Electronic Journal of Statistics
Vol. 16 (2022) 1804–1830
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS1988

High-dimensional sufficient dimension reduction through principal projections*

Eugen Pircalabelu [1] **and Andreas Artemiou** [2]

[1] *Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium*
e-mail: eugen.pircalabelu@uclouvain.be

[2] *School of Mathematics, Cardiff University, Senghennydd Road, Cardiff, CF24 4AG*
e-mail: artemioua@cardiff.ac.uk

Abstract: We develop in this work a new dimension reduction method for high-dimensional settings. The proposed procedure is based on a principal support vector machine framework where principal projections are used in order to overcome the non-invertibility of the covariance matrix. Using a series of equivalences we show that one can accurately recover the central subspace using a projection on a lower dimensional subspace and then applying an ℓ_1 penalization strategy to obtain sparse estimators of the sufficient directions. Based next on a desparsified estimator, we provide an inferential procedure for high-dimensional models that allows testing for the importance of variables in determining the sufficient direction. Theoretical properties of the methodology are illustrated and computational advantages are demonstrated with simulated and real data experiments.

MSC2020 subject classifications: Primary 62H12, 62H25, 62J02; secondary 62H15.
Keywords and phrases: Sufficient dimension reduction, support vector machines, quadratic programming, ℓ_1 penalized estimation, debiased estimator.

Received July 2021.

Contents

*The authors gratefully acknowledge the computational resources provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region. The authors would also like to thank the Data Innovation Research Institute at Cardiff University for partially funding the project.

1. Introduction

In an era where large and massive datasets are becoming the norm, one of the biggest challenges for researchers is to propose efficient procedures that address the 'small n, large p' problem, where n denotes the available sample size and p denotes the number of unknown parameters. The main difficulty with a large number of the existing methods is the need to estimate the inverse of a variance-covariance matrix $\boldsymbol{\Sigma}$ which might not per se be invertible.

The Sufficient Dimension Reduction (SDR) framework contains a class of feature extraction procedures used mainly in regression and classification settings. These procedures are most frequently used as a first step for feature extraction, in order to reduce the dimensionality of a problem before any other statistical procedure is applied. They have been developed in a number of different directions in recent years and became more popular the last decades due to the reduced cost of data collection which lead to an increase in the complexity of the data.

In a regression setting one has a response variable Y (which we assume univariate without loss of generality) and a p-dimensional predictor vector $\boldsymbol{X}$. Our efforts in SDR are concentrated in finding D functions of the predictors which contain all the information for the conditional distribution of $Y|\boldsymbol{X}$. In other words, under the linear conditional independence model

$$Y \perp \boldsymbol{X}|\boldsymbol{B}^{\mathsf{T}}\boldsymbol{X} \tag{1.1}$$

where the matrix $\boldsymbol{B}$ is a $p \times D$ matrix of unknown parameters, with $D \ll p$. Under (1.1) we extract linear functions of the predictors $\boldsymbol{X}$. The space spanned by the column vectors of $\boldsymbol{B}$ is known as the Dimension Reduction Subspace (DRS). Since there exists a large class of matrices $\boldsymbol{B}$ that satisfy (1.1), it implies as well that there exists a large number of DRSs that can be estimated. Almost always, one is interested in the one with minimum dimension D for which (1.1) is satisfied. This subspace is known as the Central Dimension Reduction Subspace (CDRS) and is denoted by $\mathcal{S}_{Y|\boldsymbol{X}}$. Although the CDRS does not always exist, the conditions for its existence are mild (Yin et al., 2008) and therefore, throughout this work we assume it exists. If the CDRS exists, then it is also unique. There is an abundance of methods used to estimate the CDRS under model (1.1) and a very short list of references includes Li (1991, 1992), Cook and Weisberg (1991), Li et al. (2005), Li and Wang (2007), Zhu et al. (2010) and many more.

Recently, there was an interest in extracting nonlinear features of the predictors and therefore, researchers introduced SDR under the nonlinear conditional

independence model:

$$Y \perp \boldsymbol{X} | \phi(\boldsymbol{X}) \tag{1.2}$$

where $\phi : \mathbb{R}^p \to \mathbb{R}^D$. Model (1.2) is more general as it allows for the extraction of both linear and nonlinear functions of the predictors. Again, there is an abundance of literature for SDR under model (1.2) including Cook (2007), Wu (2008), Fukumizu et al. (2009) and Li et al. (2011) among them. The latter introduced Support Vector Machines (SVM) in the SDR framework. This method provides a unified framework for linear and nonlinear feature extraction for SDR, among other advantages. The SVM-based SDR methodology has been extended recently by the works of Artemiou and Shu (2014), Artemiou and Dong (2016), Shin et al. (2017), Shin and Artemiou (2017), Randall et al. (2021) and Artemiou et al. (2021).

All the procedures mentioned above work only in low-dimensional settings where $n \gg p$. The literature on how to achieve SDR in 'small n, large p' settings, or as it is more generally known as the high-dimensional low sample size (HDLSS) setting is relatively thin. One of the first efforts to achieve dimension reduction without matrix inversion in the SDR framework, was the work of Cook and Li (2002) where an iterative approach is proposed to avoid the use of the inverse matrix in the iterative Hessian transformations (IHT), a method which finds directions in the Central Mean Subspace, i.e. the space estimated under the model $Y \perp \mathrm{E}(Y|\boldsymbol{X})|\boldsymbol{B}^\intercal \boldsymbol{X}$ where $\mathrm{E}(Y|\boldsymbol{X})$ denotes the conditional expectation of Y given $\boldsymbol{X}$. Recently, Lin et al. (2019) proposed the use of Lasso with the SIR algorithm and Lin et al. (2018) proposed the use of Diagonal Thresholding with SIR. In this paper we propose the use of Lasso in an SVM-based algorithm for SDR. As will be shown, our method utilizes principal projections of the covariance matrix $\boldsymbol{\Sigma} = \mathrm{var}(\boldsymbol{X})$ to establish an equivalence relationship between high and lower dimensional SVM problems, after which an ℓ_1 penalized procedure is applied to obtain sparse estimators of the underlying directions.

The paper is structured as follows: in Section 2 we introduce the background literature by revisiting the Principal Support Vector Machine (PSVM) framework and its estimation procedure. In Section 3 we motivate an ℓ_1 regularized approach to address high-dimensional problems through a novel procedure coined 'LassoPSVM' and show that by using principal component projections, one can avoid the use of the inverse of the covariance matrix. In Section 4 we present the computational algorithm used to obtain the LassoPSVM solution. In Section 5 we discuss theoretical properties of the proposed method, while in Section 6 we present a high-dimensional inferential procedure based on a desparsified estimator for LassoPSVM. In Section 7 we present numerical studies using both a controlled simulation example, as well as real data to show the performance of the method. Finally, we close with a discussion on the method and future possible extensions in Section 8.

2. Principal support vector machine (PSVM)

The SDR approach we propose in this manuscript uses the multiple index model with additive error of the form

$$Y = g(\boldsymbol{\beta}_1^\intercal \boldsymbol{X}, \boldsymbol{\beta}_2^\intercal \boldsymbol{X}, \ldots, \boldsymbol{\beta}_D^\intercal \boldsymbol{X}) + \epsilon \tag{2.1}$$

where $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal \in \mathbb{R}^p$ with $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D \in \mathbb{R}^p$, $\epsilon \perp X_j \quad \forall\, j = 1, \ldots, p$ with $\epsilon \sim N(0, \sigma^2)$ and $g : \mathbb{R}^D \to \mathbb{R}$. The link function g is an unknown many-to-one function, the vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D$ are unknown vectors of coefficients and for simplicity we assume throughout that the dimension D is fixed and known. We concatenate the vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D$ to obtain the matrix of unknown coefficients as $\boldsymbol{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_D]$ of dimension $p \times D$.

In this section we give a brief overview of the low-dimensional PSVM and in the next section we present our high-dimensional proposal. We assume without loss of generality that $\text{E}(\boldsymbol{X}) = \boldsymbol{0}$. In the population version, PSVM minimizes the following objective function:

$$\mathcal{L}(\boldsymbol{\psi}, t) = \boldsymbol{\psi}^\intercal \boldsymbol{\Sigma} \boldsymbol{\psi} + c\text{E}(1 - \tilde{Y}(\boldsymbol{\psi}^\intercal \boldsymbol{X} - t))_+ \tag{2.2}$$

where $\tilde{Y} \in \{-1, 1\}$ is a discretized version of the response Y, $\boldsymbol{\Sigma}$ is the covariance matrix of the vector $\boldsymbol{X}$, c is a positive regularization constant, t is a slack variable, $\boldsymbol{\psi}$ is a vector of unknown coefficients and $a_+ = \max\{0, a\}$. It has been shown in Li et al. (2011) that under mild conditions, if the pair $(\boldsymbol{\psi}^*, t^*)$ minimizes the objective function (2.2) then $\boldsymbol{\psi}^* \in \mathcal{S}_{Y|\boldsymbol{X}}$ which motivates the SVM approach for low-dimensional problems.

Considering $H-1$ different discretized versions of the response Y, dimension reduction is achieved by performing an eigenvalue decomposition of $\boldsymbol{V}$ defined as $\boldsymbol{V} = \sum_{h=1}^{H-1} \boldsymbol{\psi}^{*h} (\boldsymbol{\psi}^{*h})^\intercal$ where $\boldsymbol{\psi}^{*h}$ are the minimizers of equation (2.2) for each $h = 1, \ldots, H-1$, and selecting the eigenvectors corresponding to its largest D eigenvalues, due to the fact that the column space of $\boldsymbol{V}$ is the same as the column space of $\boldsymbol{B}$, denoted as $\text{col}(\boldsymbol{V}) = \text{col}(\boldsymbol{B})$ in light of Proposition 1.

Proposition 1. *Under the specification of model (1.1) and assuming* $\text{E}(\boldsymbol{X}|\boldsymbol{B}^\intercal \boldsymbol{X})$ *is a linear function of* $\boldsymbol{B}^\intercal \boldsymbol{X}$ *then* $\text{col}(\boldsymbol{V}) = \text{col}(\boldsymbol{B})$.

The proof follows directly by applying Theorem 1 from Li et al. (2011) which showed that the minimizer $\boldsymbol{\psi}^* \in \mathcal{S}_{Y|\boldsymbol{X}}$. Thus, constructing the candidate matrix $\boldsymbol{V}$ using minimizers $\boldsymbol{\psi}^h$'s $(h = 1, \ldots, H-1)$ ensures that the eigenvectors of $\boldsymbol{V}$ will span $\mathcal{S}_{Y|\boldsymbol{X}}$. Proposition 1 justifies why performing an eigenvalue decomposition of $\boldsymbol{V}$ is beneficial for SDR.

