{\theta}}\{m(\boldsymbol{\theta}, \boldsymbol{w})\} = (2\boldsymbol{\psi}^\intercal \boldsymbol{\Sigma}, 0)^\intercal - q\lambda^\dagger \boldsymbol{x}^\dagger \tilde{y}\{1 - \tilde{y}(\boldsymbol{\psi}^\intercal \boldsymbol{x} - t)\}^{q-1} I\{1 - \tilde{y}(\boldsymbol{\psi}^\intercal \boldsymbol{x} - t) > 0)\}. \tag{25}$$

Take expectation of (25). Apply Lemma 1 to get the desired result. □

To compute the derivative of expectation of a non-Lipschitz function, two additional Lemmas are needed before we prove Proposition 3. The first one is true if U and V are linearly independent and the second covers the case when they are linearly dependent. Let $D_{\epsilon=0}$ denote the operation of first taking derivative with respect to ϵ and then evaluating the derivative at $\epsilon = 0$.

Lemma 2 *Let U and V be random variables, $\boldsymbol{h}(u, v)$ be a measurable $\mathbb{R}^k$-valued function, and b be a constant. Suppose:*

1. *the joint distribution of (U, V) is dominated by the Lebesgue measure;*
2. *for each v, the function $u \mapsto (b - u + \epsilon(\eta - v))^{q-1}\boldsymbol{h}(u, v)f_{U|V}(u|v)$ is continuous, where $f_{U|V}$ denotes the conditional probability density function of U given V;*
3. *for each component $h_i(u, v)$ of $\boldsymbol{h}(u, v)$, there is a function $c_i(v) \geq 0$ such that*

$$|(b - u + \epsilon(\eta - v))^{q-1} h_i(u, v)|f_{U|V}(u|v) \leq c_i(v), \quad E\{c_i(V)\} < \infty.$$

Then, for any constant a, the function $\epsilon \mapsto E\{(b - U + \epsilon(\eta - V))^{q-1}\boldsymbol{h}(U, V)I(U + \epsilon V < a + \epsilon\eta)\}$ is differentiable at $\epsilon = 0$ with derivative

$$\begin{aligned} &D_{\epsilon=0}[E\{(b - U + \epsilon(\eta - V))^{q-1}\boldsymbol{h}(U, V)I(U + \epsilon V < a + \epsilon\eta)\}] \\ &\quad = f_U(a)E\{(\eta - V)(b - a)^{q-1}\boldsymbol{h}(U, V)|U = a\} - E\{(\eta - V)(b - U)^{q-2}h_i(U, V)I(U < a)\}. \end{aligned} \tag{26}$$

Proof. We need to show that, for each $i = 1, \ldots, k$, the limit

$$\lim_{\epsilon \to 0} \int \left\{ \epsilon^{-1} \int_a^{a+\epsilon(\eta - v)} (b - u + \epsilon(\eta - v))^{q-1} h_i(u, v) f_{U|V}(u|v) du \right\} f_V(v) dv \tag{27}$$

exists. By the mean value theorem for integration, there exists $\xi \in (0, \epsilon)$ such that

$$\begin{aligned} &\left| \epsilon^{-1} \int_a^{a+\epsilon(\eta - v)} (b - u + \epsilon(\eta - v))^{q-1} h_i(u, v) f_{U|V}(u|v) du \right| \\ =&|(b - \xi(\eta - v) - a + \epsilon(\eta - v))^{q-1}(h_i(a + \xi(\eta - v), v) f_{U|V}(a + \xi(\eta - v)|v)| \leq c(v), \end{aligned}$$

where the inequality follows from assumptions 2 and 3. By the dominated convergence theorem, the limit in (27) becomes

$$\int \lim_{\epsilon\to 0}\left\{\epsilon^{-1}\int_a^{a+\epsilon(\eta-v)}(b-u+\epsilon(\eta-v))^{q-1}h_i(u,v)f_{U|V}(u|v)du\right\}f_V(v)dv. \tag{28}$$

