Communications for Statistical Applications and Methods
2024, Vol. 31, No. 2, 235–246

https://doi.org/10.29220/CSAM.2024.31.2.235
Print ISSN 2287-7843 / Online ISSN 2383-4757

A concise overview of principal support vector machines and its generalization

Jungmin Shin[a], Seung Jun Shin[1,a]

[a]Department of Statistics, Korea University, Korea

Abstract

In high-dimensional data analysis, sufficient dimension reduction (SDR) has been considered as an attractive tool for reducing the dimensionality of predictors while preserving regression information. The principal support vector machine (PSVM) (Li *et al.*, 2011) offers a unified approach for both linear and nonlinear SDR. This article comprehensively explores a variety of SDR methods based on the PSVM, which we call principal machines (PM) for SDR. The PM achieves SDR by solving a sequence of convex optimizations akin to popular supervised learning methods, such as the support vector machine, logistic regression, and quantile regression, to name a few. This makes the PM straightforward to handle and extend in both theoretical and computational aspects, as we will see throughout this article.

Keywords: sufficient dimension reduction, principal support vector machine, principal machine, M-estimation, convex optimization

1. Introduction

Statistics is essentially about how to reduce the dimension of data without loss of parameter information of interest. There are tons of statistical methods for dimension reduction, and they are broadly categorized into two types: Feature selection and feature extraction. Feature selection, often referred to as variable selection tries to identify a subset of variables to achieve the dimension reduction. Since the seminal LASSO (Tibshirani, 1996) was proposed, penalization becomes a canonical approach for feature selection in supervised learning problems. Popular examples of penalized variable selection methods includes but not limited to Fan and Li (2001), Zou and Hastie (2005), Zou (2006), Hastie *et al.* (2009), and Zhang (2010). Feature extraction, on the other hand, seeks a transformation of variables. The feature extraction is more popular in an unsupervised learning context and principal component analysis (PCA) (Pearson, 1901) is a canonical example.

Sufficient dimension reduction (SDR) is a feature extraction method for supervised learning, which reduces the predictor dimension while preserving information about the response. Unlike the penalized approaches for the feature selection, SDR is often model-free and has gained great interest in many applications. SDR seeks a matrix $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_d) \in \mathbb{R}^{p\times d}$, or more precisely span$\{\mathbf{B}\}$, which satisfies the following assumption:

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^\top\mathbf{X}, \tag{1.1}$$

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Seongbuk-Gu, Seoul 02841, Korea. E-mail: sjshin@korea.ac.kr

Published 31 March 2024 / journal homepage: http://csam.or.kr

where $\perp\!\!\!\perp$ denotes statistical independence. span{**B**} satisfying (1.1) is referred to as dimension reduction subspace (DRS). When $d \ll p$, SDR projects the predictors to a lower dimensional subspace while preserving information of Y contained in $\mathbf{X}$. In other word, $\mathbf{B}^\top\mathbf{X}$ captures all the regression information of Y contained in $\mathbf{X}$. SDR under (1.1) is often referred to as the linear SDR since $\mathbf{B}^T\mathbf{X}$ is a linear function of $\mathbf{X}$. Note that DRS is not unique since any linear transformation of $\mathbf{B}$ satisfying (1.1) is DRS. Thereby, the intersection of all DRSes is defined as the central subspace, which is the smallest denoted by $\mathcal{S}_{Y|\mathbf{X}}$, and is the primal identifiable target of interest in SDR. The dimension of the central subspace $\dim(\mathcal{S}_{Y|\mathbf{X}}) = d$ referred to as a structural dimension is another important quantity to be estimated. Since the $\mathcal{S}_{Y|\mathbf{X}}$ exists under mild conditions (Yin *et al.*, 2008), we assume $\mathbf{B}$ in (1.1) spans $\mathcal{S}_{Y|\mathbf{X}}$.

There are numerous ways to estimate $\mathbf{B}$, the basis matrix of $\mathcal{S}_{Y|\mathbf{X}}$, and the earliest proposal is the sliced inverse regression (SIR) (Li, 1991). The idea of SIR is simple and based on the following result

$$E(\mathbf{X} \mid Y) \in \mathcal{S}_{Y|\mathbf{X}}, \tag{1.2}$$

and Li (1991) proposed to estimate the inverse regression function $E(\mathbf{X}|Y)$ by slicing Y, which explains its name, SIR. The SIR sparked numerous subsequent "inverse" SDR methods based on the conditional moment of $\mathbf{x}$ given Y, such as the slice averaged variance estimation (SAVE) (Cook and Weisberg, 1991), contour regression (Li *et al.*, 2005), directional regression (Li and Wang, 2007), cumulative slicing estimation (Zhu *et al.*, 2010), among many others. Since the inverse moment is easily obtained by slicing, the inverse methods are computationally efficient to implement. However, they require untestable assumptions on $\mathbf{X}$, such as the linearity condition and the constant variance condition.

