Received: 15 May 2019 | Revised: 15 June 2020 | Accepted: 13 July 2020

DOI: 10.1111/sjos.12484

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Sufficient dimension reduction based on distance-weighted discrimination

Hayley Randall[1] | Andreas Artemiou[1] | Xingye Qiao[2]

[1]School of Mathematics, Cardiff University, UK

[2]Department of Mathematical Sciences, Binghamton University, UK

Correspondence
Andreas Artemiou, School of Mathematics, Cardiff University, Senghennydd Road, Cardiff CF24 4AG Wales, UK.
Email: ArtemiouA@cardiff.ac.uk

Funding information
Engineering and Physical Sciences Research Council

Abstract

In this paper, we introduce a sufficient dimension reduction (SDR) algorithm based on distance-weighted discrimination (DWD). Our methods is shown to be robust on the dimension p of the predictors in our problem, and it also utilizes some new computational results in the DWD literature to propose a computationally faster algorithm than previous classification-based algorithms in the SDR literature. In addition to the theoretical results of similar methods we prove the consistency of our estimate for divergent number of p. Finally, we demonstrate the advantages of our algorithm using simulated and real datasets.

KEYWORDS

dimension reduction, feature extraction, classifiers

1 | INTRODUCTION

Sufficient dimension reduction (SDR) is a class of feature extraction techniques introduced in regression settings with high-dimensional predictors. Let $\boldsymbol{X}$ be a p-dimensional predictor vector and Y be a response variable (which is assumed univariate for the time being). In linear SDR our effort is to reduce the dimension of the predictors, $\boldsymbol{X}$, without losing information of the conditional distribution $Y|\boldsymbol{X}$. In other words we are trying to find a $p \times d$ ($d < p$) matrix $\boldsymbol{\beta}$ such that the following conditional independence model holds:

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

The space spanned by the columns of $\boldsymbol{\beta}$ is called a Dimension Reduction Subspace (DRS). The intersection of all possible DRS's, if it is itself a DRS, is called the Central Dimension Reduction Subspace (CDRS) or the Central Space (CS) and is denoted with $\mathcal{S}_{Y|\boldsymbol{X}}$. Conditions of existence of the CS are mild (see Yin, Li, & Cook, 2008) therefore we assume it exists throughout this paper. Some literature on SDR includes and is not limited to Sliced Inverse Regression (SIR) by Li (1991), Sliced Average Variance Estimation (SAVE) by Cook and Weisberg (1991), principal Hessian directions (pHd) by Li (1992), Contour Regression by Li, Zha, and Chiaromonte (2005) and Directional Regression by Li and Wang (2007) among others.

In recent years, there is an interest in nonlinear SDR, where we extract linear or nonlinear functions of the predictors. That is, we work under the nonlinear conditional independence model:

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

where $\phi : \mathbb{R}^p \to \mathbb{R}^d$ denotes linear or nonlinear functions of the predictors. Some examples include the work by Wu (2008) and by Yeh, Huang, and Lee (2009) which introduced Kernel SIR, the work by Fukumizu, Bach, and Jordan (2009) which used kernel regression and the work by Li, Artemiou, and Li (2011) who used Support Vector Machine (SVM) algorithms to achieve linear and nonlinear dimension reduction under a unified framework. The idea of using SVM and different algorithms have since been expanded in a number of directions. Artemiou and Dong (2016) used LqSVM which ensures the uniqueness of the solution, Zhou and Zhu (2016) used a minimax variation for sparse SDR, Shin and Artemiou (2017) replaced the hinge loss with a logistic loss to achieve the desired result, Shin, Wu, Zhang, and Liu (2017) used weighted SVM approach for binary responses and Artemiou and Shu (2014) and Smallman and Artemiou (2017) focused on removing the bias due to imbalance.

One of the most interesting variations of SVM was proposed by Marron, Todd, and Ahn (2007) and is known as Distance-Weighted Discrimination (DWD). The interest of DWD lies on the fact that it works much better than SVM as the dimension of the predictors $\boldsymbol{X}$ increases. This is due to the fact that SVM suffers from data piling when the dimension of the predictor space is large. Data piling occurs in high-dimensional low sample size (HDLSS) settings and it describes the tendency of the data in each class to project to a single point on a discriminant direction. In the upper plot of Figure 1, we can see that for $p = 100$, SVM and DWD project the data in a similar manner. However, the lower plot in Figure 1 shows that when $p = 500$, SVM (black points) projects almost all the points to 1 for one class and to −1 for the other class while DWD spreads the data in each class on a wider range of values. In a classification setting, data piling makes generalization of the classification results difficult.

In this paper we will investigate whether DWD has similar advantages over SVM in the SDR framework, as the ones it has in the classification framework. We will create a similar method as the one in Li et al. (2011) with the difference that the objective function of DWD will replace the objective function of SVM. We call our method Principal DWD following a similar pattern to Li et al. (2011) calling their method Principal SVM (principal support vector machines (PSVM)). Interestingly, results show that actually DWD works better than SVM for low-dimensional problems and as the dimension increases PSVM gets closer to the performance of principal distance weighted discrimination (PDWD). Thus, data piling seems to help in the dimension reduction framework in the regression setting. This observation may be explained due to the fact that in the regression setting we are more interested in a hyperplane alignment than reducing

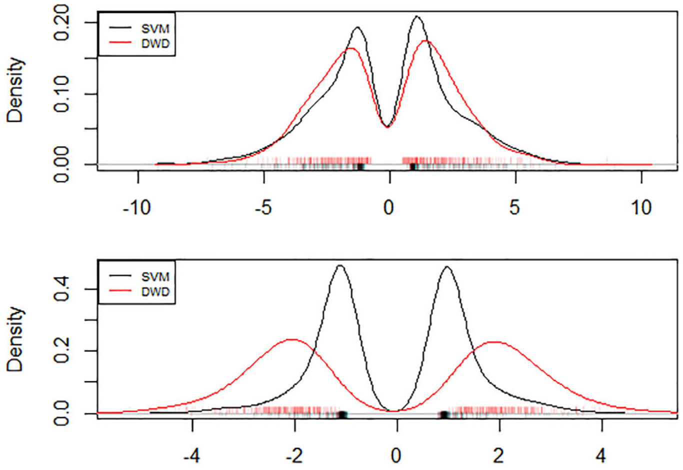


FIGURE 1 Density of projections for $n = 1{,}000$. Top panel: $p = 500$; bottom panel: $p = 1{,}000$. The datasets consists of the projection of points after discretizing the response in two slices under the model $Y = X_1 + \epsilon$ [Color figure can be viewed at wileyonlinelibrary.com]

misclassification error. Therefore, data piling may help "stabilize" the alignment of the hyperplane on the correct direction for PSVM.

The paper is constructed as follows. In Section 2 we discuss DWD and we introduce Principal DWD and in Section 3 we discuss its asymptotic properties. In Section 4 we present nonlinear feature extraction and in Section 5 our numerical studies. We close with a discussion section.

2 | PRINCIPAL DWD

In this section we develop the idea of using DWD for SDR. We discuss first DWD as it was presented by Marron et al. (2007) and then we demonstrate how it can be incorporated into the SDR framework, giving some theoretical results, a sample estimation algorithm and a method for determining the dimension of the central subspace.

2.1 | Review of DWD

Let $(\boldsymbol{X}_i,Y_i), i = 1, \dots, n$ be an iid sample of $(\boldsymbol{X},Y)$. Denote $\overline{\boldsymbol{X}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i$ and $\boldsymbol{\Sigma} = \text{var}(\boldsymbol{X})$. Now suppose Y is a binary random variable, which takes values ±1. DWD is defined by the following optimisation problem:

$$\text{minimize} \quad \sum_{i=1}^{n}\frac{1}{r_i} + \frac{\lambda}{n}\sum_{i=1}^{n}\xi_i \quad \text{among } (\boldsymbol{r}, \boldsymbol{\psi}, t, \boldsymbol{\xi}) \in \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^n,$$

$$\text{subject to} \quad r_i = Y_i[\boldsymbol{\psi}^T(\boldsymbol{X}_i - \overline{\boldsymbol{X}}) - t] + \xi_i \geq 0, \quad \xi_i \geq 0, \; i = 1, \dots, n, \quad \|\boldsymbol{\psi}\| \leq 1, \tag{3}$$

where $\boldsymbol{r}$ is a vector of all r_i's and $\boldsymbol{\xi}$ is the vector of all ξ_i's. Here $\lambda > 0$ is a tuning parameter also called the cost (or misclassification penalty) and $\boldsymbol{\xi}$ is a penalization vector where $\xi_i = 0$ for correctly classified points and $\xi_i > 0$ for misclassified points.

The above optimization problem can be written slightly differently using the following vector form (for details, see Qiao & Zhang, 2015):

$$\boldsymbol{\psi}^T\boldsymbol{\psi} + \sum_{i=1}^{n}\left[\left[Y_i[\boldsymbol{\psi}^T(\boldsymbol{X}_i - \overline{\boldsymbol{X}}) - t] + \left(\frac{1}{\sqrt{\lambda}} - Y_i[\boldsymbol{\psi}^T(\boldsymbol{X}_i - \overline{\boldsymbol{X}}) - t]\right)^{+}\right]^{-1}\right.$$

$$+\lambda\left(\frac{1}{\sqrt{\lambda}}-Y_i[\boldsymbol{\psi}^T(\boldsymbol{X}_i-\overline{\boldsymbol{X}})-t]\right)^{+}\Bigg], \tag{4}$$

where the first term comes from the constraint $||\boldsymbol{\psi}|| \leq 1$ and the rest from replacing ξ_i with the hinge loss $\left(\frac{1}{\sqrt{\lambda}}-Y_i[\boldsymbol{\psi}^T(\boldsymbol{X}_i-\overline{\boldsymbol{X}})-t]\right)^{+}$. In the dimension reduction framework we are interested to work with the population version of the above objective function. The population version is the following:

$$\boldsymbol{\psi}^T\boldsymbol{\psi}+\mathrm{E}\left[\left[Y[\boldsymbol{\psi}^T(\boldsymbol{X}-\mathrm{E}[\boldsymbol{X}])-t]+\left(\frac{1}{\sqrt{\lambda}}-Y[\boldsymbol{\psi}^T(\boldsymbol{X}-\mathrm{E}[\boldsymbol{X}])-t]\right)^{+}\right]^{-1}\right.$$
$$\left.+\lambda\left(\frac{1}{\sqrt{\lambda}}-Y[\boldsymbol{\psi}^T(\boldsymbol{X}-\mathrm{E}[\boldsymbol{X}])-t]\right)^{+}\right]. \tag{5}$$

There were a number of extensions of the DWD algorithm. Some include the weighted DWD approach by Qiao, Zhang, Liu, Todd, and Marron (2010) and the sparse DWD approach by Wang and Zou (2016). Marron et al. (2007) as well as the extensions discussed above used cone programming to solve the optimization problem in (3) (or the respective one for each extension). More recently, Wang and Zou (2018) proposed the generalized DWD algorithm which allow for faster computational algorithm. In this paper, we utilize their idea and thus our estimation algorithm is much faster than previous methodology in the SVM-based SDR framework. Another computationally fast algorithm appeared in Lam, Marron, Sun, and Toh (2018).