Apply the generalized Leibniz integral rule and (28) becomes

$$\begin{aligned}&\int(\eta-v)(b-a)^{q-1}h_i(a,v)f_{U|V}(a|v)duf_V(v)dv-(q-1)\int\int_a^{a+\epsilon(\eta-v)}(\eta-v)(b-u)^{q-2}h_i(u,v)f_{U|V}(u|v)duf_V(v)dv\\ &\quad=f_U(a)\int(\eta-v)(b-a)^{q-1}h_i(a,v)f_{V|U}(v|a)dv-(q-1)E\{(\eta-V)(b-U)^{q-2}h_i(U,V)I(U<a)\}\\ &\quad=f_U(a)E\{(\eta-V)(b-a)^{q-1}h_i(a,V)|U=a\}-(q-1)E\{(\eta-V)(b-U)^{q-2}h_i(U,V)I(U<a)\},\end{aligned}$$

and we get the desired result. □

Lemma 3 *Let U and V be linearly dependent random variables and $\boldsymbol{h}(u)$ be a measurable $\mathbb{R}^k$-valued function. Suppose*

1. *the distribution of U is dominated by the Lebesgue measure;*
2. *$(b-u+\epsilon(\eta-v))^{q-1}\boldsymbol{h}(u)f_U(u)$ is continuous.*

Then, for any constant a, the function $\epsilon\mapsto E\{(b-U+\epsilon(\eta-V))^{q-1}\boldsymbol{h}(U,V)I(U+\epsilon V<a+\epsilon\eta)\}$ is differentiable at $\epsilon=0$ with derivative equal to (26).

Proof. Suppose, without loss of generality, $V=\kappa U$ for some $\kappa>0$. We have

$$\begin{aligned}&E\{(b-U+\epsilon(\eta-V))^{q-1}h_i(U)I(U+\epsilon V<a+\epsilon\,\eta)\}\\ &\quad=\int_{-\infty}^{(a+\epsilon\eta)/(1+\epsilon\kappa)}E\{(b-(1+\epsilon\kappa)U)^{q-1}h_i(U)|U=u\}f_U(u)du.\end{aligned}$$

The generalized Leibniz integral rule leads to:

$$\begin{aligned}&D_{\epsilon=0}E\{(b-U+\epsilon(\eta-V))^{q-1}h_i(U)I(U+\epsilon V<a+\epsilon\eta)\}\\ &\quad=(\eta-\kappa a)E\{(b-U)^{q-1}h_i(U)|U=a\}f_U(a)-(q-1)E\{(\eta-kU)(b-U)^{q-2}h_i(U)I(U<a)\}.\end{aligned}$$

Under the condition that $U=a$ and $V=kU$, we have the desired result. □

Proof of Proposition 3. The first term of $D_{\boldsymbol{\theta}}[E\{m(\boldsymbol{\theta},\boldsymbol{W})\}]$, or $(2\boldsymbol{\psi}^\intercal\boldsymbol{\Sigma},0)^\intercal$, is jointly differentiable with derivative $2\mathrm{diag}(\boldsymbol{\Sigma},0)$. We focus on the second term of $D_{\boldsymbol{\theta}}[E\{m(\boldsymbol{\theta},\boldsymbol{W})\}]$.

Rewrite $E\{\boldsymbol{X}^\dagger\tilde{Y}(1-\boldsymbol{\theta}^\intercal\boldsymbol{X}^\dagger\tilde{Y})^{q-1}I(1-\boldsymbol{\theta}^\intercal\boldsymbol{X}^\dagger\tilde{Y}>0)\}$ as

$$\sum_{\tilde{y}=-1,1}P(\tilde{Y}=\tilde{y})E\{\boldsymbol{X}^\dagger\tilde{y}(1-\boldsymbol{\theta}^\intercal\boldsymbol{X}^\dagger\tilde{y})^{q-1}I(1-\boldsymbol{\theta}^\intercal\boldsymbol{X}^\dagger\tilde{Y}>0)|\tilde{Y}=\tilde{y}\}.$$