As an alternative to the inverse SDR methods, the "forward" SDR method has been proposed. The forward SDR method achieves SDR by using the conditional expectation of $\mathbf{X}$ given Y. The earliest example of the forward SDR method is the principal Hessian direction (pHd) (Li, 1992). The key idea of pHd is that the Hessian matrix of $E(Y|\mathbf{X})$ belongs to the central mean subspace (Li, 1992), denoted by $\mathcal{S}_{E(Y|\mathbf{X})}$. That is,

$$\frac{\partial^2 E(Y \mid \mathbf{X})}{\partial \mathbf{X} \partial \mathbf{X}^T} \subseteq \mathcal{S}_{E(Y|\mathbf{X})}. \tag{1.3}$$

The central mean subspace $\mathcal{S}_{E(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ is the minimal subspace of $\mathbf{X}$ spanned by $\mathbf{B}$ satisfying $E(Y|\mathbf{X}) = E(Y|\mathbf{B}^T\mathbf{X})$, often regarded as an alternative target of $\mathcal{S}_{Y|\mathbf{X}}$. One advantage of the forward approach is that it does not require additional assumptions, such as the linearity condition that is indispensable to inverse methods. There are several forward SDR methods, including a dimension reduction for a multi-index model (Hristache *et al.*, 2001), outer product of gradient method (OPG) (Xia *et al.*, 2002), minimum average variance estimation (MAVE) (Xia *et al.*, 2002), which is a computationally enhanced variant of the OPG estimator. Also, Xia (2007) proposed the dOPG method that extends the idea of the OPG to estimate $\mathcal{S}_{Y|\mathbf{X}}$ by replacing the conditional mean. It requires no strong assumptions on the covariates or the functional relation between regressors and the response variable. Thereafter, Kong and Xia (2014) proposed the adaptive composite quantile approach with OPG method (qOPG) for identifying the $\mathcal{S}_{Y|\mathbf{X}}$ directions exhaustively, and Kang and Shin (2022) recently proposed weighted OPG (wOPG) for SDR in a binary classification. We also refer to Chapter 11 of Li (2018) for a concise overview of the forward SDR methods.

Cook (2007) and Lee *et al.* (2013) generalized the linear SDR (1.1) and introduced the nonlinear SDR which seeks a function $\phi : \mathbb{R}^p \to \mathbb{R}^d$ that satisfies

$$Y \perp\!\!\!\perp \mathbf{X} \mid \phi(\mathbf{X}), \tag{1.4}$$

where $\phi(\cdot)$ is assumed to be a unique modulo injective transformation of $\mathbf{X}$ in order to guarantee its identifiability. The nonlinear SDR methods under (1.4) have been proposed by directly generalizing the idea of the linear SDR methods to nonlinear function via kernel trick. See for example, kernel SIR (Wu, 2008), and kernel canonical correlation analysis developed by Akaho (2001) and Fukumizu *et al.* (2007), to name a few.

In the meantime, Li *et al.* (2011) proposed a novel idea of SDR called principal support vector machine (PSVM) that can solve both linear and nonlinear SDR under a unified framework. The principle idea of the PSVM is simple as follows. First, the response Y is dichotomized and a pseudo response $\tilde{Y}_c$ is defined as 1 if $Y \geq c$ and -1 otherwise. Then, we train a support vector machine with respect to $(\tilde{Y}_c, \mathbf{Z})$, where $\mathbf{Z}$ is a standardized predictor. Li *et al.* (2011) showed that the separating hyperplane obtained by the SVM contains information about $\mathcal{S}_{Y|\mathbf{X}}$.

The beauty of the PSVM comes from the fact that it provides a connection between SDR and the SVM, a conventional supervised learning method. The PSVM can handle both linear and nonlinear SDR in a unified framework, as the SVM does. Moreover, the use of SVM is not essential for estimating $\mathcal{S}_{Y|\mathbf{X}}$ and one can replace the hinge loss with other popular covex loss functions. This leads to a variety of extensions of the PSVM. For example, Artemiou and Dong (2016) proposed the principal L_q-SVM with L_q hinge loss, Shin and Artemiou (2017) proposed the principal logistic loss with binomial likelihood loss, Wang *et al.* (2018) proposed the principal quantile regression with the check loss, and Artemiou *et al.* (2021) proposed the principal least square SVM with the squared loss.

The PSVM appears to be an inverse SDR method that slices data based on Y, but it can also be viewed as a forward method because it repeatedly applies the usual SVM algorithm to estimate the optimal hyperplanes. Some variants of the PSVM, such as principal weighted SVM (PWSVM) (Shin *et al.*, 2017), principal quantile regression (Wang *et al.*, 2018) do not require slicing Y, i.e., $\tilde{Y}_c$, but training a series of weighted supervised learning with original response. In that, those methods can be regarded as forward SDR approaches. In this regard, the PSVM provides a connection between the inverse and forward SDR methods.

