Statistics and Its Interface Volume 15 (2022) 415–431

Sufficient dimension reduction for spatial point processes using weighted principal support vector machines

Subha Datta and Ji Meng Loh*

We consider sufficient dimension reduction (SDR) for spatial point processes. SDR methods aim to identify a lower dimensional sufficient subspace of a data set, in a model-free manner. Most SDR results are based on independent data, and also often do not work well with binary data. [13] introduced a SDR framework for spatial point processes by characterizing point processes as a binary process, and applied several popular SDR methods to spatial point data. On the other hand, [29] proposed Weighted Principal Support Vector Machines (WPSVM) for SDR and showed that it performed better than other methods with binary data. We combine these two works and examine WPSVM for spatial point processes. We show consistency and asymptotic normality of the WPSVM estimated sufficient subspace under some conditions on the spatial process, and compare it with other SDR methods via a simulation study and an application to real data.

AMS 2000 subject classifications: Primary 62M30; secondary 62H11.

Keywords and phrases: Spatial point processes, Sufficient dimension reduction, Weighted principal support vector machine.

1. INTRODUCTION

High dimensional data pose computational challenges as well as challenges for statistical modeling and inference. One direction for addressing these challenges is Sufficent Dimension Reduction (SDR), which aims to reduce the dimensionality of the data while retaining all the relevant information.

Given $(\mathbf{X}^T, Y)^T \in \mathbb{R}^p \times \mathbb{R}$, a p-dimensional predictor and a response, linear SDR assumes that a $p \times k$ matrix $\mathbf{B}$ with $k \ll p$ exists such that

$$Y \perp \boldsymbol{X} | \mathbf{B}^\top \boldsymbol{X}, \tag{1}$$

where '$\perp$' denotes *conditional independence*, i.e. Y depends on $\mathbf{X}$ only through $\mathbf{B}^\top \boldsymbol{X}$. Under (1), SDR is achieved by estimating $\mathbf{B}$, in particular, the space $\mathcal{S}(\mathbf{B})$ spanned by $\mathbf{B}$, often referred to as the *dimension reduction subspace*. Note that $\mathbf{B}$ 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. [8] introduced the central subspace, denoted $\mathcal{S}_{Y|\mathbf{X}}$, as the intersection of all subspaces satisfying (1), and provided conditions for it to uniquely exist.

While there are many proposed SDR methods, most do not perform well with binary responses. For example, methods such as Slice Inverse Regression (SIR; [18]) and Principal Support Vector Machine (PSVM; [15]) divide $\mathbf{X}$ according to the response Y and if Y is binary, there is only one slice. [29] introduced the Weighted Principal Support Vector Machines (WPSVM) method to handle SDR with binary responses better.

Most SDR methods also deal with independent data. [13] introduced a framework for SDR with spatial point processes, including the idea of a Central Intensity Subspace (CIS) related to the intensity functions of the point process (see the Appendix). By means of a fine grid imposed on the observational window, [13] converted a spatial pattern into a binary response, and applied some popular SDR methods to the resulting binary response.

Our paper has a modest goal. We consider the WPSVM SDR method for spatial point processes. [13] is the only paper we know of that concerns SDR for spatial point data, but uses SDR methods not particularly suited for binary responses. [29] showed that WPSVM preserves the merits of PSVM while achieving SDR for binary data, but for the independent data regression setting. In this paper, we show that WPSVM can be used to perform SDR for spatial point processes, under the framework of [13]. We derive asymptotic properties of the WPSVM estimator for spatial point processes under a mixing condition. We also show, via a simulation study, that the benefit of WPSVM carries over to the spatial point process setting.