We first consider the case $\tilde{y}=1$ and verify directional differentiability of the function $(\boldsymbol{\psi},t)\mapsto E\{\boldsymbol{X}^\dagger(1+t-\boldsymbol{\psi}^\intercal\boldsymbol{X})^{q-1}I(\boldsymbol{\psi}^\intercal\boldsymbol{X}<t+1)|\tilde{Y}=1\}$. To do this we define $\boldsymbol{\psi}$ and $\boldsymbol{\delta}$ to be linearly

independent vectors in $\mathbb{R}^p$. Let η be a number. The directional derivative along $(\boldsymbol{\delta}^\intercal, \eta)^\intercal$ is the derivative of the following function with respect to ϵ at $\epsilon = 0$:

$$\begin{aligned}&E\{\boldsymbol{X}^\dagger(1+t-\boldsymbol{\psi}^\intercal\boldsymbol{X}+\epsilon(\eta-\boldsymbol{\delta}^\intercal\boldsymbol{X}))^{q-1}I(\boldsymbol{\psi}^\intercal\boldsymbol{X}+\epsilon\boldsymbol{\delta}^\intercal\boldsymbol{X}<t+1+\epsilon\eta)|\tilde{Y}=1\}\\ =&E\{(1+t-\boldsymbol{\psi}^\intercal\boldsymbol{X}+\epsilon(\eta-\boldsymbol{\delta}^\intercal\boldsymbol{X}))^{q-1}E(\boldsymbol{X}^\dagger|\boldsymbol{\psi}^\intercal\boldsymbol{X},\boldsymbol{\delta}^\intercal\boldsymbol{X},\tilde{Y}=1)I(\boldsymbol{\psi}^\intercal\boldsymbol{X}+\epsilon\boldsymbol{\delta}^\intercal\boldsymbol{X}<t+1+\epsilon\,\eta)|\tilde{Y}=1\}.\end{aligned}$$

Let $U=\boldsymbol{\psi}^\intercal\boldsymbol{X}$, $V=\boldsymbol{\delta}^\intercal\boldsymbol{X}$, $\boldsymbol{h}(U,V)=E(\boldsymbol{X}^\dagger|U,V,\tilde{Y}=1)$. Apply Lemma 2 to $P(\cdot\,|\tilde{Y}=1)$ with $b=a=t+1$. The derivative of the equation above becomes

$$\begin{aligned}&f_{\boldsymbol{\psi}^\intercal\boldsymbol{X}|\tilde{Y}}(t+1|1)E\{(\eta-V)((t+1)-(t+1))^{q-1}E(\boldsymbol{X}^\dagger|U,V,\tilde{Y}=1)|U=t+1,\tilde{Y}=1\}\\ &-(q-1)E\{(\eta-V)(t+1-\boldsymbol{\psi}^\intercal\boldsymbol{X})^{q-2}h_i(U,V)I(U<a)|\tilde{Y}=1\}.\end{aligned}\tag{29}$$

The first term is equal to 0. Since this holds for all $(\boldsymbol{\delta}^\intercal,\eta)^\intercal$, the function $(\boldsymbol{\psi},t)\mapsto E\{\boldsymbol{X}^\dagger(1+t-\boldsymbol{\psi}^\intercal\boldsymbol{X})^{q-1}I(\boldsymbol{\psi}^\intercal\boldsymbol{X}<t+1)|\tilde{Y}=1\}$ is directionally differentiable with derivative equal to $-(q-1)E\{(1-\boldsymbol{\theta}^\intercal\boldsymbol{X}^\dagger)^{q-2}\boldsymbol{X}^\dagger(\boldsymbol{X}^\dagger)^\intercal I(1-\boldsymbol{\theta}^\intercal\boldsymbol{X}^\dagger>0)|\tilde{Y}=1\}$.