2.2 | DWD for SDR

One of the tricks that many classic algorithms are using for SDR is the idea of slicing/discretizing the response which in most regression settings is a continuous random variable (e.g., Li, 1991; Li et al., 2011). When the response is discrete this step is ignored as each discrete value is considered a slice. To unify the notation though and following the idea of Li et al. (2011) we define Ω_Y to be the support of Y and A_1 and A_2 to be disjoint subsets of Ω_Y (not necessarily exhaustive subsets). Then one can define

$$\tilde{Y}=I(Y\in A_1)-I(Y\in A_2). \tag{6}$$

Replacing this into the population objective function of DWD we get the following objective function in the SDR framework:

$$L(\boldsymbol{\psi},t)=\boldsymbol{\psi}^T\boldsymbol{\Sigma}\boldsymbol{\psi}+\mathrm{E}\left[\left[\tilde{Y}[\boldsymbol{\psi}^T(\boldsymbol{X}-\mathrm{E}[\boldsymbol{X}])-t]+\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}[\boldsymbol{\psi}^T(\boldsymbol{X}-\mathrm{E}[\boldsymbol{X}])-t]\right)^{+}\right]^{-1}\right]$$
$$+\mathrm{E}\left[\lambda\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}[\boldsymbol{\psi}^T(\boldsymbol{X}-\mathrm{E}[\boldsymbol{X}])-t]\right)^{+}\right]. \tag{7}$$

Following Li et al. (2011) we note that we have also inserted $\boldsymbol{\Sigma}$ into the first term to ensure the resulting DWD estimate is unbiased and to provide the unified framework for nonlinear SDR. Assuming $E(\boldsymbol{X}) = 0$ without loss of generality, and by setting $u = \tilde{Y}[\boldsymbol{\psi}^T\boldsymbol{X} - t]$ we can simplify the above objective function to:

$$\boldsymbol{\psi}^T\boldsymbol{\Sigma}\boldsymbol{\psi} + \mathrm{E}\left[\left[u + \left(\frac{1}{\sqrt{\lambda}} - u\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - u\right)^+\right]. \tag{8}$$

The following Lemma is used to prove the convexity of the objective function. This lemma will be crucial in proving the theorem which shows that the normal vector $\boldsymbol{\psi}$ of the optimal hyperplane, developed by the PDWD, is indeed in the CS.

Lemma 1. *If* $f(u) = \left[u + \left(\frac{1}{\sqrt{\lambda}} - u\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - u\right)^+$, *then f is convex for all* $\lambda > 0$.

Proof. To prove convexity we need to show

$$f(\alpha u_1 + (1-\alpha)u_2) \leq \alpha f(u_1) + (1-\alpha)f(u_2),$$

for all $u \in \mathbb{R}$ and $\alpha \in [0,1]$. Firstly we can rewrite f as

$$f(u) = \begin{cases} \frac{1}{u} & u \geq \frac{1}{\sqrt{\lambda}} \\ 2\sqrt{\lambda} - \lambda u & u < \frac{1}{\sqrt{\lambda}} \end{cases}.$$

For $u \geq \frac{1}{\sqrt{\lambda}}$ we have $2\sqrt{\lambda} - \lambda u \leq \frac{1}{u}$ and for $u_1 \leq u_2$ we have $f(u_1) \geq f(u_2)$ since f is a decreasing function. We need to consider three cases:

(i) When $u_1 < \frac{1}{\sqrt{\lambda}}, u_2 < \frac{1}{\sqrt{\lambda}}$. Here we have $\alpha u_1 + (1-\alpha)u_2 < \frac{1}{\sqrt{\lambda}}$ and hence

$$\begin{aligned} f(\alpha u_1 + (1-\alpha)u_2) &= 2\sqrt{\lambda} - \lambda(\alpha u_1 + (1-\alpha)u_2) \\ &= \alpha(2\sqrt{\lambda} - \lambda u_1) + (1-\alpha)(2\sqrt{\lambda} - \lambda u_2) \\ &= \alpha f(u_1) + (1-\alpha)f(u_2). \end{aligned}$$

(ii) When $u_1 < \frac{1}{\sqrt{\lambda}}, u_2 \geq \frac{1}{\sqrt{\lambda}}$. Since the gradient of f is equal when approaching from the left and right of $\frac{1}{\sqrt{\lambda}}$ we can assume without loss of generality that $\alpha u_1 + (1-\alpha)u_2 < \frac{1}{\sqrt{\lambda}}$ and so

$$\begin{aligned} f(\alpha u_1 + (1-\alpha)u_2) &= 2\sqrt{\lambda} - \lambda(\alpha u_1 + (1-\alpha)u_2) \\ &= \alpha(2\sqrt{\lambda} - \lambda u_1) + (1-\alpha)(2\sqrt{\lambda} - \lambda u_2) \\ &\leq \alpha(2\sqrt{\lambda} - \lambda u_1) + \frac{(1-\alpha)}{u_2} \\ &= \alpha f(u_1) + (1-\alpha)f(u_2). \end{aligned}$$

(iii) When $u_1 \geq \frac{1}{\sqrt{\lambda}}, u_2 \geq \frac{1}{\sqrt{\lambda}}$ which gives $\alpha u_1 + (1-\alpha)u_2 \geq \frac{1}{\sqrt{\lambda}}$. In this case we can simply prove that the second derivative of $f(u) = \frac{1}{u}$ only gives positive values as follows

$$f''(u) = \frac{2}{u^3} > 0 \quad \text{since} \quad \lambda > 0.$$

Hence we have

$$f(\alpha u_1 + (1-\alpha)u_2) \leq \alpha f(u_1) + (1-\alpha)f(u_2),$$

for all $u \in \mathbb{R}$, and therefore f is convex. ■

Having verified the convexity of the objective function then one can prove the following theorem which demonstrates that the normal vector of the hyperplane is in $\mathcal{S}_{Y|\boldsymbol{X}}$. This follows directly from the proof in Li et al. (2011) due to the fact that the hinge loss in SVM is replaced with another convex function and as Li et al. (2011) claim their proof holds for every convex function.

Theorem 1. *If* $\mathrm{E}(\boldsymbol{X}|\boldsymbol{\beta}^T\boldsymbol{X})$ *is a linear function of* $\boldsymbol{\beta}^T\boldsymbol{X}$*, where* $\boldsymbol{\beta}$ *is defined as in (1) and if* $(\boldsymbol{\psi}^*, t^*)$ *minimizes the objective function (13) among all* $(\boldsymbol{\psi}, t) \in \mathbb{R}^p \times \mathbb{R}$*, then* $\boldsymbol{\psi}^* \in \mathcal{S}_{Y|\boldsymbol{X}}$.

Proof. It is important to note that under the conditions of the theorem we can write the conditional expectation

$$\mathrm{E}[\boldsymbol{X}|\boldsymbol{\beta}^T\boldsymbol{X}] = \boldsymbol{P}_{\beta}^T(\boldsymbol{\Sigma})\boldsymbol{X},$$

where $\boldsymbol{P}_{\beta}(\boldsymbol{\Sigma})$ is the projection matrix $\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T\boldsymbol{\Sigma}$.

Our objective function takes the form

$$L(\boldsymbol{\psi}, t) = \boldsymbol{\psi}^T\boldsymbol{\Sigma}\boldsymbol{\psi} + \mathrm{E}\left[\left[u + \left(\frac{1}{\sqrt{\lambda}} - u\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - u\right)^+\right].$$

Beginning with the first term we have

$$\begin{aligned}
\boldsymbol{\psi}^T\boldsymbol{\Sigma}\boldsymbol{\psi} &= \mathrm{var}[\boldsymbol{\psi}^T\boldsymbol{X}] \\
&= \mathrm{var}[\mathrm{E}[\boldsymbol{\psi}^T\boldsymbol{X}|\boldsymbol{\beta}^T\boldsymbol{X}]] + \mathrm{E}[\mathrm{var}[\boldsymbol{\psi}^T\boldsymbol{X}|\boldsymbol{\beta}^T\boldsymbol{X}]] \\
&\geq \mathrm{var}[\mathrm{E}[\boldsymbol{\psi}^T\boldsymbol{X}|\boldsymbol{\beta}^T\boldsymbol{X}]] \\
&= \mathrm{var}[\boldsymbol{\psi}^T\boldsymbol{P}_{\beta}^T\boldsymbol{X}] \\
&= (\boldsymbol{P}_{\beta}(\boldsymbol{\Sigma})\boldsymbol{\psi})^T\boldsymbol{\Sigma}(\boldsymbol{P}_{\beta}(\boldsymbol{\Sigma})\boldsymbol{\psi}).
\end{aligned}$$

Hence

$$\boldsymbol{\psi}^T\boldsymbol{\Sigma}\boldsymbol{\psi} \geq (\boldsymbol{P}_{\beta}\boldsymbol{\psi})^T\boldsymbol{\Sigma}(\boldsymbol{P}_{\beta}\boldsymbol{\psi}). \tag{9}$$

Now lets look at the second term. Again we can write

$$\mathrm{E}\left[\left[u + \left(\frac{1}{\sqrt{\lambda}} - u\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - u\right)^+\right]$$

$$= \mathrm{E}\left[\mathrm{E}\left[\left[u + \left(\frac{1}{\sqrt{\lambda}} - u\right)^{+}\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - u\right)^{+} \middle| \tilde{Y}, \boldsymbol{\beta}^{T}\boldsymbol{X}\right]\right],$$

so if we define the function f such that $f(a) = \left[a + (1/\sqrt{\lambda} - a)^{+}\right]^{-1} + \lambda(1/\sqrt{\lambda} - a)^{+}$ then this gives

$$\mathrm{E}\left[\left[u + \left(\frac{1}{\sqrt{\lambda}} - u\right)^{+}\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - u\right)^{+} \middle| \tilde{Y}, \boldsymbol{\beta}^{T}\boldsymbol{X}\right] = \mathrm{E}[f(u)|\tilde{Y}, \boldsymbol{\beta}^{T}\boldsymbol{X}].$$

Since f is a convex function, we can use Jensen's inequality as follows:

$$\begin{aligned}
\mathrm{E}[f(u)|\tilde{Y}, \boldsymbol{\beta}^{T}\boldsymbol{X}] \geq & \left[\mathrm{E}[u|\tilde{Y}, \boldsymbol{\beta}^{T}\boldsymbol{X}] + \left(\frac{1}{\sqrt{\lambda}} - \mathrm{E}[u|\tilde{Y}, \boldsymbol{\beta}^{T}\boldsymbol{X}]\right)^{+}\right]^{-1} \\
& + \lambda\left(\frac{1}{\sqrt{\lambda}} - \mathrm{E}[u|\tilde{Y}, \boldsymbol{\beta}^{T}\boldsymbol{X}]\right)^{+} \\
= & \left[\tilde{Y}(\mathrm{E}[\boldsymbol{\psi}^{T}\boldsymbol{X}|\boldsymbol{\beta}^{T}\boldsymbol{X}] - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\mathrm{E}[\boldsymbol{\psi}^{T}\boldsymbol{X}|\boldsymbol{\beta}^{T}\boldsymbol{X}] - t)\right)^{+}\right]^{-1} \\
& + \lambda\left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\mathrm{E}[\boldsymbol{\psi}^{T}\boldsymbol{X}|\boldsymbol{\beta}^{T}\boldsymbol{X}] - t)\right)^{+} \\
= & \left[\tilde{Y}(\boldsymbol{\psi}^{T}\boldsymbol{P}_{\beta}^{T}(\boldsymbol{\Sigma})\boldsymbol{X} - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\boldsymbol{\psi}^{T}\boldsymbol{P}_{\beta}^{T}(\boldsymbol{\Sigma})\boldsymbol{X} - t)\right)^{+}\right]^{-1} \\
& + \lambda\left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\boldsymbol{\psi}^{T}\boldsymbol{P}_{\beta}^{T}(\boldsymbol{\Sigma})\boldsymbol{X} - t)\right)^{+}.
\end{aligned}$$

Thus combining this with (9) we get

$$L(\boldsymbol{\psi}, t) \geq L(\boldsymbol{P}_{\beta}(\boldsymbol{\Sigma})\boldsymbol{\psi}, t). \tag{10}$$

If $\boldsymbol{\psi}$ does not belong to $\mathcal{S}_{Y|\boldsymbol{X}}$, then $\mathrm{var}[\boldsymbol{\psi}^{T}\boldsymbol{X}|\boldsymbol{\beta}^{T}\boldsymbol{X}] > 0$ and the inequality in (9) becomes strict. Hence the inequality in (10) is strict. Therefore, such $\boldsymbol{\psi}$ cannot be the minimizer of $L(\boldsymbol{\psi}, t)$. ■

2.3 | Sample estimation algorithm

Having established the theoretical properties of the minimizer of the objective function in PDWD we now look into the sample estimation algorithm of our method. Before giving the algorithm though we look at available packages in solving the optimization problem of DWD. As the available packages solve the objective function of DWD which does not include $\boldsymbol{\Sigma}$ in the first term, we demonstrate below that by standardizing the data the objective function of PDWD becomes equivalent to the objective function of DWD and therefore available packages can be used.