This article is to provide a concise but comprehensive overview of the SDR method that stems from the PSVM, which we will call the principal machine (PM). The rest of the article is organized as follows. In Section 2, we start with briefly reviewing PSVM which serves as a building block of PM. In Section 3, we introduce the principal machine, a general form of the PSVM that covers all of its variants. In Section 4, some miscellaneous including computation and large sample properties are described, respectively, both of which turn out straightforward since it is a simple convex optimization. The penalized principal machine is introduced in Section 5. Concluding remarks are followed in Section 6.

2. Principal support vector machine

Support vector machine (SVM) (Vapnik, 1999) is a well-known binary classifier that seeks an optimal hyperplane to separate two classes. Borrowing the idea of the SVM, Li *et al.* (2011) proposed the PSVM for the linear SDR (1.1) that minimizes the following objective function:

$$\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + \lambda E \left[1 - \tilde{Y}_c \left\{ \alpha + \boldsymbol{\beta}^\top (\mathbf{X} - E\mathbf{X}) \right\} \right]_+ , \tag{2.1}$$

where $\tilde{Y}_c$ denotes the dichotomized pseudo response that takes 1 if $Y > c$ and -1 otherwise, $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X})$ is the covariance matrix of $\mathbf{X}$, and $x_+ = \max(0, x)$. A positive constant λ is a cost parameter that balances the goodness of fit and the complexity of the model, and its choice is not overly sensitive for estimating $\mathcal{S}_{Y|\mathbf{X}}$. We remark that (2.1) is a population level objective function of the linear SVM for $(\tilde{Y}_c, \mathbf{Z})$ where $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - E\mathbf{X})$ is a standardized predictor.

Li *et al.* (2011) established Theorem 1 that provides the theoretical foundation of the PSVM.

Theorem 1. *Assume the linearity condition that $E(\mathbf{X}|\boldsymbol{B}^\top\mathbf{X})$ is a linear function of $\boldsymbol{B}^\top\mathbf{X}$, where $\boldsymbol{B}$ is in* (1.1). *Let $(\alpha^*, \boldsymbol{\beta}^*)$ denote the minimizer of* (2.1), *then $\boldsymbol{\beta}^* \in \mathcal{S}_{Y|\mathbf{X}}$ for any given c.*

For a given set of data $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$, the PSVM estimates $\mathcal{S}_{Y|\mathbf{X}}$ as follows by Theorem 1. Given a grid $\min Y_i < c_1 < \cdots < c_H < \max Y_i$, the PSVM solves a sequence of objective function (2.2) for different values of c_h:

$$\boldsymbol{\beta}^\top \widehat{\mathbf{\Sigma}}_n \boldsymbol{\beta} + \frac{\lambda}{n} \sum_{i=1}^{n} \left[1 - \widetilde{Y}_{i,c_h}\left(\alpha + \boldsymbol{\beta}^\top\left(\mathbf{X}_i - \overline{\mathbf{X}}_n\right)\right)\right]_+. \tag{2.2}$$

Let $(\hat{\alpha}_h, \hat{\boldsymbol{\beta}}_h)$ denotes the minimizer of (2.2), which we call the PSVM solution. Then, $\mathcal{S}_{Y|\mathbf{X}}$, is estimated by span$(\widehat{\mathbf{V}})$, where $\widehat{\mathbf{V}} = (\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_d)$ is a $(p \times d)$ matrix whose j^{th} column, $\hat{\mathbf{v}}_j$ is the j^{th} leading eigenvector of $\widehat{\mathbf{M}}_n$ in (2.3).

$$\widehat{\mathbf{M}}_n = \sum_{h=1}^{H} \hat{\boldsymbol{\beta}}_{n,h} \hat{\boldsymbol{\beta}}_{n,h}^\top. \tag{2.3}$$

Li *et al.* (2011) showed that the PSVM performs better than classical SDR methods, such as SIR, SAVE, etc.

Due to its similarity to the SVM, one can readily extend the linear PSVM to its nonlinear counter. Namely, the kernel PSVM minimizes the population objective function:

$$\langle \psi, \Sigma\psi \rangle_{\mathcal{H}} + \lambda E\left[1 - \tilde{Y}\left(\alpha + \psi(\mathbf{X}) - E\psi(\mathbf{X})\right)\right]_+, \tag{2.4}$$

where ψ being a function in a Hilbert space $\mathcal{H}$ of functions of $\mathbf{X}$, $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$, and Σ is the covariance operator. Let (α^*, ψ^*) be the minimizer of (2.4), ψ^* is necessarily a function of the sufficient predictor $\phi(\mathbf{X})$ in (1.4). A sample version of (2.4) can be readily obtained by employing the reproducing kernel Hilbert space (RKHS) (Wahba, 1999), just like the SVM.

3. Principal machines: A generalization of the PSVM

The PSVM opens up a new class of SDR approaches by connecting the conventional supervised learning algorithm to the SDR context, and achieves both linear and nonlinear SDR in a unified framework. Theorem 1 provides the theoretical foundation of the PSVM and it turns out that we only require the convexity of the hinge loss to prove Theorem 1. It leads to numerous variations of the PSVM by replacing the hinge loss with other convex loss functions.

