Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

DIMENSION REDUCTION TECHNIQUES FOR HIGH DIMENSIONAL AND ULTRA-HIGH DIMENSIONAL DATA

by
Subha Datta

This dissertation introduces two statistical techniques to tackle high-dimensional data, which is very commonplace nowadays. It consists of two topics which are inter-related by a common link, dimension reduction.

The first topic is a recently introduced classification technique, the weighted principal support vector machine (WPSVM), which is incorporated into a spatial point process framework. The WPSVM possesses an additional parameter, a weight parameter, besides the regularization parameter. Most statistical techniques, including WPSVM, have an inherent assumption of independence, which means the data points are not connected with each other in any manner. But spatial data violates this assumption. Correlation between two spatial data points increases as the distance between them decreases. However, under some conditions on the spatial point process, the WPSVM is still valid. Furthermore, through extensive simulations it has been shown that WPSVM performs better than other dimension reduction techniques. The main advantage of WPSVM comes from the fact that it can handle non-linear relationships. WPSVM is also applied to a rainforest dataset.

The second topic talks about another recently introduced technique, joint-screening. Unlike the previous method, this works for ultra-high dimensional data ($p \gg n$). Most existing variable screening methods fail to identify those marginally unimportant but jointly important genetic variables. The joint screening (JS) procedure screens all the covariates at the same time based on a criterion. In this way a subset of variables that are suspected to be highly associated with the outcome can be identified. One massive advantage of the JS procedure comes from the fact

that it is computationally simple and easy to understand. The performance of the proposed JS procedure is evaluated via simulation studies and an application to the Genetics Analysis Workshop 20 data.

DIMENSION REDUCTION TECHNIQUES FOR HIGH DIMENSIONAL AND ULTRA-HIGH DIMENSIONAL DATA

by
Subha Datta

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
and Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences, NJIT
Department of Mathematics and Computer Science, Rutgers–Newark

December 2019

APPROVAL PAGE

DIMENSION REDUCTION TECHNIQUES FOR HIGH DIMENSIONAL AND ULTRA-HIGH DIMENSIONAL DATA

Subha Datta

Dr. Ji Meng Loh, Dissertation Co-advisor — Date
Associate Professor of Mathematics, NJIT

Dr. Yixin Fang, Dissertation Co-advisor — Date
Director of GMA-Statistics, Abbvie

Dr. Sunil Dhar, Committee Member — Date
Professor of Mathematics, NJIT

Dr. Antai Wang, Committee Member — Date
Associate Professor of Mathematics, NJIT

Dr. Yang Feng, Committee Member — Date
Associate Professor of Statistics, Columbia University

BIOGRAPHICAL SKETCH

Author: Subha Datta

Degree: Doctor of Philosophy

Date: December 2019

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, & Department of Mathematics and Computer Science, Rutgers-Newark, NJ, 2019
- Master of Science in Statistics,
  University of Calcutta, Kolkata, India, 2007
- Bachelor of Science in Statistics,
  University of Calcutta, Kolkata, India, 2005

Major: Applied Statistics

Presentations and Publications:

S. Datta, Y. Fang, J.M. Loh, "Joint screening of ultra-high dimensional variables for family-based genetic studies", *BMC Proceedings*, **12**(Suppl 9):24, 2018.

S. Datta, "WPSVM for inhomogeneous spatial point processes", Presentation at Summer Program, Department of Mathematical Sciences, NJIT, July 2019.

S. Datta, "Dimension reduction techniques for high dimensional and ultra-high dimensional data", Presentation at Joint Statistical Meetings, Vancouver, July 2018.

S. Datta, "WPSVM for inhomogeneous spatial point processes", Presentation at Summer Program, Department of Mathematical Sciences, NJIT, June 2018.

S. Datta, "Joint screening of ultra-high dimensional variables for family-based genetic studies", Presentation at Summer Program, Department of Mathematical Sciences, NJIT, June 2017.

S. Datta, "Sufficient dimension reduction techniques", Presentation at Summer Program, Department of Mathematical Sciences, NJIT, August 2016.

আমার সহধর্মিনি, আত্রেয়ী ও আমাদের আদরের কণ্যা, ইলিকা কে।
To my wife, Atreyee Majumder and our beloved daughter, Ilika Datta.

ACKNOWLEDGMENT

My journey towards achieving academic fineness would not have been possible if not for certain people who have guided me through it. I take this opportunity to thank each and every one of them.

First, I would like to express my sincere gratitude to my advisors Drs. Ji Meng Loh and Yixin Fang for their patience, humility, and motivation. Their profound knowledge of statistics helped me grow as a researcher which I am sure will help me immensely in future. I could not have asked for better mentors for my PhD program. I would like to thank Dr. Yixin Fang for providing me with funding needed to attend the GAW20 workshop. It was a great learning experience for me.

Besides my advisors, I would like to thank the rest of my committee members: Drs. Sunil Dhar, Antai Wang, and Yang Feng for providing me with encouragement and constructive feedback which helped me better my research material. I would like to thank our department chair Dr. Jonathan Luke, and past and present directors of our graduate program Drs. Lou Kondic, Richard Moore, and Michael Seigel for helping make Department of Mathematical Sciences one of the best at New Jersey Institute of Technology.

My sincere thanks also go to my fellow doctoral students for their friendship and guidance through tough times. I would also like to thank our departmental staff for helping me with administrative challenges and, of course, the fun activities.

Last and by no means least, I would like to thank my wife, Dr. Atreyee Majumder for being a guiding light and accepting no less than excellence from me. I am thankful to the Almighty for bringing our daughter, Ilika Datta into our lives. I am also grateful to my family: my parents, parents-in-law, my brother for rendering moral and spiritual support throughout the duration of the PhD program and my life in general.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

With the advancement of data acquisition and storage techniques, we now frequently encounter high-dimensional data. Multidimensional problems become notoriously difficult to solve with increase in dimensionality. One obvious solution is to reduce the dimensionality and at the same time making sure we identify important features. Sufficient Dimension Reduction (SDR) has become an essential tool in dealing with problems with high-dimensional data in the past few years.

1.1 Sufficient Dimension Reduction

SDR assumes that

$$Y \perp \boldsymbol{X} | \mathbf{B}^\top \boldsymbol{X}, \tag{1.1}$$

where $\left(\boldsymbol{X}^\top, Y\right)^\top \in \mathbb{R}^p \times \mathbb{R}$ is a pair of p-dimensional predictor and response, '$\perp$' denotes *conditional independence*, and $\mathbf{B}$ is a $p \times k$-dimensional matrix. This indicates that the dependent structure between Y and $\boldsymbol{X}$ is only through $\mathbf{B}^\top \boldsymbol{X}$. Under model (1.1), SDR is achieved by estimating the matrix $\mathbf{B}$, in particular, the space spanned by it, referred to in the literature as the *dimension reduction subspace*. However, $\mathbf{B}$ itself may not be unique due to the fact that any full-rank linear combination of the columns of $\mathbf{B}$ would have the same properties. For example, if $\mathbf{B}^\top \boldsymbol{X}$ is a sufficient dimension reduction then so is $(\mathbf{BA})^\top \boldsymbol{X}$ for any $k \times k$ matrix $\mathbf{A}$ of full rank.

Model (1.1) is referred to as linear SDR. One can think of a nonlinear version of SDR introduced by Cook [10] in 2007 which assumes

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

where $\phi : \mathbb{R}^p \mapsto \mathbb{R}^k$ is an arbitrary function of $\boldsymbol{X}$. Under model (1.2), SDR is achieved by obtaining a function ϕ which need not be linear. Similar to linear SDR, the unknown function ϕ is not unique, but is assumed to be unique modulo injective transformations (see Li et al. [28]).

1.1.1 Existing SDR methods

Inverse regression techniques typically regresses $\boldsymbol{X}$ against Y. The advantage of this technique is realized when we deal with high dimensional problems. This effectively boils down to dealing with a one-dimension to one-dimension regression problem, rather than the high-dimensional regression problem. Sliced inverse regression (SIR) [32] is the most popular which has its fair share of limitations. SIR slices the data in terms of the response variable. However, this is confusing for spatial point processes due to lack of ordinality. Li [32] treats the spatial point process as a binary response and forms two slices. SIR manages to capture monotonic regression problems well but cannot estimate highly symmetric ones (see Li [32]; Cook and Weisberg [12]). Sliced average variance estimator (SAVE), proposed by Cook and Weisberg [12], on the other hand, can estimate highly symmetric relationships well. However, SAVE is less sensitive towards monotonic patterns. Li and Wang [29] developed directional regression (DR) combining the benefits of SIR and SAVE to tackle this problem. Other methods include principal Hessian direction (pHd) [8, 33], iterative Hessian transformation (IHT) [11], Fourier method [59], projection pursuit regression (PPR) [22], the alternating conditional expectation (ACE) method [3], and minimum average variance estimation (MAVE) [55]. For nonlinear SDR, several methods have been recently proposed by extending the idea of SIR (see Wu [52], Wu et al. [53], Yeh et al. [56]).

1.1.2 Dealing with binary response

Most SDR methods do not perform well when the response is binary. For example, SIR can estimate at most one direction of $\mathcal{S}_{Y|\boldsymbol{x}}$. Although, SAVE is more comprehensive than SIR, it has also been known to fail for binary data, as shown by Li and Wang [29]. In addition, SIR performs poorly in symmetric regressions with $y = f\left(\mathbf{B}^\top \boldsymbol{x}\right) + \epsilon$, where f is a symmetric function of the argument $\mathbf{B}^\top \boldsymbol{x}$ (see Li [32], Cook and Weisberg [12]). Shin et al. [41] introduced the WPSVM to deal with these problems. We apply the WPSVM to spatial point processes by characterizing the process as a binary response on a grid, similar to what Guan and Wang [23] did.

1.2 Dealing with Ultra-high Dimensionality for Mixed Models

Our second problem deals with mixed models, especially ultra-high dimensional mixed models, where the number of features p is much higher than the number of observations n. Mixed models are a useful tool for evaluating the association between an outcome variable and genetic variables from a family-based genetic study, taking into account the kinship coefficients. When there are ultra-high dimensional genetic variables (i.e., $p \gg n$), it is challenging to fit any mixed effect model. We propose a two-stage strategy, screening genetic variables in the first stage and then fitting the mixed effect model in the second stage to those variables that survive the screening. For the screening stage, we can use the sure independence screening (SIS) procedure, proposed by Fan and Lv [18]. SIS fits the mixed model to one genetic variable at a time. Since, the SIS procedure may fail to identify those marginally unimportant but jointly important genetic variables, we propose a joint screening (JS) procedure, which screens all the genetic variables at the same time.

1.2.1 Mixed models for family data

Mixed model analysis provides a general, flexible approach when dealing with correlated data (see Fitzmaurice et al. [21]). Mixed models allow a wide variety of variance-covariance structures to be explicitly modeled which makes it a useful tool to analyze a family-based data set, because subjects within a family are correlated with one another via genetic structure.

Suppose that there are n subjects from a family study and there are p genetic variables. Assume that we can relate the phenotypes with the genetic variables via the following mixed model,

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{1.3}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of observed phenotypes, $\mathbf{X}$ is an $n \times p$ design matrix of genetic variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector representing the fixed effects of genetic variables, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$ is an $n \times 1$ vector representing the random effects. We assume that $\boldsymbol{\epsilon}$ has zero-mean and $Var(\boldsymbol{\epsilon}) = \sigma_e^2 I_n$, and

$$\boldsymbol{\alpha} \sim N(0, \sigma_g^2 K),$$

where $n \times n$ matrix $K = (k_{ij})_{n \times n}$ is the kinship matrix among the n subjects from the family data. The kinship coefficent k_{ij} is a measure of genetic relatedness between two individuals i and j.

If p were small compared with n, we would estimate the unknown parameters, $\boldsymbol{\beta}$, σ_e^2 and σ_g^2, in the above mixed model and then identify those genetic variables that are significantly associated with the phenotype.

Specifically, if p were small compared with n, we could estimate the coefficient vector $\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{Y}$,

$$V = \mathrm{Var}(\boldsymbol{Y}) = \sigma_g^2 K + \sigma_e^2 I_n, \tag{1.4}$$

via the weighted least-squares (WLS),

$$\widehat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^\top \widehat{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \widehat{V}^{-1}\boldsymbol{Y}, \tag{1.5}$$

and the restricted maximum likelihood (REML),

$$\widehat{V} = \underset{\boldsymbol{\beta},\sigma_e^2,\sigma_g^2}{\operatorname{argmax}} \left\{ l_p(V) - \log|X^\top V^{-1}\mathbf{X}| \right\}, \tag{1.6}$$

where $l_p(V) = -\left\{ \log|V| + (\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top V^{-1}(\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \right\}$.

1.2.2 Curse of dimensionality

If the dimension of the genetic variables is high ($p \sim n$ or $p > n$), we can use some regularization methods. These methods simultaneously estimate parameters and perform variable selection by penalizing a loss function with the help of a sparsity inducing penalty. For example, see Lasso [43], Ridge regression [25], SCAD [17], elastic net [60], and Schelldorfer et al. [39]. However, in ultra-high dimensional cases the computation cost for these regularization methods becomes a concern.

When the dimension of the genetic variables is ultra-high ($p \gg n$) we cannot use the above estimates (1.5) and (1.6) for $\boldsymbol{\beta}$ and V, respectively. This is an example of the curse of dimensionality; the matrix under inverse in equation (1.5), $\mathbf{X}^\top \widehat{V}^{-1}\mathbf{X}$, is a $p \times p$ matrix, but its rank is at most n. There are two reasons the classical mixed model does not work. First, the matrix $\mathbf{X}^\top \widehat{V}^{-1}\mathbf{X}$ is not invertible, so the solution of the equation (1.5) is not unique. Second, when p is ultra-high, the computation of the general inverse of $\mathbf{X}^\top \widehat{V}^{-1}\mathbf{X}$ is very hard, not to mention the estimation of V in equation (1.6).

There has been a rapid development in approaches for handling ultra-high dimensional problems. Fan and Lv [18] introduced the SIS procedure which significantly reduces dimensionality in a simple manner. The screening procedure has been extended to a variety of other models, for example, to generalized linear

models (Fan and Fan [15]; Fan et al. [19]; Fan and Song [20]), additive models (Fan et al. [16]), and to adapt to conditional correlation (Barut et al. [1]).

A sufficient condition for SIS is that the marginal correlation for the relevant features should be bounded away from zero. But this assumption is often violated as predictors are often correlated. One major downside is that irrelevant features highly correlated with important predictors have a high chance of being selected which is not desirable. Contrarily, relevant features jointly correlated to the response are being filtered out. To this end, a number of papers have proposed improvements (see Hall et al. [24]; Wang [49, 50]; Cho and Fryzlewicz [4]; Wang and Leng [51]). Fan and Lv [18] in their SIS paper proposed using an iterative SIS procedure which applies SIS iteratively to the residual in a finite number of steps.

Therefore, for the situation with ultra-high dimensional genetic variables, we propose a novel joint screening procedure for mixed models. We conduct variable screening to identify a subset of genetic variables that are suspected to be associated with the outcome; choosing the subset size such that it is manageable by mixed models. We can then conduct mixed model analysis using those genetic variables that survive the screening stage. Our main area of focus is the proposed joint screening approach.

1.3 Outline of Thesis

The first chapter briefly describes the two interrelated topics, dimension reduction for high and ultra-high dimensional data, and the motivation behind them. In Chapter 2, we apply WPSVM (Shin et al. [41]) to a spatial point process framework. The WPSVM is a weighted version of principal support vector machine (PSVM), proposed by Li et al. [28] for an ordinary regression setup. Shin et al. [41] have shown that the WPSVM preserves all the merits of PSVM while achieving SDR at the same time. We explore the asymptotic properties of the WPSVM estimator for spatial point

processes. In Chapter 3, we deal with ultra-high dimensional problems and propose a novel joint screening (JS) procedure. This JS estimator for mixed models is motivated by Wang and Leng [51], who introduced a similar estimator for linear models. We show that the JS estimator works for mixed models too and has the desired sure screening properties. We performed extensive simulation studies to show that the two estimators (WPSVM and JS) perform better in comparison with a number of competitors. Application to real data examples also give favorable results.

CHAPTER 2

WPSVM FOR SPATIAL POINT PROCESSES DIRECTED BY GAUSSIAN RANDOM FIELDS

2.1 Introduction

In this chapter, we attempt to tackle the issue of high-dimensionality in spatial point patterns using a novel SDR method, called weighted principal support vector machine (WPSVM). As the name indicates, it is a weighted version of the principal support vector machine (PSVM) [28]. Shin et al. [41] introduced the WPSVM to deal with binary responses. We proceed with applying WPSVM to a spatial point process by first setting up the dimension reduction framework.

2.2 Model Setup

2.2.1 Notation

We use the same notations introduced by Guan and Wang [23]. Consider $\mathbb{X} = \{\boldsymbol{X}(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^2\}$, a p-dimensional Gaussian random field with $\boldsymbol{X}(\boldsymbol{s}) = \{X_1(\boldsymbol{s}), \ldots, X_p(\boldsymbol{s})\}^\top \in \mathbb{R}^p$. Without loss of generality, we assume $\mathbb{E}\{\boldsymbol{X}(\boldsymbol{s})\} = 0$ and $\text{cov}\{\boldsymbol{X}(\boldsymbol{s})\} = \boldsymbol{I}_p$ (see Li [32], Cook [9]), where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix. In a practical setting, the standardization can be done on the basis of the estimated mean and covariance matrix of $\boldsymbol{X}(\boldsymbol{s})$. Let $\mathcal{N}$ be a spatial point process that results from a certain stochastic mechanism conditional on $\mathbb{X}$. Let $\mathcal{N}(.)$ be the counting measure that is induced by $\mathcal{N}$, and let $\mathcal{B}_1, \ldots, \mathcal{B}_k$ be some bounded Borel sets in $\mathbb{R}^2$. The k-th order moment measure of the spatial point process (see Cressie [13]) is defined as

$$\mu_{\mathcal{N}}^{(k)}(\mathcal{B}_1 \times \cdots \times \mathcal{B}_k) = \mathbb{E}\left\{\mathcal{N}(\mathcal{B}_1) \ldots \mathcal{N}(\mathcal{B}_k) \,|\mathbb{X}\right\}. \tag{2.1}$$

For pairwise disjoint $\mathcal{B}_1, \ldots, \mathcal{B}_k$, the k-th order moment measure becomes the k-th order factorial moment measure (see Stoyan et al. [42]). If the latter is locally finite

and absolutely continuous with respect to the Lebesgue measure, then the k-th order intensity function $\lambda_k(.)$, also known as the k-th order product density, exists and satisfies

$$\mu_{\mathcal{N}}^{(k)}\left(\mathrm{d}\boldsymbol{s}_1 \times \cdots \times \mathrm{d}\boldsymbol{s}_k\right) = \lambda_k(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)\mathrm{d}\boldsymbol{s}_1 \ldots \mathrm{d}\boldsymbol{s}_k,$$

where $\mathrm{d}\boldsymbol{s}_i\,(i = 1, \ldots, k)$ are some distinct infinitesimal sets in $\mathbb{R}^2$. We assume that λ_k exists for all $k \geq 1$.

2.2.2 Central subspace (CS)

We denote the linear subspace spanned by the column vectors of an arbitrary matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$ by $\mathcal{S}(\mathbf{B})$. Similar to model (1.1) we call $\mathcal{S}(\mathbf{B})$ a *sufficient dimension reduction subspace* if, for all positive integers k and for any k bounded Borel sets $\mathcal{B}_1, \ldots, \mathcal{B}_k \subseteq \mathbb{R}^2$,

$$\{\mathcal{N}\left(\mathcal{B}_1\right) \ldots \mathcal{N}\left(\mathcal{B}_k\right)\} \perp \{\mathbb{X}(\mathcal{B}_1), \ldots, \mathbb{X}(\mathcal{B}_k)\} \,|\, \left\{\mathbf{B}^\top\mathbb{X}(\mathcal{B}_1), \ldots, \mathbf{B}^\top\mathbb{X}(\mathcal{B}_k)\right\}, \qquad (2.2)$$

where $\mathbb{X}(\mathcal{B}) = \{\boldsymbol{X}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{B}\}$. The *central subspace* denoted by $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$ is defined as the intersection of all dimension reduction subspaces satisfying equation (2.2). Under mild conditions, $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$ uniquely exists, according to Cook [7, 9] and has the lowest dimension among all dimension reduction subspaces. Throughout this dissertation, we assume that $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$ uniquely exists and has a basis given by $\mathbf{B}_0 \in \mathbb{R}^{p\times k}$, where $k = \dim\left(\mathcal{S}_{\mathcal{N}|\mathbb{X}}\right)$ is the structural dimension of $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$. The estimation of k is another important step in SDR.

2.2.3 Central intensity subspace (CIS)

It is possible to uniquely determine the probability distribution of $\mathcal{N}$ by its moment measures as defined in equation (2.1), since we can express the k-th order moment measure as a sum of integrals involving the intensity functions up to order k (see Zessin

[57]). Similarly, we can have a dimension reduction model for the intensity functions. Define $\mathcal{S}(\mathbf{B})$ to be the k-th order *sufficient intensity dimension reduction subspace* of $\mathcal{N}$ if, for some function $f_k(.)$

$$\lambda_k(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k) = f_k\left\{\mathbf{B}^\top \boldsymbol{X}(\boldsymbol{s}_1), \ldots, \mathbf{B}^\top \boldsymbol{X}(\boldsymbol{s}_k)\right\}. \tag{2.3}$$

Let us define $\mathcal{S}_k$ to be the intersection of all dimension reduction subspaces satisfying equation (2.3). If $\mathcal{S}_k$ is itself a sufficient intensity dimension reduction subspace, then it should be the smallest and we call it the k-th order CIS. We also assume that it exists.