As was mentioned above the objective function of DWD is

$$\boldsymbol{\psi}^T\boldsymbol{\psi} + \mathrm{E}_n\left[\left[\tilde{Y}(\boldsymbol{\psi}^T\boldsymbol{X} - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\boldsymbol{\psi}^T\boldsymbol{X} - t)\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\boldsymbol{\psi}^T\boldsymbol{X} - t)\right)^+\right], \qquad (11)$$

and the one for PDWD is

$$\boldsymbol{\psi}^T\boldsymbol{\Sigma}^T\boldsymbol{\psi} + \mathrm{E}_n\left[\left[\tilde{Y}(\boldsymbol{\psi}^T\boldsymbol{X} - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\boldsymbol{\psi}^T\boldsymbol{X} - t)\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\boldsymbol{\psi}^T\boldsymbol{X} - t)\right)^+\right]. \qquad (12)$$

Now if we let $\boldsymbol{\zeta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\psi}$ and $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})$, and substitute these into (12) we have

$$\boldsymbol{\zeta}^T\boldsymbol{\zeta} + \mathrm{E}_n\left[\left[\tilde{Y}(\boldsymbol{\zeta}^T\boldsymbol{Z} - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\boldsymbol{\zeta}^T\boldsymbol{Z} - t)\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\boldsymbol{\zeta}^T\boldsymbol{Z} - t)\right)^+\right], \qquad (13)$$

which we can see is of the same form as (11). Hence, as we stated above we can see that standardizing $\boldsymbol{X}$ modifies the PDWD in a desired way. We emphasize here that this fact allows us to use existing algorithms for DWD in the literature to estimate the PDWD solution. Hence, in our algorithm below we require the standardization of the data.

To solve (13) Wang and Zou (2018) proposed an algorithm to iteratively calculate the hyperplane. This is a fast algorithm that calculates the hyperplane until convergence. To calculate the solution in iteration $(m + 1)$ the following formula is used:

$$\begin{pmatrix} t^{(m+1)} \\ \boldsymbol{\zeta}^{(m+1)} \end{pmatrix} = \begin{pmatrix} t^{(m)} \\ \boldsymbol{\zeta}^{(m)} \end{pmatrix} - \frac{n}{4}\boldsymbol{P}^{-1}(\lambda)\begin{pmatrix} \boldsymbol{1}^T\boldsymbol{z} \\ \boldsymbol{Z}^T\boldsymbol{z} + 2\lambda\boldsymbol{\zeta}^{(m)} \end{pmatrix},$$

where $\boldsymbol{z} = (z_1, \ldots, z_n)^T$,

$$z_i = \begin{cases} -\frac{\tilde{y}_i}{n}, & \tilde{y}_i(\boldsymbol{\zeta}^T\boldsymbol{Z} - t) \leq \frac{1}{2} \\ -\frac{1}{n(\tilde{y}_i(\boldsymbol{\zeta}^T\boldsymbol{Z} - t))^2}\left(\frac{1}{2}\right)^2, & \tilde{y}_i(\boldsymbol{\zeta}^T\boldsymbol{Z} - t) > \frac{1}{2} \end{cases}.$$

Finally,

$$\boldsymbol{P}^{-1}(\lambda) = \begin{pmatrix} n & \boldsymbol{1}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{1} & \boldsymbol{Z}^T\boldsymbol{Z} + \frac{n\lambda}{2}\boldsymbol{I}_p \end{pmatrix}^{-1}.$$

This iterative process replaces the quadratic programming process which was used in PSVM and therefore the PDWD algorithm becomes computationally much faster. For more details the interested reader is referred to Wang and Zou (2018).

We will define two methods for generating $\tilde{Y}$, which were first proposed in Li et al. (2011). These are named left versus right (LVR) which is more appropriate when the response is continuous or discrete with a sense of ordering between the values and one versus another (OVA) which is more appropriate when the response is categorical with no sense of ordering between the values. When using LVR you choose $\tilde{Y}_i^r = I(Y_i > q_r) - I(Y_i \leq q_r)$ for $i=1, \dots, n$ and $r=1, \dots, h-1$ where h is the number of slices. OVA follows a similar method but we choose $\tilde{Y}_i^{rs} = I(q_{s-1} < Y_i \leq q_s) - I(q_{r-1} < Y_i \leq q_r)$ where $r, s = 1, \dots, h$ with $r \neq s$.

The estimation procedure is as follows:

1. Compute the sample mean $\overline{\boldsymbol{X}}$ and sample variance matrix $\hat{\boldsymbol{\Sigma}}$.
2. We find the minimizer using the algorithm in Wang and Zou (2018). In more detail: (LVR) Let q_r, $r=1, \dots h-1$, be $h-1$ dividing points and let

$$\tilde{Y}_i^r = I(Y_i > q_r) - I(Y_i \leq q_r),$$

for $i=1, \dots, n$. Then using DWD, let $(\hat{\boldsymbol{\psi}}_r, \hat{t}_r)$ be the minimizers of

$$\boldsymbol{\psi}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\psi} + \mathrm{E}\left[\left[\tilde{Y}^r(\boldsymbol{\psi}^T\boldsymbol{X} - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}^r(\boldsymbol{\psi}^T\boldsymbol{X} - t)\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}^r(\boldsymbol{\psi}^T\boldsymbol{X} - t)\right)^+\right].$$

(OVA) Apply DWD to each pair of slices from the h slices. More specifically, let $q_0 = \min\{Y_1, \dots, Y_h\}$ and $q_h = \max\{Y_1, \dots, Y_h\}$. Then for each (r, s) such that $1 \leq r < s \leq h$, let

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

Let $(\hat{\boldsymbol{\psi}}_{rs}, \hat{t}_{rs})$ be the minimizers of

$$\boldsymbol{\psi}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\psi} + \mathrm{E}\left[\left[\tilde{Y}^{rs}(\boldsymbol{\psi}^T\boldsymbol{X} - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}^{rs}(\boldsymbol{\psi}^T\boldsymbol{X} - t)\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}^{rs}(\boldsymbol{\psi}^T\boldsymbol{X} - t)\right)^+\right].$$

3. Let $\hat{\boldsymbol{v}}_1, \dots, \hat{\boldsymbol{v}}_d$ be the d leading eigenvectors of one of the matrices

$$\hat{\boldsymbol{M}}_n = \sum_{r=1}^{h-1} \hat{\boldsymbol{\psi}}_r \hat{\boldsymbol{\psi}}_r^T \quad \text{or} \quad \hat{\boldsymbol{M}}_n = \sum_{r=1}^{h-1}\sum_{s=r+1}^{h} \hat{\boldsymbol{\psi}}_{rs}\hat{\boldsymbol{\psi}}_r{s}^T. \tag{14}$$

We can now estimate $S_{Y|\boldsymbol{X}}$ using the subspace spanned by $\hat{\boldsymbol{v}} = (\hat{\boldsymbol{v}}_1, \dots, \hat{\boldsymbol{v}}_d)$.

2.4 | Order determination

Now we turn our attention to the estimation of the dimension d. Developing an effective method for determining the dimension is vital when developing methods for SDR and plays an important role in the performance of such method. For PSVM, Li et al. (2011) opted to use method based on a cross-validated bayesian information criterion. We propose to use a relatively new approach to order determination developed by Luo and Li (2016) which is called the ladle estimate.

The ladle estimator is a combination of the scree plot method and the Ye-Weiss plot developed by Ye and Weiss (2003). Let $\hat{\boldsymbol{M}}$ be defined as one of the matrices in (14) and let $\hat{\lambda}_i$ define the ith eigenvalue of $\hat{\boldsymbol{M}}$. Now since $\hat{\boldsymbol{M}}$ is a consistent estimator of $\boldsymbol{M}$ and $\boldsymbol{M}$ has rank d we can establish that $\hat{\lambda}_{d+1}$ will be much smaller than $\hat{\lambda}_d$. Using this the following function is defined

$$\phi_n : \{0, \dots, p-1\} \to \mathbb{R}, \quad \phi_n(k) = \frac{\hat{\lambda}_{k+1}}{1 + \sum_{i=0}^{p-1} \hat{\lambda}_{i+1}}. \tag{15}$$

The eigenvalues have been shifted so that ϕ_n takes a small values at $k = d$ rather than at $k = d + 1$.

Next we turn our attention to the Ye–Weiss plot. Let F be the distribution of $(\boldsymbol{X}, Y)$ and let F_n be the empirical distribution based on $S = (\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)$. Conditioning on S, let $(\boldsymbol{X}^*_{1,n}, Y^*_{1,n}), \dots, (\boldsymbol{X}^*_{n,n}, Y^*_{n,n})$ be and iid bootstrap sample from F_n. Now define $\{\hat{\lambda}_1, \dots, \hat{\lambda}_p, \hat{v}_1, \dots, \hat{v}_p\}$ and $\{\lambda^*_1, \dots, \lambda^*_p, v^*_1, \dots, v^*_p\}$ be the eigenvalues and eigenvectors of $\hat{\boldsymbol{M}}$ and $\boldsymbol{M}^*$ respectively. For each $k < p$, let

$$\hat{\boldsymbol{B}}_k = (\hat{v}_1, \dots, \hat{v}_k) \quad \boldsymbol{B}^*_k = (v^*_1, \dots, v^*_k),$$

and define the function

$$f^0_n : \{0, \dots, p-1\} \to \mathbb{R}, \quad f^0_n(k) = \begin{cases} 0 & k = 0 \\ n^{-1} \sum_{i=1}^{n} 1 - |\det(\boldsymbol{B}^T_k \boldsymbol{B}^*_{k,i})| & k = 1, \dots, p-1 \end{cases}, \tag{16}$$

where $\boldsymbol{B}^*_{k,i}$ denotes the ith bootstrap sample. From Ye and Weiss (2003), it can be established that the function $f^0_n(k)$ gives a measure of the variability of the bootstrap estimates around the full sample estimate $\hat{\boldsymbol{B}}_k$. The range of f^0_n is [0,1], where 0 indicates each $\boldsymbol{B}^*_{k,i}$ spans the same column space as $\hat{\boldsymbol{B}}_k$ and 1 occurs when $\boldsymbol{B}^*_{k,i}$ spans a space orthogonal to $\hat{\boldsymbol{B}}_k$. So if we define the function

$$f_n : \{0, \dots, p-1\} \to \mathbb{R} \quad f_n(k) = \frac{f^0_n(k)}{1 + \sum_{i=0}^{p-1} f^0_n(i)}, \tag{17}$$

Ye and Weiss (2003) determined that f_n is small for $k = d$ and larger for $k > d$.

Lastly, the ladle estimator of the rank d is defined to be

$$\hat{d} = \arg\min_k \{g_n(k) : k \in \mathcal{D}(g_n)\}, \tag{18}$$

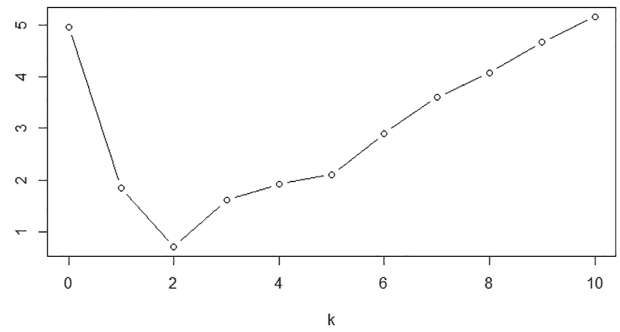
where $g_n(k) = \phi_n(k) + f_n(k)$.

Consider the regression model

$$Y = \frac{X_1}{0.5 + (X_2 + 1)^2} + \sigma\epsilon. \tag{19}$$

Choosing $n = 100$ and $p = 10$, Figure 2 shows the ladle plot for model (19). As we can see, the ladle plot correctly estimates d to be 2. We have tried this simulation many times with approximately 98% accuracy. More detailed simulations will be discussed in our numerical studies section.