Estimation under (2.2) proceeds as follows. Let $(Y_i, \boldsymbol{X}_i^\intercal)^\intercal, i = 1, \ldots, n$ be an independent sample of n observations. Divide first the data into H equal-sized slices, where H is a fixed number of slices. Let q^h, $h = 1, \ldots, H-1$ denote the cut-off points between the H slices in the range of Y. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\intercal$ and, for each h, we create the discretized version of the vector $\boldsymbol{Y}$ denoted as $\tilde{\boldsymbol{Y}}^h$ which has entries:

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

where $I(\cdot)$ denotes the indicator function attributing the value 1 if the condition is satisfied and 0, otherwise. Then, for each h, in the low-dimensional setting for PSVMs one minimizes the objective function:

$$\min_{\boldsymbol{\psi},t,\boldsymbol{\xi}^h} \quad \boldsymbol{\psi}^\intercal \boldsymbol{\Sigma} \boldsymbol{\psi} + (c/n)\mathbf{1}_n^\intercal \boldsymbol{\xi}^h \tag{2.3}$$

under the constraints

$$\tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}\boldsymbol{\psi} - \boldsymbol{t}_n) \geq \mathbf{1}_n - \boldsymbol{\xi}^h, \quad \boldsymbol{\xi}^h \geq \mathbf{0}_n,$$

where $\mathcal{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]^\intercal$ is an $n \times p$ predictor matrix and $\odot$ denotes the elementwise multiplication of two vectors. The vector $\boldsymbol{\xi}^h = (\xi_1^h, \ldots, \xi_n^h)^\intercal$ where ξ_i^h is the misclassification distance for each data point (set to 0 if the point is correctly classified) when q^h is used as a cut-off point, while $\boldsymbol{t}_n = (t, \ldots, t)^\intercal \in \mathbb{R}^n$, $\mathbf{0}_n = (0, \ldots, 0)^\intercal \in \mathbb{R}^n$ and $\mathbf{1}_n = (1, \ldots, 1)^\intercal \in \mathbb{R}^n$. Optimizing (2.3) gives rise to $H-1$ minimizers $\hat{\boldsymbol{\psi}}^h, h = 1, \ldots, H-1$ used next to create the candidate matrix $\hat{\boldsymbol{V}} = \sum_{h=1}^{H-1} \hat{\boldsymbol{\psi}}^h (\hat{\boldsymbol{\psi}}^h)^\intercal$.

To solve the quadratic programming problem in (2.3) one can use the Lagrangian approach. This is done by creating first the Lagrangian function $\mathcal{L}(\boldsymbol{\psi}, t, \boldsymbol{\xi}^h)$ defined as

$$\mathcal{L}(\boldsymbol{\psi}, t, \boldsymbol{\xi}^h) = \boldsymbol{\psi}^\intercal \boldsymbol{\Sigma} \boldsymbol{\psi} + (c/n)\mathbf{1}_n^\intercal \boldsymbol{\xi}^h - \boldsymbol{\alpha}^\intercal (\mathbf{1}_n - \boldsymbol{\xi}^h - \tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}\boldsymbol{\psi} - \boldsymbol{t}_n)) - \boldsymbol{\gamma}^\intercal \boldsymbol{\xi}^h \tag{2.4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\intercal$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\intercal$ are vectors of Lagrangian multipliers. The Karush-Kuhn-Tucker (KKT) conditions imply that an optimal solution exists when the partial derivatives of (2.4) are set to 0. Therefore:

$$\begin{aligned}
\partial \mathcal{L}(\boldsymbol{\psi}, t, \boldsymbol{\xi}^h)/\partial \boldsymbol{\psi} &= 2\boldsymbol{\Sigma}\boldsymbol{\psi} - \mathcal{X}^\intercal (\tilde{\boldsymbol{Y}}^h \odot \boldsymbol{\alpha}) = \mathbf{0}_p \\
\partial \mathcal{L}(\boldsymbol{\psi}, t, \boldsymbol{\xi}^h)/\partial t &= \boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{Y}}^h = 0 \\
\partial \mathcal{L}(\boldsymbol{\psi}, t, \boldsymbol{\xi}^h)/\partial \boldsymbol{\xi}^h &= (c/n)\mathbf{1}_n + \boldsymbol{\alpha} - \boldsymbol{\gamma} = \mathbf{0}_n,
\end{aligned}$$

which implies two very important properties needed later in Section 3. The first is that the minimizer of the objective function specified by (2.3) is given by:

$$\hat{\boldsymbol{\psi}}^h = (1/2)\hat{\boldsymbol{\Sigma}}^{-1} \mathcal{X}^\intercal (\tilde{\boldsymbol{Y}}^h \odot \boldsymbol{\alpha}) \tag{2.5}$$

where $\hat{\boldsymbol{\Sigma}}$ is an estimator of $\boldsymbol{\Sigma}$ and the second is that

$$\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{Y}}^h = 0. \tag{2.6}$$

Note that the same quadratic programme is solved for each cut-off point q^h, the only difference being that only the values of $\tilde{\boldsymbol{Y}}^h$ change (as they are a function of q^h). Note also that $\hat{\boldsymbol{\psi}}^h$ in (2.5) depends directly on the inverse of the estimator for the covariance matrix $\hat{\boldsymbol{\Sigma}}$. For low-dimensional problems this

generally does not pose problems, but for cases where $p \gg n$, this is problematic as the inverse might not always exist since the covariance matrix might not be full rank. As will be illustrated in the following section, our proposed procedure will eliminate completely the dependence on an estimator whose inverse does not exist. See Section 3 for details.

For ease of exposition, we avoid here the details on how one can use the dual problem to estimate $\boldsymbol{\alpha}$ (details on the estimation are offered in Cortes and Vapnik, 1995 and Li et al., 2011 among others) and focus on the fact that once one has the vectors $\hat{\boldsymbol{\psi}}^h$ for $h = 1, \ldots, H-1$ one can perform an eigenvalue decomposition of the estimated candidate matrix $\hat{\boldsymbol{V}} = \sum_{h=1}^{H-1} \hat{\boldsymbol{\psi}}^h (\hat{\boldsymbol{\psi}}^h)^\intercal$ in order to get the vectors that span the central subspace. This is done by selecting the first D eigenvectors $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_D$ associated with the largest D eigenvalues of $\hat{\boldsymbol{V}}$ as the vectors that span the CDRS.

3. Using principal projections for LassoPSVM

In this section we propose a procedure coined as 'LassoPSVM', that bypasses the need of inverting the covariance matrix in order to obtain estimators in the high-dimensional setting.

Our procedure uses first the fact that one can construct an optimization problem with a solution contained within the space that is spanned by the eigenvectors associated with the non-zero eigenvalues of $\boldsymbol{\Sigma}$. As such, solving this new optimization problem is equivalent to solving the original problem in $\mathbb{R}^p$. Due to the equivalence between the two problems, in a second step we use a principal component projection (Mardia et al., 1979; Zafeiriou et al., 2007) to solve the reduced problem in a lower dimensional space and project back to the original dimensional problem.

Theorem 1. *Let $\boldsymbol{\Sigma} = \operatorname{var}(\boldsymbol{X})$ be a $p \times p$ matrix of rank $r < p$ and let $\mathcal{A}$ be the space spanned by the eigenvectors corresponding to the non-zero eigenvalues of $\boldsymbol{\Sigma}$. The minimizer of the constrained objective function specified in (2.3), is equivalent to the minimizer of*

$$\min_{\boldsymbol{u}, t, \boldsymbol{\xi}^h} \quad \boldsymbol{u}^\intercal \boldsymbol{\Sigma} \boldsymbol{u} + (c/n) \mathbf{1}_n^\intercal \boldsymbol{\xi}^h$$

with $\boldsymbol{u} \in \mathcal{A}$ and under the constraints

$$\tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}\boldsymbol{u} - \boldsymbol{t}_n) \geq \mathbf{1}_n - \boldsymbol{\xi}^h, \quad \boldsymbol{\xi}^h \geq \mathbf{0}_n.$$

Proof. As $\boldsymbol{\Sigma}$ is a real, symmetric and positive semidefinite matrix in $\mathbb{R}^{p \times p}$, its eigenvectors form an orthogonal basis of $\mathbb{R}^p$. Therefore every vector $\boldsymbol{\psi} \in \mathbb{R}^p$ can be written as a linear combination of the eigenvectors of $\boldsymbol{\Sigma}$, as they form a basis in $\mathbb{R}^p$.

Let $\mathcal{A}$ be the space spanned by the eigenvectors corresponding to non-zero eigenvalues of $\boldsymbol{\Sigma}$ and $\mathcal{A}^\perp$ to be the space spanned by the eigenvectors corresponding to the zero eigenvalues of $\boldsymbol{\Sigma}$. Taking vectors $\boldsymbol{u} \in \mathcal{A}$ and $\boldsymbol{s} \in \mathcal{A}^\perp$, one

can write $\boldsymbol{\psi} \in \mathbb{R}^p$ as $\boldsymbol{\psi} = \boldsymbol{u} + \boldsymbol{s}$. To show the equivalence, note that:

$$\boldsymbol{\psi}^\intercal \boldsymbol{\Sigma} \boldsymbol{\psi} = (\boldsymbol{u} + \boldsymbol{s})^\intercal \boldsymbol{\Sigma} (\boldsymbol{u} + \boldsymbol{s}) = \boldsymbol{u}^\intercal \boldsymbol{\Sigma} \boldsymbol{u}$$

since $\boldsymbol{\Sigma}\boldsymbol{s} = \mathbf{0}_p$ due to the fact that $\boldsymbol{s} \in \mathcal{A}^\perp$ (i.e. the null space of $\boldsymbol{\Sigma}$).

Moreover, $\text{var}(\boldsymbol{s}^\intercal \boldsymbol{X}) = \boldsymbol{s}^\intercal \boldsymbol{\Sigma} \boldsymbol{s} = 0$ which implies that under the projection of $\boldsymbol{s}$, all points $\boldsymbol{x}_i \in \mathbb{R}^p$ project on the same point and therefore for any pair (i, i') where $i = 1, \ldots, n$, $i' = 1, \ldots, n$ and $i \neq i'$, we have $\boldsymbol{s}^\intercal \boldsymbol{x}_i = \boldsymbol{s}^\intercal \boldsymbol{x}_{i'} = \kappa$. Therefore, the constraints of (2.3) are equivalent to:

$$\tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}\boldsymbol{\psi} - \boldsymbol{t}_n) \geq \mathbf{1}_n - \boldsymbol{\xi}^h \Leftrightarrow$$
$$\tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}\boldsymbol{u} + \mathcal{X}\boldsymbol{s} - \boldsymbol{t}_n) \geq \mathbf{1}_n - \boldsymbol{\xi}^h.$$

Using the above equivalence, one also has that the Lagrangian in (2.4) is equivalent to the following Lagrangian:

$$\begin{aligned}
&\boldsymbol{u}^\intercal \boldsymbol{\Sigma} \boldsymbol{u} + c\mathbf{1}_n^\intercal \boldsymbol{\xi}^h - \boldsymbol{\alpha}^\intercal (\mathbf{1}_n - \boldsymbol{\xi}^h - \tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}\boldsymbol{u} + \mathcal{X}\boldsymbol{s} - \boldsymbol{t}_n)) - \boldsymbol{\gamma}^\intercal \boldsymbol{\xi}^h \Leftrightarrow \\
&\boldsymbol{u}^\intercal \boldsymbol{\Sigma} \boldsymbol{u} + c\mathbf{1}_n^\intercal \boldsymbol{\xi}^h - \boldsymbol{\alpha}^\intercal (\mathbf{1}_n - \boldsymbol{\xi}^h - \tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}\boldsymbol{u} - \boldsymbol{t}_n)) - (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}^h)^\intercal \mathcal{X}\boldsymbol{s} - \boldsymbol{\gamma}^\intercal \boldsymbol{\xi}^h \Leftrightarrow \\
&\boldsymbol{u}^\intercal \boldsymbol{\Sigma} \boldsymbol{u} + c\mathbf{1}_n^\intercal \boldsymbol{\xi}^h - \boldsymbol{\alpha}^\intercal (\mathbf{1}_n - \boldsymbol{\xi}^h - \tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}\boldsymbol{u} - \boldsymbol{t}_n)) - (\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}^h)^\intercal \boldsymbol{\kappa}_n - \boldsymbol{\gamma}^\intercal \boldsymbol{\xi}^h,
\end{aligned}$$

where $\boldsymbol{\kappa}_n = (\kappa, \ldots, \kappa)^\intercal \in \mathbb{R}^n$ is a constant vector. Due to equation (2.6), this implies that

$$(\boldsymbol{\alpha} \odot \tilde{\boldsymbol{Y}}^h)^\intercal \boldsymbol{\kappa}_n = \kappa \boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{Y}}^h = 0.$$