SDR deals with high-dimensional data in a model-free manner, and complements model-based methods such as variable selection. Analysis of spatial point data, and more specifically, of the relationship between the intensity function and available covariates is becoming more prevalent. With today's emphasis on data collection, high-dimensional spatial point data are readily available, either directly, or by

*Corresponding author.

merging point data with other data sets (e.g. Census data). Applications span many fields including ecology, telecommunications, crime analysis, earthquake analysis and transportation. For example, [3] introduces a dataset of traffic accident locations together with properties of the road (e.g. speed limit and curvature), and attributes of the accident (e.g. severity and time of day). [32, 37] analyzed the intensity function of tree species in a rainforest using covariates such as the slope, altitude, soil mineral concentrations and distances to other tree species as covariates. [37] also analyzed fast food restaurant locations as a function of census variables, zoning as well as proximity to schools. The food access analysis of [14] using Chicago supermarket data can be extended by including census and transit variables, for example, and could benefit from applying SDR. In this paper, we illustrate the use of WPSVM with a data set on burglaries in London.

The paper is organized as follows. We present the WPSVM method applied to spatial point processes and provide theoretical results (Section 2). A kernel version of WPSVM is presented in Section 3. Sections 4 and 5 contain simulation study and data analysis results respectively. Section 6 contains a brief discussion. The Appendix contains some background material on SDR methods and central intensity subspaces for point processes. Details of proofs are also in the Appendix.

2. WPSVM FOR SPATIAL POINT PROCESSES

Using notation similar to that in [13], let $\mathbb{X} = \{\boldsymbol{X}(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^2\}$ be 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 set $\boldsymbol{\Sigma_X} = \text{cov}\{\boldsymbol{X}(\boldsymbol{s})\}$.

Let $\mathcal{N}$ be a spatial point process that results from some stochastic mechanism conditional on $\mathbb{X}$. We write $\mathcal{N}(.)$ as the counting measure induced by $\mathcal{N}$. For bounded Borel sets $\mathcal{B}_1, \ldots, \mathcal{B}_k$ in $\mathbb{R}^2$, the k-th order moment measure of $\mathcal{N}$ (see [11]) is defined as

$$\mu_{\mathcal{N}}^{(k)}(\mathcal{B}_1 \times \cdots \times \mathcal{B}_k) = \mathbb{E}\{\mathcal{N}(\mathcal{B}_1) \ldots \mathcal{N}(\mathcal{B}_k) | \mathbb{X}\}. \tag{2}$$

Under certain conditions [see e.g. 30], the k-th order intensity function $\lambda_k(.)$ exists and is related to $\mu_{\mathcal{N}}^{(k)}$ as follows:

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

where $\mathrm{d}\boldsymbol{s}_i$ $(i = 1, \ldots, k)$ are distinct infinitesimal sets in $\mathbb{R}^2$. We assume that λ_k exists for all $k \geq 1$. We also assume that the central subspace for $\mathcal{N}$, denoted $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$, depends only on the first-order intensity, more specifically, that $\mathcal{S}_{\mathcal{N}|\mathbb{X}} = \mathcal{S}_1$, where $\mathcal{S}_1$ is the first-order intensity subspace. See the Appendix for a description of central intensity subspaces and their relationship to the central space. Here, intuitively, we are assuming that the central subspace depends only on the first-order intensity function, and higher-order intensity functions do not come into play.

We convert $\mathcal{N}$ to a binary field Y as follows: for $\boldsymbol{s} \in W$, the observation window, set

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

More formally, we consider $d\boldsymbol{s}$ to be a small region around $\boldsymbol{s}$, with $\mathrm{P}(\mathcal{N}(d\boldsymbol{s}) > 1) = o(|d\boldsymbol{s}|)$, so that $Y(\boldsymbol{s}) = 1$ if $\mathcal{N}(d\boldsymbol{s}) = 1$ and $Y(\boldsymbol{s}) = -1$ otherwise. [13] also used a similar characterization in their SDR framework for spatial point processes. SDR methods can then be applied to $Y(\boldsymbol{s})$. [13] applied sliced inverse regression (SIR; [18]), sliced average variance estimation (SAVE; [10]) and directional regression (DR; [16]). The Appendix contains details of the SIR, SAVE and DR methods.