2.2.4 Relationship between CS and CIS

By definition, the CS contains all information of $\mathbb{X}$ about $\mathcal{N}$. Hence, it contains all information of $\mathbb{X}$ on any summary function of $\mathcal{N}$, e.g., the intensity functions $\{\lambda_k(.) : k \geq 1\}$. Clearly,

$$\mathcal{S}_k \subseteq \mathcal{S}_{\mathcal{N}|\mathbb{X}}, \text{ for any } k \geq 1 \implies \bigcup_{k\geq 1} \mathcal{S}_k \subseteq \mathcal{S}_{\mathcal{N}|\mathbb{X}}.$$

Interestingly, a similar reverse relationship, i.e., $\mathcal{S}_{\mathcal{N}|\mathbb{X}} \subseteq \bigcup_{k\geq 1} \mathcal{S}_k$, holds too (see Guan and Wang [23]). Thus, $\mathcal{S}_{\mathcal{N}|\mathbb{X}} = \bigcup_{k\geq 1} \mathcal{S}_k$.

Although it may seem difficult, but it is possible to estimate the CS by estimating all the CISs. For simplicity we study only the first- and second-order intensity functions for spatial point processes (see Diggle [14], Moller and Waagepetersen [36]). Hence, we assume the following *coverage condition,* $\mathcal{S}_{\mathcal{N}|\mathbb{X}} = \mathcal{S}_1 \cup \mathcal{S}_2$. Thus it is essential to estimate $\mathcal{S}_1$ and $\mathcal{S}_2$. To that end, we only consider the case of estimating $\mathcal{S}_1$. Thus, we have $\mathcal{S}_{\mathcal{N}|\mathbb{X}} = \mathcal{S}_1$ if

$$\lambda_k(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k) = \lambda_1(\boldsymbol{s}_1) \ldots \lambda_1(\boldsymbol{s}_k) g_k(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k), \tag{2.4}$$

where $g_k(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)$ is free of any covariates for all $k \geq 2$.

2.3 Weighted Principal Support Vector Machines

As mentioned earlier, WPSVM is a weighted version of the PSVM. Li et al. [28] introduced PSVM which can extract the sufficient predictors for the two models given by equations (1.1) and (1.2). The basic idea is to first divide the covariates into several slices based on the response values, and then use a modified form of SVM to find optimal hyperplanes to separate them. These optimal hyperplanes are then aligned by the principal components of their normal vectors. Li et al. [28] showed that PSVM not only improves the accuracy for sufficient dimension reduction, but it can handle both linear and nonlinear SDR in a unified framework. The aligned normal vectors provide an unbiased, $\sqrt{n}$-consistent, and asymptotically normal estimator of the sufficient dimension reduction space.

However, the PSVM suffers from estimating at most one direction of $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$ for binary responses. Shin et al. [41] introduced WPSVM and showed that it can estimate more than one direction of $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$ with binary response. It also preserves the above mentioned benefits and asymptotic properties of PSVM.

2.3.1 $\tilde{\rho}$-mixing sequence of random variables

Before proceeding further, it is important to introduce a property typical of some spatial point processes, the $\tilde{\rho}$-mixing property. Suppose, $\{\xi_n\}_{n\in\mathbb{N}}$ is a sequence of random variables on a probability space $(\Omega, \mathcal{M}, \mathbb{P})$. For any $U \subset \mathbb{N}$, define $\mathcal{F}_U = \sigma\{\xi_k : k \in U\}$. Given σ-fields $\mathcal{F}, \mathcal{G} \subset \mathcal{M}$, let

$$\rho(\mathcal{F}, \mathcal{G}) = \sup\{|\text{corr}(f, g)| : f \in L_2(\mathcal{F}), g \in L_2(\mathcal{G})\}.$$

Bradley [2] defined the following coefficients of dependence, for $n \geq 0$

$$\tilde{\rho}(n) = \sup\{\rho(\mathcal{F}_U, \mathcal{F}_V)\},$$

where the supremum is taken over all pairs of nonempty finite sets $u, v \subset \mathbb{N}$ such that

$$\text{dist}(U, V) = \min_{u \in U, v \in V} |u - v| \geq n.$$

Definition 2.1. A sequence of random variables $\{\xi_n\}_{n \in \mathbb{N}}$ is said to be a $\tilde{\rho}$-mixing sequence if

$$\lim_{n \to \infty} \tilde{\rho}(n) < 1.$$

Since, $0 \leq \tilde{\rho}(n) \leq \tilde{\rho}(n-1) \leq \cdots \leq \tilde{\rho}(1) \leq 1$, the above is equivalent to

$$\tilde{\rho}(n_0) < 1, \text{ for some } n_0 \geq 1.$$

This crudely means two realizations of the sequence become asymptotically independent, as the distance between them increases. Stationary point processes directed by Gaussian sequences are typically $\tilde{\rho}$-mixing. Other examples include Neyman-Scott processes.

2.3.2 Principal support vector machine

In order to smoothly transition to WPSVM, we first briefly review PSVM. Without loss of generality assume $\mathbb{E}(\boldsymbol{X}) = 0$. Li et al. [28] developed the linear PSVM which requires minimizing the following objective function:

$$(a_0, \mathbf{b}_0) = \underset{a, \mathbf{b}}{\text{argmin}}\, \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b} + \lambda \mathbb{E}\left[1 - \widetilde{Y}_c(a + \mathbf{b}^\top \boldsymbol{X})\right]_+, \tag{2.5}$$

where $\boldsymbol{\Sigma} = \text{Var}(\boldsymbol{X}), [a]_+ = \max(a, 0), \widetilde{Y}_c = \mathbb{1}(Y \geq c) - \mathbb{1}(Y < c)$ for a given constant c, and the parameter λ is regarded as 'cost'. The objective function (2.5) is similar to that of the support vector machine (SVM; [46]) with linear kernel and this is where its name, PSVM comes from. Li et al. [28] show that $\mathbf{b}_0$ is unbiased for linear SDR.

Given a set of data $\left\{\boldsymbol{Z}_i = \left(\boldsymbol{X}_i^\top, Y_i\right)^\top : \boldsymbol{X}_i \in \mathbb{R}^p, y \in \mathbb{R}, i = 1, \ldots, n\right\}$, we can consider a sequence of cutoff points of c denoted by $c_h, h = 1, \ldots, H$ with an

associated $\widetilde{Y}_{i,c_h} = \mathbb{1}(Y_i \geq c_h) - \mathbb{1}(Y_i < c_h)$. For each c_h, $h = 1, \ldots, H$, a sample version of the objective function (2.5) is given by

$$(\hat{a}_{n,c_h}, \widehat{\mathbf{b}}_{n,c_h}) = \underset{a,\mathbf{b}}{\operatorname{argmin}} \, \mathbf{b}^\top \widehat{\boldsymbol{\Sigma}}_n \mathbf{b} + \frac{\lambda}{n} \sum_{i=1}^{n} \left[1 - \widetilde{Y}_{i,c_h}(a + \mathbf{b}^\top \boldsymbol{X}_i)\right]_+, \tag{2.6}$$

where $\widehat{\boldsymbol{\Sigma}}_n$ denotes the sample covariance matrix of $\boldsymbol{X}$. Since the linear PSVM solution is unbiased, the eigenvectors of a candidate matrix $\widehat{\mathbf{M}}_n^L$ can be used as an estimator of the basis of the CS where

$$\widehat{\mathbf{M}}_n^L = \sum_{h=1}^{H} \widehat{\mathbf{b}}_{n,c_h} \widehat{\mathbf{b}}_{n,c_h}^\top.$$

The PSVM is more accurate and robust for SDR than the popular inverse regression based methods such as SIR. However, similar to SIR, the PSVM suffers from the difficulty of estimating at most one direction for the CS in binary classification, since at most a single cut-off value of c is available if Y takes only two values.

2.3.3 Weighted support vector machine

In the standard SVM, introduced by Vapnik et al. [47], each observation is treated equally no matter which class it belongs to. This may not always be optimal. In some situations, it is desired to assign different weights to the observations from different classes; one such example is when one type of misclassification induces a larger cost than the other type of misclassification. Motivated by this, Lin et al. [34] considered the so-called weighted SVM (WSVM) by incorporating a weight parameter to adjust the imbalance between the two classes. Thereafter, Shin et al. [41] proposed a weighted version of the PSVM.

We are now ready to apply WPSVM to a spatial point process setup. However, it is important that we first fit a spatial point process to WPSVM's framework. We do this by treating the spatial point process as a binary (response) random field,

$\{Y(\boldsymbol{s})\}$, with

$$Y(\boldsymbol{s}) = \begin{cases} 1, & \text{if } \boldsymbol{s} \in \mathcal{N} \\ 0, & \text{otherwise.} \end{cases}$$

2.4 Weighted PSVM for Linear SDR

Let $Y(\boldsymbol{s}) \in \{-1, +1\}, \boldsymbol{Z}(\boldsymbol{s}) = \left\{\boldsymbol{X}(\boldsymbol{s})^\top, Y(\boldsymbol{s})\right\}^\top \in \mathbb{R}^p \times \mathbb{R}, \boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^\top)^\top$, and $f\left(\boldsymbol{X}(\boldsymbol{s}); \boldsymbol{\theta}\right) = \alpha + \boldsymbol{\beta}^\top \left[\boldsymbol{X}(\boldsymbol{s}) - \mathbb{E}\left\{\boldsymbol{X}(\boldsymbol{s})\right\}\right]$. The linear WPSVM minimizes the following objective function

$$\Lambda_\pi(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + \lambda \mathbb{E}\left\{\pi(Y(\boldsymbol{s}))[1 - Y(\boldsymbol{s}) f(\boldsymbol{X}(\boldsymbol{s}); \boldsymbol{\theta})]_+\right\}, \tag{2.7}$$

where $\pi(Y(\boldsymbol{s})) = 1 - \pi$ if $Y(\boldsymbol{s}) = 1$ and π otherwise with an associated parameter $\pi \in (0, 1)$, $\boldsymbol{\Sigma} = \text{var}(\boldsymbol{X}(\boldsymbol{s}))$, $[a]_+ = \max(a, 0)$, and λ is the tuning parameter. Let $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0^\top)^\top$ be the minimizer of $\Lambda_\pi(\boldsymbol{\theta})$ in equation (2.7) for an arbitrary $\pi \in (0, 1)$. The following theorem (see Li et al. [28]) states that $\boldsymbol{\beta}_0$ is unbiased for the linear SDR model (1.1).

Theorem 2.1. *Assume that* $\mathbb{E}\left\{\boldsymbol{X}(\boldsymbol{s}) | \mathbf{B}^\top \boldsymbol{X}(\boldsymbol{s})\right\}$ *is a linear function of* $\mathbf{B}^\top \boldsymbol{X}(\boldsymbol{s})$. *Then* $\mathcal{S}(\boldsymbol{\beta}_0) \subseteq \mathcal{S}_{\mathcal{N}|\mathbb{X}} = \mathcal{S}_1$ *under (1.1).*

The above assumption is referred to as the *linearity condition* and implies that

$$\mathbb{E}\left\{\boldsymbol{\beta}^\top \boldsymbol{X}(\boldsymbol{s}) | \mathbf{B}^\top \boldsymbol{X}(\boldsymbol{s})\right\} = \boldsymbol{\beta}^\top \mathbf{P}_{\mathbf{B}}(\boldsymbol{\Sigma}) \boldsymbol{X}(\boldsymbol{s}),$$

where $\mathbf{P}_{\mathbf{B}}(\boldsymbol{\Sigma}) = \mathbf{B}(\mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B})^{-1} \mathbf{B}^\top \boldsymbol{\Sigma}$ is a projection matrix on $\mathcal{S}_1$ with respect to $\boldsymbol{\Sigma}$ (see Cook [9]). Theorem 2.1 helps us estimate the CS, $\mathcal{S}_1$ from normals of linear WPSVM solutions $\boldsymbol{\beta}_0$ for different weight parameters.

2.4.1 Finite sample estimation

For a given set of data $\{\boldsymbol{Z}(\boldsymbol{s}_i) = (\boldsymbol{X}(\boldsymbol{s}_i), Y(\boldsymbol{s}_i)) : \boldsymbol{X}(\boldsymbol{s}_i) \in \mathbb{R}^p, Y(\boldsymbol{s}_i) \in \{-1, +1\}, i = 1, \ldots, n\}$, the sample version of the objective function (2.7) is given by

$$\widehat{\Lambda}_{n,\pi}(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\beta} + \frac{\lambda}{n} \sum_{i=1}^{n} \pi(Y(\boldsymbol{s}_i))[1 - Y(\boldsymbol{s}_i) f_n(\boldsymbol{X}(\boldsymbol{s}_i); \boldsymbol{\theta})]_+, \tag{2.8}$$

where $f_n(\boldsymbol{X}(\boldsymbol{s}_i); \boldsymbol{\theta}) = \alpha + \boldsymbol{\beta}^\top \left\{ \boldsymbol{X}(\boldsymbol{s}_i) - \overline{\boldsymbol{X}}_n \right\}$ and $\overline{\boldsymbol{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}(\boldsymbol{s}_i)$. We can show that equation (2.8) is a sample version of equation (2.7) by using the following strong law of large numbers.

2.4.1.1 Strong law of large numbers for $\tilde{\rho}$-mixing random variables. Let $f(x), g(x)$ be real positive functions defined on the same domain $[h, +\infty), \psi(x) = f(x)g(x)$, where $0 \leq h \leq 1$. Note that $f(x)$ or $g(x)$ may not be well defined at the point h, but if so, $\lim_{x \to h+0} f(x)g(x)$ exists, and we can let $\psi(h)$ be equal to the limit at this point. Assume the following conditions are satisfied:

(**i**) $f(x)$ is increasing on its domain, and $\lim\limits_{x \to +\infty} f(x) = +\infty$;

(**ii**) $\psi(x)$ is strictly increasing on $[h, +\infty)$, $\lim\limits_{x \to +\infty} \psi(x) = +\infty$, and its range is $[0, +\infty)$;

(**iii**) there exists constants $a, b \in \mathbb{R}$ such that for every $t \in \mathbb{R}, t^2 \int_{\psi^{-1}(|t|)}^{+\infty} \frac{\mathrm{d}x}{\psi^2(x)} \leq a\psi^{-1}(|t|) + b$.

Theorem 2.2 (Meng and Lin [35])**.** *Let $f(x), g(x), \psi(x)$ be functions satisfying the above conditions, and let $\{\xi_n, n \in \mathbb{N}\}$ be a sequence of $\tilde{\rho}$-mixing identically distributed random variables. Set*

$$A_n = \mathbb{E}(\xi_n I_{\{|\xi_n| < \psi(n)\}}),$$
$$B_n = \frac{1}{f(n)} \sum_{k=1}^{n} \frac{\xi_k - A_k}{g(k)}.$$

If $\mathbb{E}(\psi^{-1}(|\xi_1|)) < \infty$, *then* $B_n \xrightarrow{a.s.} 0$ *as* $n \to \infty$.

Remark 2.1. If we take $f(x) = x^{1/p}\,(0 < p < 2)$, $g(x) = 1$, $\psi(x) = f(x)g(x) = x^{1/p}$, $x \in [0, +\infty)$, then we get a Marcinkiewicz type SLLN, $B_n = \frac{1}{n^{1/p}} \sum_{k=1}^{n} (\xi_k - A_k) \to 0$ a.s. as $n \to \infty$. For $p = 1$ we get precisely the following:

$$\frac{1}{n} \sum_{k=1}^{n} \left(\xi_k - \mathbb{E}\left(\xi_k I_{\{|\xi_k| < k\}} \right) \right) \to 0 \text{ a.s. as } n \to \infty,$$

which is the standard SLLN for $\tilde{\rho}$-mixing sequences.

The above remark shows that equation (2.8) is a sample version of equation (2.7) and we proceed with the finite sample estimation. For a given grid of π, $0 < \pi_1 < \cdots < \pi_H < 1$, we minimize $\widehat{\Lambda}_{n,\pi_h}$, and let the corresponding minimizers be $(\widehat{\alpha}_{n,h}, \widehat{\boldsymbol{\beta}}_{n,h})^\top, h = 1, \ldots, H$. The candidate matrix of the linear WPSVM is given by

$$\widehat{\mathbf{M}}_n^{LW} = \sum_{h=1}^{H} \widehat{\boldsymbol{\beta}}_{n,h} \widehat{\boldsymbol{\beta}}_{n,h}^\top \tag{2.9}$$

The basis of the central subspace $\mathcal{S}_1$ is estimated by the first k leading eigenvectors of $\widehat{\mathbf{M}}_n^{LW}$ denoted by $\widehat{\mathbf{V}}_n^{LW} = (\widehat{\mathbf{v}}_1^{LW}, \ldots, \widehat{\mathbf{v}}_k^{LW})$. It is possible for $\widehat{\mathbf{M}}_n^{LW}$ to have more than one eigenvector with non-zero eigenvalue in binary classification as we have varied the weight parameter π in the above procedure.

2.4.2 Large sample properties

Let us assume without loss of generality $\mathbb{E}\{\boldsymbol{X}(\boldsymbol{s})\} = 0$ and let $\widetilde{\boldsymbol{X}(\boldsymbol{s})} = \left(1, \boldsymbol{X}(\boldsymbol{s})^\top\right)^\top$. Then we have, $f\left(\boldsymbol{X}(\boldsymbol{s}); \boldsymbol{\theta}\right) = \boldsymbol{\theta}^\top \widetilde{\boldsymbol{X}(\boldsymbol{s})}$. Let $\widehat{\boldsymbol{\theta}}_n = (\widehat{\alpha}_n, \widehat{\boldsymbol{\beta}}_n^\top)^\top$ be the minimizer of $\widehat{\Lambda}_{n,\pi}(\boldsymbol{\theta})$ in equation (2.8). The following conditions are required to proceed with the properties.

(A1) $\boldsymbol{X}(\boldsymbol{s})$ has an open and convex support and satisfies $\mathbb{E}(\|\boldsymbol{X}(\boldsymbol{s})\|^2) < \infty$.

(A2) The conditional distribution $\boldsymbol{X}(\boldsymbol{s})|Y = y$ is dominated by Lebesgue measure for $y = -1, 1$.

(A3) For an arbitrary $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, $\sum_{y\in\{-1,+1\}} P\{Y = y, \boldsymbol{X}(\boldsymbol{s}) \in \Psi(y, \boldsymbol{\theta})\} > 0$, where $\Psi(y, \boldsymbol{\theta}) = \left\{\boldsymbol{X}(\boldsymbol{s}) : \left(1 - y\boldsymbol{\theta}^\top \widetilde{\boldsymbol{X}(\boldsymbol{s})}\right)\left(1 - y\boldsymbol{\theta}_0^\top \widetilde{\boldsymbol{X}(\boldsymbol{s})}\right) < 0\right\}$.

(A4) Let U and V denote $\boldsymbol{\beta}^\top \boldsymbol{X}(\boldsymbol{s})$ and $\boldsymbol{\delta}^\top \boldsymbol{X}(\boldsymbol{s})$ respectively. Then a map $u \mapsto \mathbb{E}\{\boldsymbol{X}(\boldsymbol{s})|U = u, V = v, Y = y\} f_{U|V,Y}(u|v, y)$ is continuous for any linear independent vector $\boldsymbol{\beta}, \boldsymbol{\delta} \in \mathbb{R}^p$, $Y \in \{-1, +1\}$, and any constant $v \in \mathbb{R}$.

(A5) Given $U = u$, there exists a non-negative function $c_0(v, y)$ with $\mathbb{E}(c_0(V, Y)|Y) < \infty$ such that $\mathbb{E}\left\{\widetilde{\boldsymbol{X}(\boldsymbol{s})}|U = u, V, Y\right\} f_{U|V,Y}(U = u|V, Y) < c_0(v, y)$.

2.4.2.1 Consistency. The consistency of $\widehat{\boldsymbol{\theta}}_n$ is established in the following theorem.

Theorem 2.3. *Suppose* $Var\{\boldsymbol{X}(\boldsymbol{s})\} = \boldsymbol{\Sigma}$ *is positive definite and assumption* $(A2)$ *holds. Then,* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.

2.4.2.2 Asymptotic normality. Now we have the following theorem which gives the Bahadur representation of $\widehat{\boldsymbol{\theta}}_n$.