FIGURE 2 Ladle plot of model (19) with $n = 100$ and $p = 10$



3 | ASYMPTOTIC ANALYSIS OF PDWD

In this section we discuss the asymptotic properties of the PDWD. We find the Hessian matrix and the influence function before proving consistency. We demonstrate the consistency when p is fixed, as well as when p is not fixed and tends to infinity, although we still require it to be less than n. To make the proofs easier to read we use the following notation. Let $\boldsymbol{\theta} = (\boldsymbol{\psi}^T, t)^T, \boldsymbol{Z} = (\boldsymbol{X}^T, \tilde{Y})^T$, $\boldsymbol{X}^* = (\boldsymbol{X}^T, -1)^T$ and $\boldsymbol{\Sigma}^* = \text{diag}(\boldsymbol{\Sigma}, 0)$, then $u = \boldsymbol{\theta}^T\boldsymbol{X}^*\tilde{Y}$ and thus

$$\boldsymbol{\psi}^T\boldsymbol{\Sigma}\boldsymbol{\psi} + \left[u + \left(\frac{1}{\sqrt{\lambda}} - u\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - u\right)^+ = \boldsymbol{\theta}^T\boldsymbol{\Sigma}^*\boldsymbol{\theta}$$
$$+ \left[\boldsymbol{\theta}^T\boldsymbol{X}^*\tilde{Y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}^T\boldsymbol{X}^*\tilde{Y}\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}^T\boldsymbol{X}^*\tilde{Y}\right)^+.$$

We denote this function by $m(\boldsymbol{\theta}, \boldsymbol{Z})$. Let $\Omega_{\boldsymbol{Z}}$ be the support of $\boldsymbol{Z}$ and let $\boldsymbol{h} : \Theta \times \Omega_{\boldsymbol{Z}} \to \mathbb{R}^+$ be a function of $(\boldsymbol{\theta}, \boldsymbol{Z})$. Let $D_{\boldsymbol{\theta}}$ denote the $(p+1)$-dimensional column vector of differential operators $(\partial/\partial\theta_1, \ldots, \partial/\partial\theta_{p+1})^T$.

Before we consider the gradient of the DWD objective function, we prove that the function f is differentiable at all points.

Lemma 2. *The function f, as defined in Lemma 1, is differentiable at all points.*

Proof. We need to prove that the gradient of f as we approach $\frac{1}{\sqrt{\lambda}}$ from below is equal to the gradient as we approach from the above. We have

$$\begin{aligned} f'(a) &= -\left[a + \left(\frac{1}{\sqrt{\lambda}} - a\right)^+\right]^{-2} \\ &= \begin{cases} -a^{-2} & a \geq \frac{1}{\sqrt{\lambda}} \\ -\lambda & a < \frac{1}{\sqrt{\lambda}} \end{cases}. \end{aligned}$$

Hence $\lim_{a\downarrow\frac{1}{\sqrt{\lambda}}} f'(a) = -\lambda = \lim_{a\uparrow\frac{1}{\sqrt{\lambda}}} f'(a)$. Therefore f is differentiable everywhere. ■

The next theorem gives the gradient of the DWD objective function $\mathrm{E}[m(\boldsymbol{\theta}, \boldsymbol{Z})]$. The proof follows straight from Lemma 2 and is therefore omitted. Let D^2_{θ} denote the operator $D_{\theta}D^T_{\theta}$. Thus, $D^2_{\theta}m(\boldsymbol{\theta}, \boldsymbol{Z})$ is the $(p+1)\times(p+1)$ matrix whose (i,j)th entry is $\partial^2 m/\partial\theta_i\partial\theta_j$.

Theorem 2. *The gradient of* $\mathrm{E}(m(\boldsymbol{\theta}, \boldsymbol{z}))$ *takes the form*

$$D_{\theta}\mathrm{E}[m(\boldsymbol{\theta}, \boldsymbol{z})] = 2\boldsymbol{\Sigma}^*\boldsymbol{\theta} - \mathrm{E}\left[\boldsymbol{X}^*\tilde{Y}\left[\boldsymbol{\theta}^T\boldsymbol{X}^*\tilde{Y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}^T\boldsymbol{X}^*\tilde{Y}\right)^+\right]^{-2}\right]. \tag{20}$$

The next step is to find the Hessian matrix. Before doing so, we state some helpful results. First we use the following notation. Let $\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{z}) = D_{\theta}m(\boldsymbol{\theta}, \boldsymbol{z})$ and for each $\boldsymbol{\theta} \in \Theta$, $N_{\theta}(\boldsymbol{n})$ be the set of $\boldsymbol{x}$ for which a function $\boldsymbol{n}(\boldsymbol{z}, \cdot)$ is not differentiable at $\boldsymbol{\theta}$. That is,

$$N_{\theta}(\boldsymbol{n}) = \{\boldsymbol{z} : D_{\theta}\boldsymbol{n}(\cdot, \boldsymbol{z}) \text{ is not differentiable at } \boldsymbol{\theta}\}.$$

Lemma 3. *Suppose that* $\boldsymbol{n} : \Theta \times \Omega_Z \to \mathbb{R}$ *satisfies the following conditions*

1. *(almost surely differentiable) for each* $\boldsymbol{\theta} \in \Theta$, $\mathrm{P}[\boldsymbol{Z} \in N_{\theta}(\boldsymbol{n})] = 0$;
2. *(Lipschitz condition) there is an integrable function* $c(\boldsymbol{z})$, *independent of* $\boldsymbol{\theta}$, *such that for any* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,

$$|\boldsymbol{n}(\boldsymbol{\theta}_2, \boldsymbol{z}) - \boldsymbol{n}(\boldsymbol{\theta}_1, \boldsymbol{z})| \leq c(\boldsymbol{z})\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|.$$

Then $D_{\theta}[\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{Z})]$ *is integrable,* $\mathrm{E}[D_{\theta}\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{Z})]$ *is differentiable and*

$$D_{\theta}\mathrm{E}[\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{Z})] = \mathrm{E}[D_{\theta}\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{Z})]. \tag{21}$$

Lemma 4. *For* $c > 0$ *we have the following identity*

$$\left|\frac{(a + (c-a)^+)^2 - (b + (c-b)^+)^2}{(a + (c-a)^+)^2(b + (c-b)^+)^2}\right| \leq \frac{2}{c^3}|b - a|.$$

Now we have the necessary results which will be helpful in finding the Hessian matrix as the following theorem states.

Theorem 3. *Suppose, for each* $\tilde{y} = -1, 1$, *the distribution of* $\boldsymbol{X}|\tilde{Y} = \tilde{y}$ *is dominated by the Lebesgue measure and* $\mathrm{E}[\|\boldsymbol{X}\|^2] < \infty$. *Then*

$$D_{\theta}\mathrm{E}[\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{Z})] = 2\boldsymbol{\Sigma}^* + \mathrm{E}\left[\boldsymbol{X}^*\boldsymbol{X}^{*^T} I\left(\theta^T\boldsymbol{X}^*\tilde{Y} < \frac{1}{\sqrt{\lambda}}\right)[\theta^T\boldsymbol{X}^*\tilde{Y}]^{-3}\right]. \tag{22}$$

Proof. Let $H(\boldsymbol{\psi}, a)$ denote the hyperplane $\{\boldsymbol{x} : \boldsymbol{\psi}^T\boldsymbol{x} = a\}$. We first need to verify the two assumptions in Lemma 3. In our case,

$$\mathrm{P}[(\boldsymbol{X}, \tilde{Y}) \in N_{\theta}(n)] = \sum_{\tilde{y}\in\{-1,1\}} \mathrm{P}(\tilde{Y} = \tilde{y})\mathrm{P}\left[\boldsymbol{X} \in H\left(\boldsymbol{\psi}, t + \frac{\tilde{y}}{\sqrt{\lambda}}\right)\middle|\tilde{Y} = \tilde{y}\right].$$

Since the Lebesgue measure of $H\left(\boldsymbol{\psi}, t+\frac{\tilde{y}}{\sqrt{\lambda}}\right)$ is 0 for $\tilde{y} \in \{-1, 1\}$, so by assumption 1 of the theorem, the above probability is 0. Thus condition 1 of Lemma 3 is satisfied.

Let $\boldsymbol{n}_1(\boldsymbol{\theta}, \boldsymbol{z}) = \boldsymbol{\Sigma}^*\boldsymbol{\theta}$ and $\boldsymbol{n}_2(\boldsymbol{\theta}, \boldsymbol{z}) = -\boldsymbol{x}^*\tilde{y}\left[\boldsymbol{\theta}^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-2}$. Then $\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{z}) = 2\boldsymbol{n}_1(\boldsymbol{\theta}, \boldsymbol{z}) + \boldsymbol{n}_2(\boldsymbol{\theta}, \boldsymbol{z})$. Since $\boldsymbol{n}_1$ is nonrandom and differentiable, it obviously satisfies $\mathrm{E}[D_{\boldsymbol{\theta}}\boldsymbol{n}_1(\boldsymbol{\theta}, \boldsymbol{z})] = D_{\boldsymbol{\theta}}\mathrm{E}[\boldsymbol{n}_1(\boldsymbol{\theta}, \boldsymbol{z})]$. To verify that $\boldsymbol{n}_2$ is Lipschitz, let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathrm{R}^{p+1}$. Then

$$\begin{aligned}\|\boldsymbol{n}_2(\boldsymbol{\theta}_2, \boldsymbol{z}) - \boldsymbol{n}_2(\boldsymbol{\theta}_1, \boldsymbol{z})\| &= \left\| \boldsymbol{x}^*\tilde{y}\left[\boldsymbol{\theta}_2^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_2^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-2} - \boldsymbol{x}^*\tilde{y}\left[\boldsymbol{\theta}_1^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_1^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-2} \right\| \\ &\leq \|\boldsymbol{x}^*\| \left\| \frac{\left(\boldsymbol{\theta}_1^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_1^T\boldsymbol{x}^*\tilde{y}\right)^+\right)^2 - \left(\boldsymbol{\theta}_2^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_2^T\boldsymbol{x}^*\tilde{y}\right)^+\right)^2}{\left(\boldsymbol{\theta}_1^T\boldsymbol{x}^*\tilde{y} + \left(1\sqrt{\lambda} - \boldsymbol{\theta}_1^T\boldsymbol{x}^*\tilde{y}\right)^+\right)^2\left(\boldsymbol{\theta}_2^T\boldsymbol{x}^*\tilde{y} + \left(1\sqrt{\lambda} - \boldsymbol{\theta}_2^T\boldsymbol{x}^*\tilde{y}\right)^+\right)^2} \right\|.\end{aligned}$$

From Lemma 4 we get:

$$\begin{aligned}\|\boldsymbol{n}_2(\boldsymbol{\theta}_2, \boldsymbol{z}) - \boldsymbol{n}_2(\boldsymbol{\theta}_1, \boldsymbol{z})\| &\leq 2\lambda^{3/2}\|\boldsymbol{x}^*\|\|\boldsymbol{\theta}_1^T\boldsymbol{x}^* - \boldsymbol{\theta}_2^T\boldsymbol{x}^*\| \\ &\leq 2\lambda^{3/2}(1 + \|\boldsymbol{x}\|^2)\|\boldsymbol{\theta}_1^T - \boldsymbol{\theta}_2^T\|.\end{aligned}$$

Since $\mathrm{E}[\|\boldsymbol{X}\|^2] < \infty$,

$$\mathrm{E}[1 + \|\boldsymbol{X}\|^2] = 1 + \mathrm{E}[\|\boldsymbol{X}\|^2] < \infty.$$

This verifies condition 2 of Lemma 3.