For a binary response Y, the SIR, SAVE and DR methods use only one slice, the one corresponding to $Y = 1$. This is true for a non-spatial binary Y and also for a binary field $Y(\mathbf{s})$, like in [13]. The weighted principal support vector machine (WPSVM; [29]) has been shown to work better for independent binary responses due to its ability to construct more slices. Here we apply WPSVM to possibly correlated spatial point processes through the binary field $Y(\boldsymbol{s})$.

Note that the conversion of a spatial point pattern to 1's and 0's is a common procedure in spatial point analysis. See, for example, [1], [2]. Although on the surface it seems like a loss of information, the 1's and -1's have the covariate values $\mathbf{X}(\boldsymbol{s})$ associated with each binary response. Under the assumed model that the first-order intensity function is a function of the covariates $\mathbf{X}$, all the information for fitting this model is retained. Furthermore, the coordinates of $\boldsymbol{s}$ can be included as one of the covariates, which would allow the full spatial pattern to be reproduced if needed.

If we divide the response $Y(\boldsymbol{s})$ into slices and use a form of SVM to find optimal hyperplanes $a + \mathbf{b}^T\mathbf{X}(\boldsymbol{s})$ to separate them, we seek $(a_0, \mathbf{b}_0)$ such that

$$(a_0, \mathbf{b}_0) = \underset{a,\mathbf{b}}{\text{argmin}} \left\{ \mathbf{b}^\top \boldsymbol{\Sigma_X} \mathbf{b} + \lambda \mathbb{E}\left[1 - \widetilde{Y}_c(a + \mathbf{b}^\top \boldsymbol{X})\right]_+ \right\}, \tag{3}$$

where $\boldsymbol{\Sigma_X} = \text{Var}(\boldsymbol{X}), [a]_+ = \max(a, 0), \widetilde{Y}_c = \mathbb{1}(Y \geq c) - \mathbb{1}(Y < c)$ for a given constant c, and λ a tuning parameter. This is the Principal SVM (PSVM) method for the model in (1), and [15] showed that $\mathbf{b}_0$ is unbiased for linear SDR.

In practice, a sequence of cutoff points c denoted by c_h, $h = 1, \ldots, H$ with an associated $\widetilde{Y}_{i,c_h} = \mathbb{1}(Y_i \geq c_h) - \mathbb{1}(Y_i < c_h)$ is used. With binary responses, such as our $Y(\boldsymbol{s})$, however, there is only one slice, and PSVM suffers from estimating at most one direction of $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$.

Weighted PSVM uses a function g_π such that $g_\pi(Y) = 1 - \pi$ if $Y = 1$ and $g_\pi(Y) = \pi$ if $Y = 0$, for $\pi \in (0, 1)$, and minimizes the objective function

$$\Lambda_\pi(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} + \lambda \mathbb{E}\left\{ g_\pi(Y(\boldsymbol{s}))[1 - Y(\boldsymbol{s})(\alpha + \boldsymbol{\beta}^T \mathbf{X}(\boldsymbol{s}))]_+ \right\}, \tag{4}$$

where we have written $\boldsymbol{\theta}$ for $(\alpha,\boldsymbol{\beta}^T)^T$. Note that while in PSVM, the use of $\widetilde{Y}_c$ for different c produces multiple slices, this is achieved via g_π in (4).