Theorem 2.4. *Under assumptions* $(A1) - (A5)$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -n^{-\frac{1}{2}} \mathbf{H}_{\boldsymbol{\theta}_0}^{-1} \sum_{i=1}^{n} \mathbf{D}_{\boldsymbol{\theta}_0}(\boldsymbol{Z}_i) + o_p(1), \tag{2.10}$$

where

$$\mathbf{D}_{\boldsymbol{\theta}_0}(\boldsymbol{Z}) = (0, 2\boldsymbol{\Sigma}\boldsymbol{\beta})^\top - \lambda\left[\pi(Y)\widetilde{\boldsymbol{X}}Y\mathbb{1}\{\boldsymbol{\theta}^\top\widetilde{\boldsymbol{X}}Y < 1\}\right], \text{ and} \tag{2.11}$$

$$\mathbf{H}_{\boldsymbol{\theta}} = 2diag(0, \boldsymbol{\Sigma}) + \lambda \sum_{y=-1,1} P(Y = y)\pi(y) f_{\boldsymbol{\beta}^\top \boldsymbol{X}|Y}(y - \alpha|y)\mathbb{E}(\widetilde{\boldsymbol{X}} wide \tilde{\boldsymbol{X}}^\top | \boldsymbol{\theta}^\top \widetilde{\boldsymbol{X}} = y). \tag{2.12}$$

Now for a given π_h, let $\boldsymbol{\theta}_{0,h} = (\alpha_{0,h}, \boldsymbol{\beta}_{0,h})$ be the minimizers of $\Lambda_{\pi_h}(\boldsymbol{\theta})$ and $\mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}) = \mathbf{F}_{\boldsymbol{\theta}_{0,h}} \mathbf{D}_{\boldsymbol{\theta}_{0,h}}(\boldsymbol{Z})$ for $h = 1, \ldots, H$, where $\mathbf{F}_{\boldsymbol{\theta}_{0,h}}$ denotes the last p rows of $\mathbf{H}^{-1}_{\boldsymbol{\theta}_{0,h}}$. Note that $\mathbb{E}(\mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z})) = 0 \, \forall \, h = 1, \ldots, H$. Thus a Bahadur representation of $\widehat{\boldsymbol{\beta}}_{n,h}$ is given by

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{n,h} - \boldsymbol{\beta}_{0,h}\right) = -n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}_i) + o_p(1) \tag{2.13}$$

by theorem 2.4. Using the Bahadur representation the asymptotic normality of the candidate matrix $\widehat{\mathbf{M}}_n$ given by equation (2.9) can be established.

Theorem 2.5. *Under assumptions* $(A1) - (A5)$ *and* $rank(\mathbf{M}_0) = k$,

$$\sqrt{n} vec(\widehat{\mathbf{M}}_n - \mathbf{M}_0) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{M}}),$$

where $\mathbf{M}_0 = \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h} \boldsymbol{\beta}_{0,h}^{\top}$. *The covariance matrix* $\boldsymbol{\Sigma}_{\mathbf{M}}$ *is given by*

$$\boldsymbol{\Sigma}_{\mathbf{M}} = (\mathbf{I}_{p^2} + \mathbf{T}_{p,p}) \sum_{h=1}^{H} \sum_{h'=1}^{H} (\boldsymbol{\beta}_{0,h} \boldsymbol{\beta}_{0,h'}^{\top} \otimes \mathbb{E}(\mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}) \mathbf{S}^{\top}(\boldsymbol{\theta}_{0,h'}, \boldsymbol{Z}))) (\mathbf{I}_{p^2} + \mathbf{T}_{p,p}),$$

where $\mathbf{T}_{u,v} \in \mathbb{R}^{uv \times uv}$ *denotes a communication matrix such that* $\mathbf{T}_{u,v} vec(\mathbf{A}) = vec(\mathbf{A}^{\top})$ *for a matrix* $\mathbf{A} \in \mathbb{R}^{u \times v}$, *and* $\mathbf{I}_u$ *is a u-dimensional identity matrix. The matrix operator* $\otimes$ *denotes Kronecker product.*

Corollary 2.1. *Under assumptions* $(A1) - (A5)$ *and* $rank(\mathbf{M}_0) = k$,

$$\sqrt{n} \, vec(\widehat{\mathbf{V}}_n - \mathbf{V}_0) \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{V}}),$$

where

$$\boldsymbol{\Sigma}_{\mathbf{V}} = (\mathbf{D}^{-1} \mathbf{U}^{\top} \otimes \mathbf{I}_p) \boldsymbol{\Sigma}_{\mathbf{M}} (\mathbf{U} \mathbf{D}^{-1} \otimes \mathbf{I}_p) \tag{2.14}$$

with $\mathbf{U}$ *being a* $p \times k$ *matrix whose columns are the eigenvectors of* $\mathbf{M}_0$ *corresponding to nonzero eigenvalues and* $\mathbf{D}$ *being a* $k \times k$ *diagonal matrix with diagonal elements given by the nonzero eigenvalues.*

The proofs of theorems 2.3, 2.4, and 2.5 follow the same techniques as depicted in Li et al. [28] and Shin et al. [41] and hence has been omitted here and detailed in Appendix A.

2.4.3 Determination of structure dimensionality, k

Estimating the dimension k of the CS is another key step in linear SDR. Following Li et al. [28], we can use a cross-validated BIC (CVBIC) procedure to determine k. Their procedure is based on the asymptotic properties of the PSVM estimator. Similar asymptotic properties for the WPSVM estimator will enable us to extend CVBIC to the WPSVM. For linear WPSVM, the BIC-type criterion, as suggested by Shin et al. [41], is given by

$$G_n\left(k;\eta,\mathbf{M}\right)=\sum_{j=1}^{k}v_j-\eta\,\frac{k\log n}{\sqrt{n}}\,v_1, \tag{2.15}$$

where v_j is the jth leading eigenvalue of a candidate matrix $\mathbf{M}$ and η is a tuning parameter. Then $\widehat{k}=\underset{k\in\{1,\ldots,p\}}{\operatorname{argmax}}\,G_n\left(k;\eta,\widehat{\mathbf{M}}_n^{LW}\right)$ is a reasonable estimator of k, where $\widehat{\mathbf{M}}_n^{LW}$ is the candidate matrix of the linear WPSVM as defined in equation (2.9). The estimator $\widehat{k}$ is consistent which can be proved from the asymptotic normality of $\widehat{\mathbf{M}}_n^{LW}$ established in theorem 2.5. The proof of theorem 2.6 can be found in Appendix B.

Theorem 2.6. *Under* $(A1)-(A5)$ *and* $rank(\mathbf{M}_0)=k$, $\lim\limits_{n\to\infty}\mathbb{P}\left(\widehat{k}=k\right)=1$.

The following steps help us choose an optimal η from the data using the modified CVBIC procedure:

1. Randomly split the data into training and test sets denoted by

$$\left\{\left(\boldsymbol{X}(\boldsymbol{s})_j^{\mathrm{tr}},Y(\boldsymbol{s})_j^{\mathrm{tr}}\right):j=1,\ldots,n_{\mathrm{tr}}\right\},\ \text{and}$$
$$\left\{\left(\boldsymbol{X}(\boldsymbol{s})_{j'}^{\mathrm{ts}},Y(\boldsymbol{s})_{j'}^{\mathrm{ts}}\right):j'=1,\ldots,n_{\mathrm{ts}}(=n-n_{\mathrm{tr}})\right\}\ \text{respectively.}$$

2. Since it is not advisable to randomly split spatial data due to the inherent correlations we have used a chess board method to split the data into training and test. The method is explained in Appendix C.

3. Fit the WPSVM to the training data $\left(\boldsymbol{X}(\boldsymbol{s})_j^{\text{tr}}, Y(\boldsymbol{s})_j^{\text{tr}}\right)$ and get the corresponding candidate matrix, $\widehat{\mathbf{M}}_n^{\text{tr}}$.

4. For each η do the following:

   (a) Compute $\hat{k}_{\text{tr}} = \underset{k=1,\ldots,p}{\text{argmax}}\, G_n\left(k; \eta, \widehat{\mathbf{M}}_n^{\text{tr}}\right)$.

   (b) Transform the training predictors $\widetilde{\boldsymbol{X}(\boldsymbol{s})}_j^{\text{tr}} = \left(\widehat{\boldsymbol{V}}_n^{\text{tr}}\right)^\top \boldsymbol{X}(\boldsymbol{s})_j^{\text{tr}}$ where $\widehat{\boldsymbol{V}}_n^{\text{tr}} = \left(\widehat{\boldsymbol{v}}_1^{\text{tr}}, \ldots, \widehat{\boldsymbol{v}}_{\hat{k}_{\text{tr}}}^{\text{tr}}\right)$ are the first $\hat{k}_{\text{tr}}$ leading eigenvectors of $\widehat{\mathbf{M}}_n^{\text{tr}}$.

   (c) For each $\pi_h, h = 1, \ldots, H$, fit the WSVM to $\left\{\left(\widetilde{\boldsymbol{X}(\boldsymbol{s})}_j^{\text{tr}}, Y(\boldsymbol{s})_j^{\text{tr}}\right) : j = 1, \ldots, n_{\text{tr}}\right\}$ to predict $Y(\boldsymbol{s})_{j'}^{\text{ts}}$. Denote the prediction by $\widehat{Y(\boldsymbol{s})}_{j',h}^{\text{ts}} : j' = 1, \ldots, n_{\text{ts}}; h = 1, \ldots, H$.

   (d) Calculate the associated total cost for the test data:

$$TC(\eta) = \sum_{h=1}^{H} \left\{ \sum_{j'=1}^{n_{\text{ts}}} \pi_h \left(Y(\boldsymbol{s})_{j'}^{\text{ts}}\right) \mathbb{1}\left(\widehat{Y(\boldsymbol{s})}_{j',h}^{\text{ts}} \neq Y(\boldsymbol{s})_{j'}^{\text{ts}}\right) \right\}, \tag{2.16}$$

   where $\pi_h(1) = 1 - \pi_h$ and $\pi_h(-1) = \pi_h$.

   (e) Repeat 4(a)-(d) over an appropriately chosen grid of η and select $\hat{\eta}$ that minimizes $TC(\eta)$.

Finally, compute $\widehat{k} = \underset{k\in\{1,\ldots,p\}}{\text{argmax}}\, G_n\left(k; \widehat{\eta}, \widehat{\mathbf{M}}_n^{LW}\right)$.

2.5 Kernel Version of WPSVM for Nonlinear SDR

It is worthwhile to mention here that the strength of WPSVM comes from the fact that it can employ a kernel technique. In this section we explore the kernel WPSVM used for nonlinear problems. The corresponding objective function (see Shin et al. [41]) is

given by

$$\Lambda_\pi(\alpha, \psi) = \text{Var}\left(\psi\left(\boldsymbol{X}(\boldsymbol{s})\right)\right) + \lambda \mathbb{E}\left\{\pi(Y(\boldsymbol{s}))[1 - Y(\boldsymbol{s}) f(\boldsymbol{X}(\boldsymbol{s}); \alpha, \psi)]_+\right\}, \tag{2.17}$$

where $f(\boldsymbol{X}(\boldsymbol{s}); \alpha, \psi) = \alpha + \psi(\boldsymbol{X}(\boldsymbol{s})) - \mathbb{E}\left\{\psi(\boldsymbol{X}(\boldsymbol{s}))\right\}$. We assume the function ψ belongs to a Hilbert space denoted by $\mathcal{H}$. Now, define a bilinear mapping $b : \mathcal{H} \times \mathcal{H} \mapsto \mathbb{R}$ as $b(\psi_1, \psi_2) = \text{Cov}\left(\psi_1(\boldsymbol{X}(\boldsymbol{s})), \psi_2(\boldsymbol{X}(\boldsymbol{s}))\right)$ for $\psi_1, \psi_2 \in \mathcal{H}$. Let us assume that a mapping from $\mathcal{H}$ to $L_2(P_{\boldsymbol{X}}) = \left\{\psi : \int \psi^2 \mathbf{d} P_{\boldsymbol{X}} < \infty\right\}$ is continuous. Then there exists a bounded and self-adjoint operator $\Sigma : \mathcal{H} \mapsto \mathcal{H}$ such that $\langle \psi_1, \Sigma \psi_2 \rangle_{\mathcal{H}} = b(\psi_1 \psi_2)$ (see Conway [6]). Based on this the objective function given by equation (2.17) can be equivalently written as

$$\Lambda_\pi(\alpha, \psi) = \langle \psi_1, \Sigma \psi_2 \rangle_{\mathcal{H}} + \lambda \mathbb{E}\left\{\pi(Y(\boldsymbol{s}))[1 - Y(\boldsymbol{s}) f(\boldsymbol{X}(\boldsymbol{s}); \alpha, \psi)]_+\right\}. \tag{2.18}$$

Note that the objective function for kernel WPSVM is analogous to the linear one given by equation (2.7) in that it is a nonlinear generalization of the linear WPSVM. Notice that the linear function $\boldsymbol{\beta}^\top \boldsymbol{X}(\boldsymbol{s}))$ is replaced by an arbitrary nonlinear function $\psi(\boldsymbol{X}(\boldsymbol{s}))$ and the corresponding covariance matrix $\boldsymbol{\Sigma}$ with operator Σ. Similar to linear WPSVM, Li et al. [28] introduced the notion of unbiasedness for nonlinear SDR.

Definition 2.2. A function $\psi \in \mathcal{H}$ is unbiased for nonlinear SDR given by equation (1.2) if it has a version that is measurable $\sigma\left\{\boldsymbol{\phi}(\boldsymbol{X}(\boldsymbol{s}))\right\}$, where $\sigma\left\{\boldsymbol{\phi}(\boldsymbol{X}(\boldsymbol{s}))\right\}$ denotes the σ-field generated by $\boldsymbol{\phi}(\boldsymbol{X}(\boldsymbol{s}))$.

Theorem 2.7 states the conditions under which the kernel WPSVM solution is unbiased. For a proof of the theorem please see Li et al. [28].

Theorem 2.7. *Suppose that the mapping from $\mathcal{H}$ to $L_2(P_{\boldsymbol{X}})$ is continuous and that $\mathcal{H}$ is a dense subset of $L_2(P_{\boldsymbol{X}})$, then $\psi_0(\boldsymbol{X}(\boldsymbol{s}))$ is unbiased, where (a_0, ψ_0) is the minimizer of the kernel WPSVM objective function given by equation (2.18).*

2.5.1 Finite sample estimation

The sample version of the objective function (2.18) is a bit difficult to obtain, since $\mathcal{H}$ is an infinite dimensional space of functions. Suppose $\mathcal{H}$ is a linear space of functions spanned by $\Omega = \{\omega_1, \ldots, \omega_d\}$, i.e.,

$$\mathcal{H} = \left\{ \psi : \psi(.) = \sum_{j=1}^{d} \gamma_j \omega_j(.),\, \gamma_j \in \mathbb{R},\, j = 1, \ldots, d \right\}. \tag{2.19}$$

We will talk about an appropriate choice of Ω in the next section, but for the moment let us assume that Ω is known. The sample version of the objective function (2.18) using the basis representation (2.19) is given by

$$\widehat{\Lambda}_{n,\pi}(\alpha, \boldsymbol{\gamma}) = \boldsymbol{\gamma}^\top \boldsymbol{\Omega}^\top \boldsymbol{\Omega} \gamma + \lambda \sum_{i=1}^{n} \pi(Y(\boldsymbol{s}_i))[1 - Y(\boldsymbol{s}_i)\left\{\alpha + \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i\right\}]_+, \tag{2.20}$$

where $\boldsymbol{\Omega}$ is an $(n \times d)$-dimensional matrix with i-th row given by

$$\boldsymbol{\Omega}_i = \left\{\omega_1(\boldsymbol{X}(\boldsymbol{s}_i)) - \bar{\omega}_1, \cdots, \omega_d(\boldsymbol{X}(\boldsymbol{s}_i)) - \bar{\omega}_d\right\}^\top,$$

where $\bar{\omega}_j = \frac{1}{n}\sum_{i=1}^{n} \omega_j(\boldsymbol{X}(\boldsymbol{s}_i))$, $j = 1, \ldots, d$. The minimizer of the sample objective function given by (2.20) is obtained using theorem 2.8 (see Li et al. [28])

Theorem 2.8. *Let $\hat{\boldsymbol{\nu}} = (\hat{\nu}_1, \ldots, \hat{\nu}_n)^\top$ denote the maximizer of the following quadratic programming problem*

$$\max_{\nu_1, \cdots, \nu_n} \sum_{i=1}^{n} \nu_i - \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_i \nu_j Y(\boldsymbol{s}_i) Y(\boldsymbol{s}_j) P_{\boldsymbol{\Omega}}^{(i,j)}$$

$$\textit{subject to } 0 \leq \nu_i \leq \lambda \pi(Y(\boldsymbol{s}_i)) \textit{ for } i = 1, \ldots, n, \textit{ and } \sum_{i=1}^{n} \nu_i Y(\boldsymbol{s}_i) = 0,$$

where $P_{\boldsymbol{\Omega}}^{(i,j)}$ is the (i,j)-th element of $P_{\boldsymbol{\Omega}} = \boldsymbol{\Omega}\left(\boldsymbol{\Omega}^\top \boldsymbol{\Omega}\right)^{-1} \boldsymbol{\Omega}^\top$. Then the minimizer of the objective function (2.20) is given by

$$\hat{\boldsymbol{\gamma}}_n = \frac{\lambda}{2} \sum_{i=1}^{n} \nu_i Y(\boldsymbol{s}_i) \left\{ \left(\boldsymbol{\Omega}^\top \boldsymbol{\Omega}\right)^{-1} \boldsymbol{\Omega}_i \right\}.$$

Please see Li et al. [28] and Shin et al. [41] for a proof of theorem 2.8. For each π_h, $h = 1, \ldots, H$, we minimize the objective function (2.18) and let the corresponding minimizers be $(\alpha_{n,h}, \boldsymbol{\gamma}_{n,h})$, $h = 1, \ldots, H$. Thus, the candidate matrix of kernel WPSVM is given by

$$\widehat{\mathbf{M}}_n^{KW} = \sum_{h=1}^{H} \widehat{\boldsymbol{\gamma}}_{n,h} \widehat{\boldsymbol{\gamma}}_{n,h}^{\top}. \tag{2.21}$$

As before, the basis of the central subspace $\mathcal{S}_1$ is estimated by $\widehat{\boldsymbol{\phi}}(\mathbf{x}) = \left\{\mathbf{V}_n^{KW}\right\}^{\top} \boldsymbol{\omega}(\mathbf{x})$, where $\widehat{\mathbf{V}}_n^{KW} = \left(\widehat{\mathbf{v}}_1^{KW}, \ldots, \widehat{\mathbf{v}}_k^{KW}\right)$ are the first k leading eigenvectors of $\widehat{\mathbf{M}}_n^{KW}$ and $\boldsymbol{\omega}(\mathbf{x}) = \{\omega_1(\mathbf{x}), \ldots, \omega_d(\mathbf{x})\}^{\top}$.

2.5.2 Choosing Ω

It is necessary to choose an optimal Ω in order to estimate the sufficient predictor for kernel WPSVM. Let us denote the optimal Ω by Ω_n. Li et al. [28] recommended using eigenfunctions of the linear operator Σ_n as an estimate for Ω_n, where the operator Σ_n is defined by $\langle \psi_1, \Sigma_n \psi_2 \rangle_{\mathcal{H}} = \mathrm{Cov}\left(\psi_1(\boldsymbol{X}(\boldsymbol{s})), \psi_2(\boldsymbol{X}(\boldsymbol{s}))\right)$. Here we compute the sample covariance.

Let $\mathbf{K}_n$ be the $n \times n$ kernel matrix and $\mathbf{Q}_n = \mathbf{I}_n - \mathbf{J}_n$, where $\mathbf{I}_n$ is an n-dimensional identity matrix and $\mathbf{J}_n$ is an n-dimensional square matrix with all elements equal to one. Following proposition 2 from Li et al. [28], $P_{\boldsymbol{\Omega}}$ is given by $(\mathbf{w}_1, \ldots, \mathbf{w}_d)$, where $\mathbf{w}_j$ is the j-the leading eigenvector of $\mathbf{Q}_n \mathbf{K}_n \mathbf{Q}_n$ with corresponding eigenvalue λ_j, $j = 1, \ldots, d$. Thus the j-th basis function $\omega_j(\mathbf{x})$, $j = 1, \ldots, d$ is given by

$$\omega_j(\boldsymbol{X}(\boldsymbol{s})) = \frac{1}{\lambda_j} \mathbf{w}_j^{\top} \mathbf{k}_n(\boldsymbol{X}(\boldsymbol{s})),$$

where $\mathbf{k}_n(\boldsymbol{X}(\boldsymbol{s})) = \left\{ K(\boldsymbol{X}(\boldsymbol{s})), \boldsymbol{X}(\boldsymbol{s}_i)) - \frac{1}{n} \sum_{i=1}^{n} K(\boldsymbol{X}(\boldsymbol{s})), \boldsymbol{X}(\boldsymbol{s}_i)),\ i = 1, \ldots, n \right\}$. We have chosen $d = \frac{n}{4}$ for our simulations and the real data example.

2.6 Simulations

2.6.1 Simulation design

In order to compare WPSVM to existing dimension reduction methods we conduct a simulation study similar to the models chosen by Guan and Wang [23]. They simulated a stationary multivariate Gaussian random field $\{\boldsymbol{X}(\boldsymbol{s})\}$ over an $m \times m$ window as the covariates, where $m = 1$ or 2 and $\boldsymbol{X}(\boldsymbol{s}) = \{X_1(\boldsymbol{s}), \ldots, X_5(\boldsymbol{s})\}^\top \in \mathbb{R}^5$. For each $1 \leq j \leq 5$, $\{X_j(\boldsymbol{s})\}$ is a stationary univariate Gaussian random field with $\mathbb{E}\{X_j(\boldsymbol{s})\} = 0$, $\mathrm{Var}\{X_j(\boldsymbol{s})\} = 1$, and the covariance is given by

$$\mathrm{Cov}\left\{X_{j_1}(\boldsymbol{s}_1), X_{j_2}(\boldsymbol{s}_2)\right\} = 0.5^{|j_1 - j_2|} \exp\left(-\frac{\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|}{\gamma}\right),$$

where $\gamma = 0.1$ or 0.2. $\{\epsilon(\boldsymbol{s})\}$ is also independently simulated as a stationary univariate Gaussian random field having the same mean, variance, and covariance structure as $\{X_j(\boldsymbol{s})\}$.