There are two types of principal machines (PM) depending on how to generalize the idea of the PSVM. One is response-based principal machine (RPM) and the other is loss-based principal machine (LPM).

3.1. Response-based principal machine (RPM)

Note that SVM computes multiple solutions to estimate $\mathcal{S}_{Y|\mathbf{X}}$ by generating multiple pseudo responses $\tilde{Y}_c$ for different values of c. Variants of RPM are constructed by simply replacing the hinge loss of the PSVM with other convex loss functions.

Table 1: Different types of the convex loss functions for response-based principal machine along with the corr esponding methods

Method Loss	PSVM	L$_q$PSVM	PLR	PLSSVM
$L(\tilde{Y}_c, f)$	$\left[1-\tilde{Y}_c f\right]_+$	$\left(\left[1-\tilde{Y}_c f\right]_+\right)^q$	$\ln\left\{1+e^{(-\tilde{Y}_c f)}\right\}$	$\left(1-\tilde{Y}_c f\right)^2$

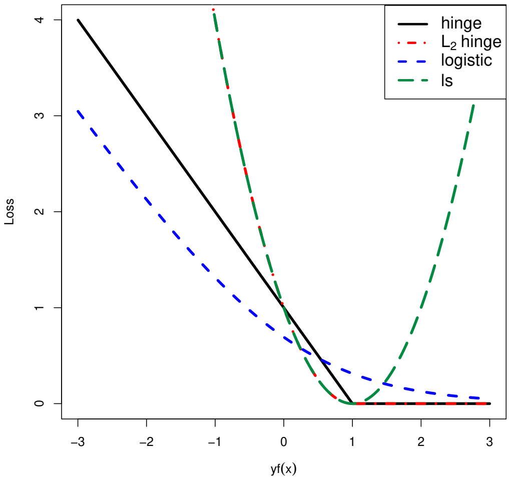

Figure 1: *SVM(hinge), L_2hinge (colored in red), logistic (binomial log-likelihood, colored in blue), and least squares (ls, colored in green) loss functions are visualized. The loss functions are smooth, convex, and continuously differentiable functions across the support, except for the hinge loss (solid black).*

Assuming $\mathbf{X}$ is centered without loss of generality, RPM minimizes the following objective function:

$$\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + \lambda E\left\{L\left(\tilde{Y}_c, \alpha + \boldsymbol{\beta}^\top \mathbf{X}\right)\right\}, \tag{3.1}$$

where $L(Y, f)$ denotes a convex loss function for a linear model $f = \alpha + \boldsymbol{\beta}^\top \mathbf{X}$ to learn Y. We note that (3.1) yields different minimizers as the (pseudo) response $\tilde{Y}_c$ changes while the loss function L is fixed, like the name of RPM. Some examples of RPM are presented below and each corresponding loss functions are summarized in Table 1 and visualized in Figure 1.

The PSVM (Li *et al.*, 2011) is a canonical method of RPM. It adopts the hinge loss function for $L(Y, f)$ as we discussed earlier in (2.1). By noticing that the solution of SVM may not be unique with respect to α, Artemiou and Dong (2016) proposed principal L$_q$PSVM ($q > 1$) in both linear and nonlinear manners. The loss function of L$_q$SVM is found in Table 1. The objective function of principal L$_q$SVM is strictly convex for both $(\alpha, \boldsymbol{\beta})$, which guarantees the uniqueness of the solution.

The advantages of the logistic regression over SVM are obvious since its loss function is smooth and strictly convex. Not only does it entail simpler asymptotic results under less stringent conditions, but it is also computationally stable due to the convexity of the loss. As such, Shin and Artemiou (2017) proposed principal logistic regression (PLR) by applying a binomial log-likelihood (or logistic) loss function.

Table 2: Different types of the convex loss functions for loss-based principal machine along with the corresponding methods

Method Loss	WPSVM	PWLR	PWLSSVM	PQR	PALS
$L_c(Y, f)$	$\pi(Y; c)\,[(1 - Yf)]_+$	$\pi(Y; c)\ln\{1 + e^{(-Yf)}\}$	$\pi(Y; c)\,(1 - Yf)^2$	$\rho_c\,(Y - f)$	$\rho_c^A\,(Y - f)$

*$\rho_c(u) = u\{c - I(u < 0)\}$, $\rho_c^A(u) = (1 - c)u^2 I(u < 0) + cu^2 I(u \geq 0)$.

Recently, Artemiou *et al.* (2021) proposed a principal least squares SVM (PLSSVM) that uses a least squares loss function. PLSSVM enjoys better estimation of $\mathcal{S}_{Y|\mathbf{X}}$ and it also can be used in case of streamed data for an instant real-time update.