If $\boldsymbol{\theta}_0 \equiv (\alpha_0, \boldsymbol{\beta}_0^T)^T$ is the minimizer of Λ_π, for a particular π, [15] showed that the span of $\boldsymbol{\beta}_0$, $\mathcal{S}(\boldsymbol{\beta}_0)$, is a subset of the central subspace, $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$ in our case:

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).*

The above *linearity condition* assumption implies that

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

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

Given a spatial point process observed in W, let $\boldsymbol{s}_i$, $i = 1, \ldots, n$, be the locations of these points. Further, let $\boldsymbol{s}_i$, $i = n+1, \ldots, n+n_d \equiv N$, be additional locations in W where no points are observed. Set

$$\widehat{\Lambda}_{n,\pi}(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}_N \boldsymbol{\beta} + \frac{\lambda}{N} \sum_{i=1}^{N} g_\pi(Y(\boldsymbol{s}_i))[1 - Y(\boldsymbol{s}_i)(\alpha + \boldsymbol{\beta}^T \boldsymbol{X}(\boldsymbol{s}_i))]_+, \tag{5}$$

where $Y(\boldsymbol{s}_i) = 1$ for $i = 1, \ldots, n$, $Y(\boldsymbol{s}_i) = -1$ for $i = n+1, \ldots, N$, and $\boldsymbol{X}(\boldsymbol{s}_i)$ are the centered covariates.

In practice, the instances where $Y = 1$ are obtained from the observations of the spatial point process. From the literature, the instances where $Y = -1$ can be obtained in two ways. The first is to define a fine grid so that each cell contains 0 or 1 data point. Cells with no data points are assigned $Y = -1$. [1], however, suggests that the choice of grid size might pose complicating issues. Also, an extremely large number of -1's will usually be produced. The second way is to randomly generate a set of dummy points and use these locations for $Y = -1$ (see [2]). In this paper we use the latter method, and obtain the dummy points using the `default.dummy` function in the spatstat R package to generate a quasi-random regular pattern. The same procedure can be repeated with multiple sets of dummy points to assess the impact of using a random set of dummy points, but we do not pursue that here.

Lemma 1. *Suppose that the spatial point process* $\mathcal{N}$ *is* $\tilde{\rho}$ *mixing. Then (5) is a sample version of (4).*

The definition of $\tilde{\rho}$ mixing and the proof of Lemma 1 is given in the Appendix. Briefly, for a $\tilde{\rho}$ mixing sequence of random variables, disjoint subsets of these random variables become less correlated as the separation between the subsets increases. Point processes directed by Gaussian sequences are typically $\tilde{\rho}$ mixing. Neyman–Scott processes are another example of $\tilde{\rho}$ mixing spatial point processes. The proof Lemma 1 makes use of a Theorem in [21] giving a SLLN for $\tilde{\rho}$ mixing sequences.

Following [29], for finite sample estimation, we use a grid of H values, $0 < \pi_1 < \cdots < \pi_H < 1$, and minimize $\widehat{\Lambda}_{n,\pi_h}$. Let the corresponding minimizers be $(\widehat{\alpha}_{n,h}, \widehat{\boldsymbol{\beta}}_{n,h})^\top, h = 1, \ldots, H$. Set the candidate matrix of the linear WPSVM to be

$$\widehat{\mathbf{M}}_n^{LW} = \sum_{h=1}^{H} \widehat{\boldsymbol{\beta}}_{n,h} \widehat{\boldsymbol{\beta}}_{n,h}^\top, \tag{6}$$

and 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 $\widehat{\mathbf{V}}_n^{LW} = (\hat{\mathbf{v}}_1^{LW}, \ldots, \hat{\mathbf{v}}_k^{LW})$. By using different values for the weight parameter π, it is possible for $\widehat{\mathbf{M}}_n^{LW}$ to have more than one eigenvector with non-zero eigenvalue in binary classification.

We assume the following:

(A0) $\mathcal{N}$ exists in $\mathbb{R}^d$ and is $\tilde{\rho}$ mixing, and the observation window $W_{\mathcal{N}} \subset \mathbb{R}^d$ is such that $|W_{\mathcal{N}}| \to \infty$. Hence $n \to \infty$. Furthermore we assume that the number of dummy points n_d is chosen such that $n_d \sim n$.