Four inhomogeneous spatial Poisson processes are generated, conditional on the above simulated $\{X_j(\boldsymbol{s})\}$ and $\{\epsilon(\boldsymbol{s})\}$, with the following first-order intensity functions:

(I) $\lambda_1(\boldsymbol{s}) = \alpha \exp\left\{X_1(\boldsymbol{s}) + X_2(\boldsymbol{s}) + 0.4\,\epsilon(\boldsymbol{s})\right\}$,

(II) $\lambda_1(\boldsymbol{s}) = \alpha \exp\left\{\frac{X_1^2(\boldsymbol{s})}{4} + 0.4\,\epsilon(\boldsymbol{s})\right\}$,

(III) $\lambda_1(\boldsymbol{s}) = \alpha \exp\left\{\frac{X_1(\boldsymbol{s})}{0.5 + \{1.5 + X_2(\boldsymbol{s})\}^2} + 0.4\,\epsilon(\boldsymbol{s})\right\}$,

where the constant $\alpha > 0$ is chosen in such a way that the expected number of events is 200 for the 1×1 window and 800 for the 2×2 window.

2.6.2 Linear WPSVM (LWPSVM)

We generate 500 spatial point processes for each model and compare our method with the SIR, SAVE, and DR estimation methods. We also test our CVBIC procedure in order to determine the structure dimensionality, k. Let $\mathcal{S}\left(\widetilde{\mathbf{B}}\right)$ denote an estimated

CS. We measure the estimation error by (see Li et al. [30], Xia [54])

$$\Delta\left(\mathbf{B}_0, \widetilde{\mathbf{B}}\right) = \left\| \mathbf{B}_0 \left(\mathbf{B}_0^\top \mathbf{B}_0\right)^{-1} \mathbf{B}_0^\top - \widetilde{\mathbf{B}} \left(\widetilde{\mathbf{B}}^\top \widetilde{\mathbf{B}}\right)^{-1} \widetilde{\mathbf{B}}^\top \right\|_{\max}, \tag{2.22}$$

where $\|\mathbf{A}\|_{\max}$ is the maximum absolute singular value of an arbitrary matrix $\mathbf{A}$, and $0 \leq \|\mathbf{A}\|_{\max} \leq 1$. Analogous to any distance measure, smaller values are better.

Tables 2.1, 2.2, and 2.3 capture the sample mean and standard deviation of $\Delta\left(\mathbf{B}_0, \widetilde{\mathbf{B}}\right)$ for each of the models respectively based on 500 simulation replications.

Table 2.1 Mean (Standard Deviation) of $\Delta\left(\mathbf{B}_0, \widetilde{\mathbf{B}}\right)$ for Model I

γ	p	Window	Results for:			
			SIR	SAVE	DR	LWPSVM
0.1	5	1×1	0.23 (0.03)	0.67 (0.12)	0.24 (0.03)	0.18 (0.07)
		2×2	0.30 (0.02)	0.52 (0.20)	0.32 (0.02)	0.08 (0.03)
	10	1×1	0.34 (0.03)	0.94 (0.09)	0.37 (0.03)	0.31 (0.07)
		2×2	0.30 (0.02)	1.00 (0.00)	0.32 (0.02)	0.20 (0.05)
	20	1×1	0.45 (0.03)	0.99 (0.01)	0.51 (0.03)	0.43 (0.07)
		2×2	0.34 (0.02)	1.00 (0.00)	0.37 (0.02)	0.26 (0.04)
0.2	5	1×1	0.25 (0.07)	0.82 (0.19)	0.30 (0.08)	0.21 (0.11)
		2×2	0.29 (0.02)	0.81 (0.09)	0.32 (0.02)	0.16 (0.05)
	10	1×1	0.35 (0.05)	0.90 (0.03)	0.45 (0.04)	0.27 (0.07)
		2×2	0.25 (0.03)	1.00 (0.00)	0.28 (0.03)	0.17 (0.04)
	20	1×1	0.42 (0.05)	0.98 (0.02)	0.47 (0.05)	0.60 (0.07)
		2×2	0.40 (0.03)	0.99 (0.00)	0.47 (0.03)	0.29 (0.04)

Model I is a linear model with $\mathbf{B}_0 = (1, 1, 0, 0, 0)^\top \in \mathbb{R}^5$. Here the intensity function $\lambda_1(\boldsymbol{s})$ is monotonic in $\mathbf{B}_0^\top \boldsymbol{X}(\boldsymbol{s})$ and hence, SIR is expected to perform well. DR performs reasonably since the properties of SIR are embedded in it. However, LWPSVM manages to outperform the existing methods with respect to the distance measure.

For Model II, the true structural dimension is still $k = 1$, but $\mathbf{B}_0 = (1, 0, 0, 0, 0)^\top \in \mathbb{R}^5$. The intensity function is symmetric in this case and hence, the performance of SIR is poor. However, SAVE and DR are sensitive to symmetric

Table 2.2 Mean (Standard Deviation) of $\Delta\left(\mathbf{B}_0, \widetilde{\mathbf{B}}\right)$ for Model II

γ	p	Window	Results for:			
			SIR	SAVE	DR	LWPSVM
0.1	5	1×1	0.46 (0.16)	0.33 (0.07)	0.33 (0.06)	0.64 (0.17)
		2×2	0.90 (0.13)	0.32 (0.03)	0.32 (0.03)	0.88 (0.12)
	10	1×1	0.65 (0.16)	0.43 (0.07)	0.41 (0.06)	0.86 (0.09)
		2×2	0.83 (0.11)	0.40 (0.05)	0.40 (0.05)	0.91 (0.09)
	20	1×1	0.87 (0.07)	0.52 (0.06)	0.56 (0.06)	0.95 (0.03)
		2×2	0.98 (0.03)	0.55 (0.04)	0.55 (0.04)	0.99 (0.02)
0.2	5	1×1	0.90 (0.12)	0.56 (0.21)	0.57 (0.21)	0.88 (0.11)
		2×2	0.52 (0.10)	0.44 (0.04)	0.41 (0.04)	0.65 (0.08)
	10	1×1	0.86 (0.13)	0.80 (0.14)	0.78 (0.14)	0.84 (0.12)
		2×2	0.95 (0.06)	0.49 (0.09)	0.50 (0.09)	0.94 (0.06)
	20	1×1	0.89 (0.09)	0.88 (0.09)	0.85 (0.10)	0.91 (0.07)
		2×2	0.94 (0.04)	0.55 (0.08)	0.56 (0.08)	0.93 (0.05)

direction and hence, perform reasonably well. LWPSVM performs poorly due to the symmetricity. However, the strength of WPSVM comes from the fact that a kernel technique can be employed as we will see in Section 2.6.3.

Model III is two-dimensional with $\mathbf{B}_0 = \left\{(1,0,0,0,0)^\top, (0,1,0,0,0)^\top\right\} \in \mathbb{R}^{5\times 2}$. Since, SIR can extract only one direction it performs poorly. The numbers for LWPSVM are comparable but the improvement is relatively small. Kernel WPSVM would also work in this case but has been omitted due to repetitiveness.

2.6.2.1 Structural dimensionality. The CVBIC procedure as formulated in Section 2.4.3 helps us determine the value of k. In this section we check the performance of the procedure for the three models against the true value of k. We test the procedure for Window 1×1. Table 2.4 contains the empirical probabilities (percentage) of correctly estimating true k based on 100 independent simulations.

Table 2.3 Mean (Standard Deviation) of $\Delta\left(\mathbf{B}_0, \widetilde{\mathbf{B}}\right)$ for Model III

γ	p	Window	Results for:			
			SIR	SAVE	DR	LWPSVM
0.1	5	1×1	1.00 (0.00)	0.83 (0.16)	0.80 (0.17)	0.69 (0.20)
		2×2	1.00 (0.00)	0.44 (0.20)	0.34 (0.09)	0.40 (0.18)
	10	1×1	1.00 (0.00)	0.92 (0.08)	0.91 (0.09)	0.81 (0.08)
		2×2	1.00 (0.00)	0.62 (0.14)	0.58 (0.13)	0.82 (0.08)
	20	1×1	1.00 (0.00)	0.97 (0.04)	0.97 (0.04)	0.82 (0.08)
		2×2	1.00 (0.00)	0.90 (0.11)	0.86 (0.13)	0.78 (0.11)
0.2	5	1×1	1.00 (0.00)	0.77 (0.18)	0.73 (0.19)	0.63 (0.21)
		2×2	1.00 (0.00)	0.58 (0.09)	0.57 (0.11)	0.77 (0.14)
	10	1×1	1.00 (0.00)	0.82 (0.09)	0.83 (0.10)	0.53 (0.12)
		2×2	1.00 (0.00)	0.61 (0.09)	0.59 (0.10)	0.69 (0.15)
	20	1×1	1.00 (0.00)	0.99 (0.02)	0.98 (0.02)	0.84 (0.07)
		2×2	1.00 (0.00)	0.91 (0.10)	0.93 (0.06)	0.69 (0.09)

Table 2.4 Empirical Probabilities (in Percentage) of Correctly Estimating True k Based on 100 Independent Simulations for Window 1×1

Model	γ	Results for:		
		$p = 5$	$p = 10$	$p = 20$
I	0.1	0.40	0.64	0.51
	0.2	0.19	0.31	0.11
II	0.1	0.58	0.29	0.10
	0.2	0.31	0.22	0.08
III	0.1	0.42	0.30	0.19
	0.2	0.42	0.26	0.18

It can be observed that the percentages are not that high for most of the models but it is not of much concern as the true values of k have the highest percentages among all the other possible values of k. This means, when we apply this to a real data we can work with the value of k having the highest empirical probability.

2.6.3 Kernel WPSVM (KWPSVM)

The strength of the weighted PSVM comes from the fact that it can employ a kernel technique for non-linear problems. As we saw in Section 2.6.2, the LWPSVM failed

to separate the classes for Model II, which is a non-linear model. In this section, we test our procedure by employing a kernel version of WPSVM.

Note that we cannot use the distance measure given by equation (2.22) to measure the performance. Commonly used techniques include computing the correlation between the response and the estimated sufficient predictor, and Hotelling T^2 statistics as used by Shin et al. [41]. Using correlation is inappropriate as we are dealing with binary response. The same is true for Hotelling T^2 as in this case the underlying assumptions of normality and independence are violated.

As an alternative, we try the Wilcoxon rank sum test in order to measure the performance. In particular, it is a non-parametric test involving two data samples. We are trying to find out whether the two estimated sufficient predictors come from distinct populations and do not affect each other. We compute the p-values for 500 independent simulations and compare them. The lesser the p-values, the more evidence toward class separation. Table 2.5 summarizes the results.

Table 2.5 Mean (Standard Deviation) of p-values from Wilcoxon Rank Sum Test Based on 500 Simulation Replications for Model II

γ	p	Window	Results for:			
			SIR	SAVE	DR	KWPSVM
0.1	5	1×1	0.09 (0.13)	0.17 (0.22)	0.17 (0.22)	0.00 (0.00)
		2×2	0.20 (0.20)	0.40 (0.28)	0.40 (0.27)	0.00 (0.00)
	10	1×1	0.03 (0.06)	0.27 (0.27)	0.26 (0.27)	0.00 (0.00)
		2×2	0.06 (0.07)	0.45 (0.26)	0.45 (0.26)	0.00 (0.00)
	20	1×1	0.00 (0.00)	0.18 (0.22)	0.14 (0.19)	0.00 (0.00)
		2×2	0.00 (0.01)	0.46 (0.27)	0.46 (0.27)	0.00 (0.00)
0.2	5	1×1	0.12 (0.15)	0.56 (0.27)	0.53 (0.28)	0.00 (0.00)
		2×2	0.00 (0.00)	0.00 (0.01)	0.00 (0.01)	0.00 (0.00)
	10	1×1	0.03 (0.05)	0.43 (0.29)	0.34 (0.30)	0.00 (0.00)
		2×2	0.02 (0.04)	0.40 (0.28)	0.39 (0.28)	0.00 (0.00)
	20	1×1	0.00 (0.00)	0.53 (0.27)	0.40 (0.30)	0.00 (0.00)
		2×2	0.00 (0.00)	0.21 (0.25)	0.18 (0.25)	0.00 (0.00)

It is clear that KWPSVM outperforms the existing methods in terms of separation of the classes when the decision curve is non-linear. In conclusion, WPSVM works reasonably well for the three models and hence is our suggested approach.

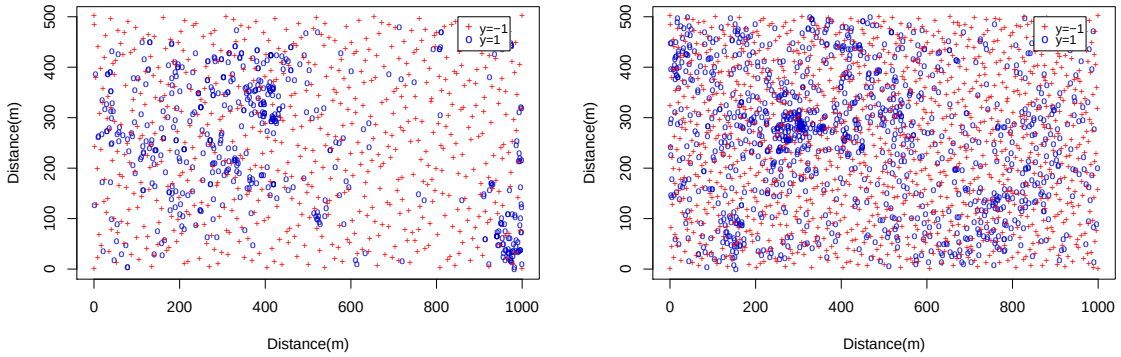
2.7 Application to Rainforest Data

For the purpose of validation, we apply WPSVM to rainforest data collected as part of the BCI forest dynamics research. The BCI forest dynamics research project was founded by S.P. Hubbell and R.B. Foster and is now managed by R. Condit, S. Lao, and R. Perez under the Center for Tropical Forest Science and the Smithsonian Tropical Research in Panama. Numerous organizations have provided funding, principally the U.S. National Science Foundation, and hundreds of field workers have contributed. Please see Hubbell et al. [26], Condit [5], Hubbell et al. [27] for a reference of their work.

We consider 503 trees of the species *Laetia thamnia* and 1132 trees of the species *Cassipourea elliptica* that were recorded in a 2015 census in part of the Barro Colorado Island plot. We have considered some soil variables as features for the model. Estimates for concentration of the soil nutrients were downloaded from `http://ctfs.si.edu/webatlas/datasets/bci/soilmaps/BCIsoil.html`. We acknowledge the principal investigators who were responsible for collecting and analysing the soil maps (Jim Dallin, Robert John, Kyle Harms, Robert Stallard and Joe Yavitt), the funding sources (National Science Foundation grants DEB021104, 021115, 0212284 and 0212818 and Office of International Science and Engineering grant 0314581, the Smithsonian Tropical Research Institute soils initiative and the Center for Tropical Forest Science) and field assistants (Paolo Segre and Juan Di Trani).

We considered soil content of aluminium, potassium, phosphorus and mineralized nitrogen and soil acidity level pH as features. Figure 2.1 shows the point patterns for the two species along with the dummy points used. In each case, a

quasi-random pattern of dummy points are generated which includes the four corner points in the window.



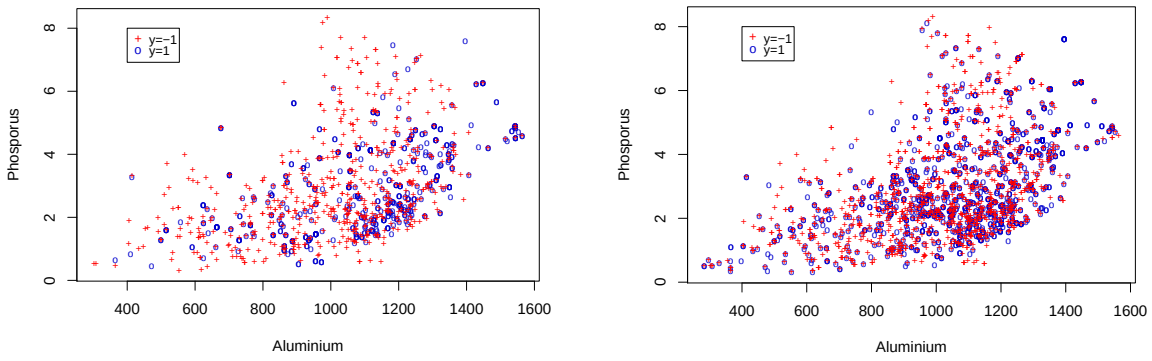
(a) (b)

Figure 2.1 Location of (a) 503 *Laetia thamnia* and (b) 1132 *Cassipourea elliptica* trees along with dummy points.

Figure 2.2 illustrates the difficulty of separating the trees from the dummy points based on soil concentrations of Aluminium and Phosphorus. As such, our main motive is to test the performance of the linear WPSVM and the kernel WPSVM in classifying trees from dummy points while reducing dimension at the same time.

2.7.1 Laetia thamnia

The first step is to find the optimal value of k, the structural dimension. For the linear WPSVM, we have considered twenty values of π equally spaced between 0 and 1 and λ is taken to be 1. We employ a two-fold CVBIC approach as described in Section 2.4.3. In order to minimize the total cost given by equation (2.16), we choose an appropriate grid for η, in particular, ten equally spaced points in each interval given by $(10^j, 10^{j+1}]$, $j = -3, \ldots, 2$. This gives us a grid of 64 equally spaced points over the interval $[10^{-3}, 10^3]$. The optimal η is 0.002 and 0.001 for the two folds and the corresponding $G_n(k; \eta, \mathbf{M})$ given by equation (2.15) is maximized when $k = 2$.



(a) (b)

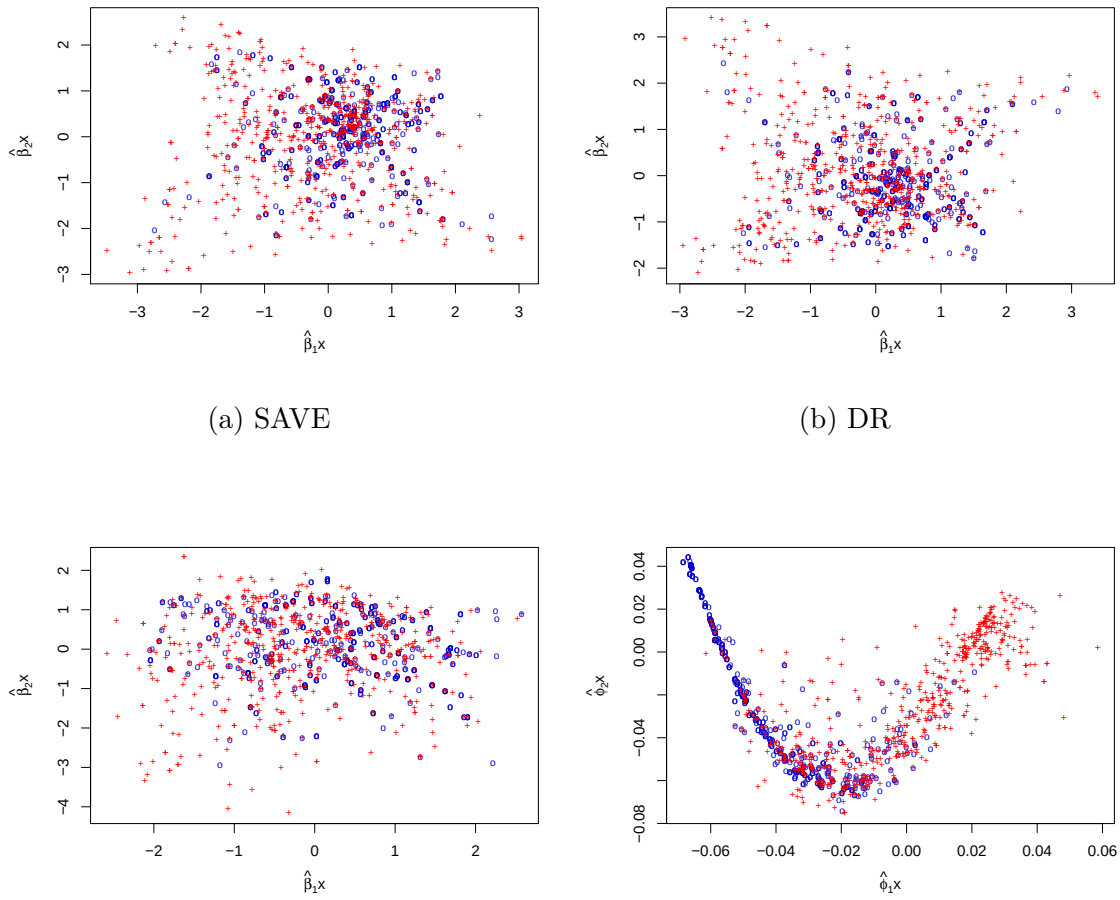
Figure 2.2 Scatter plots of trees along with dummy points for (a) *Laetia thamnia* and (b) *Cassipourea elliptica* based on soil concentrations of Aluminium and Phosphorus.