Therefore, we have shown that the Lagrangian in (2.4) is equivalent to:

$$\boldsymbol{u}^\intercal \boldsymbol{\Sigma} \boldsymbol{u} + (c/n)\mathbf{1}_n^\intercal \boldsymbol{\xi}^h - \boldsymbol{\alpha}^\intercal (\mathbf{1}_n - \boldsymbol{\xi}^h - \tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}\boldsymbol{u} - \boldsymbol{t}_n)) - \boldsymbol{\gamma}^\intercal \boldsymbol{\xi}^h,$$

which implies that the optimal hyperplane $\boldsymbol{\psi}$ depends *only* on $\boldsymbol{u}$ and not on $\boldsymbol{s}$. It also implies that $\boldsymbol{s}$ can be any arbitrary vector from $\mathcal{A}^\perp$. We have thus shown that the original problem formulated in (2.3) can be solved in the space $\mathcal{A}$, rather than in the entire space $\mathbb{R}^p$. □

Assume now that r, the rank of $\boldsymbol{\Sigma}$, is such that $r \geq D$ and $r < n$. We construct next the matrix $\boldsymbol{P}$ of dimension $p \times r$. The columns of $\boldsymbol{P}$ are formed by the r non-null eigenvectors of $\boldsymbol{\Sigma}$, that is, the eigenvectors corresponding to the non-zero eigenvalues. The columns of $\boldsymbol{P}$ are an orthogonal basis in $\mathbb{R}^r$ therefore, one can create the mapping $m : \mathbb{R}^r \to \mathcal{A}$ which is a one-to-one mapping from $\mathbb{R}^r$ to $\mathcal{A}$ and where for any $\boldsymbol{w} \in \mathbb{R}^r$ we have that $m(\boldsymbol{w}) = \boldsymbol{P}\boldsymbol{w} = \boldsymbol{u}$. Using next the mapping $m(\boldsymbol{w})$ and replacing $\boldsymbol{u}$ in Theorem 1 by $\boldsymbol{P}\boldsymbol{w}$, one can show the following corollary holds.

Corollary 1. *The minimizer of the objective function specified in (2.3), is equivalent to the minimizer of*

$$\min_{\boldsymbol{w}, t, \boldsymbol{\xi}^h} \quad \boldsymbol{w}^\intercal \boldsymbol{\Sigma}^\dagger \boldsymbol{w} + (c/n)\mathbf{1}_n^\intercal \boldsymbol{\xi}^h \tag{3.1}$$

under the constraints

$$\tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}^\dagger \boldsymbol{w} - \boldsymbol{t}_n) \geq \mathbf{1}_n - \boldsymbol{\xi}^h, \quad \boldsymbol{\xi}^h \geq \mathbf{0}_n,$$

where $\mathcal{X}^\dagger = \mathcal{X}\boldsymbol{P} = [\boldsymbol{X}_1^\dagger, \ldots, \boldsymbol{X}_n^\dagger]^\intercal$ *is the matrix of dimension* $n \times r$*, formed by the projected vectors of observations, i.e.* $\boldsymbol{X}_i^\dagger = \boldsymbol{P}^\intercal \boldsymbol{X}_i$ *and* $\boldsymbol{\Sigma}^\dagger = \boldsymbol{P}^\intercal \boldsymbol{\Sigma} \boldsymbol{P}$ *is the covariance matrix of dimension* $r \times r$*, of the projected vectors* $\boldsymbol{X}^\dagger \in \mathbb{R}^r$*.*

Rather than optimizing (3.1), we propose to optimize

$$\min_{\boldsymbol{w}, t, \boldsymbol{\xi}^h} \quad n\boldsymbol{w}^\intercal \boldsymbol{\Sigma}^\dagger \boldsymbol{w} + c\mathbf{1}_n^\intercal \boldsymbol{\xi}^h.$$

which coupled with the constraints in Corolloary 1 leads to the Lagrangian:

$$n\boldsymbol{w}^\intercal \boldsymbol{\Sigma}^\dagger \boldsymbol{w} + c\mathbf{1}_n^\intercal \boldsymbol{\xi}^h - \boldsymbol{\alpha}^\intercal (\mathbf{1}_n - \boldsymbol{\xi}^h - \tilde{\boldsymbol{Y}}^h \odot (\mathcal{X}^\dagger \boldsymbol{w} - \boldsymbol{t}_n)) - \boldsymbol{\gamma}^\intercal \boldsymbol{\xi}^h,$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are the Lagrangian multipliers. Using KKT conditions as in Section 2, one can show that for each h, the estimator for $\boldsymbol{w}$ is:

$$\hat{\boldsymbol{w}}^h = \frac{1}{2n} (\hat{\boldsymbol{\Sigma}}^\dagger)^{-1} (\mathcal{X}^\dagger)^\intercal \tilde{\boldsymbol{Y}}^h \odot \boldsymbol{\alpha}, \tag{3.2}$$

where $\hat{\boldsymbol{\Sigma}}^\dagger$ is an estimator for the covariance matrix of the projected vectors.

In light of Theorem 1, LassoPSVM uses the candidate matrix $\hat{\boldsymbol{V}}_{\boldsymbol{u}}$ defined as

$$\hat{\boldsymbol{V}}_{\boldsymbol{u}} = \sum_{h=1}^{H-1} \hat{\boldsymbol{u}}^h (\hat{\boldsymbol{u}}^h)^\intercal = \sum_{h=1}^{H-1} \boldsymbol{P} \hat{\boldsymbol{w}}^h (\hat{\boldsymbol{w}}^h)^\intercal \boldsymbol{P}^\intercal$$

where the second equality comes from using Corollary 1. Plugging in the expression for $\hat{\boldsymbol{V}}_{\boldsymbol{u}}$ the estimator $\hat{\boldsymbol{w}}^h$ defined in (3.2), one obtains

$$\hat{\boldsymbol{V}}_{\boldsymbol{u}} = \frac{1}{4n^2} \boldsymbol{P} (\hat{\boldsymbol{\Sigma}}^\dagger)^{-1} (\mathcal{X}^\dagger)^\intercal \sum_{h=1}^{H-1} (\tilde{\boldsymbol{Y}}^h \odot \boldsymbol{\alpha})(\tilde{\boldsymbol{Y}}^h \odot \boldsymbol{\alpha})^\intercal \mathcal{X}^\dagger (\hat{\boldsymbol{\Sigma}}^\dagger)^{-1} \boldsymbol{P}^\intercal. \tag{3.3}$$

The advantage of the estimator expressed in (3.3) is the fact that $\hat{\boldsymbol{\Sigma}}^{-1}$ is not needed, as it is replaced by the inverse of $\hat{\boldsymbol{\Sigma}}^\dagger$ which always exists. We have thus removed the major hurdle one has encountered in the development presented in Section 2.

Let now $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_D > 0$ be the ordered eigenvalues of $\hat{\boldsymbol{V}}_{\boldsymbol{u}}$ and let $\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_D$ be the associated eigenvectors. For any couple $(\hat{\lambda}_d,\ \hat{\boldsymbol{\eta}}_d)$ with $d = 1, \ldots, D$, one now has that:

$$\begin{aligned}
\hat{\lambda}_d \hat{\boldsymbol{\eta}}_d =& \hat{\boldsymbol{V}}_{\boldsymbol{u}} \hat{\boldsymbol{\eta}}_d \\
=& \frac{1}{4n^2} \boldsymbol{P} (\hat{\boldsymbol{\Sigma}}^\dagger)^{-1} (\mathcal{X}^\dagger)^\intercal \sum_{h=1}^{H-1} (\tilde{\boldsymbol{Y}}^h \odot \boldsymbol{\alpha})(\tilde{\boldsymbol{Y}}^h \odot \boldsymbol{\alpha})^\intercal \mathcal{X}^\dagger (\hat{\boldsymbol{\Sigma}}^\dagger)^{-1} \boldsymbol{P}^\intercal \hat{\boldsymbol{\eta}}_d
\end{aligned}$$

$$=\frac{1}{4}\boldsymbol{P}((\mathcal{X}^\dagger)^\intercal\mathcal{X}^\dagger)^{-1}(\mathcal{X}^\dagger)^\intercal\sum_{h=1}^{H-1}(\tilde{\boldsymbol{Y}}^h\odot\boldsymbol{\alpha})(\tilde{\boldsymbol{Y}}^h\odot\boldsymbol{\alpha})^\intercal\mathcal{X}^\dagger((\mathcal{X}^\dagger)^\intercal\mathcal{X}^\dagger)^{-1}\boldsymbol{P}^\intercal\hat{\boldsymbol{\eta}}_d$$

where the third equality used $\hat{\boldsymbol{\Sigma}}^\dagger=\frac{1}{n}(\mathcal{X}^\dagger)^\intercal\mathcal{X}^\dagger$.

We set now

$$\tilde{y}_d^\dagger=\frac{1}{4\hat{\lambda}_d}\sum_{h=1}^{H-1}(\tilde{\boldsymbol{Y}}^h\odot\boldsymbol{\alpha})(\tilde{\boldsymbol{Y}}^h\odot\boldsymbol{\alpha})^\intercal\mathcal{X}^\dagger((\mathcal{X}^\dagger)^\intercal\mathcal{X}^\dagger)^{-1}\boldsymbol{P}^\intercal\hat{\boldsymbol{\eta}}_d$$

and as such, by substituting $\tilde{y}_d^\dagger$ in the above equation, one now has that

$$\hat{\boldsymbol{\eta}}_d=\boldsymbol{P}((\mathcal{X}^\dagger)^\intercal\mathcal{X}^\dagger)^{-1}(\mathcal{X}^\dagger)^\intercal\tilde{y}_d^\dagger.$$

As shown in Section 2 in Proposition 1, with PSVM the basic identity is that the column space of $\boldsymbol{V}$ is the same as the column space of $\boldsymbol{B}$. As such, and with $\boldsymbol{\eta}_d$ denoting the population version of $\hat{\boldsymbol{\eta}}_d$ which corresponds to the d-th eigenvector of $\boldsymbol{V}_{\boldsymbol{u}}=\sum_{h=1}^{H-1}\boldsymbol{u}^h(\boldsymbol{u}^h)^\intercal$, we have that:

$$\begin{aligned}\boldsymbol{\eta}_d\propto\boldsymbol{\beta}_d\Leftrightarrow\\ \boldsymbol{P}((\mathcal{X}^\dagger)^\intercal\mathcal{X}^\dagger)^{-1}(\mathcal{X}^\dagger)^\intercal\tilde{y}_d^\dagger\propto\boldsymbol{\beta}_d\Leftrightarrow\\ ((\mathcal{X}^\dagger)^\intercal\mathcal{X}^\dagger)^{-1}(\mathcal{X}^\dagger)^\intercal\tilde{y}_d^\dagger\propto\boldsymbol{P}^\intercal\boldsymbol{\beta}_d\Leftrightarrow\\ (\mathcal{X}^\dagger)^\intercal\tilde{y}_d^\dagger\propto(\mathcal{X}^\dagger)^\intercal\mathcal{X}^\dagger\boldsymbol{P}^\intercal\boldsymbol{\beta}_d\Leftrightarrow\\ \boldsymbol{P}(\mathcal{X}^\dagger)^\intercal\tilde{y}_d^\dagger\propto\boldsymbol{P}(\mathcal{X}^\dagger)^\intercal\mathcal{X}^\dagger\boldsymbol{P}^\intercal\boldsymbol{\beta}_d\end{aligned}$$

where the second result is obtained by approximating $\boldsymbol{\eta}_d\approx\hat{\boldsymbol{\eta}}_d$. We can set now $\boldsymbol{P}(\mathcal{X}^\dagger)^\intercal=(\tilde{\mathcal{X}}^\dagger)^\intercal$ and therefore the above relation becomes

$$(\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{y}_d^\dagger\propto(\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger\boldsymbol{\beta}_d.$$

To obtain now a sparse estimator $\hat{\boldsymbol{\beta}}_d$ we propose to minimize the ℓ_1 penalized quadratic loss function for a regularization parameter $\mu>0$:

$$\mathcal{L}(\boldsymbol{\beta}_d)=\frac{1}{2n}\|\tilde{y}_d^\dagger-\tilde{\mathcal{X}}^\dagger\boldsymbol{\beta}_d\|_2^2+\mu\|\boldsymbol{\beta}_d\|_1,\tag{3.4}$$

hence the name 'LassoPSVM' for the proposed procedure. Here $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the usual ℓ_1 and ℓ_2 vector norms.