If $\boldsymbol{\delta}$ and $\boldsymbol{\psi}$ are linearly dependent vectors in $\mathbb{R}^p$, then $\boldsymbol{\psi}^\intercal\boldsymbol{X}$ and $\boldsymbol{\psi}^\intercal\boldsymbol{X}$ are linearly dependent random variables. We apply Lemma 3 in the similar fashion to arrive at the same directional derivative. The case for $\tilde{y}=-1$ can be proved similarly. Hence the directional derivative of $D_{\boldsymbol{\theta}}[E\{m(\boldsymbol{\theta},\boldsymbol{W})\}]$ is given by equation (14). □

Proof of Theorem 2. The proof is similar to Jiang, Zhang and Cai (2008), and is thus omitted. A different approach to this can be seen also in Koo et al (2008). □

Proof of Theorem 3. The proof is parallel to Theorem 7 in Li, Artemiou and Li (2011), and is thus omitted. □

Proof of Theorem 4. The proof is similar to Theorem 8 in Li, Artemiou and Li (2011), and is thus omitted. □

Proof of Proposition 4. This is parallel to proof of Proposition 1, and is thus omitted. □

Appendix B

With $q=1$, $\Lambda(\boldsymbol{\psi},t)$ in (5) becomes

$$\Lambda(\boldsymbol{\psi},t)=\boldsymbol{\psi}^\intercal\boldsymbol{\Sigma}\boldsymbol{\psi}+\lambda E[1-\tilde{Y}\{\boldsymbol{\psi}^\intercal(\boldsymbol{X}-E\boldsymbol{X})-t\}]^+.\tag{30}$$

We now prove that under specific circumstances, $\Lambda(\boldsymbol{\psi},t)$ in (30) does not have a unique minimizer. First let's define the following disjoint sets: $\mathcal{I}_1=\{(\boldsymbol{X},\tilde{Y}):\tilde{Y}=1,\boldsymbol{\psi}^\intercal\boldsymbol{X}-t>1\}$, $\mathcal{I}_2=\{(\boldsymbol{X},\tilde{Y}):\tilde{Y}=1,\boldsymbol{\psi}^\intercal\boldsymbol{X}-t=1\}$, $\mathcal{I}_3=\{(\boldsymbol{X},\tilde{Y}):\tilde{Y}=1,\boldsymbol{\psi}^\intercal\boldsymbol{X}-t<1\}$, $\mathcal{I}_4=\{(\boldsymbol{X},\tilde{Y}):\tilde{Y}=-1,\boldsymbol{\psi}^\intercal\boldsymbol{X}-t<-1\}$, $\mathcal{I}_5=\{(\boldsymbol{X},\tilde{Y}):\tilde{Y}=-1,\boldsymbol{\psi}^\intercal\boldsymbol{X}-t=-1\}$, and $\mathcal{I}_6=\{(\boldsymbol{X},\tilde{Y}):\tilde{Y}=-1,\boldsymbol{\psi}^\intercal\boldsymbol{X}-t>-1\}$, which essentially cover all possible cases where a data point might fall. It can have either a positive or a negative $\tilde{Y}$. If it is positive(negative) it divides the sets in $\mathcal{I}_1(\mathcal{I}_4)$ which implies correct classification, or in $\mathcal{I}_2$ ($\mathcal{I}_5$) which implies correct classification of a data point on a support vector, or in $\mathcal{I}_3$ ($\mathcal{I}_6$) which implies incorrect classification.

Denote indicator function $I_j=I\{(\boldsymbol{X},\tilde{Y})\in\mathcal{I}_j\}$ and define $P_j=P(I_j)$. Assume there exists a unique minimizer $\boldsymbol{\psi}$. Since the first term in (30) is not affected by the value of t we ignore it in this development. Assume $E(\boldsymbol{X})=0$ without loss of generality. We focus on the the second term of (30), which is equal to $\sum_{i=1}^6\lambda P_iE[\{1-(\boldsymbol{\psi}^\intercal\boldsymbol{X}-t)\}^+|(\boldsymbol{X},\tilde{Y})\in\mathcal{I}_i]$. Use the fact that $\{1-\tilde{Y}(\boldsymbol{\psi}^\intercal\boldsymbol{X}-t)\}^+=0$