Finally, by direct calculations we find that, for $\boldsymbol{z} \notin N_{\boldsymbol{\theta}}(\boldsymbol{n})$,

$$\begin{aligned}D_{\boldsymbol{\psi}}[\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{z})] &= 2\boldsymbol{\Sigma} + 2\boldsymbol{x}^*\boldsymbol{x}^T\mathrm{I}\left(\tilde{y}(\boldsymbol{\psi}^T\boldsymbol{x} - t) \geq \frac{1}{\sqrt{\lambda}}\right)[\tilde{y}(\boldsymbol{\psi}^T\boldsymbol{x} - t)]^{-3} \\ D_t[\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{z})] &= -2\boldsymbol{x}^*\mathrm{I}\left(\tilde{y}(\boldsymbol{\psi}^T\boldsymbol{x} - t) \geq \frac{1}{\sqrt{\lambda}}\right)[\tilde{y}(\boldsymbol{\psi}^T\boldsymbol{x} - t)]^{-3}.\end{aligned}$$

Hence

$$D_{\boldsymbol{\theta}}[\boldsymbol{n}(\boldsymbol{\theta}, \boldsymbol{z})] = 2\boldsymbol{\Sigma}^* + 2\boldsymbol{x}^*\boldsymbol{x}^{*^T}\mathrm{I}\left(\boldsymbol{\theta}^T\boldsymbol{x}^*\tilde{y} \geq \frac{1}{\sqrt{\lambda}}\right)[\boldsymbol{\theta}^T\boldsymbol{x}^*\tilde{y}]^{-3}.$$

The theorem follows now from Lemma 3. ■

The following theorem proves the consistency of our estimator. A similar result in the SVM literature can be found in Jiang, Zhang, and Cai (2008).

Theorem 4. *Let $\boldsymbol{\theta}_0 = (\boldsymbol{\psi}_0^T, t_0)^T$ be the minimizer of $\mathrm{E}[m(\boldsymbol{\theta}, \boldsymbol{Z})]$. Suppose, for each $\tilde{y} = -1, 1$, the distribution of $\boldsymbol{X}|\tilde{Y} = \tilde{y}$ is dominated by the Lebesgue measure and $\mathrm{E}[\|\boldsymbol{X}\|^2] < \infty$. Then*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = -n^{-1}\boldsymbol{H}^{-1}\sum_{i=1}^{n}\mathrm{B}_i(\boldsymbol{z}) + o_p(n^{-1/2}), \qquad (23)$$

where $\mathrm{B}_i(\boldsymbol{z}) = 2\boldsymbol{\Sigma}^*\boldsymbol{\theta}_0 - \boldsymbol{x}_i^*\tilde{y}_i\left[\boldsymbol{\theta}_0^T\boldsymbol{x}_i^*\tilde{y}_i + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_0^T\boldsymbol{x}_i^*\tilde{y}_i\right)^*\right]^{-2}$ *and H is the Hessian matrix defined previously.*

Proof. Let $\boldsymbol{a} = (\boldsymbol{\psi_a}^T, t_{\boldsymbol{a}})^T$ and now we write

$$
\begin{aligned}
m(\boldsymbol{z}, \boldsymbol{\theta}_0 + \boldsymbol{a}) - m(\boldsymbol{z}, \boldsymbol{\theta}_0) &= (\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{\Sigma}^*(\boldsymbol{\theta}_0 + \boldsymbol{a}) - \boldsymbol{\theta}_0^T\boldsymbol{\Sigma}^*\boldsymbol{\theta} \\
&\quad + \left[(\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - (\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-1} \\
&\quad - \left[\boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-1} \\
&\quad + \lambda\left(\frac{1}{\sqrt{\lambda}} - (\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y}\right)^+ - \lambda\left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y}\right)^+ \\
&= \boldsymbol{a}^T\boldsymbol{\Sigma}^*\boldsymbol{a} + 2\boldsymbol{a}^T\boldsymbol{\Sigma}^*\boldsymbol{\theta} + \left[(\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - (\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-1} \\
&\quad - \left[\boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-1} + \lambda\left(\frac{1}{\sqrt{\lambda}} - (\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y}\right)^+ \\
&\quad - \lambda\left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y}\right)^+ \\
&= \boldsymbol{a}^T D_{\boldsymbol{\theta}_0} m(\boldsymbol{z}, \boldsymbol{\theta}_0) + R(\boldsymbol{z}, \boldsymbol{a}),
\end{aligned}
$$

where

$$
\begin{aligned}
R(\boldsymbol{z}, \boldsymbol{a}) &= \boldsymbol{a}^T\boldsymbol{\Sigma}^*\boldsymbol{a} + \left[(\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - (\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-1} \\
&\quad - \left[\boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-1} \\
&\quad + \lambda\left(\frac{1}{\sqrt{\lambda}} - (\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y}\right)^+ - \lambda\left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y}\right)^+ \\
&\quad - \boldsymbol{a}^T\boldsymbol{x}^*\tilde{y}\left[\boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-2}, \\
D_{\boldsymbol{a}}R(\boldsymbol{z}, \boldsymbol{a}) &= 2\boldsymbol{\Sigma}^*\boldsymbol{a} - \boldsymbol{x}^*\tilde{y}\left[(\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - (\boldsymbol{\theta}_0 + \boldsymbol{a})^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-2} \\
&\quad + \boldsymbol{x}^*\tilde{y}\left[\boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y} + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_0^T\boldsymbol{x}^*\tilde{y}\right)^+\right]^{-2},
\end{aligned}
$$

and

$$D_{\boldsymbol{a}}[D_{\boldsymbol{a}}R(\boldsymbol{z},\boldsymbol{a})] = 2\boldsymbol{\Sigma}^{*} + 2\boldsymbol{x}\boldsymbol{x}^{T}I\left((\boldsymbol{\theta}_0+\boldsymbol{a})^{T}\boldsymbol{x}^{*}\tilde{y} \geq \frac{1}{\sqrt{\lambda}}\right)[(\boldsymbol{\theta}_0+\boldsymbol{a})^{T}\boldsymbol{x}^{*}\tilde{y}]^{-3},$$

This gives, $R(\boldsymbol{z},\boldsymbol{0})=0$, $D_{\boldsymbol{a}}R(\boldsymbol{z},\boldsymbol{0})=\boldsymbol{0}$ and $\mathrm{E}[D_{\boldsymbol{a}}[D_{\boldsymbol{a}}R(\boldsymbol{z},\boldsymbol{a})]] = \boldsymbol{H}$. By definition we also have $\mathrm{E}[D_{\boldsymbol{\theta}_0}m(\boldsymbol{z},\boldsymbol{\theta}_0)] = \boldsymbol{0}$. Hence

$$\mathrm{E}[m(\boldsymbol{z},\boldsymbol{\theta}_0+\boldsymbol{a}) - m(\boldsymbol{z},\boldsymbol{\theta}_0)] = \mathrm{E}[R(\boldsymbol{z},\boldsymbol{a})] = \frac{\boldsymbol{a}^{T}\boldsymbol{H}\boldsymbol{a}}{2} + o(\|\boldsymbol{a}\|^{2}), \tag{24}$$

and since $\boldsymbol{H}$ is the Hessian of a convex function we can establish that it is symmetric and positive definite.

Now let $\boldsymbol{s} = (\boldsymbol{\psi_s}^{T}, t_{\boldsymbol{s}})^{T}$ and $A_n(\boldsymbol{s}) = \sum_{i=1}^{n}\{m(\boldsymbol{z}_i,\boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{s}) - m(\boldsymbol{z}_i,\boldsymbol{\theta}_0)\}$. We can see that $A_n(\boldsymbol{s})$ is convex with respect to $\boldsymbol{s}$ and is therefore minimized by $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Now we can write

$$\begin{aligned} A_n(\boldsymbol{s}) &= \sum_{i=1}^{n}\{n^{-1/2}\boldsymbol{s}^{T}\mathrm{B}(\boldsymbol{z}_i) + \mathrm{R}(\boldsymbol{z}_i, n^{-1/2}\boldsymbol{s}) - \mathrm{E}[\mathrm{R}(\boldsymbol{z}_i, n^{-1/2}\boldsymbol{s})]\} + n\mathrm{E}[\mathrm{R}(\boldsymbol{z}, n^{-1/2}\boldsymbol{s})] \\ &= n^{-1/2}\sum_{i=1}^{n}\boldsymbol{s}^{T}\mathrm{B}(\boldsymbol{z}_i) + \frac{1}{2}\boldsymbol{s}^{T}\boldsymbol{H}\boldsymbol{s} + r_{n,0}(\boldsymbol{s}) + r_{n,1}(\boldsymbol{s}), \end{aligned}$$

where $r_{n,0}(\boldsymbol{s}) = o(\|\boldsymbol{s}\|^{2}) \to 0$ for fixed $\boldsymbol{s}$ and $r_{n,1}(\boldsymbol{s}) = \sum_{i=1}^{n}\mathrm{R}(\boldsymbol{z}_i, n^{-1/2}\boldsymbol{s}) - \mathrm{E}[\mathrm{R}(\boldsymbol{z}_i, n^{-1/2}\boldsymbol{s})] \to 0$ in probability since it has mean zero and variance $o(\|\boldsymbol{s}\|^{2})$.

Since $\boldsymbol{H}$ is positive definite, and the covariance matrix $\mathrm{var}[\boldsymbol{X}]$ is finite, it follows from the basic corollary of Hjort and Pollard (1993) that (23) holds. ■

Let $\boldsymbol{\theta}_{0r} = (\boldsymbol{\psi}_{0r}^{T}, t_{0r})^{T}$ be the minimizer of $\mathrm{E}[m(\boldsymbol{\theta},\boldsymbol{Z}^{r})]$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{r}} = (\hat{\boldsymbol{\psi}}_{r}^{T}, t_r)^{T}$ be the minimizer of $\mathrm{E}_n[m(\boldsymbol{\theta},\boldsymbol{Z}^{r})]$. Let $\boldsymbol{H}_r$ be the Hessian matrix of $\mathrm{E}[m(\boldsymbol{\theta},\boldsymbol{Z}^{r})]$ and let $\boldsymbol{F}_r$ be the first p rows of $\boldsymbol{H}_r^{-1}$. By the last theorem we have

$$\hat{\boldsymbol{\psi}}_r = \boldsymbol{\psi}_{0r} - n^{-1}\boldsymbol{F}_r\sum_{i=1}^{n}\tilde{\mathrm{B}}_i(\boldsymbol{z}) + o_p(n^{-1/2}), \tag{25}$$

where $\tilde{\mathrm{B}}_i(\boldsymbol{z}) = 2\boldsymbol{\Sigma}\boldsymbol{\psi}_0 - \boldsymbol{x}_i\tilde{y}_i\left[\boldsymbol{\theta}_{0r}^{T}\boldsymbol{x}^{*}{}_i\tilde{y}_i + \left(\frac{1}{\sqrt{\lambda}} - \boldsymbol{\theta}_{0r}^{T}\boldsymbol{x}^{*}{}_i\tilde{y}_i\right)^{+}\right]^{-2}$. Now let

$$\hat{\boldsymbol{M}}_n = \sum_{r=1}^{h-1}\boldsymbol{\psi}_r\boldsymbol{\psi}_r^{T} \quad\text{and}\quad \boldsymbol{M}_0 = \sum_{r=1}^{h-1}\boldsymbol{\psi}_{0r}\boldsymbol{\psi}_{0r}^{T}. \tag{26}$$

Then it can be shown that

$$\hat{\boldsymbol{M}}_n = \boldsymbol{M}_0 + \sum_{r=1}^{h-1}\{\boldsymbol{\psi}_{0r}^{T}\mathrm{D}(\boldsymbol{\theta}_{0r},\boldsymbol{z}) + \mathrm{D}^{T}(\boldsymbol{\theta}_{0r},\boldsymbol{z})\boldsymbol{\psi}_{0r} + \mathrm{D}(\boldsymbol{\theta}_{0r},\boldsymbol{z})\mathrm{D}^{T}(\boldsymbol{\theta}_{0r},\boldsymbol{z})\}, \tag{27}$$

where $\mathrm{D}(\boldsymbol{\theta}_{0r},\boldsymbol{z}) = -n^{-1}\boldsymbol{F}_r\sum_{i=1}^{n}\tilde{\mathrm{B}}_i(\boldsymbol{z}) + o_p(n^{-1/2})$.