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

(A2) Given a realization of $\mathcal{N}$, the conditional distribution of $\mathbf{X}(\mathbf{s})$ given $\mathbf{s} \in \mathcal{N}$ is dominated by Lebesgue measure. Similarly, for a set $\mathcal{D}$ of dummy points, the conditional distribution of $\mathbf{X}(\mathbf{s})$ given $\mathbf{s} \in \mathcal{D}$ is dominated by Lebesgue measure.

(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 $\Psi(y,\boldsymbol{\theta}) = \{\boldsymbol{X}(\boldsymbol{s}) : (1 - y\boldsymbol{\theta}^\top \widetilde{\boldsymbol{X}}(\boldsymbol{s}))(1 - y\boldsymbol{\theta}_0^\top \widetilde{\boldsymbol{X}}(\boldsymbol{s})) < 0\}$, where $\widetilde{\boldsymbol{X}}(\boldsymbol{s}) = (1, \boldsymbol{X}(\boldsymbol{s})^T)^T$.

(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

$$u \mapsto \mathbb{E}\left\{\boldsymbol{X}(\boldsymbol{s})|U = u, V = v, Y = y\right\} f_{U|V,Y}(u|v,y)$$

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 $\mathbb{E}\left\{\widetilde{\boldsymbol{X}}(\boldsymbol{s})|U = u, V, Y\right\} f_{U|V,Y}(U = u|V,Y) < c_0(v,y)$.

Assumption (A0) is a common assumption in spatial statistics under the increasing domain asymptotic regime: the correlation of the spatial process is limited; the observation window increases, with corresponding increase in the number of observed points of the process. With $\tilde{\rho}$ mixing and an increasing observation region, the effect of spatial correlation becomes smaller and smaller even as the number of observations increase. (A1) and (A2) are standard assumptions used in the SDR literature, and basically ensures that $\mathbf{X}$ is well-behaved. (A3)–(A5) are regularity conditions for $\Lambda_\pi(\boldsymbol{\theta})$.

With the above assumptions, we have consistency and a Bahadur representation for $\widehat{\boldsymbol{\theta}}_n$.

Theorem 2.2 (Consistency of $\widehat{\boldsymbol{\theta}}_n$). *Suppose $Var\{\boldsymbol{X}(\boldsymbol{s})\} = \boldsymbol{\Sigma}_{\boldsymbol{X}}$ is positive definite and assumption (A2) holds. Then, $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.*

Theorem 2.3 (Bahadur representation of $\widehat{\boldsymbol{\theta}}_n$). *Under assumptions (A1)–(A5),*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -n^{-\frac{1}{2}}\mathbf{H}_{\boldsymbol{\theta}_0}^{-1}\sum_{i=1}^{n}\mathbf{D}_{\boldsymbol{\theta}_0}(\boldsymbol{Z}_i) + o_p(1), \tag{7}$$

where

$$\mathbf{D}_{\boldsymbol{\theta}_0}(\boldsymbol{Z}) = (0, 2\boldsymbol{\Sigma}\boldsymbol{\beta})^\top - \lambda\left[\pi(Y)\widetilde{\boldsymbol{X}}Y\mathbb{1}\{\boldsymbol{\theta}^\top\widetilde{\boldsymbol{X}}Y < 1\}\right], \tag{8}$$

$$\mathbf{H}_{\boldsymbol{\theta}} = 2diag(0, \boldsymbol{\Sigma}) + \lambda\sum_{y=-1,1}\Big[P(Y=y)\pi(y)\times f_{\boldsymbol{\beta}^\top\boldsymbol{X}|Y}(y-\alpha|y)\mathbb{E}(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top|\boldsymbol{\theta}^\top\widetilde{\boldsymbol{X}} = y)\Big]. \tag{9}$$

The above results are for the minimizer of $\widehat{\Lambda}_{n,\pi}$. The next result concerns asymptotic normality of the matrix $\widehat{\mathbf{M}}_n$ in (6) from which the basis of the central subspace is obtained.