3.2. Loss-based principal machine (LPM)

We have just seen how RPM is formulated along with some examples. When Y is binary, however, the possible choice of pseudo response $\tilde{Y}_c$ is one. Thus RPM cannot recover $\mathcal{S}_{Y|\mathbf{X}}$ exhaustively if $\dim(\mathcal{S}_{Y|\mathbf{X}}) > 1$. To tackle this issue, Shin *et al.* (2017) proposed the principal weighted support vector machine (PWSVM) which gets multiple solutions by changing the shape of the loss function via imposing class weights while keeping the response Y fixed. Again, the hinge loss is not the only choice, and LPM refers to PMs that follow this loss-changing idea.

LPM minimizes the following objective function:

$$\boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta} + \lambda E\left\{L_c\left(Y, \alpha + \boldsymbol{\beta}^\top \mathbf{X}\right)\right\}, \tag{3.2}$$

where $L_c(y, f)$ denotes the loss function with an additional paramter c that controls its shape. One standard way to define L_c with binary Y is to introduce class weight as follows.

$$L_c(Y, f) = \pi(Y; c)L(Y, f), \quad c \in (0, 1),$$

where $\pi(Y; c)$ denotes class weight which takes c if $y = 1$ and $1 - c$ otherwise.

One can extend this idea to a regression problem with a continuous response by employing asymmetric loss functions, such as the check loss function for a quantile regression and the asymmetric squared loss function. Below are some examples of LPM, and their loss functions are summarized in Table 2.

The weighted principal support vector machine (PWSVM) (Shin *et al.*, 2017) is an SDR method for a binary classification, which borrows the idea of weighted SVM. It adopts the hinge loss and imposes the class weights differently by the change of the cutoff value c. Shin *et al.* (2017) showed a fisher consistency of the PWSVM ensures that an estimated hyperplane optimally separates the two classes. Also, by applying weighted logistic loss and weighted least square loss, Kim and Shin (2019) and Jang *et al.* (2023) suggested the principal weighted logistic regression (PWLR) and principal weighted least square SVM (PWLSSVM), respectively. Those methods effectively estimate $\mathcal{S}_{Y|\mathbf{X}}$ for a binary classification.

For a continuous response $Y \in \mathbb{R}$, LPM is also demonstrated by applying an asymmetric loss function. Quantile regression (Koenker and Bassett, 1978) is a well-known method for estimating different quantile functions by changing the angle of the check loss over the given quantile τ. As such, Wang *et al.* (2018) proposed principal quantile regression (PQR) for SDR. PQR chooses the check loss, $\rho_c(u) = u\{c - I(u < 0)\}$ for each cutoff value, $c \in (0, 1)$. PQR echoes the advantages of quantile regression in that it efficiently tackles heteroscedasticity in data and is robust to outliers.

Also, the principal asymmetric L_2 regression for SDR is proposed by Soale and Dong (PALS, 2022). PALS adopts the asymmetric expectile loss function $\rho_c^A(u) = (1-c)u^2 I(u < 0) + cu^2 I(u \geq 0)$. It performs well in the presence of heteroscedasticity by synthesizing different expectile levels. Other than the above examples, we can imagine any type of LPM by varying the loss function following the cutoff value. It illustrates the extensibility of the PSVM.

4. Miscellaneous

As for a computational aspect, several functions in R programming, such as `ipop` in kernlab package and `quadprog` in quadprog package are often used to solve SVM via quadratic programming. These functions are also useful for solving the PSVM, which is regarded to be a sequence of SVMs.

However, when dealing with high-dimensional datasets, quadratic programming is computationally expensive and often unstable. Considering the convexity of the objective function of the PSVM and its variants, the gradient descent (GD) algorithm (including its stochastic version) proves itself a simple and reliable alternative. When coupled with a convex loss function, the GD offers several advantages, including a guaranteed convergence, improved stability, and computational efficiency (Boyd and Vandenberghe, 2004).

Furthermore, the GD for the PSVM is a flexible solver in that the GD only requires a derivative of the objective function, which can be easily calculated if the loss function is specified in advance. Or the derivative of any arbitrary convex function is easily obtained numerically, as long as it retains differentiability and convexity within its support. This flexibility facilitates the development of a unified algorithm for principal sufficient dimension reduction methods.

The convexity of the objective function of the PSVM motivates a direct application of the standard M-estimation theory (Van der Vaart, 2000) and existing asymptotic results of the SVM (Jiang *et al.*, 2008) can mostly be applied in the realm of the PSVM. Consistency and asymptotic normality of the estimates $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ in (2.2) are well-addressed in Section 7 of Li *et al.* (2011) along with proofs and regularity conditions.

Estimating the dimension d in (1.1) of the central subspace, referred to as structural dimension is also a crucial component of sufficient dimension reduction. In this regard, Li *et al.* (2011) proposed the BIC-type criterion for estimating d as below:

$$\hat{d} = \underset{d \in \{1,\ldots,p\}}{\operatorname{argmax}} \sum_{j=1}^{d} v_j - \rho \frac{d \log n}{\sqrt{n}} v_1, \tag{4.1}$$

where $v_1 \geq \cdots \geq v_p (p > d)$ are eigenvalues of $\widehat{\mathbf{M}}_n$ in (2.3). Following the lines of Theorem 8 of Li *et al.* (2011) and Theorem 4 of Shin *et al.* (2017), the BIC-type criterion in (4.1) is proofed as a consistent estimator by showing $\lim_{n \to \infty} P(\hat{d} = d) = 1$. A tuning parameter ρ can be achieved by a cross validation.