Having shown consistency when p is fixed, we now turn into demonstrating consistency when both n and p tend to infinity We show that as long as $p < n$ we get a consistent estimator even when p diverges.

Theorem 5. *Let $\boldsymbol{\theta}_0 = (\boldsymbol{\psi}_0^T, t_0)^T$ be the minimizer of $E[(\boldsymbol{\theta}, \boldsymbol{Z})]$. Suppose for each $\tilde{y} = -1, 1$, the distribution of $\boldsymbol{X}|\tilde{Y} = \tilde{y}$ is dominated by the Lebesgue measure and $E \cdot [||\boldsymbol{X}||^2] < \infty$. Then $\hat{\boldsymbol{\theta}}$ is a consistent estimate of $\boldsymbol{\theta}_0$ as long as $p < n$ as p and n tend to infinity.*

Proof. To begin, we first state the following identity:

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \leq \sqrt{p} \max_i |[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]_i|. \tag{28}$$

Using this and (23) we can write

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \leq \sqrt{p} \max_i \left| n^{-1} \boldsymbol{H}_i^{-1} \sum_{j=1}^{n} B_j(\boldsymbol{z}) \right| + o_p(\sqrt{p/n}). \tag{29}$$

We know the first term on the right tends to 0 as $n \to \infty$, by the consistency of sample mean. Therefore, $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$ if $o_p(\sqrt{p/n}) \to 0$ as $n \to \infty$. Hence we require p to remain less than n. ■

4 | NONLINEAR PDWD

In this section we turn our attention to the extension of this method to the nonlinear case. Let $\mathcal{H}$ be a Hilbert space of functions of $\boldsymbol{X}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Similar to the linear case, the objective function, $\Lambda(\psi, t) : \mathcal{H} \times \mathbb{R} \mapsto \mathbb{R}^+$, takes the form

$$\begin{aligned}\Lambda(\psi, t) &= \mathrm{var}(\psi(\boldsymbol{X})) + \mathrm{E}\left[\left[\tilde{Y}(\psi(\boldsymbol{X}) - \mathrm{E}[\psi(\boldsymbol{X})] - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\psi(\boldsymbol{X}) - \mathrm{E}[\psi(\boldsymbol{X})] - t) \right)^+ \right]^{-1} \right. \\ &\quad \left. + \lambda \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\psi(\boldsymbol{X}) - \mathrm{E}[\psi(\boldsymbol{X})] - t) \right)^+ \right],\end{aligned} \tag{30}$$

where $\tilde{Y}$ is defined as in (6). Now define $\langle f_1, \Sigma f_2 \rangle_{\mathcal{H}} = \mathrm{cov}[f_1(\boldsymbol{X}), f_2(\boldsymbol{X})]$, for any $f_1, f_2 \in \mathcal{H}$, where $\Sigma : \mathcal{H} \mapsto \mathcal{H}$ is the covariance operator. Therefore (30) can be rewritten as

$$\begin{aligned}\Lambda(\psi, t) = \langle \psi, \Sigma \psi \rangle_{\mathcal{H}} + \mathrm{E} &\left[\left[\tilde{Y}(\psi(\boldsymbol{X}) - \mathrm{E}[\psi(\boldsymbol{X})] - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\psi(\boldsymbol{X}) - \mathrm{E}[\psi(\boldsymbol{X})] - t) \right)^+ \right]^{-1} \right. \\ &\left. + \lambda \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}(\psi(\boldsymbol{X}) - \mathrm{E}[\psi(\boldsymbol{X})] - t) \right)^+ \right].\end{aligned} \tag{31}$$

Lemma 5. *Suppose the mapping $\mathcal{H} \to L_2(P_{\boldsymbol{X}}), f \mapsto f$ is continuous. Then for each fixed t in $\mathbb{R}$, the function $\psi \mapsto \Lambda(\psi, t)$ is continuous with respect to the $L_2(P_{\boldsymbol{X}})$-norm.*

Proof. Let ψ_1 and ψ_2 be two members of $L_2(P_{\boldsymbol{X}})$. Then

$$|\Lambda(\psi_2, t) - \Lambda(\psi_1, t)| \leq \Big| \mathrm{var}[\psi_2(\boldsymbol{X})] - \mathrm{var}[\psi_1(\boldsymbol{X})]|$$

$$
\begin{aligned}
&+\mathrm{E}\left|\left[\tilde{Y}(\psi_2(\boldsymbol{X})-t)+\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}(\psi_2(\boldsymbol{X})-t)\right)^{+}\right]^{-1}\right.\\
&-\left[\tilde{Y}(\psi_1(X)-t)+\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}(\psi_1(\boldsymbol{X})-t)\right)^{+}\right]^{-1}\\
&\left.+\lambda\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}(\psi_2(\boldsymbol{X})-t)\right)^{+}-\lambda\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}(\psi_1(\boldsymbol{X})-t)\right)^{+}\right|.
\end{aligned}
$$

We start by considering the first term on the right-hand side. This gives

$$
\begin{aligned}
|\mathrm{var}[\psi_2(\boldsymbol{X})]-\mathrm{var}[\psi_1(\boldsymbol{X})]| &= |\mathrm{var}[\psi_2(\boldsymbol{X})-\psi_1(\boldsymbol{X})+\psi_1(\boldsymbol{X})]-\mathrm{var}[\psi_1(\boldsymbol{X})]|\\
&= |\mathrm{var}[\psi_2(\boldsymbol{X})-\psi_1(\boldsymbol{X})]+2\mathrm{cov}[\psi_2(\boldsymbol{X})-\psi_1(\boldsymbol{X}),\psi_1(\boldsymbol{X})]|\\
&\le |\mathrm{var}[\psi_2(\boldsymbol{X})-\psi_1(\boldsymbol{X})]|+2|\mathrm{var}[\psi_2(\boldsymbol{X})-\psi_1(\boldsymbol{X})]\mathrm{var}[\psi_1(\boldsymbol{X})]|^{1/2}\\
&\le \|\psi_2-\psi_1\|_{L_2P_X}^2+2\|\psi_2-\psi_1\|_{L_2P_X}\|\psi_1\|_{L_2P_X}.
\end{aligned}
$$

Before we consider the remaining terms of the above equation, we first note, for $a, b \in \mathbb{R}$ and $c > 0$ we have

$$
|[b+(c-b)^+]^{-1}-[a+(c-a)^+]^{-1}+c^{-2}(c-b)^+-c^{-2}(c-a)^+|\le c^{-2}|a-b|.
$$

Therefore, last terms become

$$
\begin{aligned}
|\Lambda(\psi_2,t)-\Lambda(\psi_1,t)| &\le \|\psi_2-\psi_1\|_{L_2P_X}^2+2\|\psi_2-\psi_1\|_{L_2P_X}\|\psi_1\|_{L_2P_X}\\
&\quad+\lambda\mathrm{E}|\tilde{Y}(\psi_2(\boldsymbol{X})-t)-\tilde{Y}(\psi_1(\boldsymbol{X})-t)|\\
&= \|\psi_2-\psi_1\|_{L_2P_X}^2+2\|\psi_2-\psi_1\|_{L_2P_X}\|\psi_1\|_{L_2P_X}+\lambda\mathrm{E}|\psi_2(\boldsymbol{X})-\psi_1(\boldsymbol{X})|\\
&\le \|\psi_2-\psi_1\|_{L_2P_X}^2+2\|\psi_2-\psi_1\|_{L_2P_X}\|\psi_1\|_{L_2P_X}+\lambda\|\psi_2-\psi_1\|_{L_2P_X}\\
&= \|\psi_2-\psi_1\|_{L_2P_X}(\|\psi_2-\psi_1\|_{L_2P_X}+2\|\psi_1\|_{L_2P_X}+\lambda).
\end{aligned}
$$

Therefore $|\Lambda(\psi_2,t)-\Lambda(\psi_1,t)|\to 0$ as $\|\psi_2-\psi_1\|\to 0$. ■

Following the definition in Li et al. (2011) we say that a function $\psi \in \mathcal{H}$ is unbiased for nonlinear SDR if it has a version that is measurable $\psi\{\boldsymbol{X}\}$. Using this then we prove the following theorem which proves that the minimizer of the objective function (31) estimates the CS.

Theorem 6. *Suppose the mapping $\mathcal{H} \to L_2(P_X), f \mapsto f$ is continuous and*

1. *$\mathcal{H}$ is a dense subset of $L_2(P_X)$;*
2. *$Y \perp\!\!\!\perp \boldsymbol{X}|\phi(\boldsymbol{X})$*

If (ψ^*, t^*) minimizes (31) among all $(\psi, t) \in \mathcal{H} \times \mathbb{R}$, then $\psi^*(\boldsymbol{X})$ is unbiased.

Proof. Similar to Theorem 1, we have

$$
\mathrm{E}\left[\left[\tilde{Y}(\psi(\boldsymbol{X})-\mathrm{E}[\psi(\boldsymbol{X})]-t)+\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}(\psi(\boldsymbol{X})-\mathrm{E}[\psi(\boldsymbol{X})]-t)\right)^{+}\right]^{-1}\right.
$$

$$+\lambda\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}(\psi(\boldsymbol{X})-\mathrm{E}[\psi(\boldsymbol{X})]-t)\right)^{+}\Bigg]$$
$$\geq\left[\tilde{Y}(\mathrm{E}[\psi(\boldsymbol{X})-\mathrm{E}[\psi(\boldsymbol{X})]|\phi(\boldsymbol{X})]-t)+\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}\mathrm{E}[\psi(\boldsymbol{X})-\mathrm{E}[\psi(\boldsymbol{X})]|\phi(\boldsymbol{X})]-t)\right)^{+}\right]^{-1}$$
$$+\lambda\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}(\mathrm{E}[\psi(\boldsymbol{X})-\mathrm{E}[\psi(\boldsymbol{X})]|\phi(\boldsymbol{X})]-t)\right)^{+},$$

and

$$\mathrm{var}[\psi(\boldsymbol{X})]\geq\mathrm{var}[\mathrm{E}[\psi(\boldsymbol{X})|\phi(\boldsymbol{X})]],$$

where the second equation becomes strict if $\mathrm{E}[\mathrm{var}[\psi(\boldsymbol{X})|\phi(\boldsymbol{X})]]>0$. The equality $\mathrm{E}[\mathrm{var}[\psi(\boldsymbol{X})|\phi(\boldsymbol{X})]]=0$ is equivalent to there being a version of ψ that is measurable with respect to $\sigma\phi(\boldsymbol{X})$.