Finally, we apply kernel WPSVM to the rainforest data using the Gaussian kernel $K\left(\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s})^{\top}\right) = \exp\left(-\left\|\boldsymbol{X}(\boldsymbol{s}) - \boldsymbol{X}(\boldsymbol{s})^{\top}\right\|^2 / 2\sigma^2\right)$ with bandwidth parameter σ equal to the median of pairwise Euclidean distances between the two classes. The same set of values of π and λ were used. Figure 2.3 shows a comparison of the different methods employed.

2.7.2 Cassipourea elliptica

In the case of *Cassipourea elliptica*, the optimal η is 0.001 and 0.002 for the two folds and the corresponding cost function given by equation (2.15) is maximized when $k = 2$. Figure 2.4 gives a comparison of the different methods employed.

It is evident from the scatter plots that kernel WPSVM manages to separate the two classes and does it much better than the competing methods chosen. However, it is possible to further improve the performance by controlling the way the dummy points are chosen in the window. One such way to achieve this is by thinning the points.



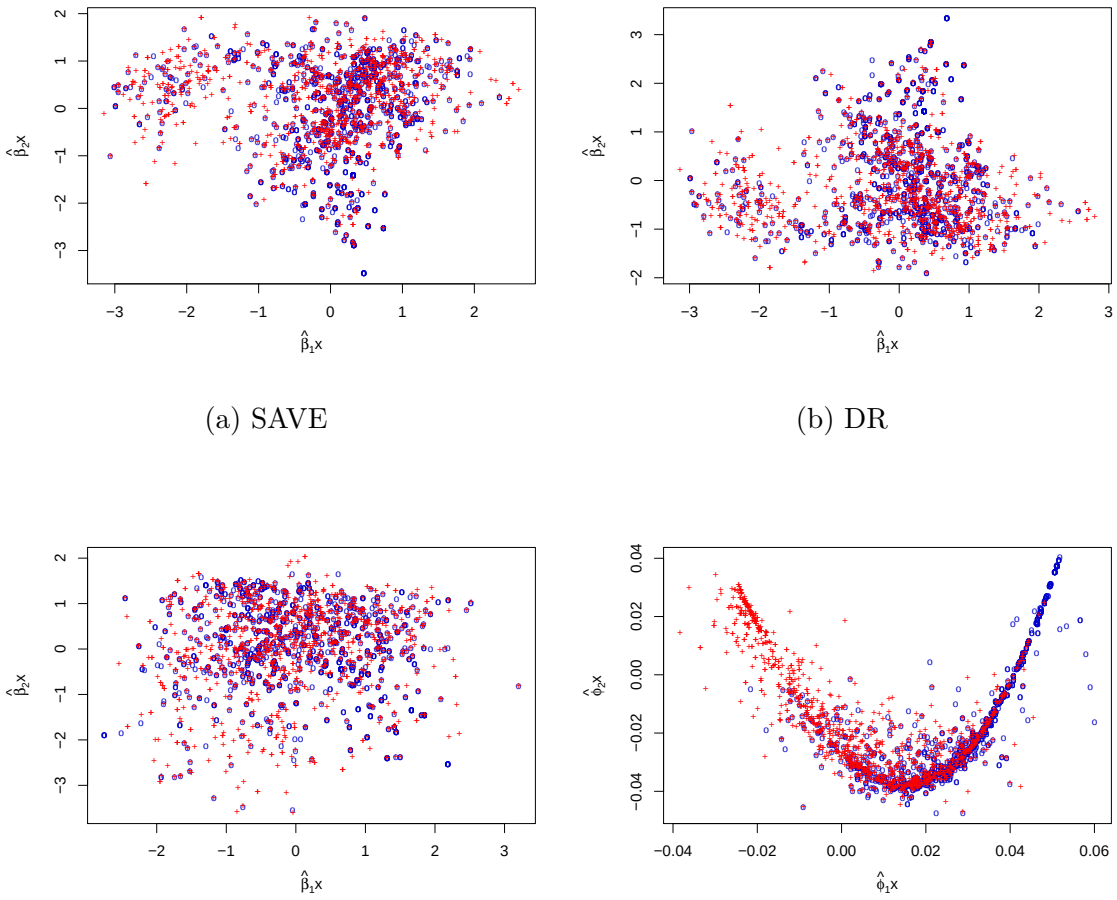
(a) SAVE

(b) DR

(c) LWPSVM

(d) KWPSVM

Figure 2.3 Scatter plots of the first two sufficient predictors as estimated by the following: (a) SAVE, (b) DR, (c) LWPSVM, and (d) KWPSVM for the *Laetia thamnia* species.



(a) SAVE

(b) DR

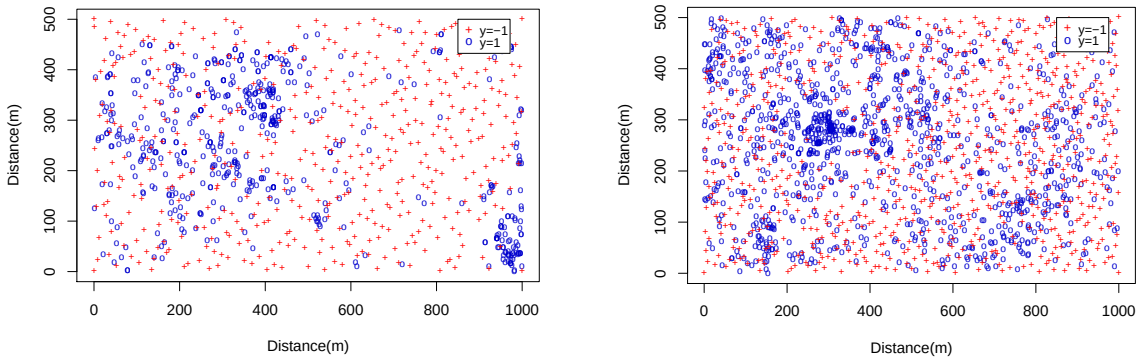
(c) LWPSVM

(d) KWPSVM

Figure 2.4 Scatter plots of the first two sufficient predictors as estimated by the following: (a) SAVE, (b) DR, (c) LWPSVM, and (d) KWPSVM for the *Cassipourea elliptica* species.

2.7.3 Thinning

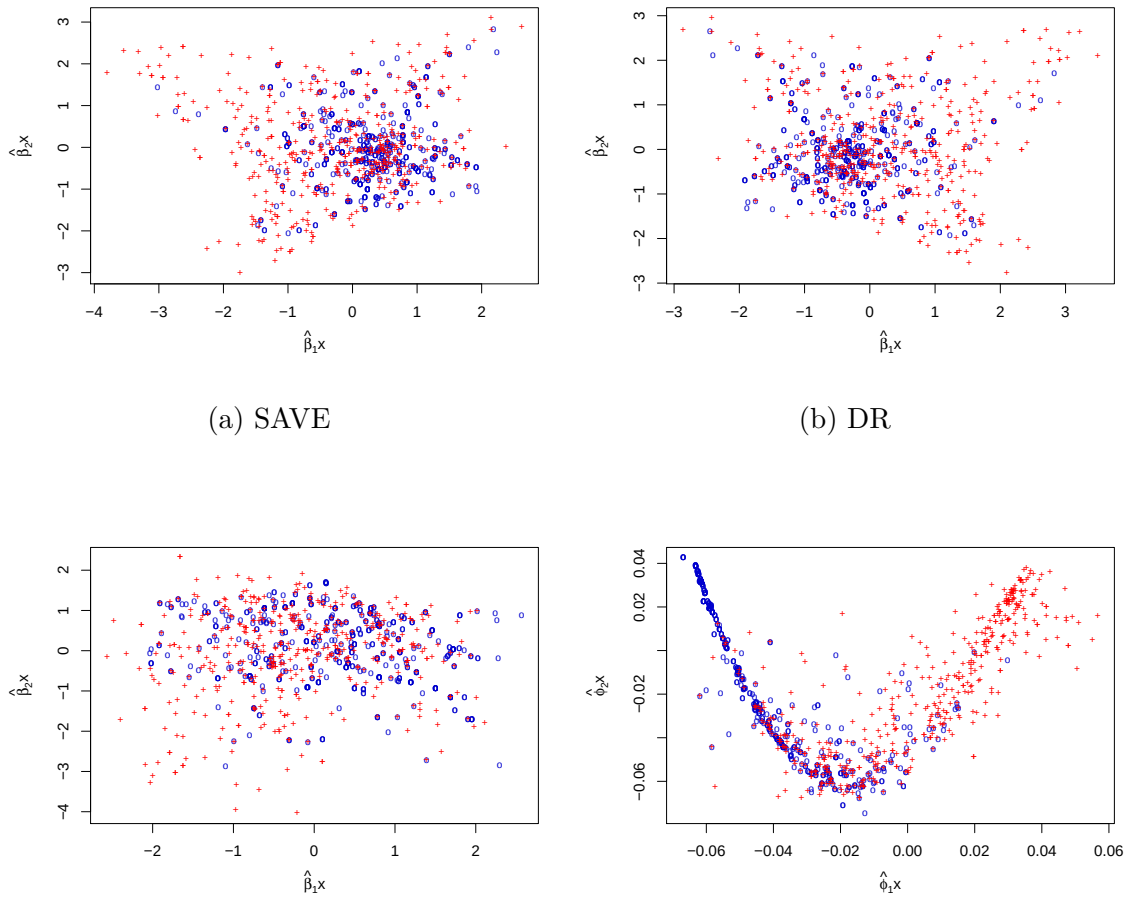
In spatial data, thinning is usually used to reduce effects of sampling bias while still retaining most of the information. If done effectively, it can reduce data redundancy and will improve analysis quality. We explore the method here by controlling the intensity of the dummy points. We make sure that the probability of a dummy point falling in a region with high intensity of data points is low. This is achieved by controlling the retention probabilities, i.e., the probability that each existing dummy point will be retained. Lower retention probabilities indicate that an existing dummy point will most likely be deleted. We perform independent random thinning which means that the retention/deletion of each dummy point is independent of other points. Figure 2.5 shows the point patterns along with the thinned dummy points for the two species considered. Notice that there are fewer dummy points when compared with Figure 2.1.



(a) (b)

Figure 2.5 Location of (a) 503 *Laetia thamnia* and (b) 1132 *Cassipourea elliptica* trees along with thinned dummy points.

Figures 2.6 and 2.7 show the same comparison as before of the different methods employed, but using the thinned data. The kernel WPSVM is achieves better classification as is evident from the plots.



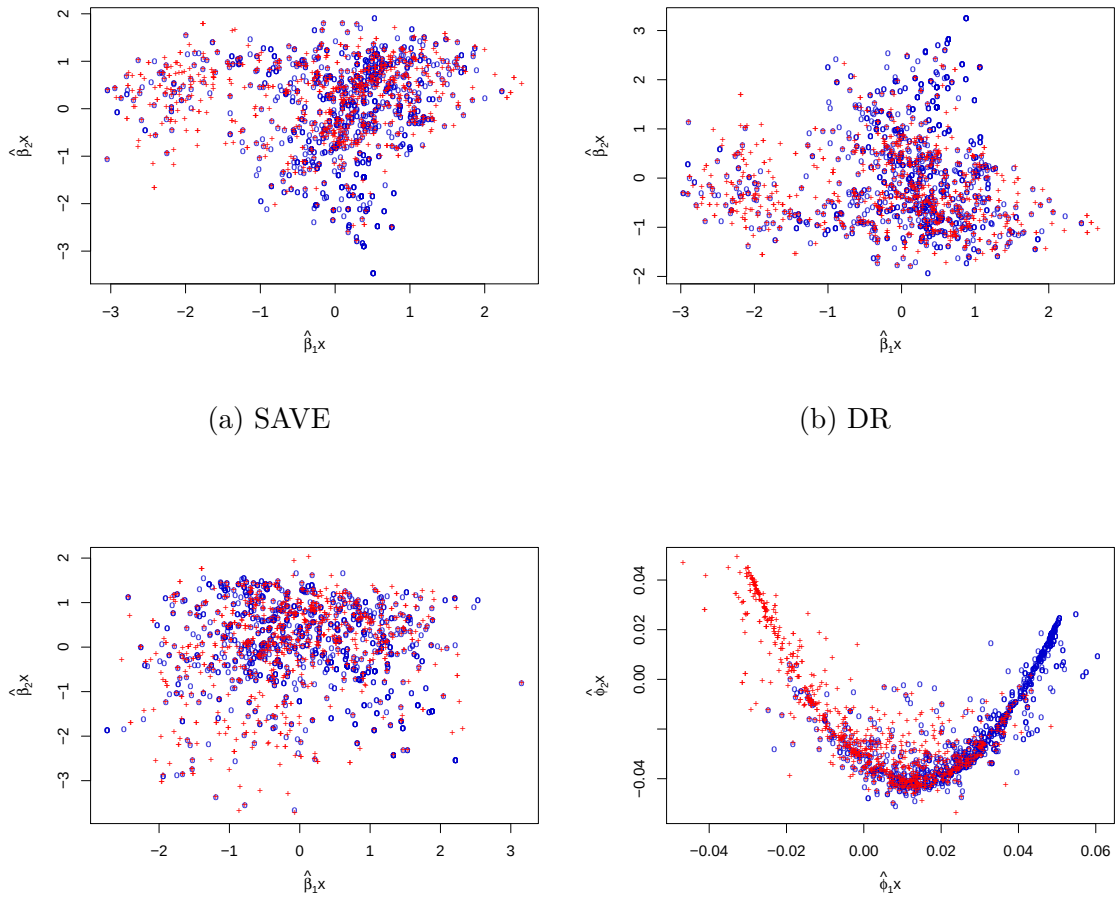
(a) SAVE

(b) DR

(c) LWPSVM

(d) KWPSVM

Figure 2.6 Scatter plots of the first two sufficient predictors as estimated by the following: (a) SAVE, (b) DR, (c) LWPSVM, and (d) KWPSVM for the *Laetia thamnia* species using thinned data.



(a) SAVE

(b) DR

(c) LWPSVM

(d) KWPSVM

Figure 2.7 Scatter plots of the first two sufficient predictors as estimated by the following: (a) SAVE, (b) DR, (c) LWPSVM, and (d) KWPSVM for the *Cassipourea elliptica* species using thinned data.

2.8 Discussion

We conclude this chapter with remarks about the proposed method for dimension reduction. As is evident from the results, WPSVM has an advantage over the other methods for spatial point processes. The reason comes from the fact that it works very well for SDR in binary classification where existing methods falter. Another advantage is that the kernel technique can be applied for nonlinear problems which other methods lack and we saw that KWPSVM managed to classify very complex data. Application to the rainforest data also gave favorable results.

In most cases of SDR estimation, the interpretability is lost. However, interpretation is not the major focus of this paper. Many existing papers have employed the use of a sparse SDR in order to improve the interpretability.

Also, WPSVM fails to work when the number of predictors, p is larger than the number of observations, n. One possible and popular way to tackle this is to use a penalty term for $\boldsymbol{\beta}$. However, this adds an extra level of complexity, in the form of a tuning parameter, to an already complex problem.

Another way, which is the focus of the next chapter, is to use joint screening. This method is catered towards reducing dimension in ultra-high dimensional cases. However, one caveat is that screening methods are heavily dependent on a model assumption whereas SDR is model free. Nonparametric methods have been explored for variable screening and its applicability towards WPSVM can be investigated in the future.

CHAPTER 3

JOINT SCREENING OF ULTRA-HIGH DIMENSIONAL VARIABLES FOR MIXED MODELS

3.1 Introduction

In this chapter, we propose a novel joint screening (JS) procedure in order to filter and obtain a subset of relevant features from an ultra-high dimensional problem. The proposed estimator is computationally efficient and does not have a marginal correlation assumption, unlike SIS [18]. We evaluate the performance of the proposed JS procedure via simulation studies and an application to a real data.

3.2 A Novel Joint Screening Procedure

Our joint screening procedure for mixed models is motivated by the joint screening procedure for linear models proposed by Wang and Leng [51]. The JS procedure is called High-dimensional Ordinary Least-squares Projection (HOLP) and is for the following linear model,

$$\widetilde{\boldsymbol{Y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\epsilon}}, \tag{3.1}$$

where $\widetilde{\boldsymbol{Y}}$ is an $n \times 1$ vector of observed phenotypes, $\widetilde{\mathbf{X}}$ is an $n \times p$ design matrix of genetic variables, and $\boldsymbol{\beta}$ is a $p \times 1$ vector representing the fixed effects of genetic variables. We assume that $\widetilde{\boldsymbol{\epsilon}}$ has zero-mean and $Var(\widetilde{\boldsymbol{\epsilon}}) = \sigma_e^2 I_n$. Note that the subjects are independent under the linear model given by equation (3.1) while the subjects are correlated via the kinship coefficient matrix under the mixed model given by equation (1.3).

3.2.1 HOLP for linear model

We first describe the HOLP procedure for the linear model. Under linear model (3.1), if dimension p were small compared with sample size n, we could consider the

following least-squares (LS) estimate,

$$\widetilde{\boldsymbol{\beta}}_{LS} = (\widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\top}\widetilde{\boldsymbol{Y}}. \tag{3.2}$$

But for the setting where $p \gg n$, the LS estimate is not applicable due to the aforementioned curse of dimensionality. To overcome this problem, the HOLP procedure simply rearranges the positions of design matrix $\widetilde{\mathbf{X}}$ in equation (3.2) and uses the following estimate,

$$\widetilde{\boldsymbol{\beta}}_{JS} = \widetilde{\mathbf{X}}^{\top}(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\top})^{-1}\widetilde{\boldsymbol{Y}}. \tag{3.3}$$

The equations (3.2) and (3.3) are commonly known as "dual equations"; see for example Shawe-Taylor and Cristianini [40]. Equation (3.3) not only solves the problem of non-uniqueness of the solution to (3.2) when the dimensional of variables is high but also, more importantly, provides some ranking for those variables. That is, based on $\widetilde{\boldsymbol{\beta}}_{JS}$, we can conduct joint screening, using the following subset of variables for the second stage analysis,

$$\widetilde{\mathcal{M}}_k = \{j : |\widetilde{\beta}_j| \text{ is among the top } k \text{ of all } |\widetilde{\beta}_j|\}. \tag{3.4}$$

To derive the sure screening consistency of the proposed JS procedure for linear models, Wang and Leng [51] assumed that the true coefficient vector $\boldsymbol{\beta}$ is sparse; that is, many of the components of $\boldsymbol{\beta}$ are exactly equal to zero. Let

$$\mathcal{M}_* = \{j : \beta_{*j} \neq 0,\, 1 \leq j \leq p\},$$

where $\boldsymbol{\beta}_* = (\beta_{*1}, \ldots, \beta_{*p})^{\top}$ is the true coefficient vector in equation (3.1). Wang and Leng [51] showed that, under some standard conditions on the design matrix $\tilde{X}$ and some weak condition on k, $P(\mathcal{M}_* \subseteq \widetilde{\mathcal{M}}_k) \to 1$ as $n \to \infty$ and p diverges with n. Furthermore, under some condition on k, $P(\widetilde{\mathcal{M}}_k = \mathcal{M}_*) \to 1$ as $n \to \infty$ and p diverges with n.

3.2.2 HOLP for mixed model

Now we are ready to describe our joint screening procedure for mixed models. Assume for the moment that the covariance matrix V given by equation (1.4) is known. Under the transformation $\widetilde{\boldsymbol{Y}} = V^{-1/2}\boldsymbol{Y}$, mixed model (1.3) becomes

$$\widetilde{\boldsymbol{Y}} = V^{-1/2}\mathbf{X}\boldsymbol{\beta} + V^{-1/2}(\boldsymbol{\alpha} + \boldsymbol{\epsilon}) = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\epsilon}},$$

which is equivalent to linear model given by equation (3.1). Therefore, motivated by the idea of HOLP in equation (3.3), we propose the joint screening estimate for mixed model as

$$\widetilde{\boldsymbol{\beta}}_{JS} = \widetilde{\mathbf{X}}^{\top}(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\top})^{-1}\widetilde{\boldsymbol{Y}},$$

where $\widetilde{\boldsymbol{Y}} = V^{-1/2}\boldsymbol{Y}$,and $\widetilde{\mathbf{X}} = V^{-1/2}\mathbf{X}$. Now, if we plug in the transformations back into the above equation, we have

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_{JS} &= \mathbf{X}^{\top}V^{-1/2}\left(V^{-1/2}\mathbf{X}\mathbf{X}^{\top}V^{-1/2}\right)^{-1}V^{-1/2}\boldsymbol{Y} \\
&= \mathbf{X}^{\top}V^{-1/2}V^{1/2}\left(\mathbf{X}\mathbf{X}^{\top}\right)^{-1}V^{1/2}V^{-1/2}\boldsymbol{Y} \\
&= \mathbf{X}^{\top}\left(\mathbf{X}\mathbf{X}^{\top}\right)^{-1}\boldsymbol{Y}.
\end{aligned}$$

Therefore, under mixed model (1.3), the joint screening estimate is

$$\widehat{\boldsymbol{\beta}}_{JS} = \mathbf{X}^{\top}\left(\mathbf{X}\mathbf{X}^{\top}\right)^{-1}\boldsymbol{Y}. \tag{3.5}$$

For the rest of the dissertation we denote the joint screening estimate for the mixed model given by equation (1.3) by $\widehat{\boldsymbol{\beta}}_{JS}$ in order to differentiate it from the linear model one given by equation (3.3). It is important to note that the JS screening estimate (3.5) does not depend on unknown matrix V. Thus, we avoid the computationally difficult problem of estimating V via the REML (1.6). Because the matrix under inverse in equation (3.5), $\mathbf{X}\mathbf{X}^{\top}$, is an $n \times n$ matrix, its computation is fast for the settings where $p \gg n$. The estimate (3.5) has a computational complexity of $O(n^2 p)$.