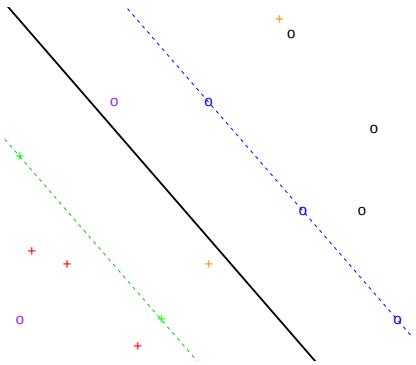

Figure 2. All circles correspond to $\tilde{Y} = 1$ and all crosses correspond to $\tilde{Y} = -1$. The black circles, the blue circles and the purple circles belong to $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$ respectively. The red crosses, the green crosses and the orange crosses belong to $\mathcal{I}_4$, $\mathcal{I}_5$ and $\mathcal{I}_6$ respectively. The dashed blue line, the solid black line and the dashed green line correspond to $\boldsymbol{\psi}^\intercal \boldsymbol{X} - t = 1$, $\boldsymbol{\psi}^\intercal \boldsymbol{X} - t = 0$ and $\boldsymbol{\psi}^\intercal \boldsymbol{X} - t = -1$ respectively.

for $(\boldsymbol{X}, \tilde{Y}) \in \{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_4, \mathcal{I}_5\}$, and $\{1 - \tilde{Y}(\boldsymbol{\psi}^\intercal \boldsymbol{X} - t)\}^+ = 1 - \tilde{Y}(\boldsymbol{\psi}^\intercal \boldsymbol{X} - t)$ for $(\boldsymbol{X}, \tilde{Y}) \in \{\mathcal{I}_3, \mathcal{I}_6\}$. The second term becomes

$$\lambda P_3 E\{1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) | (\boldsymbol{X}, \tilde{Y}) \in \mathcal{I}_3\} + \lambda P_6 E\{1 + (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) | (\boldsymbol{X}, \tilde{Y}) \in \mathcal{I}_6\}. \tag{31}$$

Now define $s = \min\{s_1, s_2\}$, where $s_1 = \min\{1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t)$ for $(\boldsymbol{X}, \tilde{Y}) \in \mathcal{I}_3\}$, and $s_2 = \min\{-1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t)$ for $(\boldsymbol{X}, \tilde{Y}) \in \mathcal{I}_4\}$. According to Figure 1, the value of s will be either the minimum distance between the purple circles to the blue dash line, or the minimum distance of the red crosses to the green dash line. Instead of the original separating hyperplane $\boldsymbol{\psi}^\intercal \boldsymbol{X} - t = 0$, we now consider the new hyperplane $\boldsymbol{\psi}^\intercal \boldsymbol{X} - t' = 0$, where $t' = t - s$. Note that $s > 0$, $1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t') = 1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) - s$, and $1 + (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t') = 1 + (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) + s$. With the new separating hyperplane, we observe

1. All the points that were in $\mathcal{I}_1$ satisfies $1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) < 0$ and $\tilde{Y} = 1$. Thus $1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t') < 0$, and these points will still be correctly classified.
2. All the points that were in $\mathcal{I}_2$ satisfies $1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) = 0$ and $\tilde{Y} = 1$. Thus $1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t') < 0$, and these points will still be correctly classified.
3. All the points that were in $\mathcal{I}_3$ satisfies $1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) > 0$ and $\tilde{Y} = 1$. Because $s \leq s_1 = \min\{1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t)$ for $(\boldsymbol{X}, \tilde{Y}) \in \mathcal{I}_3\}$, $1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t') \geq 1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) - s_1 \geq 0$. These points will now either continue to be incorrectly classified or become correctly classified as a point on the support vector. The latter happens if $1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) = s$.
4. All the points that were in $\mathcal{I}_4$ satisfies $1 + (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) < 0$ and $\tilde{Y} = -1$. Because $s \leq s_2 = \min\{-1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t)$ for $(\boldsymbol{X}, \tilde{Y}) \in \mathcal{I}_4\}$, we have $1 + (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t') \leq 1 + (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) + s_2 \leq 0$. These points will either continue to be correctly classified as non-support points or become correctly classified as a point on the support vector. The latter happens if $-1 - (\boldsymbol{\psi}^\intercal \boldsymbol{X} - t) = s$.