5. Sparse principal machines via penalization

Interpreting the estimates of $\mathbf{B}$ in (1.1) is not straightforward. This difficulty arises from the fact that a reduced dimension $\mathbf{B}^\top \mathbf{X}$ represents a linear combination of all original variables. This obscures the underlying relationship between regressors and response.

To illustrate this issue more clearly, let's consider a simple model $Y = \exp(-\boldsymbol{\beta}_0^\top \mathbf{X}) + 0.1\epsilon$, where $\mathbf{X} \in \mathbb{R}^6$, and all predictors and the ϵ are independent standard normal variables. Assume that the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is spanned by $\boldsymbol{\beta}_0 = (1, -1, 1, 0, 0, 0)/\sqrt{3}$. However, SDR methods, such as the PSVM, often estimate a vector $\hat{\boldsymbol{\beta}}$ that includes all predictors, even when some may not be truly relevant. For example, even if the PSVM can estimated $\hat{\boldsymbol{\beta}} = (0.933, -0.912, 0.891, 0.021, 0.009, 0.003)^\top$ based on finite random samples, the last three coefficients are comparatively tiny, but they still contribute to the estimates, blurring the fact that $\boldsymbol{\beta}_0$ only contains the first three predictors. This demonstrates the difficulty of interpreting SDR results due to the mixing of information across all dimensions (Li, 2007).

In the meantime, sparse SDR approaches aim to achieve both dimension reduction and variable selection simultaneously by providing a sparse representation of the basis for the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. In this regard, several sparse SDR methods have been proposed. See, for example, Li (2007), Bondell and Li (2009), Chun and Keleş (2010), and Wu and Li (2011).

Sparse SDR assumes a unique partition $\mathbf{X}^\top = (\mathbf{X}_+^\top, \mathbf{X}_-^\top)$ that satisfies

$$Y \perp \mathbf{X}_- \mid \mathbf{X}_+ , \tag{5.1}$$

where $\mathbf{X}_+ \in \mathbb{R}^q$ and $\mathbf{X}_- \in \mathbb{R}^{p-q}$ for some $q(\ll p)$ (Bondell and Li, 2009; Cook, 2004). We call $\mathbf{X}_+$ and $\mathbf{X}_-$ relevant and irrelevant variables, respectively. This partitioning leads to conditional independence. Here, we assume the first q predictors (of a total of p) constitute the relevant subset. This assumption and the conditional independence imply that the last $p-q$ rows of the matrix $\mathbf{B}$ are zero, which makes $\mathbf{B}$ sparse, but also enforces identical sparsity patterns across its columns.

Likewise, a strategy for a variable selection in a regression problem (Among many others, Tibshirani, 1996; Fan and Li, 2001; Zou and Hastie, 2005; Hastie *et al.*, 2009), the sparsity of $\hat{\mathbf{B}}$ can also be achieved by a penalization.

Given a set of data $(\mathbf{X}_i, Y_i)$, a sample level objective function of sparse SDR based on RPM is given by (5.2).

$$\sum_{h=1}^{H} \left\{ \boldsymbol{\beta}_h^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_h + \frac{\lambda}{n} \sum_{i=1}^{n} L\left(\tilde{Y}_{i,h}, \alpha + \boldsymbol{\beta}^\top \mathbf{X}_i\right) \right\} + \sum_{j=1}^{p} p_\gamma \left(\max_{1 \le h \le H} \left|\beta_{jh}\right| \right), \tag{5.2}$$

where assumed $\mathbf{X}_i$ is centered, $p_\gamma(\cdot)$ denotes a non-convex penalty function, such as SCAD (Fan and Li, 2001) and group lasso (Yuan and Lin, 2006), which are modulated by a tuning parameter γ that controls the level of the sparsity of the solution. Here Y_h and $\boldsymbol{\beta}_h^\top = (\beta_{1h}, \beta_{2h}, \ldots, \beta_{ph})$ are the pseudo response and the corresponding coefficient vector with a given cutoff value $c_h, h = 1, 2, \ldots, H$, respectively. By penalizing the maximum of $|\beta_{jh}|$ over $h = 1, 2, \ldots, H$ for a given $j = 1, 2, \ldots, p$, the penalty forces the entire elements within the same row of $\mathbf{B}$ to shrink towards zero, achieving the desired sparsity structure. The SCAD penalty has gained popularity in variable selection contexts, due to its provable oracle property (Fan and Li, 2001).