We close this section by commenting on the importance and benefits of Theorem 1 and Corollary 1. In light of this theorem we highlight that the crucial information to recover the CDRS is contained in the eigenvectors corresponding to the non-zero eigenvalues of $\boldsymbol{V}_{\boldsymbol{u}}$. However, these eigenvectors are still contained in $\mathbb{R}^p$. As such, this result is not useful enough as one still needs the inverse of a $p\times p$ covariance matrix, which might not exist. Corollary 1 on the other hand, allows us to improve on the dimensionality of the problem by projecting the crucial information to recover the CDRS in the space $\mathbb{R}^r$. As r can

be much smaller than n, the estimator $\hat{\boldsymbol{\Sigma}}^{\dagger}$ we propose is always invertible. One other possibility would be to plug in an estimator of $\boldsymbol{\Sigma}^{-1}$ in (3.2) as is proposed in Pircalabelu and Artemiou (2021) for the graph-based procedures, however an extra Gaussianity assumption is needed to justify the approach. Alternatively, based on the KKT conditions an estimator as proposed in Cai et al. (2011) could be used, but this would introduce additional computational complexity. We argue that principal projections on a lower dimensional subspace offer an elegant and natural solution in the proposed framework.

4. A sufficient dimension reduction algorithm

In this section we present schematically a sufficient dimension reduction algorithm that allows for the estimation of $\boldsymbol{B}$. Based on the development presented in Section 3 one can put forward the following estimation procedure:

- ▷ Calculate the sample mean and sample covariance of the predictors $\boldsymbol{X}_i$, $i = 1, \dots n$, denoted as $\bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$.
- ▷ Centre the $\boldsymbol{X}_i$'s to obtain $\boldsymbol{X}_i^c = \boldsymbol{X}_i - \bar{\boldsymbol{X}}$.
- ▷ Perform an eigenvalue decomposition of the matrix $\hat{\boldsymbol{\Sigma}}$, to find the r non-zero eigenvalues $\hat{\tau}_1, \dots, \hat{\tau}_r$ and the corresponding eigenvectors $\hat{\boldsymbol{\zeta}}_1, \dots, \hat{\boldsymbol{\zeta}}_r$.
- ▷ Using the vectors $\hat{\boldsymbol{\zeta}}_1, \dots, \hat{\boldsymbol{\zeta}}_r$ as columns, create the $p \times r$ matrix $\hat{\boldsymbol{P}}$.
- ▷ Project the original centred data onto $\mathbb{R}^r$, to obtain $\hat{\boldsymbol{X}}_i^{\dagger} = \hat{\boldsymbol{P}}^{\intercal} \boldsymbol{X}_i^c$ and construct its empirical covariance matrix $\hat{\boldsymbol{\Sigma}}^{\dagger}$.
- ▷ Divide the range of Y in H slices by taking $H-1$ cut-off points $q^h, h = 1, \dots, H-1$ and create the vectors $\tilde{\boldsymbol{Y}}^h = (\tilde{Y}_1^h, \dots, \tilde{Y}_n^h)^{\intercal}$.
- ▷ For each $\tilde{\boldsymbol{Y}}^h$, $h = 1, \dots, H-1$ optimize (3.1) under the specified constraints, where population parameters are replaced by their sample estimators, to find $\hat{\boldsymbol{w}}^h \in \mathbb{R}^r$.
- ▷ Construct $\hat{\boldsymbol{u}}^h = \hat{\boldsymbol{P}} \hat{\boldsymbol{w}}^h$, $h = 1, \dots, H-1$ and use these vectors to create $\hat{\boldsymbol{V}}_{\boldsymbol{u}} = \sum_{h=1}^{H-1} \hat{\boldsymbol{u}}^h (\hat{\boldsymbol{u}}^h)^{\intercal}$.
- ▷ Find the D largest eigenvalues and corresponding eigenvectors of $\hat{\boldsymbol{V}}_{\boldsymbol{u}}$ and use them to calculate each vector $\tilde{y}_d^{\dagger}$, $d = 1, \dots, D$.
- ▷ Run a Lasso algorithm with $\tilde{y}_d^{\dagger}$ as the response vector and $\tilde{\mathcal{X}}^{\dagger} = \boldsymbol{P}\boldsymbol{P}^{\intercal}\boldsymbol{X}^{\intercal}$ as predictor matrix to find each of the d (sparse) columns of the matrix $\boldsymbol{B}$.

In the first step of the algorithm, one needs an estimator of the covariance matrix $\boldsymbol{\Sigma}$. In principle, the simplest estimator that one can use is the sample covariance matrix $\boldsymbol{S}$. However, when $p \gg n$ it is well known that the performance of this estimator relative to the true unknown $\boldsymbol{\Sigma}$ is poor. See for example, the works of Ledoit and Wolf (2004), Bickel and Levina (2008), Cai et al. (2010), Fan et al. (2011) or Bien and Tibshirani (2011) among many others. For these reasons, we propose to use the thresholded estimator of Bickel and Levina (2008) due to computational simplicity, generality and good theoretical properties.

5. Theoretical results

Let $\boldsymbol{\Sigma}_0$ denote the true covariance matrix of the vector $\boldsymbol{X}$ and $\boldsymbol{V}_0$ the true candidate matrix for which $\text{col}(\boldsymbol{V}_0) = \text{col}(\boldsymbol{B}_0)$, where $\boldsymbol{B}_0$ is the true matrix of unknown parameters. Let $S_d = \{k : \boldsymbol{\beta}_{d,0,k} \neq 0\}$ be the support of $\boldsymbol{\beta}_{d,0}$, S_d^c the complement of S_d and $s_d = |S_d|$ the cardinality of S_d.

Regularity conditions: Assume

(a) $\text{rank}(\boldsymbol{\Sigma}_0)=r$ and $\text{rank}(\boldsymbol{V}_0)=D$ with $r \geq D$,
(b) there exist constants C_1 and C_2 such that

$$C_1 \geq \tau_1 \geq \ldots \geq \tau_r > 0$$
$$C_2 \geq \lambda_1 \geq \ldots \geq \lambda_D > 0$$

where $\tau_1, \ldots, \tau_r$ and $\lambda_1, \ldots, \lambda_D$ are the ordered eigenvalues of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{V}_0$.
(c) $\tilde{\mathcal{X}}^\dagger$ satisfies a compatibility condition with constant $\phi_0 > 0$ with respect to the true support for each $d = 1, \ldots, D$ i.e.

$$\frac{1}{n} \mid\mid \tilde{\mathcal{X}}^\dagger \boldsymbol{v} \mid\mid_2^2 \geq \frac{\phi_0^2}{s_d} \mid\mid \boldsymbol{v}_{S_d} \mid\mid_1 \quad \forall \boldsymbol{v} \in \mathbb{R}^p \quad \text{such that}$$
$$\mid\mid \boldsymbol{v} \mid\mid_1 > 0 \text{ and } \mid\mid \boldsymbol{v}_{S_d^c} \mid\mid_1 \leq 3 \mid\mid \boldsymbol{v}_{S_d} \mid\mid_1 .$$

Condition (a) ensures the problem is well-posed, in the sense that if $r < D$ then one cannot recover the total number of directions directly from $\boldsymbol{\Sigma}_0$. Condition (b) ensures that both $\boldsymbol{\Sigma}_0$ and $\boldsymbol{V}_0$ are well-behaving with bounded eigenvalues. Condition (c) is a standard technical Lasso condition on the design matrix, which is generally regarded not as strict as the 'Restricted eigenvalue' condition present in the works of Bickel et al. (2009), Raskutti et al. (2010) and Lin et al. (2019) (for sufficient dimension purposes) among many others. We argue the above conditions are mild conditions which ensure a standard application of Lasso techniques and results in the PSVM framework when $D = 1$.

Lemma 1. *Let $\hat{\boldsymbol{\beta}}_d$ be the LassoPSVM solution of (3.4) when $\mu = A\sigma\sqrt{\frac{\log p}{n}}$, $D = 1$ and $\tilde{\boldsymbol{\beta}}_{d,0}$ be a vector proportional to the true vector that satisfies (2.1). For a sufficiently large constant A, we have with high probability that*

$$\mid\mid \frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{\mathcal{X}}^\dagger (\hat{\boldsymbol{\beta}}_d - \tilde{\boldsymbol{\beta}}_{d,0}) \mid\mid_\infty \leq \mu.$$

Proof. Let $\mathcal{B}$ be the event $\mathcal{B} = \{\mid\mid (\tilde{\mathcal{X}}^\dagger)^\intercal \boldsymbol{\epsilon}/n \mid\mid_\infty \leq \mu\}$. By the basic Lasso inequality and Lemma 6.2 in Bühlmann and Van De Geer (2011) we know this event holds with high probability. We work further on this event. Now from the KKT conditions for Lasso, we have that $\hat{\boldsymbol{\beta}}_d$ is a solution of (3.4) if $\mathbf{0} \in \partial\mathcal{L}(\boldsymbol{\beta}_d)|_{\hat{\boldsymbol{\beta}}_d}$ implying that

$$\mu\boldsymbol{\nu} = \frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal(\tilde{y}_d^\dagger - \tilde{\mathcal{X}}^\dagger \hat{\boldsymbol{\beta}}_d),$$

where $\boldsymbol{\nu}$ is a vector such that $\mid\mid \boldsymbol{\nu} \mid\mid_\infty \leq 1$. This happens if and only if

$$\mid\mid \frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal(\tilde{y}_d^\dagger - \tilde{\mathcal{X}}^\dagger \hat{\boldsymbol{\beta}}_d) \mid\mid_\infty \leq \mu. \tag{5.1}$$

Working on the event $\mathcal{B}$ when $\hat{\boldsymbol{\beta}}_d$ is a LassoPSVM solution, also implies that $||\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{y}_d^\dagger||_\infty \leq \mu$. This bound is interesting in itself as it informs us that in an ℓ_∞ sense, the components of this vector are well controlled when p grows. To see why this inequality holds we work by contradiction. Suppose $||\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{y}_d^\dagger||_\infty > \mu$. As $\hat{\boldsymbol{\beta}}_d$ is a Lasso solution we argued (5.1) must hold. On the other hand, due to the triangle inequality, we have that

$$|| \frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal(\tilde{y}_d^\dagger - \tilde{\mathcal{X}}^\dagger\hat{\boldsymbol{\beta}}_d) ||_\infty \leq || \frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{y}_d^\dagger||_\infty + ||\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger\hat{\boldsymbol{\beta}}_d ||_\infty > \mu$$

since $||\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{y}_d^\dagger||_\infty > \mu$ (by our supposition) and since $||\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger\hat{\boldsymbol{\beta}}_d ||_\infty \geq 0$. We have arrived at a contradiction, hence the supposition is false.