5. All the points that were in $\mathcal{I}_5$ satisfies $1+(\boldsymbol{\psi}^\top \boldsymbol{X}-t)=0$ and $\tilde{Y}=-1$. Thus $1+(\boldsymbol{\psi}^\top \boldsymbol{X}-t')>0$, and these points will become incorrectly classified.
6. All the points that were in $\mathcal{I}_6$ satisfies $1+(\boldsymbol{\psi}^\top \boldsymbol{X}-t)>0$ and $\tilde{Y}=-1$. Thus $1+(\boldsymbol{\psi}^\top \boldsymbol{X}-t')>0$, and these points will continue to be incorrectly classified.

Note that for $(\boldsymbol{X},\tilde{Y})\in\{\mathcal{I}_1,\mathcal{I}_2,\mathcal{I}_4\}$, we have $\{1-\tilde{Y}(\boldsymbol{\psi}^\top \boldsymbol{X}-t')\}^+=0$. With t replaced by t', the second term of (30) becomes $\lambda P_3 E[\{1-\tilde{Y}(\boldsymbol{\psi}^\top \boldsymbol{X}-t')\}^+|(\boldsymbol{X},\tilde{Y})\in\mathcal{I}_3]+\lambda P_5 E[\{1-\tilde{Y}(\boldsymbol{\psi}^\top \boldsymbol{X}-t')\}^+|(\boldsymbol{X},\tilde{Y})\in\mathcal{I}_5]+\lambda P_6 E[\{1-\tilde{Y}(\boldsymbol{\psi}^\top \boldsymbol{X}-t')\}^+|(\boldsymbol{X},\tilde{Y})\in\mathcal{I}_6]$. After some algebra, it becomes

$$\begin{aligned}&\lambda P_3 E\{1-(\boldsymbol{\psi}^\top \boldsymbol{X}-t)|(\boldsymbol{X},\tilde{Y})\in\mathcal{I}_3\}+\lambda P_6 E\{1+(\boldsymbol{\psi}^\top \boldsymbol{X}-t)|(\boldsymbol{X},\tilde{Y})\in\mathcal{I}_6\}\\&\quad-\lambda s(P_3-P_5-P_6).\end{aligned}\tag{32}$$

Subtract (32) from (31) and we get $\lambda s(P_3-P_5-P_6)$, which is 0 if $P_3=P_5+P_6$.

For $\Lambda(\boldsymbol{\psi},t)$ defined in (30), we have shown that it is possible to have $\Lambda(\boldsymbol{\psi},t)=\Lambda(\boldsymbol{\psi},t')$ for some $t\neq t'$. If we define $t'=t-s$ with $s=\min\{s_3,s_4\}$, where $s_3=\min\{1+(\boldsymbol{\psi}^\top \boldsymbol{X}-t)$ for $(\boldsymbol{X},\tilde{Y})\in\mathcal{I}_6\}$ and $s_4=\min\{-1+(\boldsymbol{\psi}^\top \boldsymbol{X}-t)$ for $(\boldsymbol{X},\tilde{Y})\in\mathcal{I}_1\}$. Following similar arguments, we can show that $\Lambda(\boldsymbol{\psi},t)-\Lambda(\boldsymbol{\psi},t')=0$ if $P_6=P_2+P_3$. Also note that $\Lambda(\boldsymbol{\psi},t)$ has unique minimizer $\boldsymbol{\psi}_0$ according to Theorem 1 in Li, Artemiou and Li (2011).

Finally, for the development described in this section to hold we need to note that there is an underlying assumption on the two conditional distributions of $\boldsymbol{X}|\tilde{Y}$. For example, if $\boldsymbol{X}|\tilde{Y}$ has a continuous distribution on the whole real p dimensional space then it is obvious that $P_2=P_5=0$ and also $s=0$ which eliminates the issue of multiple solutions to the objective function in the population level. This is a rather technical assumption that is not very strong in the sense that when we apply SVM on a dataset we are using the empirical distribution, which is discrete and usually not dense enough so that we can always find a value $\delta>0$ such that $s>\delta$. For the theorem to be true at the population level we have to add the necessary condition to ensure that $s>0$. We summarize these findings in the next result.