For example, Shin and Artemiou (PPLR, 2017) proposed penalized principal logistic regression. This method utilizes the max-SCAD penalty (Fan and Li, 2001) applied to the PLR method described in section 3.1. The oracle property of the PPLR is addressed with detailed explanation and proof.

6. Discussion

Principal support vector machine (PSVM) (Li *et al.*, 2011) offers a novel and powerful solution to a sufficient dimension reduction. We briefly review various aspects of the PSVM, including its linear and nonlinear models, the sample level estimation process, and some theoretical properties. Also, the extensibility of the PSVM is addressed via LPM and RPM frameworks. Sparse SDR applies a penalization method to perform dimension reduction and feature selection simultaneously to improve the interpretability of the SDR result. We also discussed the computational usefulness of the gradient descent algorithm.

This paper investigates the application of the PSVM for high-dimensional data analysis, based on the scenario where the data dimension p can be large but fixed. Researchers will encounter significant analytical challenges as p diverges ($p \to \infty$). Estimating the covariance matrix of predictors under the PSVM scheme becomes difficult, necessitating stricter conditions to assure the validity of covariance matrix estimators as p increases (Wu and Li, 2011; Yin and Hilafu, 2015). This issue can be addressed by combining the PSVM with a large covariance matrix estimation method (Bickel and Levina, 2008).

The vast size of datasets inevitably entails a computational burden. One possible solution is to exploit the distributed computing (Forero *et al.*, 2010; Jin *et al.*, 2019). In the case where the data is collected in a stream-wise manner, typically done online, the question will be asked as to how an existing estimator can be effectively updated as new data streams become available. In response to this, Artemiou *et al.* (2021) proposed the principal least squares support vector machine (PLSSVM), which enables real-time sufficient dimension reduction particularly useful for online learning and streaming data scenarios. This opens further research topics of SDR for large-scale datasets.

Reducing high-dimensional data to lower dimensions through SDR aims to improve the accuracy and efficiency of regression or classification tasks. However, a common pitfall lies in treating the estimated sufficient predictors as the true ones, leading to overestimated confidence levels and biased results. Addressing these troubles, several studies including Kim *et al.* (2020) have explored post-dimension reduction inference. Despite these efforts, existing methods may lack effectiveness across all situations. Further researches are also needed on this issue.

In addition, we provide several potential areas for exploration and improvement in PSVM. First, PSVM primarily relies on data, but incorporating prior knowledge about the relationship between variables could potentially improve performance as Reich *et al.* (2011) proposed SDR via Bayesian mixture modeling and Power and Dong (2021) suggested Bayesian model averaging sliced inverse regression. Second, extending PSVM to handle ordinal or mixed data types would broaden its applicability (Bura *et al.*, 2022; Quach and Li, 2023). Lastly, following a trend of broad application of deep neural networks (DNN), the DNN framework can also be applied to the principal sufficient dimension reduction methods. Banijamali *et al.* (2018) proposed a deep variational approach for sufficient dimension reduction, and Kapla *et al.* (2022) suggested a fusing SDR with neural networks.

Like many existing dimension reduction approaches, PSVM is a new method that aims to solve challenging high-dimensional problems. So far, it has been difficult to find literature on its application to real-world problems. However, tailoring PSVM to specific domains such as bioinformatics (Li, 2010), text analysis (Kim *et al.*, 2005), or survey analysis (Weng and Young, 2017) could significantly improve its impact on related research areas.

Acknowledgement

This work is funded by the National Research Foundation of Korea (NRF) grants (2023R1A2C100658 7, 2022M3J6A1063595) and Korea University (K2302021).

References

Akaho S (2001). A kernel method for canonical correlation analysis, Available from: arXiv preprint cs/0609071

Artemiou A and Dong Y (2016). Sufficient dimension reduction via principal L q support vector machine, *Electronic Journal of Statistics*, **10**, 783–805.

Artemiou A, Dong Y, and Shin SJ (2021). Real-time sufficient dimension reduction through principal least squares support vector machines, *Pattern Recognition*, **112**, 107768.

Banijamali E, Karimi A-H, and Ghodsi A (2018). Deep variational sufficient dimensionality reduction, Available from: arXiv preprint arXiv:1812.07641

Bickel PJ and Levina E (2008). Regularized estimation of large covariance matrices, *The Annals of Statistics*, **36**, 199–227.

Bondell HD and Li L (2009). Shrinkage inverse regression estimation for model-free variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 287–299.

Boyd SP and Vandenberghe L (2004). *Convex Optimization*, Cambridge University Press, Cambridge.

Bura E, Forzani L, Arancibia RG, Llop P, and Tomassi D (2022). Sufficient reductions in regression with mixed predictors, *The Journal of Machine Learning Research*, **23**, 4377–4423.

Chun H and Keleş S (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **72**, 3–25.

Cook RD (2004). Testing predictor contributions in sufficient dimension reduction, *The Annals of Statistics*, **32**, 1062–1092.

Cook RD (2007). Fisher lecture: Dimension reduction in regression, *Statistical Science*, **22**, 1–26.