Now since,

$$\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger(\hat{\boldsymbol{\beta}}_d - \tilde{\boldsymbol{\beta}}_{d,0}) = \frac{1}{n}((\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger\hat{\boldsymbol{\beta}}_d - (\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger\tilde{\boldsymbol{\beta}}_{d,0})$$

and using that

$$\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{y}_d^\dagger \propto \frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger\tilde{\boldsymbol{\beta}}_{d,0}$$

for which we can guarantee a good control i.e. $||\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{y}_d^\dagger||_\infty \leq \mu$ implies that the term in (5.1) can be rewritten as

$$||\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger\tilde{\boldsymbol{\beta}}_{d,0} - \frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger\hat{\boldsymbol{\beta}}_d||_\infty = ||\frac{1}{n}(\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger(\tilde{\boldsymbol{\beta}}_{d,0} - \hat{\boldsymbol{\beta}}_d)||_\infty$$

as stated in the lemma. □

Proposition 2. *Let $\hat{\boldsymbol{\beta}}_d$ be the LassoPSVM solution of (3.4) when $\mu = A\sigma\sqrt{\frac{\log p}{n}}$, $D = 1$ and $\tilde{\boldsymbol{\beta}}_{d,0}$ be a vector proportional to the true vector that satisfies (2.1). Assume $\tilde{\mathcal{X}}^\dagger$ satisfies a compatibility condition with respect to the true support S_d for a compatibility constant $\phi_0 > 0$. For sufficiently large constants A, A' and A'' one has with high probability that*

$$|| \hat{\boldsymbol{\beta}}_d - \tilde{\boldsymbol{\beta}}_{d,0} ||_1 \leq \frac{A' s_d}{\phi_0^2}\sqrt{\frac{\log p}{n}}$$
$$|| \tilde{\mathcal{X}}^\dagger(\hat{\boldsymbol{\beta}}_d - \tilde{\boldsymbol{\beta}}_{d,0}) ||_2^2 \leq \frac{A''\sigma^2 s_d \log p}{n\phi_0^2}.$$

Proof. Under the compatibility assumption we analyse the vector $\boldsymbol{v} = \hat{\boldsymbol{\beta}}_d - \tilde{\boldsymbol{\beta}}_{d,0}$ implying

$$\frac{\phi_0^2}{s_d} || \boldsymbol{v} ||_1^2 \leq \frac{1}{n} || \tilde{\mathcal{X}}^\dagger\boldsymbol{v} ||_2^2 \leq \frac{1}{n} || (\tilde{\mathcal{X}}^\dagger)^\intercal\tilde{\mathcal{X}}^\dagger\boldsymbol{v} ||_\infty || \boldsymbol{v} ||_1 .$$

Using Lemma 1 we can bound $\frac{\phi_0^2}{s_d} || \boldsymbol{v} ||_1^2 \leq \mu || \boldsymbol{v} ||_1$. Now,

$$\mu || \boldsymbol{v} ||_1 = \mu(|| \boldsymbol{v}_{S_d^c} ||_1 + || \boldsymbol{v}_{S_d} ||_1) \leq 4\mu || \boldsymbol{v}_{S_d} ||_1$$

due to the constraint under the compatibility condition. As such,

$$|| \boldsymbol{v} ||_1^2 \leq 4\mu \frac{s_d}{\phi_0^2} || \boldsymbol{v}_{S_d} ||_1 \leq 4\mu \frac{s_d}{\phi_0^2} || \boldsymbol{v}_{S_d} + \boldsymbol{v}_{S_d^c} ||_1 = 4\mu \frac{s_d}{\phi_0^2} || \boldsymbol{v} ||_1$$

implying that $|| \boldsymbol{v} ||_1 \leq 4\mu \frac{s_d}{\phi_0^2}$ since $|| \boldsymbol{v} ||_1 > 0$. The first result directly follows by plugging in $\mu = A\sigma\sqrt{\frac{\log p}{n}}$.

The second result follows from

$$\frac{1}{n} || \tilde{\mathcal{X}}^{\dagger} \boldsymbol{v} ||_2^2 \leq 4\mu || \boldsymbol{v}_{S_d} ||_1 \leq 4\mu || \boldsymbol{v} ||_1 \leq 16\mu^2 \frac{s_d}{\phi_0^2}$$

by plugging in $\mu = A\sigma\sqrt{\frac{\log p}{n}}$. □

The conditions $D = 1$ and $\tilde{\boldsymbol{\beta}}_{d,0}$ satisfies (2.1) in Lemma 1 and Proposition 2 put us in the single index model framework. If one considers a stronger restricted eigenvalue condition on the design one can obtain the bound

$$\begin{aligned} &|| \hat{\boldsymbol{\beta}}_d (\hat{\boldsymbol{\beta}}_d^{\intercal} \hat{\boldsymbol{\beta}}_d)^{-1} \hat{\boldsymbol{\beta}}_d^{\intercal} - \boldsymbol{\beta}_{d,0} (\boldsymbol{\beta}_{d,0}^{\intercal} \boldsymbol{\beta}_{d,0})^{-1} \boldsymbol{\beta}_{d,0}^{\intercal} ||_F \\ &\leq \frac{C}{||\tilde{\boldsymbol{\beta}}_{d,0}||_2} \sqrt{\frac{s_d \log p}{n\phi_1^2}} \leq \frac{C'}{||\boldsymbol{\beta}_{d,0}||_2} \sqrt{\frac{s_d \log p}{n\phi_1^2}}, \end{aligned}$$

where $\phi_1 > 0$ is now the constant for which the restricted eigenvalue condition is satisfied and C' is a constant arbitrary large. Note that now the restricted eigenvalue condition is much stronger since we require it to hold for any vector of the form $\boldsymbol{v} = \hat{\boldsymbol{\beta}}_d - \tilde{\boldsymbol{\beta}}_{d,0}$ where $\tilde{\boldsymbol{\beta}}_{d,0}$ satisfies (2.1).

Proposition 3. *Let $\hat{\boldsymbol{\beta}}_d$ be the LassoPSVM solution of (3.4) when $\mu = A\sigma\sqrt{\frac{\log p}{n}}$ and $D > 1$. Let $\boldsymbol{B}_0$ be the true matrix of unknown coefficients. Assume $\tilde{\mathcal{X}}^{\dagger}$ satisfies a restricted eigenvalue condition for a constant $\phi_1 > 0$ with respect to the true support for each $d = 1, \ldots, D$:*

$$\begin{aligned} &\frac{1}{n} || \tilde{\mathcal{X}}^{\dagger} \boldsymbol{v} ||_2^2 \geq \phi_1^2 || \boldsymbol{v} ||_2^2 \quad \forall \boldsymbol{v} \in \mathbb{R}^p \quad \textit{such that} \\ &|| \boldsymbol{v} ||_1 > 0 \textit{ and } || \boldsymbol{v}_{S_d^c} ||_1 \leq 3 || \boldsymbol{v}_{S_d} ||_1 . \end{aligned}$$

For sufficiently large A and with $||\boldsymbol{\Sigma}_0 \tilde{\boldsymbol{\beta}}_{d,0}||_{\infty} > \mu$, $\forall d = 1, \ldots, D$ one has with high probability

$$|| \hat{\boldsymbol{B}} (\hat{\boldsymbol{B}}^{\intercal} \hat{\boldsymbol{B}})^{-1} \hat{\boldsymbol{B}}^{\intercal} - \boldsymbol{B}_0 (\boldsymbol{B}_0^{\intercal} \boldsymbol{B}_0)^{-1} \boldsymbol{B}_0^{\intercal} ||_F \leq C'' \sqrt{\frac{\max_d (s_d) \log p}{n\phi_1^2}}$$

where C'' is an arbitrary large constant.

Proof. First note that $\forall d \neq d' \; \exists \; \iota \in (0,1)$ such that $|\cos(\angle(\hat{\boldsymbol{\beta}}_d; \hat{\boldsymbol{\beta}}_{d'}))| \leq \iota$ implying that any two estimated vectors are not orthogonal. This can easily be shown by contradiction.

Consider without loss of generality that $D = 2$. Then $\cos(\theta) = \frac{\hat{\boldsymbol{\beta}}_1^\intercal \hat{\boldsymbol{\beta}}_2}{||\hat{\boldsymbol{\beta}}_1||_2 ||\hat{\boldsymbol{\beta}}_2||_2} = \boldsymbol{a}^\intercal \boldsymbol{b}$, where θ is the angle between the two component vectors $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ and $\boldsymbol{a}$ and $\boldsymbol{b}$ are the unit vectors obtained after normalization. Suppose there exist vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ with $\boldsymbol{a} \neq \boldsymbol{b}$ such that $||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1$ and $\boldsymbol{a}^\intercal \boldsymbol{b} = 1$ which implies the two vectors are orthogonal. This implies

$$\begin{aligned}
\sqrt{a_1^2 + a_2^2} = 1 &\equiv a_1^2 = 1 - a_2^2 \\
\sqrt{b_1^2 + b_2^2} = 1 &\equiv b_1^2 = 1 - b_2^2 \\
a_1 b_1 + a_2 b_2 = 1 &\equiv a_1 = \frac{1 - a_2 b_2}{b_1}.
\end{aligned}$$

These relations also imply that

$$\begin{aligned}
1 - a_2^2 - b_2^2 + a_2^2 b_2^2 &= 1 + a_2^2 b_2^2 - 2 a_2 b_2 \\
0 &= (a_2 - b_2)^2
\end{aligned}$$

which holds as long as $a_2 = b_2$. Similarly one can arrive at $a_1 = b_1$, however this is not allowed (due to our supposition), hence the contradiction.

Secondly, also by contradiction one can show that if $||\boldsymbol{\Sigma}_0 \tilde{\boldsymbol{\beta}}_{d,0}||_\infty > \mu$ then necessarily $||\hat{\boldsymbol{\beta}}_d||_\infty > 0$ with high probability for it to be a solution of (3.4). This implies further that the lengths of the vectors $\hat{\boldsymbol{\beta}}_d;\ d = 1, \ldots, D$ are bounded away from zero.

Suppose there exists a vector $||\hat{\boldsymbol{\beta}}_d||_\infty = 0$ which is a LassoPSVM solution when $||\boldsymbol{\Sigma}_0 \tilde{\boldsymbol{\beta}}_{d,0}||_\infty > \mu$. In light of Lemma 1 this happens iff

$$\begin{aligned}
|| \frac{1}{n} (\tilde{\mathcal{X}}^\dagger)^\intercal (\tilde{y}_d^\dagger - \tilde{\mathcal{X}}^\dagger \hat{\boldsymbol{\beta}}_d) \, ||_\infty &\leq \mu \\
||\frac{1}{n} (\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{\mathcal{X}}^\dagger \tilde{\boldsymbol{\beta}}_{d,0} - \frac{1}{n} (\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{\mathcal{X}}^\dagger \hat{\boldsymbol{\beta}}_d||_\infty &\leq \mu \\
||\frac{1}{n} (\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{\mathcal{X}}^\dagger \tilde{\boldsymbol{\beta}}_{d,0}||_\infty &\leq \mu.
\end{aligned}$$

However when $n \to \infty$, the LHS term approaches $\boldsymbol{\Sigma}_0 \tilde{\boldsymbol{\beta}}_{d,0}$ for which $||\boldsymbol{\Sigma}_0 \tilde{\boldsymbol{\beta}}_{d,0}||_\infty > \mu$ by assumption, whereas μ approaches 0. Hence the contradiction.