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 $h = 1, \dots, H$, where $\mathbf{F}_{\boldsymbol{\theta}_{0,h}}$ denotes the last p rows of $\mathbf{H}_{\boldsymbol{\theta}_{0,h}}^{-1}$. Note that $\mathbb{E}(\mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z})) = 0, \forall\, h = 1, \dots, H$. Thus, by Theorem 2.3, a Bahadur representation of $\widehat{\boldsymbol{\beta}}_{n,h}$ is given by

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{n,h} - \boldsymbol{\beta}_{0,h}\right) = -n^{-\frac{1}{2}}\sum_{i=1}^{n}\mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}_i) + o_p(1). \tag{10}$$

Using the Bahadur representation the asymptotic normality of the candidate matrix $\widehat{\mathbf{M}}_n$ given by (6) can be established.

Theorem 2.4 (Asymptotic normality). *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

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

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.

Theorem 2.4 allows a corollary result for $\widehat{\mathbf{V}}_n$, the estimated leading eigenvectors of $\widehat{\mathbf{M}}_n$:

Corollary 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 $\boldsymbol{\Sigma}_{\mathbf{V}} = (\mathbf{D}^{-1}\mathbf{U}^\top \otimes \mathbf{I}_p)\boldsymbol{\Sigma}_{\mathbf{M}}(\mathbf{U}\mathbf{D}^{-1} \otimes \mathbf{I}_p)$, $\mathbf{U}$ a $p\times k$ matrix with columns being the eigenvectors of $\mathbf{M}_0$ corresponding to nonzero eigenvalues and $\mathbf{D}$ a $k\times k$ diagonal matrix with elements given by the nonzero eigenvalues.

Proofs of Theorems 2.2, 2.3, and 2.4 follow those in [15], [29] and are relegated to the Appendix.

2.1 Determination of the structural dimension, k

In practice, the dimension of the central subspace, i.e. the structural dimension k is not known. To estimate k in PSVM, [15] used a cross-validated BIC (CVBIC) procedure based on asymptotic properties of the PSVM estimator.

We also use CVBIC for the linear WPSVM described above. Here, the BIC-type criterion is

$$G_n(k; \eta, \mathbf{M}) = \sum_{j=1}^{k}v_j - \eta\,\frac{k\log n}{\sqrt{n}}\,v_1, \tag{11}$$

where v_j is the j-th leading eigenvalue of a candidate matrix $\mathbf{M}$ and η is a tuning parameter ([29]). Hence, with $\widehat{\mathbf{M}}_n^{LW}$ the candidate matrix from linear WPSVM in (6), a reasonable estimator of k is $\widehat{k} = \underset{k\in\{1,\dots,p\}}{\operatorname{argmax}}\; G_n\left(k; \eta, \widehat{\mathbf{M}}_n^{LW}\right)$. Asymptotic normality of $\widehat{\mathbf{M}}_n^{LW}$ from Theorem 2.4 ensures that $\widehat{k}$ is consistent:

Theorem 2.5. *Under (A1)–(A5) and $rank(\mathbf{M}_0) = k$, $\lim_{n\to\infty}\mathbb{P}\left(\widehat{k} = k\right) = 1$.*

Please refer to the Appendix for the proof.

In order to use (11), a value of η is needed and we suggest to do the following:

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

$$\left\{\left(\boldsymbol{X}(\boldsymbol{s})_j^{\mathrm{tr}}, Y(\boldsymbol{s})_j^{\mathrm{tr}}\right) : j = 1, \dots, n_{\mathrm{tr}}\right\}$$

and

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

We use a chess board method to split the data ([26]) in order to preserve the spatial correlation.