Cook RD and Weisberg S (1991). Discussion of "sliced inverse regression for dimension reduction", *Journal of the American Statistical Association*, **86**, 28–33.

Fan J and Li R (2001). Variable section via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

Forero PA, Cano A, and Giannakis GB (2010). Consensus-based distributed linear support vector machines, In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, Stockholm, 35–46.

Fukumizu K, Bach FR, and Gretton A (2007). Statistical consistency of kernel canonical correlation analysis, *Journal of Machine Learning Research*, **8**, 361–383.

Hastie T, Tibshirani R, Friedman J, and Friedman JH (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Hristache M, Juditsky A, Polzehl J, and Spokoiny V (2001). Structure adaptive approach for dimension reduction, *Annals of Statistics*, **29**, 1537–1566.

Jang HJ, Shin SJ, and Artemiou A (2023). Principal weighted least square support vector machine: An online dimension-reduction tool for binary classification, *Computational Statistics & Data Analysis*, **187**, 107818.

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.

Jin J, Ying C, and Yu Z (2019). Distributed estimation of principal support vector machines for sufficient dimension reduction, Available from: arXiv preprint arXiv:1911.12732

Kang J and Shin SJ (2022). A forward approach for sufficient dimension reduction in binary classification, *The Journal of Machine Learning Research*, **23**, 9025–9055.

Kapla D, Fertl L, and Bura E (2022). Fusing sufficient dimension reduction with neural networks, *Computational Statistics & Data Analysis*, **168**, 107390.

Kim B and Shin SJ (2019). Principal weighted logistic regression for sufficient dimension reduction in binary classification, *Journal of the Korean Statistical Society*, **48**, 194–206.

Kim H, Howland P, Park H, and Christianini N (2005). Dimension reduction in text classification with support vector machines, *Journal of Machine Learning Research*, **6**, 37–53.

Kim K, Li B, Zhou Y, and Li L (2020). On post dimension reduction statistical inference, *The Annals of Statistics*, **48**, 1567–1592.

Koenker R and Bassett G (1978). Regression quantiles, *Econometrica*, **46**, 33–50.

Kong E and Xia Y (2014). An adaptive composite quantile approach to dimension reduction, *The Annals of Statistics*, **42**, 1657–1688.

Lee KY, Li B, and Chiaromonte F (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation, *The Annals of Statistics*, **41**, 221–249.

Li B (2018). *Sufficient Dimension Reduction: Methods and Applications with R*, CRC Press, Boca Raton, FL.

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

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

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

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

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**, 1025–1039.

Li L (2007). Sparse sufficient dimension reduction, *Biometrika*, **94**, 603–613.

Li L (2010). Dimension reduction for high-dimensional data, *Statistical Methods in Molecular Biology*, **620**, 417–434.

Pearson K (1901). On lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**, 559–572.

Power MD and Dong Y (2021). Bayesian model averaging sliced inverse regression, *Statistics & Probability Letters*, **174**, 109103.

Quach H and Li B (2023). On forward sufficient dimension reduction for categorical and ordinal responses, *Electronic Journal of Statistics*, **17**, 980–1006.

Reich BJ, Bondell HD, and Li L (2011). Sufficient dimension reduction via Bayesian mixture modeling, *Biometrics*, **67**, 886–895.

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

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

Soale A-N and Dong Y (2022). On sufficient dimension reduction via principal asymmetric least squares, *Journal of Nonparametric Statistics*, **34**, 77–94.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of Royal Statistical Society, series B*, **58**, 267–288.

Van der Vaart AW (2000). *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Vapnik V (1999). *The Nature of Statistical Learning Theory*, Springer Science & Business Media, New York.

Wahba G (1999). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV, *Advances in Kernel Methods-Support Vector Learning*, **6**, 69–87.

Wang C, Shin SJ, and Wu Y (2018). Principal quantile regression for sufficient dimension reduction with heteroscedasticity, *Electronic Journal of Statistics*, **12**, 2114–2140.

Weng J and Young DS (2017). Some dimension reduction strategies for the analysis of survey data, *Journal of Big Data*, **4**, 1–19.

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

Wu Y and Li L (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors, *Statistica Sinica*, **2011**, 707.

Xia Y (2007). A constructive approach to the estimation of dimension reduction directions, *The Annals of Statistics*, **35**, 2654–2690.

Xia Y, Tong H, Li WK, and Zhu L (2002). An adaptive estimation of dimension reduction space, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **64**, 363–410.

Yin X and Hilafu H (2015). Sequential sufficient dimension reduction for large p, small n problems, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **77**, 879–892.

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

Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **68**, 49–67.

Zhang C-H (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**, 894–942.

Zhu L-P, Zhu L-X, and Feng Z-H (2010). Dimension reduction in regressions through cumulative slicing estimation, *Journal of the American Statistical Association*, **105**, 1455–1466.

Zou H (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **67**, 301–320.

Received January 19, 2024; Revised February 16, 2024; Accepted February 20, 2024