Theorem 5 *Denote $\boldsymbol{\psi}_0$ as the unique vector that minimizes $\Lambda(\boldsymbol{\psi},t)$ in (30). Let t_0 be the value of t that minimize $\Lambda(\boldsymbol{\psi},t)$. Then either of the following two is a sufficient condition for the non-unique value of t_0.*

1. *$P_3=P_5+P_6$ and there exists $\delta>0$ such that $P(1-\delta<\boldsymbol{\psi}_0^\top \boldsymbol{X}-t_0<1|\tilde{Y}=1)=0$ and $P(-1-\delta<\boldsymbol{\psi}_0^\top \boldsymbol{X}-t_0<-1|\tilde{Y}=-1)=0$.*
2. *$P_6=P_2+P_3$ and there exists $\delta>0$ such that $P(1<\boldsymbol{\psi}_0^\top \boldsymbol{X}-t_0<1+\delta|\tilde{Y}=1)=0$ and $P(-1<\boldsymbol{\psi}_0^\top \boldsymbol{X}-t_0<-1+\delta|\tilde{Y}=-1)=0$.*

References

[1] Abe S. (2002). Analysis of support vector machines. In *Neural Networks for Signal Processing XII - Proceedings of the 2002 IEEE Signal Processing Society Workshops*, 89–98.

[2] Artemiou, A. and Shu, M. (2014). A cost based reweighed scheme of principal support vector machine. In *Topics in Nonparametric Statistics, Springer Proceedings in Mathematics & Statistics*, **74**, 1–12.

[3] Bura, E. and Pfeiffer, R. (2008). On the distribution of the left singular vectors of a random matrix and its applications. *Statistics and Probability Letters*, **78**, 2275–2280.

[4] Burges, C. J. C. and Crisp, S. J. (1999). Uniqueness of the SVM solution. In *Proceedings of Neural Information Processing Systems*, **12**, 223–229.
[5] Cook, R. D. (1998a). *Regression Graphics: Ideas for Studying Regressions through Graphics.* New York: Wiley.
[6] Cook, R. D. (1998b). Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association*, **93**, 84–100.
[7] Cook, R. D. (2004). Testing predictors contributions in sufficient dimension reduction. *The Annals of Statistics*, **32**, 1062–1092.
[8] Cook, R. D. (2007). Fisher lecture: dimension reduction in regression. *Statistical Science*, **22**, 1–40.
[9] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association.* **86**, 316–342.
[10] Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 1–25.
[11] Fukumizu, Bach, and Jordan (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, **37**, 1871 – 1905.
[12] Jiang, B., Zhang, X., and Cai, T. (2008). Estimating the confidence interval for prediction errors of support vector machine classifiers. *Journal of Machine Learning Research*, **9**, 521 –540.
[13] Koo, J.-Y., Lee, Y., Kim, Y. and Park, C. (2008). A Bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, **9**, 1343–1368.
[14] Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machine for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, **39**, 3182–3210
[15] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**, 997–1008.
[16] Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, **33**, 1580–1616.
[17] Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316–342.
[18] Shin, S. J., Wu, Y., Zhang, H. H. and Liu, Y. (2014). Probability-enhanced sufficient dimension reduction for binary classification. *Biometrics*, **70**, 546-555.
[19] Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D. and Vandewalle, J. (2002). Least Squares Support Vector Machines. World Scientific Pub. Co., Singapore.
[20] Vapnik, N. V. (1998). *Statistical Learning Theory.* John Wiley & Sons, Inc.
[21] Wang, Q. and Yin, X. (2008). Sufficient dimension reduction and variable selection for regression mean function with categorical predictors. *Statistics and Probability Letters*, **78**, 2798–2803.
[22] Wu, H. M. (2008). Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics*, **17**, 590–610.
[23] Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of optimal regression subspace. *Journal of the Royal Statistical Society, Series B.*, **64**, 363–410.
[24] Ye, Z. and Weiss, R. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, **98**, 968–979.
[25] Yeh, I. C. (2007). Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, **29**, 474–480.
[26] Yeh, Y. R., Huang, S. Y. and Lee, Y. Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1590–1603.
[27] Yin, X., Li, B. and Cook, R.D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis.* **99**, 1733–1757.

[28] Zhu, L. X., Miao, B. and Peng, H. (2006). On sliced inverse regression with large dimensional covariates. *Journal of the American Statistical Association*, **101**, 630–643.