2. Apply WPSVM to the training data $\left(\boldsymbol{X}(\boldsymbol{s})^{\rm tr}_{j}, Y(\boldsymbol{s})^{\rm tr}_{j}\right)$ to obtain the candidate matrix, $\widehat{\mathbf{M}}^{\rm tr}_n$.
3. Then for each η in a grid of values
   - (a) Compute
   $$\widehat{k}_{\rm tr} = \underset{k=1,\ldots,p}{\operatorname{argmax}}\, G_n\left(k;\eta,\widehat{\mathbf{M}}^{\rm tr}_n\right)$$
   and use the first $\widehat{k}_{\rm tr}$ leading eigenvectors, $\widehat{\boldsymbol{V}}^{\rm tr}_n = \left(\widehat{\boldsymbol{v}}^{\rm tr}_1,\ldots,\widehat{\boldsymbol{v}}^{\rm tr}_{\widehat{k}_{\rm tr}}\right)$, of $\widehat{\mathbf{M}}^{\rm tr}_n$ to transform the training predictors $\widetilde{\boldsymbol{X}}(\boldsymbol{s})^{\rm tr}_j = \left(\widehat{\boldsymbol{V}}^{\rm tr}_n\right)^{\top}\boldsymbol{X}(\boldsymbol{s})^{\rm tr}_j$.
   - (b) For each $\pi_h, h = 1,\ldots,H$, use weighted SVM on $\left\{\left(\widetilde{\boldsymbol{X}}(\boldsymbol{s})^{\rm tr}_j, Y(\boldsymbol{s})^{\rm tr}_j\right) : j = 1,\ldots,n_{\rm tr}\right\}$ to predict $Y(\boldsymbol{s})^{\rm ts}_{j'}$. Let $\widehat{Y(\boldsymbol{s})}^{\rm ts}_{j',h} : j' = 1,\ldots,n_{\rm ts},\ h = 1,\ldots,H$ be the predicted values.
   - (c) Calculate the associated total cost for the test data:
   $$TC(\eta) = \sum_{h=1}^{H}\left\{\sum_{j'=1}^{n_{\rm ts}} g_{\pi_h}\left(Y(\boldsymbol{s})^{\rm ts}_{j'}\right)\mathbb{1}\left(\widehat{Y(\boldsymbol{s})}^{\rm ts}_{j',h} \neq Y(\boldsymbol{s})^{\rm ts}_{j'}\right)\right\}, \tag{12}$$
   where $g_{\pi_h}(1) = 1-\pi_h$ and $g_{\pi_h}(-1) = \pi_h$.
   - (d) select $\widehat{\eta}$ that minimizes $TC(\eta)$.
4. Finally, compute $\widehat{k} = \underset{k\in\{1,\ldots,p\}}{\operatorname{argmax}}\, G_n\left(k;\widehat{\eta},\widehat{\mathbf{M}}^{LW}_n\right)$.

We note here that there are other possible methods to determine the structural dimension, such as using the eigenvalue ratio [20], which could be less computationally demanding. Comparing the effectiveness of various methods for determining the structural dimension could be a direction of future research.

3. A KERNEL VERSION OF WPSVM FOR NONLINEAR SDR

In Section 2, we considered linear WPSVM, where a hyperplane is used to separate the two classes of points corresponding to $Y = 1$ and $Y = -1$. This is under the assumption that the $\mathcal{S}_{\mathcal{N}|\mathbb{X}} = S_1$, so that the presence or absence of points can be explained by $\mathbb{X}$. However, there are situations where non-linear SDR can outperform linear SDR.

An advantage of WPSVM is that it can easily be applied using a kernel in place of the linear function, yielding non-linear SDR within a similar framework. Setting $f(\boldsymbol{X}(\boldsymbol{s});\alpha,\psi) = \alpha + \psi(\boldsymbol{X}(\boldsymbol{s})) - \mathbb{E}\left\{\psi(\boldsymbol{X}(\boldsymbol{s}))\right\}$ where ψ belongs to a Hilbert space, $\mathcal{H}$, the corresponding objective function ([29]) is given by