Using further Gram-Schmidt orthogonalisation, we can argue as in Lin et al. (2019) that

$$\begin{aligned}
&|| \, \hat{\boldsymbol{B}} (\hat{\boldsymbol{B}}^\intercal \hat{\boldsymbol{B}})^{-1} \hat{\boldsymbol{B}}^\intercal - \boldsymbol{B}_0 (\boldsymbol{B}_0^\intercal \boldsymbol{B}_0)^{-1} \boldsymbol{B}_0^\intercal \, ||_F \\
=&|| \sum_{d=1}^{D} \hat{\boldsymbol{\beta}}_d (\hat{\boldsymbol{\beta}}_d^\intercal \hat{\boldsymbol{\beta}}_d)^{-1} \hat{\boldsymbol{\beta}}_d^\intercal - \sum_{d=1}^{D} \boldsymbol{\beta}_{d,0} (\boldsymbol{\beta}_{d,0}^\intercal \boldsymbol{\beta}_{d,0})^{-1} \boldsymbol{\beta}_{d,0}^\intercal \, ||_F \\
=&|| \sum_{d=1}^{D} \hat{\boldsymbol{\beta}}_d (\hat{\boldsymbol{\beta}}_d^\intercal \hat{\boldsymbol{\beta}}_d)^{-1} \hat{\boldsymbol{\beta}}_d^\intercal - \sum_{d=1}^{D} \tilde{\boldsymbol{\beta}}_{d,0} (\tilde{\boldsymbol{\beta}}_{d,0}^\intercal \tilde{\boldsymbol{\beta}}_{d,0})^{-1} \tilde{\boldsymbol{\beta}}_{d,0}^\intercal \, ||_F
\end{aligned}$$

$$\leq \sum_{d=1}^{D} ||\hat{\boldsymbol{\beta}}_d(\hat{\boldsymbol{\beta}}_d^{\intercal}\hat{\boldsymbol{\beta}}_d)^{-1}\hat{\boldsymbol{\beta}}_d^{\intercal} - \tilde{\boldsymbol{\beta}}_{d,0}(\tilde{\boldsymbol{\beta}}_{d,0}^{\intercal}\tilde{\boldsymbol{\beta}}_{d,0})^{-1}\tilde{\boldsymbol{\beta}}_{d,0}^{\intercal}||_F$$
$$\leq D \max_d(||\hat{\boldsymbol{\beta}}_d(\hat{\boldsymbol{\beta}}_d^{\intercal}\hat{\boldsymbol{\beta}}_d)^{-1}\hat{\boldsymbol{\beta}}_d^{\intercal} - \tilde{\boldsymbol{\beta}}_{d,0}(\tilde{\boldsymbol{\beta}}_{d,0}^{\intercal}\tilde{\boldsymbol{\beta}}_{d,0})^{-1}\tilde{\boldsymbol{\beta}}_{d,0}^{\intercal}||_F)$$
$$\leq C'' \sqrt{\frac{\max_d(s_d)\log p}{n\phi_1^2}}.$$

since D is fixed. □

6. An inferential procedure based on desparsification

Following the steps of van de Geer et al. (2014), Zhang and Zhang (2014) and Javanmard and Montanari (2014) one can also construct approximate confidence intervals for the parameters $\tilde{\boldsymbol{\beta}}_{d,0}$ using the LassoPSVM procedure.

From the KKT conditions for Lasso, we have that $\hat{\boldsymbol{\beta}}_d$ satisfies

$$\mu\boldsymbol{\nu} = \frac{1}{n}((\tilde{\mathcal{X}}^{\dagger})^{\intercal}\tilde{y}_d^{\dagger} - (\tilde{\mathcal{X}}^{\dagger})^{\intercal}\tilde{\mathcal{X}}^{\dagger}\hat{\boldsymbol{\beta}}_d)$$
$$\mu\boldsymbol{\nu} \propto \frac{1}{n}(\tilde{\mathcal{X}}^{\dagger})^{\intercal}\tilde{\mathcal{X}}^{\dagger}\tilde{\boldsymbol{\beta}}_{d,0} - \frac{1}{n}(\tilde{\mathcal{X}}^{\dagger})^{\intercal}\tilde{\mathcal{X}}^{\dagger}\hat{\boldsymbol{\beta}}_d$$
$$\mu\boldsymbol{\nu} \propto \tilde{\boldsymbol{S}}^{\dagger}(\tilde{\boldsymbol{\beta}}_{d,0} - \hat{\boldsymbol{\beta}}_d)$$
$$\frac{1}{n}\hat{\boldsymbol{\Theta}}((\tilde{\mathcal{X}}^{\dagger})^{\intercal}\tilde{y}_d^{\dagger} - (\tilde{\mathcal{X}}^{\dagger})^{\intercal}\tilde{\mathcal{X}}^{\dagger}\hat{\boldsymbol{\beta}}_d) + \hat{\boldsymbol{\beta}}_d \propto \tilde{\boldsymbol{\beta}}_{d,0}$$

where $\hat{\boldsymbol{\Theta}}$ is a relaxed version of the inverse of $\tilde{\boldsymbol{S}}^{\dagger}$. This suggests the desparsified estimator

$$\hat{\boldsymbol{b}}_d = \frac{1}{n}\hat{\boldsymbol{\Theta}}\Bigg(\frac{1}{4\hat{\lambda}_d}(\tilde{\mathcal{X}}^{\dagger})^{\intercal}\sum_{h=1}^{H-1}(\tilde{\boldsymbol{Y}}^h \odot \boldsymbol{\alpha})(\tilde{\boldsymbol{Y}}^h \odot \boldsymbol{\alpha})^{\intercal}\mathcal{X}^{\dagger}((\mathcal{X}^{\dagger})^{\intercal}\mathcal{X}^{\dagger})^{-1}\boldsymbol{P}^{\intercal}\hat{\boldsymbol{\eta}}_d - (\tilde{\mathcal{X}}^{\dagger})^{\intercal}\tilde{\mathcal{X}}^{\dagger}\hat{\boldsymbol{\beta}}_d\Bigg) + \hat{\boldsymbol{\beta}}_d.$$

The above expression used the approximation $(\tilde{\mathcal{X}}^{\dagger})^{\intercal}\tilde{y}_d^{\dagger} \propto (\tilde{\mathcal{X}}^{\dagger})^{\intercal}\tilde{\mathcal{X}}^{\dagger}\boldsymbol{\beta}_d$. Equation (3.4) characterizes this approximation, in the sense that the model implicitly assumes $\tilde{y}_d^{\dagger} = \tilde{\mathcal{X}}^{\dagger}\tilde{\boldsymbol{\beta}}_d + \varphi$ where $\varphi \sim N(0, \sigma_{\varphi}^2\boldsymbol{I})$. From this perspective, Theorems 2.1 and 2.2 from van de Geer et al. (2014) under their specified conditions on the design, apply directly. As such, a hypothesis test for assessing the relevance of variables in determining the sufficient direction can be set as follows. For each component $\tilde{\beta}_{d,j}$ of the vector $\tilde{\boldsymbol{\beta}}_d$, where $j = 1, \ldots, p$ we test:

$$H_0^j : \tilde{\beta}_{d,j} = 0 \quad \text{versus}$$
$$H_a^j : \tilde{\beta}_{d,j} \neq 0.$$

Under H_0 we have that $|\sqrt{n}\,\hat{b}_{d,j}/s_j| \leq \Phi^{-1}(1-\alpha/2)$ with probability $1-\alpha$, with $s_j^2 = (\frac{1}{n}\sigma_\varphi^2 \hat{\boldsymbol{\Theta}}(\tilde{\mathcal{X}}^\dagger)^\intercal \tilde{\mathcal{X}}^\dagger \hat{\boldsymbol{\Theta}}^\intercal)_{j,j}$ (i.e. the j-th element on the matrix diagonal) and $\Phi(\cdot)$ the standard normal cumulative distribution function. As σ_φ^2 is unknown, we replace it by a consistent estimator.

7. Numerical studies

In this section we present the performance of our proposed procedure on simulated and two gene expression real datasets as well as a phonetic dataset.

7.1. Simulation study: Frobenius loss performance

The performance of the LassoPSVM procedure has been investigated in a controlled simulation study. We have benchmarked our procedure against LassoSIR (Lin et al., 2019), as this procedure has been observed in practice to produce competitive results and most importantly, it was designed for high-dimensional data as it produces sparse estimators for the coefficient matrix $\boldsymbol{B}$. Due to these similarities, we argue that the two methods are directly comparable.

Performance for both competitors has been measured by a Frobenius loss defined as:

$$\text{Loss} = || \hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^\intercal \hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{B}}^\intercal - \boldsymbol{B}_0(\boldsymbol{B}_0^\intercal \boldsymbol{B}_0)^{-1}\boldsymbol{B}_0^\intercal ||_F$$

where $\hat{\boldsymbol{B}}$ represents the estimated coefficient matrix, $\boldsymbol{B}_0$ represents the true coefficient matrix and $||\cdot||_F$ is the Frobenius norm. Smaller values for the loss, denote a better performance.

Three data generating models have been used throughout the section, defined as:

Model 1: $Y = X_1 + X_2 + \ldots + X_k + \sigma\epsilon$, where $k = \lfloor 2\%p \rfloor$;
Model 2: $Y = X_1/(.5 + (X_2 + 1)^2) + \sigma\epsilon$;
Model 3: $Y = X_1(X_1 + X_2 + 1) + \sigma\epsilon$;

Model 1 investigates the case of a linear, low or high-dimensional model (depending on the value of p, the number of predictors), where the sparsity coefficient (i.e. the number of active components) is set at 2% of the total number of predictors. Models 2 and 3 investigate the very sparse, non-linear case. The difference between model 1 and models 2 and 3 lies in the fact that model 1 needs one effective direction to model the central subspace, whereas models 2 and 3 need two distinct directions. The difference between models 2 and 3 is that X_1 affects both directions in model 3, while in model 2 there are two different variables to define the two directions.

In each model $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = (\sigma_{ab})$ obeys a Toeplitz structure with $\sigma_{ab} = \rho^{|a-b|}$ and $\rho \in \{0, .5, .8\}$. This gave rise to three different scenarios which range from uncorrelated to highly correlated predictors. The errors ϵ followed $\epsilon \sim N(0, 1)$ and $\sigma \in \{.5, 1\}$. From each model we sampled $n = 300$ independent observations and p, the number of components,

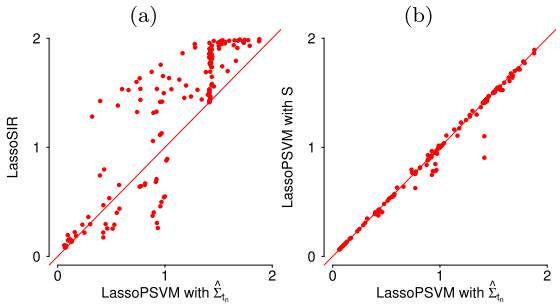

FIG 1. *Simulation data. Frobenius loss (smaller is better) for LassoPSVM with* $\hat{\boldsymbol{\Sigma}}_{t_n}$ *(x-axis) and LassoSIR (y-axis) in panel (a) and LassoPSVM with* $\boldsymbol{S}$ *(y-axis) in panel (b). Values above the main diagonal show a better performance of LassoPSVM with* $\hat{\boldsymbol{\Sigma}}_{t_n}$*. Each symbol represents the empirical median of 300 repetitions from a scenario and the results of 144 different scenarios are plotted.*

was set to $p \in \{100, 500, 1000, 2000\}$, while the number of slices was set to $H \in \{5, 20\}$. Using each distinct value of the different data generating parameters, 144 different simulation scenarios have been created and for each scenario 300 independent repetitions from the same generating process were performed.