3.3 Sure Screening Properties

It is important to check whether the proposed joint screening procedure for mixed models can filter relevant features with overwhelming probability, with increase in sample size. We explore the sure screening properties of the method in this section.

Let $\mathbf{X} = \left(\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top\right)^\top$ denote the design matrix. Without loss of generality, we assume X_j, $j = 1, \ldots, p$ have mean 0 and standard deviation 1. Let $\Sigma = \mathrm{Cov}(\boldsymbol{X})$. We define,

$$\mathbf{Z} = \mathbf{X}\Sigma^{-1/2}, \text{and } \boldsymbol{Z} = \Sigma^{-1/2}\boldsymbol{X}.$$

Note that $\mathbf{X}$ and $\mathbf{Z}$ are $p \times p$ matrices, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are p-dimensional vectors. The tail behavior of the random error $\boldsymbol{\epsilon}$ is of particular interest here since it controls the screening performance. We present the following tail condition to characterize the tail behavior of different distribution families as depicted in Vershynin [48].

Definition 3.1. A zero mean distribution F is said to have a q-exponential tail, if any $N \geq 1$ independent random variables $\epsilon_i \sim F$ satisfy that for any N constants a_i with $\sum_{i=1}^{N} a_i^2 = 1$, the following inequality holds,

$$P\left(\left|\sum_{i=1}^{N} a_i \epsilon_i\right| > t\right) \leq \exp(1 - q(t)),$$

for any $t > 0$ and some function $q(.)$.

This characterization of the tail behavior is very general. As shown in Vershynin [48], $q(t) = O(t^2/D^2)$ for some constant D depending on F if F is sub-Gaussian including Gaussian, Bernoulli, and any bounded random variables. Also, $q(t) = O(\min\{t/D, t^2/D^2\})$ if F is sub-exponential including exponential, Poisson, and χ^2 distribution. Moreover, as shown in Zhao and Yu [58], $q(t) = 2d \log t + O(1)$ if F has bounded $2d$-th moments for some positive integer d.

Throughout the rest of this dissertation, $\lambda_{\max}$ and $\lambda_{\min}$ denote respectively the largest and the smallest eigenvalues of a matrix, and d, D, d_i, and D_i denote absolute constants independent of n and p. We make the following assumptions.

(B1) The transformed $\boldsymbol{Z}$ has a spherically symmetric distribution and there exists some $d_1 > 1$ and $D_1 > 0$ such that

$$P\left(\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^\top) > d_1 \text{ or } \lambda_{\min}(p^{-1}\mathbf{Z}\mathbf{Z}^\top) < d_1^{-1}\right) < \exp(-D_1 n).$$

Assume $p > d_0 n$ for some $d_0 > 1$.

(B2) The random error ϵ has mean zero and standard deviation σ_e. The standardized error ϵ/σ_e has q-exponential tails with $q(t)$ independent of $\boldsymbol{X} = \boldsymbol{x}$.

(B3) For some $\kappa \geq 0, \nu \geq 0, \tau \geq 0$, and $d_2, d_3, d_4 > 0$,

$$\min_{j \in \mathcal{M}_*} |\beta_{*j}| \geq \frac{d_2}{n^\kappa}, \quad s = |\mathcal{M}_*| \leq d_3 n^\nu, \text{and } \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \leq d_4 n^\tau.$$

We now state the following theorems.

Theorem 3.1 (Screening property)**.** *Under assumptions* $(B1) - (B3)$*, if we choose* γ_n *such that*

$$\frac{p\gamma_n}{n^{1-\tau-\kappa}} \to 0, \textit{and } \frac{p\gamma_n\sqrt{\log n}}{n^{1-\tau-\kappa}} \to \infty,$$

then

$$P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\gamma_n}\right) = 1 - O\left\{\exp\left(\frac{-D_1 n^{1-5\tau-2\kappa-\nu}}{\log n}\right)\right\} - s\exp\left\{1 - q\left(\frac{\sqrt{D_1} n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right\}.$$

Note that we do not make any assumption on p as long as $p > d_0 n$. Under some further mild conditions on p for ultra-high dimensional problems we state the following screening consistency.

Theorem 3.2 (Screening consistency)**.** *Under assumptions* $(B1)-(B3)$*, if* p *satisfies*

$$\log p = o\left(\min\left\{\frac{n^{1-5\tau-2\kappa-\nu}}{2\log n}, q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\log n}\right)\right\}\right), \tag{3.6}$$

then for the same γ_n *as defined in theorem 3.1, we have*

$$\begin{aligned} &P\left(\min_{j\in\mathcal{M}_*}|\hat{\beta}_j| > \gamma_n > \max_{j\notin\mathcal{M}_*}|\hat{\beta}_j|\right) \\ &\quad = 1 - O\left\{\exp\left(-\frac{D_1 n^{1-5\tau-2\kappa-\nu}}{\log n}\right) + \exp\left(1 - \frac{1}{2}q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right)\right\}. \end{aligned}$$

Alternatively, we can choose a submodel $\mathcal{M}_k$ *with* $k \asymp n^{\iota}$ *for some* $\iota \in (\nu, 1]$*, such that*

$$\begin{aligned} &P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_k\right) \\ &\quad = 1 - O\left\{\exp\left(-\frac{D_1 n^{1-5\tau-2\kappa-\nu}}{\log n}\right) + \exp\left(1 - \frac{1}{2}q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right)\right\}. \end{aligned}$$

The proofs of theorems 3.1 and 3.2 are given in Appendix D.

3.3.1 Determination of k

The first part of theorem 3.2 shows that if p satisfies condition (3.6), the relevant and irrelevant features are separable with probability tending to one by thresholding the projection estimator. The second part states that as long as we choose a submodel with dimension larger than that of the true model, we are guaranteed to retain the important features with probability tending to one. If we choose $k = s$, then the proposed screening estimator selects the true model with an overwhelming probability.

The determination of k is an important issue. Here we describe two common approaches. One approach is that we use a conservatively large k initially, say $k = n$. Then, based on the top k genetic variables, we apply some penalized mixed model, say the l_1-penalized mixed model; see Schelldorfer et al. [39] along with 10-fold cross-validation, to select a subset of k' genetic variables, where $k' < k$. Another approach is that we simply use $k = \lfloor n/\log n \rfloor$. This approach was first considered

by Fan and Lv [18], where they proposed the Sure Independence Screening (SIS) procedure. We consider the second approach to determine the value of k; that is, $k = \lfloor n/\log n \rfloor$.

3.4 Simulation Studies

3.4.1 Screening accuracy

We conduct simulations study to evauate the performance of the proposed JS procedure for mixed models. Our motive is to show that it is robust to the familial effects. Consider the following model,

$$\boldsymbol{y}_{ij} = \alpha_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \epsilon_{ij}, i = 1, \cdots, M; j = 1, \cdots, J.$$

Let $N = M \times J$ which is the total number of observations. The values of the parameters are taken to be:

(**i**) There are M families; each has $J = 5$ subjects,

(**ii**) $(p, N) = (2000, 200)$ or $(p, N) = (2000, 400)$,

(**iii**) $\alpha_i \sim N(0, K)\, \forall i$, where K is a block diagonal matrix which we have generated randomly, and

(**iv**) $\epsilon_{ij} \sim N(0, 1)\, \forall i, j$.

We now consider the following scenarios.

Scenario 1: $\boldsymbol{x}_{ij} \sim MVN(0, I_N)$ with $\text{Cov}(x_{ij}, x_{ih}) = \rho$ for any j and h. We set $\rho = 0.3, 0.6$ or 0.9 and $\mathcal{M}_* = \{1, 2, 3, 4, 5\}$ with $\boldsymbol{\beta}_{\mathcal{M}_*} = (5, 5, 5, 5, -20\rho)^\top$.

Scenario 2: Similar to scenario 1 but $\mathcal{M}_* = \{1, \ldots, 15\}$ with $\boldsymbol{\beta}_{\mathcal{M}_*} = (1_{14}^\top, -1.6)^\top$.

Scenario 3: $\boldsymbol{x}_{ij} \sim MVN(0, I_N)$ with $\text{Cov}(x_{ij}, x_{ih}) = \rho^{|j-h|}$ for any j and h. We set $\rho = 0.3, 0.6$ or 0.9 and $\mathcal{M}_* = \{1, 2, 3, 4, 5\}$ with $\boldsymbol{\beta}_{\mathcal{M}_*} = (2, 2, 2, 2, -3.65)^\top$.

We replicate each scenario 100 times. In scenario 1 there are a small number of non-zero features with large effect sizes, whereas in scenario 2 we have considered more non-zero features with smaller effect sizes. Both scenarios 1 and 2 have an equal correlation structure. For scenario 3 we have adopted a first-order autoregressive correlation structure. All of these scenarios mimic the ones from Wang and Leng [51].

In order to validate the performance of the screening procedure, we first evaluate the minimum model size, i.e., the smallest number of features required to include all of the important ones in $\mathcal{M}_*$. To that end we compute the median and interquartile range of the minimum model size. Ideally, we would want any screening procedure to have a smaller minimum model size. Second, we calculate the probability of including the true model which is a proportion, out of 100 replications, that all of the features in $\mathcal{M}_*$ are selected by a submodel $\mathcal{M}_k$ of size given by $k = \lfloor n/\log n \rfloor$. We denote this probability by $\mathbb{P}_{\text{all}}$. Though not proposed for mixed models, we compare the JS procedure with SIS [18] and robust rank correlation based screening (RRCS; Li et al. [31]). Table 3.1 contains the simulation results.

The performance of the proposed joint screening procedure is better than that of SIS and RRCS in case of scenario 1. But the performance of the methods suffer for scenario 2. However, SIS, being a marginal screening procedure, selects a substantial number of unimportant features while JS selects fewer unimportant features. The performance of JS for scenario 3, which has an auto-regressive correlation structure is comparable to that of SIS and RRCS. Furthermore, for the submodel size calculated based on the model parameters, the JS procedure tends to select the truly relevant features with a higher probability. Overall, we can conclude that the joint screening method outperforms SIS and RRCS. This also shows us that the proposed estimator (3.5) is insensitive towards the covariance structure of the random effects.

Table 3.1 Screening Accuracy Results Based on 100 Independent Simulations

Model	ρ	Method	$(p, N) = (2000, 200)$			$(p, N) = (2000, 400)$		
			Median	IQR	$\mathbb{P}_{\text{all}}$	Median	IQR	$\mathbb{P}_{\text{all}}$
S1	0.3	JS	5	0	1	5	0	1
		SIS	2000	0	0	2000	0	0
		RRCS	2000	0	0	2000	0	0
	0.6	JS	6	4	0.97	5	0	1
		SIS	2000	0	0	2000	0	0
		RRCS	2000	0	0	2000	0	0
	0.9	JS	50	148.5	0.43	5	2	0.99
		SIS	2000	0	0	2000	0	0
		RRCS	2000	0	0	2000	0	0
S2	0.3	JS	347.5	414	0.01	41.5	57.75	0.68
		SIS	1986	61	0	2000	3	0
		RRCS	1980.5	86.5	0	1999.5	5	0
	0.6	JS	687	683.5	0	157.5	217.5	0.18
		SIS	1959.5	190.75	0	1998	28	0
		RRCS	1913	264	0	1995	35.25	0
	0.9	JS	1688.5	441.75	0	1244.5	689.75	0
		SIS	1919.5	217	0	1987	54.5	0
		RRCS	1909	240.75	0	1976.5	105.5	0
S3	0.3	JS	5	1	0.99	5	0	1
		SIS	5	1	0.96	5	0	1
		RRCS	5	3	0.92	5	0	1
	0.6	JS	7	12.75	0.86	5	0	1
		SIS	60.5	214.25	0.43	8.5	30.25	0.85
		RRCS	77	280.25	0.38	14.5	46.25	0.79
	0.9	JS	8	37	0.74	638	948	0.12
		SIS	5	0	1	5	0	1
		RRCS	5	0	1	5	0	1

3.4.2 Screening consistency

Theorem 3.2 states that the proposed joint screening procedure manages to separate the important features from the unimportant ones with overwhelming probability, thus guaranteeing its effectiveness. Here we test the verity of this claim. We consider all three scenarios from previous simulations with $p = 1000, \rho = 0.5$. We vary N from 50 to 500 with an increment of 50 and replicate this 50 times for each scenario. We are trying to find out whether $P\left(\min_{j\in\mathcal{M}_*}\left|\widehat{\beta}_j\right| > \max_{j\notin\mathcal{M}_*}\left|\widehat{\beta}_j\right|\right)$ increases with sample size. Figure 3.1 shows the plot of the probability against sample size for the scenarios.

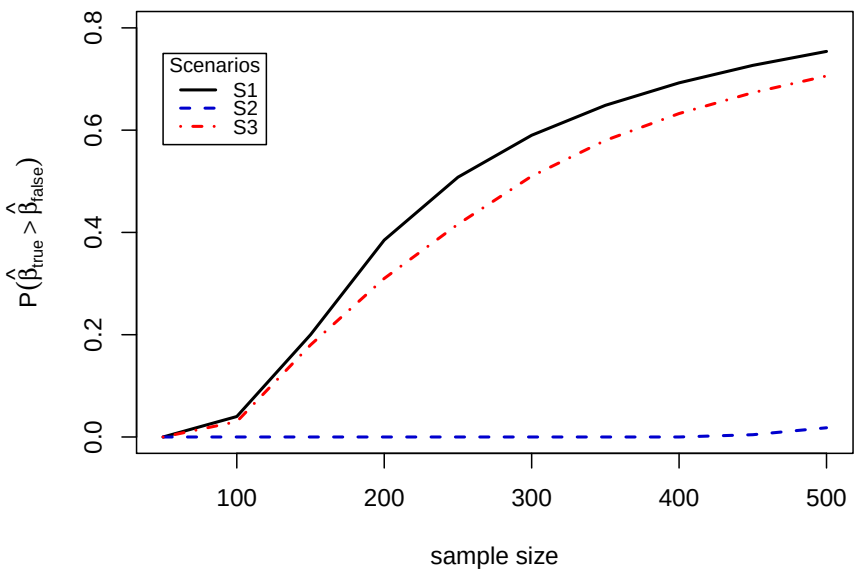


Figure 3.1 Plot showing $P\left(\min_{j\in\mathcal{M}_*}\left|\widehat{\beta}_j\right| > \max_{j\notin\mathcal{M}_*}\left|\widehat{\beta}_j\right|\right)$ versus sample size.

We can see an increasing trend of the selection probability with increase in sample size except for scenario 2. This is expected because of the small effect sizes. Nevertheless, we can conclude that the proposed screening procedure correctly identifies important features with probability tending to one as sample size increases. Having explored the joint screening properties of the JS estimator we proceed towards applying it to a real dataset.

3.5 Application to a Real Dataset

The Genetics Analysis Workshop 20 (GAW20) provided a unique opportunity for us to analyze the real data from the Genetics and Lipid Lowering Drugs and Diet Network (GOLDN) study, and the simulated data based upon it as well. The GOLDN study was funded by NIH R01 HL104135 and HL091357 (Arnett). We are thankful to GOLDN families for their contribution. We also acknowledge the GAW grant, R01 GM031575. The GAW20 data consists of cytosine-guanine dinucleotides (CpGs) variables, whose sizes are much larger than the number of subjects. Second, the subjects are not independent. Instead, the subjects are correlated within families.

We apply the proposed JS procedure to the representative simulated dataset provided by GAW20. In the representative dataset, there were 717 subjects in pedigrees, and subjects already on any lipid-lowering medication were taken off drug for a "washout period". At visit 1 (after the washout), subjects were measured after an overnight fast with a standard lipid profile. The next day, they returned to clinic, again fasting, for a second, repeat lipid profile. All subjects were then given the genomethate drug for a 3-week treatment period, after which they returned to the clinic for two consecutive days of lipid profiling (visits 3 and 4, both with overnight fasting), to assess the response to treatment. We considered the difference in the TG level between visit 4 and visit 2 as outcome variable. There are $n = 680$ subjects with the observed outcome.

Accordingly, we also consider the difference of the CpGs between visit 4 and visit 2 as the predictors, since both the TG level and the CpG value change as time goes by. That is, we consider

$$\begin{aligned} Y &= TGL_4 - TGL_2, \\ X_j &= CpG_4 - CpG_2, j = 1, \cdots, p. \end{aligned} \tag{3.7}$$

Side-by-side boxplots of the outcome variable (the difference in TGL between visit 4 and visit 2) within 13 pedigrees are displayed in Figure 3.2, which demonstrates the heterogeneity of the outcome variable.

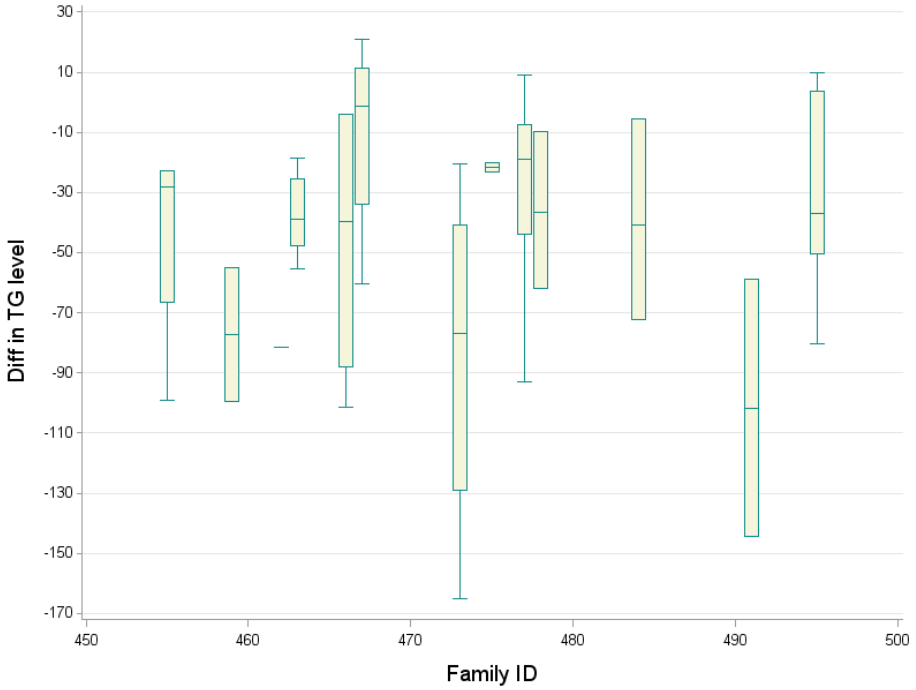


Figure 3.2 Boxplots of TGL by GPEDID.

We compute the JS estimate (3.5), using the GAW20 representative simulated dataset with $n = 680$ observations and $p = 463,995$ CpGs. We specify $d = \lfloor 680/\log(680) \rfloor = 104$ and we obtain the select subset (3.4). We observe from Figure 3.3 that among the 10 truly significant CpGs used in generating the simulated data, *cg00001261*, *cg00045910*, *cg12598270*, *cg00000363*, *cg00703276*, and *cg11736230* passed the screening.

We can conclude that the proposed procedure is computationally efficient and application to the GAW20 data shows that the proposed procedure performs well.

3.6 Discussion

Mixed models are a useful tool for analyzing family data. But when the dimension of the genetic variables is ultra-high, it is computationally difficult to fit mixed

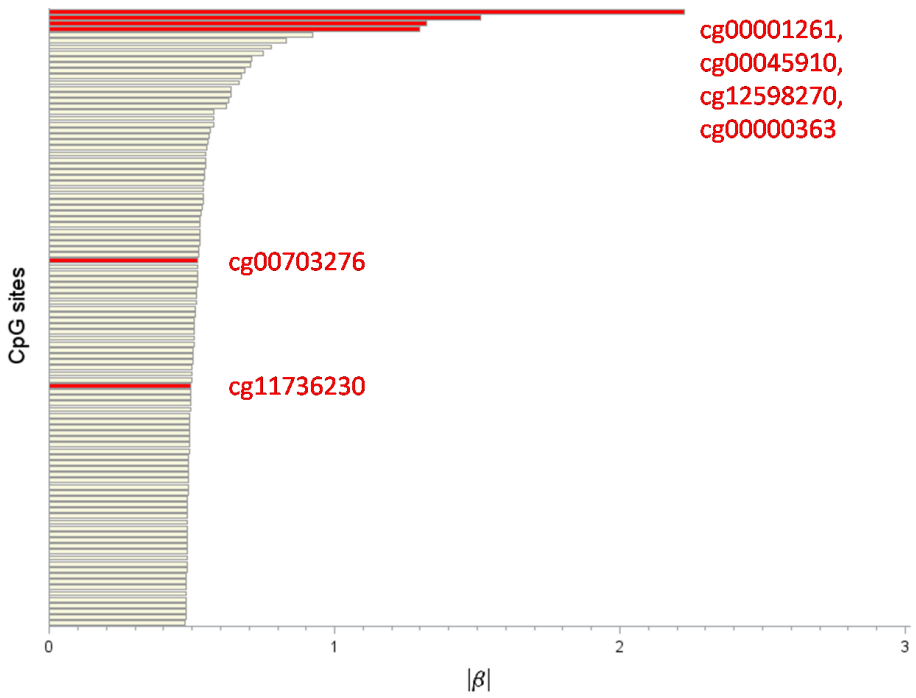


Figure 3.3 $\widehat{\boldsymbol{\beta}}_{JS}$ estimates from the joint screening procedure under model (3.7).

models, and the results from any fitted mixed model will be unstable. To overcome this problem, we can consider a joint screening strategy, which performs dimension reduction and renders the remaining data manageable for further analysis.