Hence if there is no version of ψ that is measurable with respect to $\sigma\{\boldsymbol{X}\}$, then

$$\Lambda(\psi,t)>\Lambda(\mathcal{L}(\psi),t),$$

where $\mathcal{L}(\psi)$ denotes the function $\mathrm{E}[\psi(\boldsymbol{X})-\mathrm{E}[\psi(\boldsymbol{X})]|\phi(\boldsymbol{X})]$. Since $\mathcal{H}\subset L_2(P_{\boldsymbol{X}})$, ψ belongs to $L_2(P_{\boldsymbol{X}})$, for any $\epsilon>0$, there is a $\psi_1\in\mathcal{H}$ such that

$$||\psi_1-\mathcal{L}(\psi)||_{L_2(P_{\boldsymbol{X}})}<\epsilon.$$

By Lemma 5, we can choose ϵ to be sufficiently small so that $\Lambda(\psi,t)>\Lambda(\psi_1,t)$, which means ψ cannot be ψ^*. ■

4.1 | Estimation algorithm

Let $\mathcal{H}$ be a linear space of functions from $\Omega_{\boldsymbol{X}}$ to $\mathbb{R}$ spanned by $\mathcal{F}_n=\{\psi_1,\ \dots\ ,\psi_n\}$. These functions are chosen, such that, $\mathrm{E}_n[\phi_i(\boldsymbol{X})]=0$. Let

$$\boldsymbol{\Psi}=\begin{pmatrix}\psi_1(\boldsymbol{X}_1) & \dots & \psi_1(\boldsymbol{X}_n)\\ \vdots & \ddots & \vdots\\ \psi_n(\boldsymbol{X}_1) & \dots & \psi_n(\boldsymbol{X}_n)\end{pmatrix}.$$

Hence the sample version of (31) becomes

$$\hat{\Lambda}(\boldsymbol{c})=\boldsymbol{c}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{c}+\frac{1}{n}\sum_{i=1}^{n}\left[\left[\tilde{Y}_i(\boldsymbol{\Psi}_i^T\boldsymbol{c}-t)+\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}_i(\boldsymbol{\Psi}_i^T\boldsymbol{c}-t)\right)^{+}\right]^{-1}+\lambda\left(\frac{1}{\sqrt{\lambda}}-\tilde{Y}_i(\boldsymbol{\Psi}_i^T\boldsymbol{c}-t)\right)^{+}\right],\tag{32}$$

where $\boldsymbol{c} \in \mathbb{R}^n$ and $\boldsymbol{\Psi}_i$ the ith column of $\boldsymbol{\Psi}$. This problem differs from the kernel objective function, given in Wang and Zou (2018), where $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ is replaced by the kernel matrix $\boldsymbol{K}_n = \{\kappa(i,j) : i,j = 1, \dots , n\}$ for some positive definite bivariate mapping $\kappa : \Omega_{\boldsymbol{X}} \times \Omega_{\boldsymbol{X}} \to \mathbb{R}$. For the function class $\mathcal{H}$, the reproducing kernel Hilbert space is based on the mapping κ. Many choices of κ exist, some of the more popular choices are the radial basis kernel, the polynomial kernel and many more. To be more exact Wang and Zou (2018) proposed the use of the following equation iteratively until convergence for solving their objective function in the classification framework:

$$\begin{pmatrix} t^{(m+1)} \\ \boldsymbol{c}^{(m+1)} \end{pmatrix} = \begin{pmatrix} t^{(m)} \\ \boldsymbol{c}^{(m)} \end{pmatrix} - \frac{n}{4}\boldsymbol{P}^{-1}(\lambda) \begin{pmatrix} \boldsymbol{1}^T\boldsymbol{z} \\ \boldsymbol{K}_n^T\boldsymbol{z} + 2\lambda\boldsymbol{K}_n\boldsymbol{c}^{(m)} \end{pmatrix},$$

where $\boldsymbol{z} = (z_1, \dots , z_n)^T$,

$$z_i = \begin{cases} -\frac{\tilde{y}_i}{n}, & \tilde{y}_i(t + \boldsymbol{K}\boldsymbol{c}^{(m)}) \le \frac{1}{2} \\ -\frac{1}{(n\tilde{y}_i(t+\boldsymbol{K}\boldsymbol{c}^{(m)}))^2}\left(\frac{1}{2}\right)^2, & \tilde{y}_i(t + \boldsymbol{K}\boldsymbol{c}^{(m)}) > \frac{1}{2} \end{cases}.$$

Finally,

$$\boldsymbol{P}^{-1}(\lambda) = \begin{pmatrix} n & \boldsymbol{1}^T\boldsymbol{K}_n \\ \boldsymbol{K}_n^T\boldsymbol{1} & \boldsymbol{K}_n^T\boldsymbol{K}_n + \frac{n\lambda}{2}\boldsymbol{K}_n \end{pmatrix}^{-1}.$$

This of course was to address the classification problem that DWD is proposed for. Since we are interested for dimension reduction, our objective function is different and takes the form (32). By replacing $\boldsymbol{K}_n$ with $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ we have the formulas for the dimension reduction framework to be:

$$\begin{pmatrix} t^{(m+1)} \\ \boldsymbol{c}^{(m+1)} \end{pmatrix} = \begin{pmatrix} t^{(m)} \\ \boldsymbol{c}^{(m)} \end{pmatrix} - \frac{nq}{4}\boldsymbol{P}^{-1}(\lambda) \begin{pmatrix} \boldsymbol{1}^T\boldsymbol{z} \\ \boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{z} + 2\lambda\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{c}^{(m)} \end{pmatrix},$$

where $\boldsymbol{z} = (z_1, \dots , z_n)^T$,

$$z_i = \begin{cases} -\frac{\tilde{y}_i}{n}, & \tilde{y}_i(t + \boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{c}^{(m)}) \le \frac{1}{2} \\ -\frac{1}{(n\tilde{y}_i(t+\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{c}^{(m)}))^2}\left(\frac{1}{2}\right)^2, & \tilde{y}_i(t + \boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{c}^{(m)}) > \frac{1}{2} \end{cases}.$$

Finally,

$$\boldsymbol{P}^{-1}(\lambda) = \begin{pmatrix} n & \boldsymbol{1}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi} \\ \boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{1} & \boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\Psi} + \frac{n\lambda}{2}\boldsymbol{\Psi}^T\boldsymbol{\Psi} \end{pmatrix}^{-1}.$$

As we can see the solution now does not depend on $\boldsymbol{K}_n$ but rather on $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$. Now, let $\boldsymbol{Q}_n = \boldsymbol{I}_n - \boldsymbol{J}_n/n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{J}_n$ is an $n \times n$ matrix with entries 1. The following Proposition was used in Li et al. (2011) to show that the eigenfucntions of the operator Σ_n can be instead estimated by using the eigenvectors of $\boldsymbol{Q}_n\boldsymbol{K}_n\boldsymbol{Q}_n$.

Proposition 1. *Let* $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)$, $\psi_{\boldsymbol{\omega}} = \sum \omega_i[\kappa(\boldsymbol{x}, \boldsymbol{X}_i) - \mathrm{E}_n[\kappa(\boldsymbol{x}, \boldsymbol{X})]]$. *The following statements are equivalent:*

1. $\boldsymbol{\omega}$ *is an eigenvector of the matrix* $\boldsymbol{Q}_n\boldsymbol{K}_n\boldsymbol{Q}_n$ *with eigenvalue* λ.
2. $\psi_{\boldsymbol{w}}$ *is an eigenfunction of the operator* $\boldsymbol{\Sigma}_n$ *with eigenvalue* λ/n.

If $\lambda \neq 0$, *then either statement implies* $(\psi_{\boldsymbol{\omega}}(\boldsymbol{X}_1), \dots, \psi_{\boldsymbol{\omega}}(\boldsymbol{X}_n)) = \lambda\boldsymbol{\omega}^T$.

As we mentioned above in the derivations for our problem we need $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$, where one can estimate it using the above proposition as $\boldsymbol{\Psi} = \boldsymbol{W} = (\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_n)$. Since $\boldsymbol{\omega}_i$ is an eigenvector of $\boldsymbol{Q}_n\boldsymbol{K}_n\boldsymbol{Q}_n$, $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ becomes the identity matrix. Therefore the objective function in (32) becomes independent of $\boldsymbol{X}$. For this reason we propose a slight modification that does not affect our theoretical results and we replace $\boldsymbol{\Psi}$ in (32) with $\tilde{\boldsymbol{\Psi}} = \boldsymbol{K}_n^{1/2}\boldsymbol{W}$ to reintroduce the dependence of the problem and its solution on $\boldsymbol{X}$. Therefore the objective function we try to solve is:

$$\tilde{\hat{\Lambda}}(\boldsymbol{c}) = \boldsymbol{c}^T\tilde{\boldsymbol{\Psi}}^T\tilde{\boldsymbol{\Psi}}\boldsymbol{c} + \frac{1}{n}\sum_{i=1}^{n}\left[\left[\tilde{Y}_i(\tilde{\boldsymbol{\Psi}}_i^T\boldsymbol{c} - t) + \left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}_i(\tilde{\boldsymbol{\Psi}}_i^T\boldsymbol{c} - t)\right)^+\right]^{-1}\right.$$
$$\left. + \lambda\left(\frac{1}{\sqrt{\lambda}} - \tilde{Y}_i(\tilde{\boldsymbol{\Psi}}_i^T\boldsymbol{c} - t)\right)^+\right]. \tag{33}$$

Notice that, with this modification in DWD, we achieve two things in comparison with the PSVM algorithm. One is that we remove one tuning parameter by not having to estimate k (less than n) basis functions $\boldsymbol{\psi}(\boldsymbol{X})$, that is our $\boldsymbol{\Psi}$ matrix is an $n \times n$ matrix and not a $k \times n$ matrix. The second is the ability to estimate directly the sufficient predictors therefore removing one step in the algorithm.

Therefore, the kernel PDWD estimation algorithm is as follows:

1. (Optional) Marginally standardize $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. This step can be omitted if the components of $\boldsymbol{X}_i$ have similar variances.
2. Choose a kernel κ and create the kernel matrix $\boldsymbol{K}$. Find the eigenvectors $\boldsymbol{\omega}_1, \dots, \omega_n$ of $\boldsymbol{Q}_n\boldsymbol{K}_n\boldsymbol{Q}_n$.Calculate $\tilde{\boldsymbol{\Psi}} = \boldsymbol{K}_n^{1/2}\boldsymbol{W}$.
3. Divide the sample according to LVR or OVA. For each set of slices compute $\boldsymbol{c}_1, \dots, \boldsymbol{c}_{\tilde{h}}$ by solving (33) using the kernel DWD algorithm with $\boldsymbol{K}$ replaced with $\tilde{\boldsymbol{\Psi}}^T\tilde{\boldsymbol{\Psi}}$. For LVR $\tilde{h} = h - 1$ and for OVA $\tilde{h} = \binom{h}{2}$.
4. The sufficient predictors are equivalent to the first d eigenvectors $v_1, \dots, v_d$ of the matrix $\sum_{i=1}^{\tilde{h}} \boldsymbol{c}_i\boldsymbol{c}_i^T$.

5 | NUMERICAL STUDIES

In this section we demonstrate the advantages of PDWD over PSVM through a simulation study and through a real data experiment.

5.1 | Simulation studies

We use the following three synthetic models:

$$\begin{aligned} &\text{Model I:} && Y = X_1 + X_2 + \sigma\epsilon; \\ &\text{Model II:} && Y = \tfrac{X_1}{0.5+(X_2+1)^2} + \sigma\epsilon; \\ &\text{Model III:} && Y = X_1(X_1 + X_2 + 1) + \sigma\epsilon; \end{aligned}$$

where $X \sim \mathrm{N}(\mathbf{0}, I_p)$, $\epsilon \sim \mathrm{N}(0, 1)$ and $\sigma = 0.2$. We choose $n = 100$ and $h = 20$ unless stated otherwise. Although all models use only two predictors we add noise to the data by introducing appropriate number of predictors such that p takes the values $p = 20, 30, 50, 100$. Also, notice that for the first model the effective dimension $d = 1$ and for the other two models, $d = 2$.

We will use the distance method defined in Li et al. (2005) to estimate the performance of the algorithms. Let $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ denote the basis of the central space and let $\hat{\boldsymbol{\beta}}$ be its estimator. Then we estimate the performance of $\hat{\boldsymbol{\beta}}$ as with the following distance measure

$$\mathrm{dist}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \|\boldsymbol{P}_{\beta} - \boldsymbol{P}_{\hat{\beta}}\|,$$

where $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$, that is the projection matrix, and $\|\cdot\|$ is the Frobenius norm.

We compare our method with PSVM and the results are shown in Table 1. The results show that PDWD and PSVM have similar performance for values of p close to n or close to 0 but for values in between PDWD has a clear advantage. In the classification literature (see Marron et al., 2007) it was shown that DWD clearly outperforms SVM for larger p due to the SVM suffering from data piling. The fact that here the two methods are equivalent as p tends to n we believe is due to the different nature of the problem. Remember that while in classification the performance of the classifier is measured on the percentage of correctly classified points, which will be hindered by data piling, here we are interested for dimension reduction through hyperplane alignment. It seems that in the dimension reduction framework data piling actually “hinders” the performance of both PSVM and PDWD by causing them to overfit the data and that’s why the performance of the two algorithms is becoming equivalent as p gets closer to n.