For both competitors a 10-fold cross-validation scheme has been used for selecting the tuning parameter μ that dictates the sparsity level of the coefficients $\hat{\boldsymbol{\beta}}$ and both competitors have been given information about the true number of directions needed to model the data.

As pointed out at the end of Section 4, in the first step of the algorithm one needs an estimator of the covariance matrix $\boldsymbol{\Sigma}$. We have investigated further in this section two estimators: (i) the sample covariance matrix $\boldsymbol{S}$ and (ii) the thresholded estimator of Bickel and Levina (2008). For the thresholded estimator of the $\boldsymbol{\Sigma}$ matrix, i.e. $\hat{\boldsymbol{\Sigma}}_{t_n}$, the thresholding level was set to $t_n = M\sqrt{\log p/n}$ where $M = .1$ was used. A cross-validation strategy can be used here as well to select M, but from our experiments we have observed that this value performed satisfactorily.

Figure 1 presents the obtained results. In the plot each symbol represents the empirical median over the 300 repetitions for each of the 144 scenarios and it suggests that LassoPSVM provided similar or better results in a large majority of different scenarios when compared to the LassoSIR procedure. In only a small number of scenarios, the performance seemed to be slightly worse than that of LassoSIR. Panel (b) illustrates that using the thresholded estimator or the empirical estimator leads to similar conclusions, probably due to the choice of the thresholding level t_n.

To illustrate the benefits of using the thresholded estimator rather than the empirical covariance matrix, we plot in Figure 2 the difference between $\hat{\boldsymbol{\Sigma}}_{t_n}$ and the true $\boldsymbol{\Sigma}$, and between $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ in Frobenius norm. The difference is plotted as a function of M for a particular configuration where $n = 300$, $p = 100$, $H = 5$, $\sigma = 0.5$, $\rho = .8$ and Model 3 is used as a data generating process,

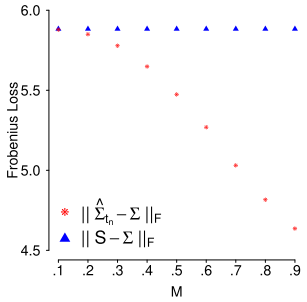

FIG 2. *Simulation data. Frobenius loss (smaller is better) of the difference between* $\hat{\boldsymbol{\Sigma}}_{t_n}$ *and the true* $\boldsymbol{\Sigma}$ *and between* $\boldsymbol{S}$ *and* $\boldsymbol{\Sigma}$*. Each symbol represents the empirical median of 300 repetitions from the same scenario where only the value* M *changes.*

which in principle favors the empirical covariance estimator, as the example is low-dimensional and enjoys a high signal to noise ratio. As the figure illustrates, for higher values of the thresholding parameter, the accuracy of reconstructing the true underlying $\boldsymbol{\Sigma}$ increases for $\hat{\boldsymbol{\Sigma}}_{t_n}$ relative to the empirical covariance estimator. Similar trends have been observed for other different configurations of parameters, but are not reported here.

We 'zoom in' further on the performance of the methods as a function of the data generating parameters in Figure 3. The figure suggests that under the non-linear models 2 and 3, the LassoPSVM is better equipped at identifying the effective directions in the data than the LassoSIR. The same holds for the settings where $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $p = 100$ for which LassoPSVM always provided comparable or better performance. For all other values of the data generating parameters, the performance of the LassoPSVM is generally better (by which we mean that the number of settings where it outperforms, is larger than the number of settings where it underperforms) than that of LassoSIR, but for all these cases we could identify a few settings where the LassoSIR is better performing than LassoPSVM.

To conclude this section, we remark that the simulation study suggests that none of the techniques is uniformly better than the other one, but in general LassoPSVM provided better results for more settings than LassoSIR. LassoPSVM proved thus to be a worthy competitor to LassoSIR in a controlled, finite sample setting.

7.2. Simulation study: Type I error and power

In this section we evaluate the Type I error rate and power of the hypothesis test proposed in Section 6 with significance level $\alpha = 5\%$. We consider the models:

Model 4: $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \sigma\epsilon$,
Model 5: $Y = X_1(X_1 + X_2 + X_3 + X_4 + X_5 + 0.5) + \sigma\epsilon$,

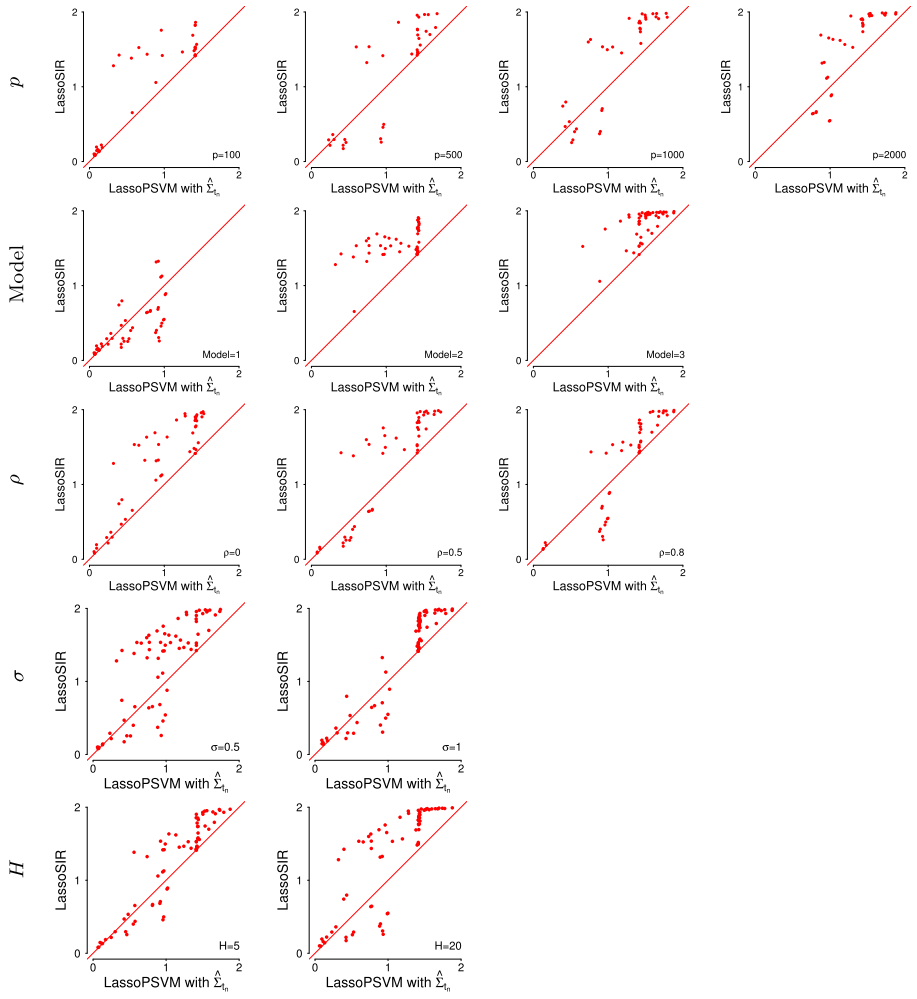

FIG 3. *Simulation data. Results of the 144 scenarios split by values of p, model, values of ρ, σ and H.*

where the total number of predictors p takes values in the set $\{500, 1000\}$, $\sigma \in \{.5, 1\}$ and $\epsilon \sim N(0, 1)$. The vector $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\sigma_{ab} = \rho^{|a-b|}$ with $\rho \in \{0, .5, .8\}$. We sampled $n = 300$ independent observations, while the number of slices was set to $H = 5$ and 20.

We evaluate the performance of the test by the average empirical power and average empirical Type I error defined as:

$$\text{Power}_{S_0} = \frac{1}{|S_0|} \sum_{k \in S_0} P(\text{Reject } H_0^k | H_a^k)$$

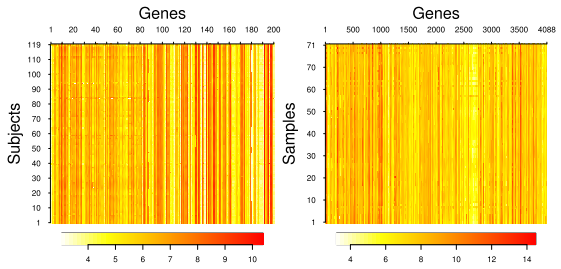

FIG 4. *Heatmap of the predictor gene expression levels for the Scheetz et al. (2006) dataset (left panel) and for the Bühlmann et al. (2014) dataset (right panel).*

$$\text{Type I}_{S_0^c} = \frac{1}{p - |S_0|} \sum_{k \in S_0^c} P(\text{Reject } H_0^k | H_0^k)$$

where S_0 and S_0^c are the sets of active/non-active components.

Model 4 investigates the case of a linear, high-dimensional model where five components of the total number of predictors are active and where one sufficient direction is needed to model the underlying signal. Model 5 investigates the sparse, non-linear case where one needs two distinct directions to model the underlying signal. For model 5 only the first component is active for the first direction and the first 5 components are active for the second direction. The competitor procedures used in this section are the desparsified LassoPSVM with $\boldsymbol{S}$ and $\hat{\boldsymbol{\Sigma}}_{t_n}$ and the desparsified Lasso (van de Geer et al., 2014). For LassoPSVM with $\hat{\boldsymbol{\Sigma}}_{t_n}$ a data splitting approach as suggested in Bickel and Levina (2008) was repeated 10 times on a grid of 10 different values in order to select an optimal thresholding level t_n.

Table 1 presents the obtained results and it suggests that both competitors control well the average type I error rate at the target level of 5%. For model 4, the average power with respect to the specified alternative seems to decrease as ρ and H increase when $p = 500$, but seems to be comparable to that of Lasso for $p = 1000$. For the non-linear model 5, LassoPSVM captures much better the important variable for the first direction, while being slightly less powerful than the Lasso with respect to the components active on the second direction. As the model is highly non-linear the obtained powers are lower than in the case of model 4 and increasing the correlation between predictors severely reduces the performance of the tests. Overall, using LassoPSVM with $\boldsymbol{S}$ or $\hat{\boldsymbol{\Sigma}}_{t_n}$ provided very close results.