While marginal screening procedures fit a mixed model for one feature at a time, the proposed joint screening procedure considers all the features simultaneously. Since high-dimensional data tend to have correlated predictors, marginal screening procedures may select unimportant variables that have a high degree of association to important predictors. Likewise, these procedures may fail to select truly important variables which are jointly correlated but have no marginal association to the response. The proposed joint screening procedure is efficient in detecting both marginally and jointly significant variables. It also retains the desired sure screening properties.

We have performed extensive simulation studies to confirm that the JS estimator for mixed models performs well under different scenarios and is very competitive. It is also evident from the results that the proposed method is insensitive towards the

variance-covariance structure of the random effects, which is our main goal. We have applied the joint screening method to the GAW20 data where we performed screening using the outcome variables as defined by equation (3.7) and selected a subset of 104 genetic variables. Since, the TG level values are skewed, it is advisable to do a log-transformation so that normality assumption is not violated. Contrary to this fact, the JS screening procedure performs well under non-normality of outcome variable. We have shown that screening using equation (3.7) performs well as 6 out of the 10 truly significant variables pass the screening.

Using the screened features for a more refined second stage analysis would mean that the same data is used for screening and testing. The reader should be cautioned that it may inflate the family-wise error (see Van Steen et al. [45]). If the dataset is large, we could divide the data into two halves, one for screening and one for testing. The impact of this two-stage strategy on the family-wise error is not investigated here.

APPENDIX A

LARGE SAMPLE PROPERTIES OF LINEAR WPSVM

A.1 Consistency

We give a proof of theorem 2.3 here.

Proof. We refer to the following theorem:

Theorem A.1 (Newey and McFadden [37]). *If there is a function $Q_0(\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that*

*(***i***) $Q_0(\boldsymbol{\theta})$ is uniquely minimized at $\boldsymbol{\theta}_0$;*

*(***ii***) $\boldsymbol{\Theta}$ is compact;*

*(***iii***) $Q_0(\boldsymbol{\theta})$ is continuous;*

*(***iv***) $\widehat{Q}_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$ i.e., $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\widehat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xrightarrow{P} 0$,*

then $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.

Looking at the objective function (2.7), we observe that the first quadratic term of $\Lambda_\pi(\boldsymbol{\theta})$ is strictly convex, since $\boldsymbol{\Sigma}$ is positive definite and $(a+b)_+ \leq a_+ + b_+$, $\forall a, b \in \mathbb{R}$. Thus, $\Lambda_\pi(\boldsymbol{\theta})$ is strictly convex and has a unique minimizer, $\boldsymbol{\theta}_0$. Similarly, the sample version of the objective function given by (2.8) is convex too by the same logic. Since, $\widehat{\boldsymbol{\Sigma}}_n \xrightarrow{P} \boldsymbol{\Sigma}$ and using theorem 2.2 we have that $\widehat{\Lambda}_{n,\pi}(\boldsymbol{\theta})$ converges to $\Lambda_\pi(\boldsymbol{\theta})$ pointwise. Now we have the following lemma:

Lemma A.1 (Pollard [38]). *Suppose $A_n(s)$ is a sequence of convex random functions defined on an open convex set $\mathscr{S} \in \mathbb{R}^p$, which converges in probability to some $A(s)$, for each s. Then $\sup_{s\in K} |A_n(s) - A(s)| \xrightarrow{P} 0$, for each compact subset K of $\mathscr{S}$.*

By the above lemma, pointwise convergence $\implies$ uniform convergence. Since, all 4 conditions of theorem A.1 hold, $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$. □

A.2 Bahadur Representation of Linear WPSVM Solution

We give a proof of theorem 2.4 here.

Proof. Let $m_\pi(\boldsymbol{\theta}, \boldsymbol{Z}) = \boldsymbol{\theta}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{\theta} + \lambda \pi(Y)[1 - Y\boldsymbol{\theta}^\top \tilde{\boldsymbol{X}}]_+$, where $\tilde{\boldsymbol{\Sigma}} = diag(0, \boldsymbol{\Sigma})$. From equation (2.7) we can see that $\Lambda_\pi(\boldsymbol{\theta}) = \mathbb{E}(m_\pi(\boldsymbol{\theta}, \boldsymbol{Z}))$. The proof of the theorem depends on the following three claims, the reason being stated later.

(a) $m_\pi(\boldsymbol{\theta}, \boldsymbol{Z})$ satisfies the Lipschitz condition with respect to $\boldsymbol{\theta}$. That is, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ there exists an integrable function $Q(\boldsymbol{Z})$ such that

$$|m_\pi(\boldsymbol{\theta}_1, \boldsymbol{Z}) - m_\pi(\boldsymbol{\theta}_2, \boldsymbol{Z})| \leq Q(\boldsymbol{Z}) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \tag{A.1}$$

Note that the first term $\boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta}$ is a continuous and deterministic function with respect to $\boldsymbol{\theta}$. Thus, it is enough to check the Lipschitz condition of the second term. Let $\tilde{m}_\pi(\boldsymbol{\theta}, \boldsymbol{Z}) = \pi(Y)[1 - Y\boldsymbol{\theta}^\top \tilde{\boldsymbol{X}}]_+$. Then for any $\boldsymbol{\theta}_i = (\alpha_i, \boldsymbol{\beta}_i) \in \boldsymbol{\Theta}, i = 1, 2$, we have

$$\begin{aligned}
\tilde{m}_\pi(\boldsymbol{\theta}_1, \boldsymbol{Z}) - \tilde{m}_\pi(\boldsymbol{\theta}_2, \boldsymbol{Z}) &= \pi(Y)[1 - Y(\alpha_1 + \boldsymbol{\beta}_1^\top \boldsymbol{X})]_+ - \pi(Y)[1 - Y(\alpha_2 + \boldsymbol{\beta}_2^\top \boldsymbol{X})]_+ \\
&\leq \pi(Y)|(\alpha_2 - \alpha_1 + \boldsymbol{X}^\top(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)|, \\
&\qquad \text{since} |u_+ - v_+| \leq |u - v|, \forall\, u, v \in \mathbb{R} \\
&\leq \pi(Y)(1 + \|\boldsymbol{X}\|^2)^{\frac{1}{2}} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|
\end{aligned}$$

Also, $\mathbb{E}[\pi(Y)(1 + \|\boldsymbol{X}\|^2)^{\frac{1}{2}}] \leq \mathbb{E}[(1 + \|\boldsymbol{X}\|^2)^{\frac{1}{2}}] \leq (1 + \mathbb{E}\|\boldsymbol{X}\|^2)^{\frac{1}{2}} < \infty$ by $(A1)$. Thus, $m_\pi(\boldsymbol{\theta}, \boldsymbol{Z})$ satisfies the Lipschitz condition.

(b) For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}, m_\pi(\boldsymbol{\theta}, \boldsymbol{Z})$ is differentiable for almost every $\boldsymbol{Z}$.

The first term is differentiable and once again it is enough to show that $\tilde{m}_\pi(\boldsymbol{\theta}, \boldsymbol{Z})$ is almost surely differentiable. Let $N_{\boldsymbol{\theta}}(\tilde{m}_\pi) = \{\boldsymbol{z} : \tilde{m}_\pi(., \boldsymbol{z})$ is not differentiable at $\boldsymbol{\theta}\}$, then $P[\boldsymbol{Z} \in N_{\boldsymbol{\theta}}(\tilde{m}_\pi)] = \sum_{y=-1,1} P(Y = y) P(\boldsymbol{X} \in \{\boldsymbol{x} : \alpha + \boldsymbol{\beta}^\top \boldsymbol{x} = y\} | Y = y) = 0$ by $(A2)$. Thus, $m_\pi(\boldsymbol{\theta}, \boldsymbol{Z})$ is almost surely differentiable with respect to any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

(c) $\Lambda_\pi(\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$ with Hessian matrix $\mathbf{H}_{\boldsymbol{\theta}}$ given by equation (2.12).

We use the following lemmas:

Lemma A.2 (Lemma 2 from Li et al. [28]). *Suppose that $m : \boldsymbol{\Theta} \times \boldsymbol{\Omega}_{\boldsymbol{Z}} \to \mathbb{R}$ satisfies the following conditions*

*(**i**) (almost surely differentiable) for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $P[\boldsymbol{Z} \in N_{\boldsymbol{\theta}}(m)] = 0$;*

*(**ii**) (Lipschitz condition) there is an integrable function $c(\boldsymbol{z})$, independent of $\boldsymbol{\theta}$, such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, $|m(\boldsymbol{\theta}_2, \boldsymbol{z}) - m(\boldsymbol{\theta}_1, \boldsymbol{z})| \leq c(\boldsymbol{z})\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|$.*

Then $\mathbf{D}_{\boldsymbol{\theta}}(m(\boldsymbol{\theta}, \boldsymbol{Z}))$ is integrable, $\mathbb{E}(m(\boldsymbol{\theta}, \boldsymbol{Z}))$ is differentiable, and $\mathbf{D}_{\boldsymbol{\theta}}\mathbb{E}(m(\boldsymbol{\theta}, \boldsymbol{Z})) = \mathbb{E}(\mathbf{D}_{\boldsymbol{\theta}} m(\boldsymbol{\theta}, \boldsymbol{Z}))$.

Lemma A.3 (Lemma 3 from Li et al. [28]). *Suppose that U and V are linearly dependent random variables and $\boldsymbol{h}(u)$ is a measurable $\mathbb{R}^k$-valued function, and*

*(**i**) the joint distribution of (U, V) is dominated by the Lebesgue measure;*

*(**ii**) for each v, the function $u \mapsto \boldsymbol{h}(u, v) f_{U|V}(u|v)$ is continuous, where $f_{U|V}$ denotes the conditional probability density function of U given V;*

*(**iii**) for each component $h_i(u, v)$ of $\boldsymbol{h}(u, v)$, there is a function $c_i(v) \geq 0$ such that $|h_i(u, v)| f_{U|V}(u|v) \leq c_i(v)$, and $\mathbb{E}(c_i(V)) < \infty$.*

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

$$D_{\epsilon=0}\mathbb{E}[\boldsymbol{h}(U, V)\mathbb{1}(U + \epsilon V < a + \epsilon\eta)] = f_U(a)\mathbb{E}[(\eta - V)\boldsymbol{h}(U, V)|U = a]. \quad \text{(A.2)}$$

Lemma A.4 (Lemma 4 from Li et al. [28]). *Suppose that U and V are linearly dependent random variables and $\boldsymbol{h}(u)$ is a measurable $\mathbb{R}^k$-valued function, and*

*(**i**) the distribution of U is dominated by the Lebesgue measure;*

*(**ii**) $\boldsymbol{h}(u)f_U(u)$ is continuous.*

Then, for any constant a, the function $\epsilon \mapsto \mathbb{E}[\boldsymbol{h}(U)\mathbb{1}(U + \epsilon V < a + \epsilon\eta)]$ is differentiable at $\epsilon = 0$ with derivative given by equation (A.2)

Already having established (a) and (b) above we apply lemma A.2 and show

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}\Lambda_\pi(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}(m_\pi(\boldsymbol{\theta}, \boldsymbol{Z})) \\ &= \mathbb{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}}m_\pi(\boldsymbol{\theta}, \boldsymbol{Z})\right) \\ &= 2\tilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \lambda\mathbb{E}[\pi(Y)\tilde{\boldsymbol{X}}Y\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}}Y < 1\}],\end{aligned}$$

where $\tilde{\boldsymbol{\Sigma}} = diag(0, \boldsymbol{\Sigma})$. Therefore we have the second derivative given by

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Lambda_\pi(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\left(2\tilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \lambda\mathbb{E}[\pi(Y)\tilde{\boldsymbol{X}}Y\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}}Y < 1\}]\right) \\ &= 2\tilde{\boldsymbol{\Sigma}} - \lambda\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}[\pi(Y)\tilde{\boldsymbol{X}}Y\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}}Y < 1\}] \\ &= 2\tilde{\boldsymbol{\Sigma}} - \lambda\sum_{y=-1,1}P(Y = y)\pi(y)\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}[\tilde{\boldsymbol{X}}y\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}}y < 1\}|Y = y]\end{aligned} \tag{A.3}$$

If we let $A_y(\boldsymbol{\theta}) = \mathbb{E}[\tilde{\boldsymbol{X}}y\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}}y < 1\}|Y = y]$, then we only need to prove the differentiability of $A_y(\boldsymbol{\theta})$. First for $Y = +1$,

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}A_{+1}(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}[\tilde{\boldsymbol{X}}\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}} < 1\}] \\ &= -f_{\boldsymbol{\beta}^\top\boldsymbol{X}|Y}(1 - \alpha|1)\mathbb{E}[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top|\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}} = 1]\end{aligned} \tag{A.4}$$

by applying lemmas A.3 and A.4 and under the assumptions $(A2) - (A5)$. Similarly, for $Y = -1$,

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}A_{-1}(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}[\tilde{\boldsymbol{X}}\mathbb{1}\{-\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}} < 1\}] \\ &= -f_{\boldsymbol{\beta}^\top\boldsymbol{X}|Y}(-1 - \alpha| - 1)\mathbb{E}[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top|\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}} = -1]\end{aligned} \tag{A.5}$$

We plug (A.4) and (A.5) into (A.3) and get the second derivative of $\Lambda_\pi(\boldsymbol{\theta})$ denoted by $\mathbf{H}_{\boldsymbol{\theta}}$ in equation (2.12).

Under the consistency established in theorem 2.3, equation (2.10) is a consequence of theorem 5.23 of van der Vaart [44], given $(a)-(c)$ are true. □

A.3 Asymptotic Normality of the Candidate Matrix

We give a proof of theorem 2.5 here.

Proof. Let $\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}) = n^{-1}\sum_{i=1}^{n}\mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}_i)$, sample average of $\mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z})$. From equation (2.13) we have,

$$
\begin{aligned}
&vec(\widehat{\mathbf{M}}_n - \mathbf{M}_0)\\
&= \sum_{h=1}^{H}\widehat{\boldsymbol{\beta}}_{n,h}\otimes\widehat{\boldsymbol{\beta}}_{n,h} - \sum_{h=1}^{H}\boldsymbol{\beta}_{0,h}\otimes\boldsymbol{\beta}_{0,h}\\
&= \sum_{h=1}^{H}\Big(\boldsymbol{\beta}_{0,h} - \bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}) + o_p(n^{-\frac{1}{2}})\Big)\otimes\Big(\boldsymbol{\beta}_{0,h} - \bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}) + o_p(n^{-\frac{1}{2}})\Big)\\
&\qquad - \sum_{h=1}^{H}\boldsymbol{\beta}_{0,h}\otimes\boldsymbol{\beta}_{0,h}\\
&= -\sum_{h=1}^{H}\big(\boldsymbol{\beta}_{0,h}\otimes\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}) + \bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z})\otimes\boldsymbol{\beta}_{0,h}\big) + \sum_{h=1}^{H}\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z})\otimes\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z})\\
&\qquad + o_p(n^{-\frac{1}{2}})\\
&= -\sum_{h=1}^{H}\big(\boldsymbol{\beta}_{0,h}\otimes\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}) + \bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z})\otimes\boldsymbol{\beta}_{0,h}\big) + o_p(n^{-\frac{1}{2}})
\end{aligned}
$$

We use the following properties of the matrix $\mathbf{T}$.

$$
\begin{aligned}
&\mathbf{T}_{i_1,i_2} = \mathbf{T}_{i_2,i_1}^{\top}\\
&\mathbf{A}\otimes\mathbf{B} = \mathbf{T}_{i_1,i_3}(\mathbf{B}\otimes\mathbf{A})\mathbf{T}_{i_4,i_2},\ \text{for } \mathbf{A}\in\mathbb{R}^{i_1\times i_2} \text{ and } \mathbf{B}\in\mathbb{R}^{i_3\times i_4}.
\end{aligned}
$$

Thus,

$$\sqrt{n}\{vec(\widehat{\mathbf{M}}_n) - vec(\mathbf{M}_0)\} = -n^{-\frac{1}{2}} \sum_{i=1}^{n} \left((\mathbf{I}_{p^2} + \mathbf{T}_{p,p}) \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h} \otimes \bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}_i) \right) + o_p(1)$$

and the result follows from the Central Limit Theorem. □

APPENDIX B

CONSISTENCY OF STRUCTURAL DIMENSIONALITY

Proof. We have $\hat{k} = \underset{k\in\{1,\ldots,p\}}{\operatorname{argmax}} G_n\left(k;\eta,\widehat{\mathbf{M}}_n\right)$, where $\widehat{\mathbf{M}}_n$ is the candidate matrix of the linear WPSVM as defined in equation (2.9). Now,

$$\begin{aligned} & G_n\left(\hat{k};\eta,\widehat{\mathbf{M}}_n\right) - G_n\left(k;\eta,\widehat{\mathbf{M}}_n\right) \\ &= \sum_{j=1}^{\hat{k}} \hat{\nu}_j - \sum_{j=1}^{k} \hat{\nu}_j - \eta\,\frac{\hat{k}\log n}{\sqrt{n}}\,\nu_1 + \eta\,\frac{k\log n}{\sqrt{n}}\,\nu_1 \\ &= \sum_{j=1}^{\hat{k}} \nu_j - \sum_{j=1}^{k} \nu_j - \eta\,\frac{(\hat{k}-k)\log n}{\sqrt{n}}\,\nu_1 + O_p\left(n^{-\frac{1}{2}}\right), \end{aligned} \tag{B.1}$$

where ν_i and $\hat{\nu}_i$ are the j-th leading eigenvalues of $\mathbf{M}_0$ and $\widehat{\mathbf{M}}_n$ respectively. The last part of equation (B.1) is due to the fact that

$$\sum_{j=1}^{d} \hat{\nu}_i = \sum_{j=1}^{d} \nu_i + O_p\left(n^{-\frac{1}{2}}\right), \forall\, d = 1,\ldots,p,$$

which can be derived as a consequence of theorem 2.5 and continuous mapping theorem.

Let us suppose $\hat{k} \neq k$. Thus, we have the following two cases:

Case 1: $\hat{k} < k$: With increase in sample size, we can see that the equation (B.1) converges to a negative value, since $rank(\mathbf{M}_0) = k$ and $\sum_{j=1}^{\hat{k}} \nu_j - \sum_{j=1}^{k} \nu_j < 0$. This leads to a contradiction.

Case 2: $\hat{k} > k$: Similarly, consider a large n and we have

$$G_n\left(\hat{k};\eta,\widehat{\mathbf{M}}_n\right) - G_n\left(k;\eta,\widehat{\mathbf{M}}_n\right) = -\eta\,\frac{(\hat{k}-k)\log n}{\sqrt{n}}\,\nu_1 + O_p\left(n^{-\frac{1}{2}}\right) < 0.$$

This leads to a contradiction too.

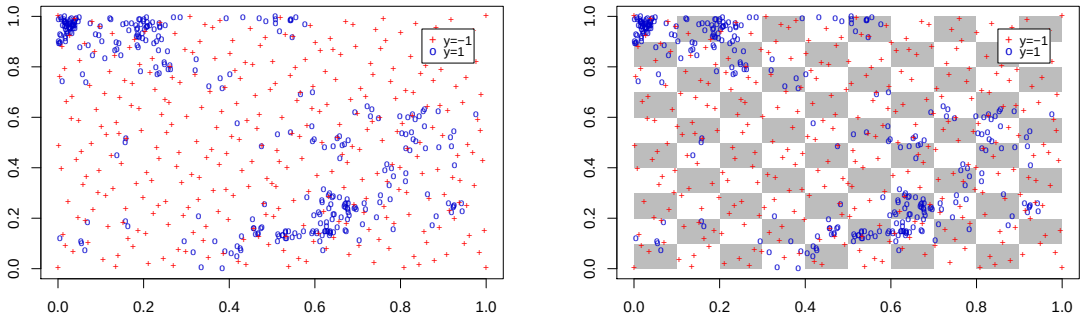
The desired result follows. □

APPENDIX C

CROSS-VALIDATION USING CHESS BOARD METHOD

Due to the inherent correlated structure of spatial data, splitting the data randomly for model validation is not recommended. In this section we explain a chess board method to split a spatial data into training and test. The two-fold CVBIC is explained in Figure C.1.

The figure to the left shows the location of the data and dummy points and the figure to the right has the chess board superimposed onto it. First we consider the points falling on a grey square as training set, the rest as test data and vice-versa. We perform model validation using both the sets and choose an optimal η.



(a) (b)

Figure C.1 Location of (a) data and dummies and (b) data and dummies with a chess board overlayed.