5.2 | Computational time

As was mentioned earlier using a newly developed algorithm for DWD by Wang and Zou (2018) there is a computational advantage as the computation of Principal DWD is much less than the one for PSVM. We emphasize here that when Li et al. (2011) proposed PSVM they identified that the fact that PSVM needs quadratic programming leads to higher computational cost and that was probably the only disadvantage of PSVM over earlier methods which were based on inverse moments. As Figure 3 indicates there is a huge difference in time as n increases (and p is constant) while the difference stays relatively the same as p increases (and n is constant).

5.3 | Order determination

We consider the three models we discussed before. As was mentioned earlier, Model I has effective dimension 1 and Models II and III have effective dimension 2. We run 1,000 simulation

TABLE 1 Comparison of estimation performance between PDWD and Principal supporting vector machine (PSVM). The table reports the mean performance of 100 iterations (SEs in parenthesis) for the two methods

Model	p	PSVM	PDWD
I	20	0.20 (0.040)	0.17 (0.047)
	30	0.29 (0.053)	0.24 (0.046)
	50	0.45 (0.078)	0.38 (0.071)
	100	1.33 (0.082)	1.31 (0.092)
II	20	0.99 (0.199)	0.95 (0.184)
	30	1.35 (0.092)	1.17 (0.133)
	50	1.54 (0.138)	1.45 (0.090)
	100	1.95 (0.034)	1.95 (0.035)
III	20	1.43 (0.281)	1.29 (0.219)
	30	1.64 (0.142)	1.47 (0.156)
	50	1.87 (0.073)	1.72 (0.114)
	100	1.97 (0.022)	1.97 (0.022)

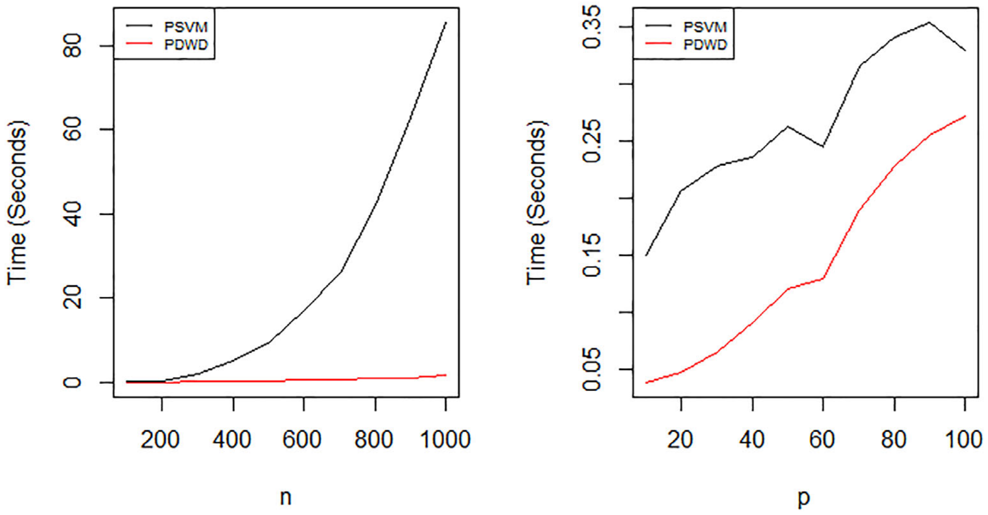


FIGURE 3 Left panel: time of two algorithms as n increases ($p = 100$); right panel: time of two algorithms as p increases ($n = 1{,}000$) [Color figure can be viewed at wileyonlinelibrary.com]

experiments with $n = 300$, $\sigma = 0.2$ and $H = 20$. Table 2 shows the proportion of correct estimates as p increases. This is a very promising result as it demonstrates that the performance of the algorithm does not suffer a lot when the dimension is increased, instead we can see that as p increase the number of correct estimates for Models II and III decreases slightly but remains high.

TABLE 2 Proportion of correct estimations of d in 1,000 simulations using the ladle estimator for the two models

	p		
Model	**10**	**30**	**50**
I	1.000	1.000	1.000
II	0.999	0.982	0.969
III	0.999	0.983	0.971

TABLE 3 Comparison of estimation performance between KPSVM and KPDWD. The table reports the mean performance of 100 iterations (standard errors in parenthesis) for the two methods

Model	***p***	**KPSVM**	**KPDWD**
III	10	0.91 (0.020)	0.97 (0.009)
	20	0.86 (0.026)	0.97 (0.011)
	30	0.82 (0.039)	0.96 (0.008)
IV	10	0.90 (0.027)	0.92 (0.023)
	20	0.82 (0.034)	0.93 (0.017)
	30	0.76 (0.037)	0.93 (0.016)

5.4 | Kernel PDWD

We consider models III and)

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

where $X \sim \mathrm{N}(\mathbf{0}, I_p)$, $\epsilon \sim \mathrm{N}(0,1)$. In this section we compare kernel PDWD (KPDWD) with kernel PSVM (KPSVM). In the same format as in the nonlinear comparisons in Li et al. (2011) we will use the absolute value of Spearman's correlation to measure the closeness of the predictors to the true predictors.

We choose $n = 100$, $\lambda = 1$, $p = 10, 20, 30$, and $h = 20$. For Spearman's correlation, the numbers are between 0 and 1, where larger numbers indicate a higher performance. Using the Gaussian kernel basis, Table 3 shows that kernel PDWD outperforms kernel PSVM for both models. It is also clear that the performance of kernel PDWD remains good as p increases.

5.5 | Real dataset: Concrete slump test

We now turn our attention to real data analysis. Our aim is to assess the effect of introducing random variables to the data. Consider the Concrete slump data analyzed in Yeh (1998). We have evaluated the response variable Compressive Strength. There are seven predictor variables called cement, slag, fly ash, water, superplasticizer (SP), coarse aggregate, and fine aggregate. The data

TABLE 4 Distances as extra predictors are added in the dataset. Each column adds a different number of predictors and we report the distance of the estimated Central Space (CS) from the "oracle" CS, that is, the one when only the original predictors are used

	3	10	30	50	90
PDWD	0.007	0.039	0.101	0.111	0.195
PSVM	0.16	0.293	0.349	0.369	0.373
Compared	0.274	0.027	0.064	0.047	0.151

Abbreviation: PSVM, principle support vector machine.

consists of 103 samples and we fix $\lambda = 0.1$ and $H = 20$. We first run the two methods and we calculate

$$\hat{\boldsymbol{\beta}}^T_{\text{PDWD}} = (0.01, -0.001, 0.009, -0.024, 0.048, -0.005, -0.003),$$
$$\hat{\boldsymbol{\beta}}^T_{\text{PSVM}} = (0.013, 0.002, 0.01, -0.02, 0.033, -0.003, -0.001),$$

which span the CS estimated by each method. Then we add extra predictors in the dataset, which are randomly distributed from a standard Normal distribution, and calculate the new $\boldsymbol{\beta}$'s that span the Central Space using the two methods. We calculate the distance of the new vector from the original one, that is the one that was calculated based on the original predictors. Table 4 shows the distances between the estimator and the original estimator for each of the two methods, PDWD and PSVM, and for different number of added predictors (3, 10, 30, 50, 90). We can see that the estimator of the PDWD moves a lot less than the PSVM predictor. The third line of results in Table 4 which is labeled "Compared" shows the distance between the estimated PDWD and PSVM directions. It is clear that the two directions start far away for $p = 3$ and they get closer (meaning they are estimating similar directions) as p increases. Then it breaks down again when $p = 90$.

6 | DISCUSSION

In this paper we propose a different classification-based algorithm for SDR. The newly proposed algorithm is based on DWD proposed by Marron et al. (2007). The main advantages of the new method are, first its performance is not affected by extra noninformative predictors in our dataset and second is computationally faster than previously proposed SVM-based algorithms. The theoretical properties of the new method are studied in detail. We are able to prove that asymptotic theory holds for fixed p and for diverging p as long as $p < n$.

While DWD was proposed in the classification framework to address the problem of data piling that SVM is suffering from in very high dimensions, we can see in this paper that the same cannot be said for the SDR framework the PDWD and PSVM were proposed for. Instead we can see that PDWD outperforms PSVM for low-dimensional problems while the estimation of the two algorithms comes closer as the dimension of the problem increases. Also, the advantage in computational cost using PDWD as the number of observations increases it can be crucial in an era where massive datasets are becoming increasingly popular.

ACKNOWLEDGEMENTS

The authors would like to thank an Associate Editor and a reviewer for their useful comments which improved the presentation of the manuscript. The work of H. Randall has been partially funded by EPSRC.

ORCID

Andreas Artemiou https://orcid.org/0000-0002-7501-4090

REFERENCES

Artemiou, A., & Dong, Y. (2016). Principal L*q* support vector machines for sufficient dimension reduction. *Electronic Journal of Statistics*, *10*, 783–805.

Artemiou, A., & Shu, M. (2014). *A cost based reweighted scheme of principal support vector machine*. In *Topics in Nonparametric Statistics* (Vol. *74*, pp. 1–22). New York, NY: Springer.

Cook, R. D., & Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, *86*, 316, MR1137117–342.

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, *37*, 1871, MR2533474–1905.

Hjort, N. L., & Pollard, D. (1993). *Asymptotics for minimizers of convex processes: Technical report*. New Haven, CO: Department of Statistics Preprint, Yale University.

Jiang, B., Zhang, X., & Cai, T. (2008). Estimating the confidence interval for prediction errors of support vector machine classifiers. *Journal of Machine Learning Research*, *9*, 521, MR2417245–540.

Lam, X. Y., Marron, J. S., Sun, D., & Toh, K. C. (2018). Fast algorithms for large-scale generalized distance weighted discrimination. *Journal of Computational and Graphical Statistics*, *27*(2), 368–379.

Li, B., Artemiou, A., & Li, L. (2011). Principal support vector machine for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, *39*, 3182, MR3012405–3210.

Li, B., & Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, *102*, 997, MR2354409–1008.

Li, B., Zha, H., & Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, *33*, 1580, MR2166556–1616.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, *86*, 316, MR1137117–342.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, *87*, 1025, MR1209564–1039.

Luo, W., & Li, B. (2016. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, *103*(4), 875–887.

Marron, J. S., Todd, M. J., & Ahn, J. (2007). Distance weighted discrimination. *Journal of the American Statistical Association.*, *102*, 1267–1271.

Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J., & Marron, J. S. (2010). Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statiestical Association*, *105*, 401–414.

Qiao, X., & Zhang, L. (2015). Flexible high-dimensional classification machines and their asymptotic properties. *Journal of Machine Learning Research*, *16*, 1547–1572.

Shin, S. J., & Artemiou, A. (2017). Penalized principal logistic regression for sparse sufficient dimension reduction. *Computational Statistics and Data Analysis*, *111*, 48–58.

Shin, S. J., Wu, Y., Zhang, H. H., & Liu, Y. (2017). Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*, *104*, 67–81.

Smallman, L., & Artemiou, A. (2017). A study on imbalance support vector machine algorithms for sufficient dimension deduction. *Communications in Statistics, Theory and Methods*, *46*, 2751–2763.

Wang, B., & Zou, H. (2016). Sparse distance weighted discrimination. *Journal of Computational and Graphical Statistics*, *25*, 826–838.

Wang, B., & Zou, H. (2018). Another look at distance-weighted discrimination. *Journal of the Royal Statistical Society, Series B*, *80*, 177–198.

Wu, H. M. (2008). Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics*, *17*, 590, MR2528238–610.

Ye, Z., & Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, *98*, 968–979.

Yeh, I.-C. (1998). Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research*, *28*(12), 1797–1808.

Yeh, Y.-R., Huang, S.-Y., & Lee, Y.-Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1590–1603.

Yin, X., Li, B., & Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, *99*, 1733–1757.

Zhou, J., & Zhu, L. (2016). Principal minimax support vector machine for sufficient dimension reduction with contaminated data. *Computational Statistics and Data Analysis*, *94*, 33–48.

How to cite this article: Randall H, Artemiou A, Qiao X. Sufficient dimension reduction based on distance-weighted discrimination. *Scand J Statist*. 2020;1–26. https://doi.org/10.1111/sjos.12484