7.3. Applications to real data: the continuous case

In this section we discuss the application of the LassoPSVM procedure to two real datasets. The first dataset we analyse is a simplified version of the 'Eye' gene expression data from Scheetz et al. (2006). The data contain information about the expression level of $p = 200$ genes (predictors) for a total of $n = 120$

TABLE 1. *Simulation data. $Power_{S_0}$ (top panels) and Type $I_{S_0^c}$ (bottom panels) for desparsified LassoPSVM with $\boldsymbol{S}$ and $\hat{\boldsymbol{\Sigma}}_{t_n}$ and desparsified Lasso.*

			Model 4					Model 5								
			$\boldsymbol{S}$		$\hat{\boldsymbol{\Sigma}}_{t_n}$		Desp.Lasso	$\boldsymbol{S}$				$\hat{\boldsymbol{\Sigma}}_{t_n}$				Desp.Lasso
								$d=1$		$d=2$		$d=1$		$d=2$		
p	σ	ρ \ H	5	20	5	20		5	20	5	20	5	20	5	20	
500	.5	0	100	100	100	100	100	69	69	19	19	70	69	19	19	23
500	.5	.5	99	94	99	94	100	40	44	12	13	40	44	12	13	18
500	.5	.8	56	33	56	33	100	22	22	8	8	21	22	8	8	11
500	1	0	100	100	100	100	100	65	65	19	19	65	64	20	19	23
500	1	.5	100	96	100	96	100	37	39	11	12	37	39	12	12	17
500	1	.8	86	37	86	36	99	20	21	7	7	20	22	7	7	11
1000	.5	0	100	100	100	100	100	69	70	20	20	68	70	20	19	22
1000	.5	.5	100	100	100	100	100	40	40	13	13	38	39	13	13	16
1000	.5	.8	98	98	98	98	100	20	21	9	9	20	21	8	9	11
1000	1	0	100	100	100	100	100	69	70	19	19	70	69	19	19	23
1000	1	.5	100	100	100	100	100	40	41	13	13	41	41	13	13	18
1000	1	.8	95	98	94	98	100	23	23	9	9	22	23	9	9	11
500	.5	0	4.9	4.9	4.9	5.1	4.9	5.0	5.0	5.0	5.0	5.1	5.0	5.1	5.0	5.1
500	.5	.5	4.8	4.6	4.8	4.7	4.9	4.9	4.9	4.9	4.9	5.0	5.0	5.0	5.0	5.1
500	.5	.8	4.7	4.7	4.7	4.7	4.9	4.9	5.0	4.9	5.0	4.9	5.0	4.9	5.0	5.0
500	1	0	4.8	4.8	4.9	4.9	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
500	1	.5	4.9	4.6	5.0	4.6	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
500	1	.8	4.8	4.7	4.8	4.7	4.9	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
1000	.5	0	4.9	4.9	5.0	5.0	5.0	4.9	4.9	4.9	4.9	4.9	4.9	4.9	4.9	5.0
1000	.5	.5	4.9	4.9	5.0	5.0	5.0	5.0	5.1	5.0	5.1	5.0	5.1	5.0	5.1	5.0
1000	.5	.8	4.9	5.0	4.9	5.0	5.0	4.9	4.9	4.9	4.9	4.9	4.9	4.9	4.9	5.1
1000	1	0	5.0	4.9	5.1	5.1	5.0	5.0	4.9	5.0	4.9	5.0	4.9	5.0	4.9	5.0
1000	1	.5	5.0	4.9	5.0	4.9	5.0	4.9	5.0	4.9	5.0	5.0	5.0	5.0	5.0	5.0
1000	1	.8	4.9	4.9	5.0	4.9	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0

TABLE 2
LOO and Sd_{LOO} for LassoPSVM and LassoSIR for the 'Eye' and 'Riboflavin' datasets.

			$H = 3$	$H = 5$	$H = 10$
Eye	LassoPSVM	$d = 1$	70.56 (1.87)	70.56 (1.87)	70.56 (1.87)
		$d = 2$	70.56 (1.87)	70.56 (1.87)	70.56 (1.87)
	LassoSIR	$d = 1$	70.41 (1.85)	70.22 (1.87)	69.07 (1.82)
		$d = 2$	70.46 (1.86)	69.88 (1.91)	67.56 (1.85)
Riboflavin	LassoPSVM	$d = 1$	52.10 (13.92)	52.08 (13.92)	52.13 (13.94)
		$d = 2$	52.11 (13.93)	52.07 (13.92)	52.12 (13.93)
	LassoSIR	$d = 1$	52.07 (13.92)	52.08 (13.90)	51.88 (13.80)
		$d = 2$	52.11 (13.95)	52.26 (13.95)	52.28 (14.08)

rats, while the response is represented by the expression level of the TRIM32 gene for all rats. The second dataset we analyse is the 'Riboflavin' dataset from Bühlmann et al. (2014) which contains information regarding the riboflavin production by Bacillus subtilis. Information for $n = 71$ observations on $p = 4088$ predictors (gene expressions) and a one-dimensional response (riboflavin production) is recorded. Both datasets are freely provided in the `R`-based packages `flare` and `hdi`. A visual inspection of the information in these two datasets is offered in Figure 4.

We evaluate LassoPSVM and LassoSIR with respect to leave-one-out average mean square prediction error and average standard deviation defined as

$$\text{LOO} = \frac{1}{D}\sum_{d=1}^{D}\Big(\frac{1}{n}\sum_{i=1}^{n}\text{Err}_{i,d}\Big)$$

$$\text{Sd}_{\text{LOO}} = \frac{1}{D}\sum_{d=1}^{D}\Big(\frac{1}{n-1}\sum_{i=1}^{n}(\text{Err}_{i,d} - \overline{\text{Err}}_{i,d})^2\Big)$$

where $\text{Err}_{i,d} = (y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{d,(-i)})^2$ and $\overline{\text{Err}}_{i,d} = \frac{1}{n}\sum_{i=1}^{n}\text{Err}_{i,d}$. The values y_i and $\boldsymbol{x}_i$ represent the observed values for the response and the vector of gene expression levels for the i-th case and $\hat{\boldsymbol{\beta}}_{d,(-i)}$ is the d-th estimated direction, when the i-th case is excluded from the training sample.

Table 2 presents the obtained results when D (the number of directions) takes values in the set $\{1, 2\}$ and H (the number of slices) takes values in the set $\{3, 5, 10\}$ for both the LassoPSVM and LassoSIR techniques. It suggests that both techniques provide very similar results for the two datasets, thus showing that the proposed procedure is a worthy competitor to the LassoSIR procedure. Moreover, for LassoPSVM a BIC-type criterion as proposed in Artemiou and Dong (2016) suggests that $D = 1$ provides best results for both datasets and as such allowing for a larger number of directions does not sensibly improve its performance, which is confirmed by the results seen in the table.

To illustrate the usefulness of the testing procedure proposed in Section 6, we present in Table 3 the selected variables when a Bonferroni correction is applied to maintain a FWER of 10% and when LassoPSVM is compared to

TABLE 3
Selected variables by desparsified LassoPSVM, desparsified Lasso and Kernel HD SIM for the 'Eye' dataset. All procedures use a Bonferroni correction to maintain the FWER at 10%.

Variable Label	Desp. LassoPSVM			Desp. Lasso	Kernel HD SIM
	$H = 3$	$H = 5$	$H = 10$		
87		x	x	x	
96	x	x	x		
102					x
140			x		x
153		x	x	x	x
174	x				
180				x	
200		x		x	

the desparsified Lasso and the desparsified kernel-based procedure for high-dimensional linear single index models of Gueuning and Claeskens (2016). The Eye dataset is used as example. The table illustrates that variables 87, 140, 153 and 200 are selected by multiple techniques, while variables 96, 102, 174 and 180 are selected by one technique, but not the others. Moreover, variable 153 is deemed important by all three competitors.

7.4. Applications to real data: the discrete case

In this section we discuss the application of the LassoPSVM procedure to the dataset of Tsanas et al. (2014) from the UC Irvine repository available at `https://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation`. The dataset consists of $n = 126$ samples and $p = 309$ features, and the aim is to assess whether voice rehabilitation treatment leads to phonations considered 'acceptable' or 'unacceptable' (a binary classification problem).

We evaluate LassoPSVM and LassoSIR with respect to Sensitivity, Specificity and F_1 score defined as: Sensitivity $= A/(A+C)$; Specificity $= D/(B+D)$; and $F_1 = 2PR/(P+R)$, where $P = A/(A+B)$, $R =$ Sensitivity and the counts A, B, C and D come from the below classification table:

Predicted \ Observed	Acceptable	Unacceptable
Acceptable	A	B
Unacceptable	C	D

The predicted class is obtained for each case in a leave-one-out scheme when the training set uses all cases but the i-th case to make a prediction for the i-th case. To maintain comparability, we use the same strategy as proposed in the work of Lin et al. (2019): we first standardize each feature, apply LassoSIR and LassoPSVM to identify the directions $\hat{\boldsymbol{B}}$ and the corresponding components, followed by a logistic regression model on the training dataset, after which the

TABLE 4
Sensitivity, Specificity and F_1 values for LassoPSVM and LassoSIR for the 'LSVT Voice Rehabilitation' dataset.

	LassoPSVM		LassoSIR
	$\boldsymbol{S}$	$\hat{\boldsymbol{\Sigma}}_{t_n}$	
Sensitivity	.80	.88	.80
Specificity	.85	.83	.83
F_1	.76	.79	.76

probability for the left out case to belong to the acceptable/unacceptable class is calculated. Table 4 presents the obtained results when $D = 1$ and $H = 2$ for both the LassoPSVM and LassoSIR techniques. It illustrates that both techniques provide very similar results for the classification problem, thus showing again that the proposed procedure is a worthy competitor to LassoSIR.

8. Discussion

In this paper we have introduced a new method for sufficient dimension reduction which allows handling 'small n, large p' problems. The proposed method is based on an SVM framework for SDR, in conjunction with the use of a principal projections framework in order to avoid the singularity of the covariance matrix. The method outputs sparse directions and this is achieved by penalizing a transformed objective function with an ℓ_1 penalty. The method's performance is assessed on simulated and real datasets, where it provides in the majority of cases similar or better results than a state-of-the-art, direct competitor.

The method is flexible enough that it admits extensions in multiple directions, but probably the most interesting extension would be towards non-linear settings as framed in model (1.2). At first sight, connections with the Kernel PCA methods are directly exploitable, but a thorough treatment of the subject is topic for future research.

References

Artemiou, A. and Dong, Y. (2016). Sufficient dimension reduction via principal L_q support vector machine. *Electronic Journal of Statistics*, 10(1):783–805. MR3486417

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

Artemiou, A. and Shu, M. (2014). A cost based reweighted scheme of principal support vector machine. In Akritas, M. G., Lahiri, S. N., and Politis, D. N., editors, *Topics in Nonparametric Statistics*, pages 1–12. Springer. MR3333330

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604. MR2485008

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732. MR2533469

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820. MR2860325

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278. MR3432840

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer. MR2807761

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607. MR2847973

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144. MR2676885

Cook, R. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474. MR1902895

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26. MR2408655

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, 86(414):328–332. MR1137117

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356. MR3012410

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905. MR2533474

Gueuning, T. and Claeskens, G. (2016). Confidence intervals for high-dimensional partially linear single-index models. *Journal of Multivariate Analysis*, 149:13–29. MR3507312

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909. MR3277152

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411. MR2026339

Li, B., Artemiou, A., and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210. MR3012405

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008. MR2354409

Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616.

MR2166556

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86(414):316–327. MR1137117

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(420):1025–1039. MR1209564

Lin, Q., Zhao, Z., and Liu, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics*, 46(2):580–610. MR3782378

Lin, Q., Zhao, Z., and Liu, J. S. (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, 114(528):1726–1739. MR4047295

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Academic Press. MR0560319

Pircalabelu, E. and Artemiou, A. (2021). Graph informed sliced inverse regression. *Computational Statistics & Data Analysis*, 164:107302. MR4287759

Randall, H., Artemiou, A., and Qiao, X. (2021). Sufficient dimension reduction based on distance weighted discrimination. *Scandinavian Journal of Statistics*, 48:1186–1211. MR4377354

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259. MR2719855

Scheetz, T., Kim, K.-Y., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T., Sheffield, V., and Stone, E. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434.

Shin, S. J. and Artemiou, A. (2017). Penalized principal logistic regression for sparse sufficient dimension reduction. *Computational Statistics & Data Analysis*, 111:48–58. MR3630217

Shin, S. J., Wu, Y., Zhang, H. H., and Liu, Y. (2017). Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*, 104(1):67–81. MR3626475

Tsanas, A., Little, M. A., Fox, C., and Ramig, L. O. (2014). Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):181–190.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202. MR3224285

Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610. MR2528238

Yin, X., Li, B., and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of*

Multivariate Analysis, 99(8):1733–1757. MR2444817

Zafeiriou, S., Tefas, A., and Pitas, I. (2007). Minimum class variance support vector machines. *IEEE Transactions on Image Processing*, 16:2551–2564. MR2467785

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 76(1):217–242. MR3153940

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(492):1455–1466. MR2796563