APPENDIX D

SURE SCREENING PROPERTIES OF THE JS ESTIMATOR FOR MIXED MODEL

We follow the framework of Fan and Lv in their SIS paper [18], with some modifications. Recall that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$. The proposed JS estimator is given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{JS} &= \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \boldsymbol{Y} \\ &= \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \mathbf{X}\boldsymbol{\beta} + \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} (\boldsymbol{\alpha} + \boldsymbol{\epsilon}) \\ &\coloneqq \boldsymbol{\xi} + \boldsymbol{\phi}.\end{aligned}$$

D.1 Property of $\boldsymbol{\xi}$

We consider the singular value decomposition of $\mathbf{Z}$ as $\mathbf{Z} = PDR^\top$, where $P \in \mathcal{O}(n), R \in P_{n,p}$, and D is an $n \times n$ diagonal matrix. Here $\mathcal{O}(n)$ is the set of all $n \times n$ orthogonal matrices and $P_{n,p} = \{R \in \mathbb{R}^{p\times n} : R^\top R = \mathbf{I}_n\}$. Thus we have $\mathbf{X} = \mathbf{Z}\Sigma^{1/2} = PDR^\top\Sigma^{1/2}$. The projection matrix can be written as

$$\begin{aligned}\mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \mathbf{X} &= \Sigma^{1/2} PDR^\top (PDR^\top \Sigma RDP^\top)^{-1} PDR^\top \Sigma^{1/2} \\ &= \Sigma^{1/2} P (P^\top \Sigma P)^{-1} P^\top \Sigma^{1/2} \\ &\coloneqq HH^\top,\end{aligned}$$

where $H = \Sigma^{1/2} P (P^\top \Sigma P)^{-1/2}$ satisfying $H^\top H = \mathbf{I}_n$. Thus $\boldsymbol{\xi} = HH^\top \boldsymbol{\beta}$.

Let $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ denote the i-th natural base in the p dimension space. Following the proofs of lemmas 4 and 5 in Wang and Leng [51], we derive the following lemmas.

Lemma D.1. *Under assumptions* $(B1)-(B3)$*, for any* $D>0$ *and for any fixed vector* $\boldsymbol{v}$ *with* $\|\boldsymbol{v}\|=1$*, there exists constants* d_1', d_2' *with* $0<d_1'<1<d_2'$ *such that*

$$P\left(\boldsymbol{v}^\top HH^\top \boldsymbol{v} < \frac{d_1' n^{1-\tau}}{p} \text{ or } \boldsymbol{v}^\top HH^\top \boldsymbol{v} > \frac{d_2' n^{1+\tau}}{p}\right) < 4\exp(-Dn).$$

In particular for $\boldsymbol{v}=\boldsymbol{\beta}_*$*, whose norm is not* 1 *though, a similar inequality holds for one side with* $d_2'>1$ *(same as previous* d_2'*; if not, the maximum of the two is used in both the inequalities) as*

$$P\left(\boldsymbol{\beta}_*^\top HH^\top \boldsymbol{\beta}_* > \frac{d_2' n^{1+\tau}}{p}\right) < 2\exp(-Dn).$$

Lemma D.2. *Under assumptions* $(B1)-(B3)$*, for any* $D>0$*, there exists constants* $d_3', d_4'>0$ *such that for any* $i\in\mathcal{M}_*$*,*

$$P\left(\left|\boldsymbol{e}_i HH^\top \boldsymbol{\beta}_*\right| < \frac{d_3' n^{1-\tau-\kappa}}{p}\right) \le O\left\{\exp\left(\frac{-Dn^{1-5\tau-2\kappa-\nu}}{2\log n}\right)\right\},$$

and for any $i\notin\mathcal{M}_*$*,*

$$P\left(\left|\boldsymbol{e}_i HH^\top \boldsymbol{\beta}_*\right| > \frac{d_4' n^{1-\tau-\kappa}}{p\sqrt{\log n}}\right) \le O\left\{\exp\left(\frac{-Dn^{1-5\tau-2\kappa-\nu}}{2\log n}\right)\right\}.$$

Applying lemmas D.1 and D.2 to all $i\in\mathcal{M}_*$, we have

$$P\left(\min_{i\in\mathcal{M}_*}|\xi_i| < \frac{d_3' n^{1-\tau-\kappa}}{p}\right) = O\left\{s\exp\left(\frac{-Dn^{1-5\tau-2\kappa-\nu}}{2\log n}\right)\right\}. \tag{D.1}$$

D.2 Property of $\boldsymbol{\phi}$

In order to derive the property of $\boldsymbol{\phi}$ we follow the proof of lemma 6 from Wang and Leng [51], which we state here for convenience.

Lemma D.3. *Under assumptions* $(B1)-(B3)$*, we have for any* $i\in\{1,\ldots,n\}$*,*

$$P\left(|\eta_i| > \frac{\sigma_e\sqrt{D_1 d_1 d_2' d_4}\, n^{1-\kappa-\tau}}{p\sqrt{\log n}}\right) < \exp\left\{1-q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right\} + 3\exp(-D_1 n),$$

where d_2' is the same as defined in lemma D.1; if not, the maximum of the two is used.

Recall the random variable $\phi_i = \boldsymbol{e}_i^\top \boldsymbol{\phi} = \boldsymbol{e}_i^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \boldsymbol{\phi}$. Let us define

$$\boldsymbol{a} = \frac{\boldsymbol{e}_i^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1}}{\left\| \boldsymbol{e}_i^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \right\|_2},$$

then $\boldsymbol{a}$ is free of $\boldsymbol{\phi}$ and

$$\phi_i = \left\| \boldsymbol{e}_i^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \right\|_2 V^{1/2} w,$$

where w is a standardized random variable such that $w = \boldsymbol{a}^\top \boldsymbol{\alpha}/V^{1/2}$ and V is given by equation 1.4.

We investigate the bound of the squared norm as follows.

$$\begin{aligned}
\left\| \boldsymbol{e}_i^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \right\|_2^2 &= \boldsymbol{e}_i^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-2} \mathbf{X}\boldsymbol{e}_i \\
&= \boldsymbol{e}_i^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top\right)^{-1/2} \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \left(\mathbf{X}\mathbf{X}^\top\right)^{-1/2} \mathbf{X}\boldsymbol{e}_i \\
&\leq \lambda_{\max} \left\{ \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \right\} \boldsymbol{e}_i^\top H H^\top \boldsymbol{e}_i \\
&= \lambda_{\max} \left\{ \left(\mathbf{Z}\Sigma\mathbf{Z}^\top\right)^{-1} \right\} \boldsymbol{e}_i^\top H H^\top \boldsymbol{e}_i. \qquad \text{(D.2)}
\end{aligned}$$

We investigate the first term in equation (D.2) further.

$$\begin{aligned}
\lambda_{\max} \left\{ \left(\mathbf{Z}\Sigma\mathbf{Z}^\top\right)^{-1} \right\} &= \left\{ \lambda_{\min} \left(\mathbf{Z}\Sigma\mathbf{Z}^\top\right) \right\}^{-1} \\
&\leq \left\{ \lambda_{\min} \left(\mathbf{Z}\mathbf{Z}^\top\right) \right\}^{-1} \left\{ \lambda_{\min}(\Sigma) \right\}^{-1} \\
&= p^{-1} \left\{ \lambda_{\min} \left(p^{-1}\mathbf{Z}\mathbf{Z}^\top\right) \right\}^{-1} \left\{ \lambda_{\min}(\Sigma) \right\}^{-1}.
\end{aligned}$$

Now since the trace of Σ is p, $\lambda_{\max}(\Sigma) \geq 1$. By assumption $(B3)$ we have,

$$\lambda_{\min}(\Sigma) \geq \frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)} > \frac{1}{d_4 n^\tau}.$$

Thus we have

$$\lambda_{\max}\left\{\left(\mathbf{Z}\Sigma\mathbf{Z}^\top\right)^{-1}\right\} < \frac{d_4 n^\tau}{p}\left\{\lambda_{\min}\left(p^{-1}\mathbf{Z}\mathbf{Z}^\top\right)\right\}^{-1}. \tag{D.3}$$

According to assumption $(B1)$

$$P\left(\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^\top) > d_1 \text{ or } \lambda_{\min}(p^{-1}\mathbf{Z}\mathbf{Z}^\top) < d_1^{-1}\right) < \exp(-D_1 n),$$

which along with equation (D.3) gives us

$$\begin{aligned} P\left(\lambda_{\max}\left\{\left(\mathbf{Z}\Sigma\mathbf{Z}^\top\right)^{-1}\right\} > \frac{d_1 d_4 n^\tau}{p}\right) &< P\left(\frac{d_4 n^\tau}{p}\left\{\lambda_{\min}\left(p^{-1}\mathbf{Z}\mathbf{Z}^\top\right)\right\}^{-1} > \frac{d_1 d_4 n^\tau}{p}\right) \\ &= P\left(\lambda_{\min}\left(p^{-1}\mathbf{Z}\mathbf{Z}^\top\right) < d_1\right) \\ &< \exp(-D_1 n). \end{aligned} \tag{D.4}$$

Combining equation (D.4) with lemma D.1 and using the same $D_1 > 0$ we have

$$P\left(\left\|\boldsymbol{e}_i^\top \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\right\|_2^2 > \frac{d_1 d_2' d_4 n^{1+2\tau}}{p^2}\right) < 3\exp(-D_1 n). \tag{D.5}$$

For w, according to q-exponential tail definition,

$$P\left(\left|\sum_{i=1}^{n}\frac{a_i\phi_i}{V^{1/2}}\right| > t\right) \leq \exp(1-q(t)).$$

If we choose $t = \frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}$ we have,

$$P\left(|w| > \frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right) < \exp\left\{1 - q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right\}.$$

Combining the above with equation (D.5) and taking the union bound, we have

$$\begin{aligned} &P\left(|\phi_i| > \frac{V^{1/2}\sqrt{D_1 d_1 d_2' d_4}n^{1-\kappa-\tau}}{p\sqrt{\log n}}\right) \\ &\quad < \exp\left\{1 - q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right\} + 3\exp(-D_1 n). \end{aligned}$$

Applying this to all $i \in \mathcal{M}_*$, we have

$$P\left(\max_{i\in\mathcal{M}_*}|\phi_i| > \frac{V^{1/2}\sqrt{D_1 d_1 d_2' d_4}n^{1-\kappa-\tau}}{p\sqrt{\log n}}\right)$$
$$= s\exp\left\{1 - q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right\} + 3s\exp(-D_1 n). \tag{D.6}$$

D.3 Proof of the Theorems

We are now ready to prove the theorems.

Proof of theorem 3.1: Since $s = d_3 n^\nu$, if M is large enough, we have from equation (D.1)

$$P\left(\min_{i\in\mathcal{M}_*}|\xi_i| < \frac{d_3' n^{1-\tau-\kappa}}{p}\right) = O\left\{\exp\left(\frac{-Dn^{1-5\tau-2\kappa-\nu}}{\log n}\right)\right\}.$$

Now if we choose γ_n such that

$$\frac{p\gamma_n}{n^{1-\tau-\kappa}} \to 0, \text{ and } \frac{p\gamma_n\sqrt{\log n}}{n^{1-\tau-\kappa}} \to \infty,$$

we have

$$\begin{aligned}
P\left(\min_{i\in\mathcal{M}_*}|\widehat{\beta}_i| < \gamma_n\right) &= P\left(\min_{i\in\mathcal{M}_*}|\xi_i + \phi_i| < \gamma_n\right) \\
&\le P\left(\min_{i\in\mathcal{M}_*}|\xi_i| < \frac{d_3' n^{1-\tau-\kappa}}{p}\right) \\
&+ P\left(\max_{i\in\mathcal{M}_*}|\phi_i| > \frac{V^{1/2}\sqrt{D_1 d_1 d_2' d_4}n^{1-\kappa-\tau}}{p\sqrt{\log n}}\right) \\
&= O\left\{\exp\left(\frac{-Dn^{1-5\tau-2\kappa-\nu}}{2\log n}\right)\right\} + s\exp\left\{1 - q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right\}.
\end{aligned}$$

This completes the proof of theorem 3.1. □

Proof of theorem 3.2: From lemma D.2, for any $i \notin \mathcal{M}_*$, and any $D > 0$, there exists an d_4' such that

$$P\left(\left|\boldsymbol{e}_i^\top H H^\top \boldsymbol{\beta}\right| > \frac{d_4' n^{1-\tau-\kappa}}{p\sqrt{\log n}}\right) \le O\left\{\exp\left(\frac{-Dn^{1-5\tau-2\kappa-\nu}}{2\log n}\right)\right\}.$$

Using Bonferroni's inequality, we have

$$P\left(\min_{i\notin\mathcal{M}_*}|\xi_i| > \frac{d_4' n^{1-\tau-\kappa}}{p\sqrt{\log n}} >\right) < O\left\{p\exp\left(\frac{-Dn^{1-5\tau-2\kappa-\nu}}{2\log n}\right)\right\}.$$

Applying Bonferroni's inequality again to the result for $\boldsymbol{\phi}$, we have

$$\begin{aligned}P&\left(\max_{i\in\mathcal{M}_*}|\phi_i| > \frac{V^{1/2}\sqrt{D_1 d_1 d_2' d_4}\, n^{1-\kappa-\tau}}{p\sqrt{\log n}}\right)\\ &< p\exp\left\{1-q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right\} + 3p\exp(-D_1 n).\end{aligned}$$

Now, if p satisfies

$$\log p = o\left(\min\left\{\frac{n^{1-5\tau-2\kappa-\nu}}{2\log n}, q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\log n}\right)\right\}\right),$$

we have

$$P\left(\min_{i\notin\mathcal{M}_*}|\xi_i| > \frac{d_4' n^{1-\tau-\kappa}}{p\sqrt{\log n}} >\right) < O\left\{\exp\left(\frac{-D_1 n^{1-5\tau-2\kappa-\nu}}{2\log n}\right)\right\}, \text{ and}$$

$$\begin{aligned}P&\left(\max_{i\in\mathcal{M}_*}|\phi_i| > \frac{V^{1/2}\sqrt{D_1 d_1 d_2' d_4}\, n^{1-\kappa-\tau}}{p\sqrt{\log n}}\right)\\ &< O\left\{\exp\left(1-\frac{1}{2}q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right) + 3\exp\left(-\frac{D_1 n}{2}\right)\right\}.\end{aligned}$$

For the same γ_n as chosen in theorem 3.1, we have

$$P\left(\max_{i\notin\mathcal{M}_*}\left|\widehat{\beta}_i\right| > \gamma_n\right) < O\left\{\exp\left(\frac{-D_1 n^{1-5\tau-2\kappa-\nu}}{2\log n}\right) + \exp\left(1-\frac{1}{2}q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right)\right\}.$$

Combining this with theorem 3.1 and the fact that $s < p$, we have

$$\begin{aligned}P&\left(\min_{i\in\mathcal{M}_*}\left|\widehat{\beta}_i\right| > \gamma_n > \max_{i\notin\mathcal{M}_*}\left|\widehat{\beta}_i\right|\right)\\ &= 1 - O\left\{\exp\left(\frac{-D_1 n^{1-5\tau-2\kappa-\nu}}{2\log n}\right) + \exp\left(1-\frac{1}{2}q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right)\right\}.\end{aligned}$$

Now, if we choose a submodel with size $k \geq s$, we have

$$P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_k\right) = 1 - O\left\{\exp\left(\frac{-D_1 n^{1-5\tau-2\kappa-\nu}}{2\log n}\right) + \exp\left(1 - \frac{1}{2}q\left(\frac{\sqrt{D_1}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right)\right\}.$$

This completes the proof of theorem 3.2. □

BIBLIOGRAPHY

[1] E. Barut, J. Fan, and A. Verhasselt. Conditional sure independence screening. Technical report, Princeton University, Princeton, NJ, 2012.

[2] R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2:107–144, 2005.

[3] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Am. Stat. Assoc.*, 80:580–619, 1985.

[4] H. Cho and P. Fryzlewicz. High-dimensional variable selection via tilting. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, 74:593–622, 2012.

[5] R. Condit. *Tropical Forest Census Plots*. Springer-Verlag, Berlin, Germany, and Georgetown, Texas, 1998.

[6] J. Conway. *A Course in Functional Analysis*. Springer, New York, NY, 2nd edition, 1990.

[7] R. Cook. Graphics for regressions with a binary response. *J. Am. Stat. Assoc.*, 91:983–992, 1996.

[8] R. Cook. Principal hessian directions revisited. *J. Am. Stat. Assoc.*, 93:84–94, 1998.

[9] R. Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York, NY, 1998.

[10] R. Cook. Fisher lecture: Dimension reduction in regression. *Stat. Sci.*, 22:1–26, 2007.

[11] R. Cook and B. Li. Dimension reduction for conditional mean in regression. *Ann. Stat.*, 30:455–474, 2002.

[12] R. Cook and S. Weisberg. Discussion of "sliced inverse regression for dimension reduction". *J. Am. Stat. Assoc.*, 86:28–33, 1991.

[13] N. Cressie. *Statistics for Spatial Data*. Wiley, New York, NY, 2nd edition, 1993.

[14] P. J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Oxford University Press, New York, NY, 2003.

[15] J. Fan and Y. Fan. High-dimensional classification using features annealed independence rules. *Ann. Stat.*, 36:2605–2637, 2008.

[16] J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Am. Stat. Assoc.*, 116:544–557, 2011.

[17] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96:1348–1360, 2001.

[18] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, 70:849–911, 2008.

[19] J. Fan, R. J. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.*, 10:1829–1853, 2009.

[20] J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Stat.*, 6:3567–3604, 2010.

[21] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied Longitudinal Analysis*. John Wiley & Sons, Inc., Hoboken, NJ, 2nd edition, 2011.

[22] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Am. Stat. Assoc.*, 76:817–823, 1981.

[23] Y. Guan and H. Wang. Sufficient dimension reduction for spatial point processes directed by gaussian random fields. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, 72:367–387, 2010.

[24] P. Hall and H. Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Stat.*, 18:533–550, 2009.

[25] A. E. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[26] S. P. Hubbell, R. Condit, and R. B. Foster. Barro Colorado Forest Census Plot Data. `http://ctfs.si.edu/webatlas/datasets/bci` (accessed on 03/26/2019), 2010.

[27] S. P. Hubbell, R. B. Foster, S. T. O'Brien, K. E. Harms, R. Condit, B. Wechsler, S. J. Wright, and S. Loo de Lao. Light gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science*, 283:554–557, 1999.

[28] B. Li, A. Artemiou, and L. Li. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann. Stat.*, 39:3182–3210, 2011.

[29] B. Li and S. Wang. On directional regression for dimension reduction. *J. Am. Stat. Assoc.*, 102:997–1008, 2007.

[30] B. Li, H. Zha, and F. Chiaromonte. Contour regression: a general approach to dimension reduction. *Ann. Stat.*, 33:1580–1616, 2005.

[31] G. Li, H. Peng, J. Zhang, and Zhu L. Robust rank correlation based screening. *Ann. Stat.*, 40:1846–1877, 2012.

[32] K.-C. Li. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.*, 86:316–327, 1991.

[33] K.-C. Li. On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. *J. Am. Stat. Assoc.*, 87:1025–1039, 1992.

[34] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Mach. Learn.*, 46:191–202, 2004.

[35] Y.-J. Meng and Z.-Y. Lin. Strong laws of large numbers for $\tilde{\rho}$-mixing random variables. *J. Math. Anal. Appl.*, 365:711–717, 2010.

[36] J. Moller and R. Waagepetersen. Modern statistics for spatial point processes. *Scand. J. Stat.*, 4:643–684, 2007.

[37] W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics IV*, pages 2113–2245, 1994.

[38] D. Pollard. Aymptotics for least absolute deviation regression estimator. *Econom. Theory*, 7:186–199, 1991.

[39] J. Schelldorfer, P. Bühlmann, and S. van de Geer. Estimation for high-dimensional linear mixed-effects models using l_1-penalization. *Scand. J. Stat.*, 38:197–214, 2010.

[40] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, New York, NY, 2004.

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

[42] D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic Geometry and Its Applications*. Wiley, New York, NY, 1995.

[43] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, 58:267–288, 1996.

[44] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, New York, NY, 1998.

[45] K. Van Steen, M. B. McQueen, A. Herbert, B. Raby, H. Lyon, D. L. Demeo, A. Murphy, J. Su, S. Datta, C. Rosenow, M. Christman, E. K. Silverman, N. M. Laird, S. T. Weiss, and C. Lange. Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.*, 37:683–691, 2005.

[46] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, 1995.

[47] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Autom. Remote Control*, 24:774–780, 1963.

[48] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv e-prints*, page arXiv:1011.3027, 2010.

[49] H. Wang. Forward regression for ultra-high dimensional variable screening. *J. Am. Stat. Assoc.*, 104:1512–1524, 2009.

[50] H. Wang. Factor profiled sure independence screening. *Biometrika*, 99:15–28, 2012.

[51] X. Wang and C. Leng. High-dimensional ordinary least-squares projection for screening variables. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, 78(3):589–611, 2016.

[52] Q. Wu. Kernel sliced inverse regression with applications to classification. *J. Comput. Graph. Stat.*, 17:590–610, 2008.

[53] Q. Wu, F. Liang, and S. Mukherjee. Regularized sliced inverse regression for kernel models. Technical report, Duke University, Durham, NC, 2008.

[54] Y. Xia. A constructive approach to the estimation of dimension reduction directions. *Ann. Stat.*, 35:2654–2690, 2007.

[55] Y. Xia, H. Tong, W. K. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, 64:363–410, 2002.

[56] Y.-R. Yeh, S.-Y. Huang, and Y.-Y. Lee. Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Trans. Knowl. Data Eng.*, 21:1590–1603, 2009.

[57] H. Zessin. The method of moments for random measures. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62:395–409, 1983.

[58] P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2567, 2006.

[59] Y. Zhu and P. Zeng. Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Am. Stat. Assoc.*, 101:1638–1651, 2006.

[60] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, 67:301–320, 2005.