$$\begin{aligned}\Lambda_\pi(\alpha,\psi) &= \operatorname{Var}\left(\psi\left(\boldsymbol{X}(\boldsymbol{s})\right)\right)\\ &\quad + \lambda\mathbb{E}\left\{g_\pi(Y(\boldsymbol{s}))[1 - Y(\boldsymbol{s})f(\boldsymbol{X}(\boldsymbol{s});\alpha,\psi)]_+\right\}\\ &= \langle\psi,\Sigma\psi\rangle_{\mathcal{H}}\\ &\quad + \lambda\mathbb{E}\left\{g_\pi(Y(\boldsymbol{s}))[1 - Y(\boldsymbol{s})f(\boldsymbol{X}(\boldsymbol{s});\alpha,\psi)]_+\right\},\end{aligned} \tag{13}$$

where $\Sigma : \mathcal{H} \mapsto \mathcal{H}$ is a bounded, self-adjoint operator such that $\langle\psi_1,\Sigma\psi_2\rangle_{\mathcal{H}} = \operatorname{Cov}\left(\psi_1(\boldsymbol{X}(\boldsymbol{s})),\psi_2(\boldsymbol{X}(\boldsymbol{s}))\right)$ for $\psi_1,\psi_2 \in \mathcal{H}$ (see [7]). Note the similarity between (13) and (4).

[15] showed that if (a_0,ψ_0) is the minimizer of the kernel WPSVM objective function (13), then $\psi_0(\mathbf{X}(\mathbf{s}))$ is unbiased for non-linear SDR. To obtain the sample version of objective function (13), suppose $\mathcal{H}$ is spanned by $\Omega = \{\omega_1,\ldots,\omega_d\}$, i.e.,

$$\mathcal{H} = \left\{\psi : \psi(.) = \sum_{j=1}^{d}\gamma_j\omega_j(.),\ \gamma_j \in \mathbb{R},\ j = 1,\ldots,d\right\}. \tag{14}$$

The sample version of (13) using the basis representation (14) is given by

$$\begin{aligned}\widehat{\Lambda}_{n,\pi}(\alpha,\boldsymbol{\gamma}) &= \boldsymbol{\gamma}^\top\boldsymbol{\Omega}^\top\boldsymbol{\Omega}\gamma\\ &\quad + \lambda\sum_{i=1}^{n}\pi(Y(\boldsymbol{s}_i))[1 - Y(\boldsymbol{s}_i)\left\{\alpha + \boldsymbol{\gamma}^\top\boldsymbol{\Omega}_i\right\}]_+,\end{aligned} \tag{15}$$

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

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

and $\bar{\omega}_j = \sum_{i=1}^{n}\omega_j(\boldsymbol{X}(\boldsymbol{s}_i))/n, j = 1,\ldots,d$. We consider choosing an appropriate Ω in Section 3.1.

[15] and [29] show how to obtain the minimizer of (15) by solving a quadratic programming problem. We refer the reader to [15] and [29] for details.

Then, for each π_h, $h = 1,\ldots,H$, we minimize (15) and let the corresponding minimizers be $(\widehat{\alpha}_{n,h},\widehat{\boldsymbol{\gamma}}_{n,h})$, $h = 1,\ldots,H$. The candidate matrix of kernel WPSVM is given by

$$\widehat{\mathbf{M}}^{KW}_n = \sum_{h=1}^{H}\widehat{\boldsymbol{\gamma}}_{n,h}\widehat{\boldsymbol{\gamma}}^\top_{n,h}. \tag{16}$$

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

3.1 Choosing Ω

In order to choose an optimal Ω for estimating the sufficient predictor for kernel WPSVM, we follow the suggestion in [15] to use the eigenfunctions of the linear operator Σ, where Σ is such that $\langle \psi_1, \Sigma_n \psi_2 \rangle_{\mathcal{H}} = \mathrm{Cov}\left(\psi_1(\boldsymbol{X}(\boldsymbol{s})), \psi_2(\boldsymbol{X}(\boldsymbol{s}))\right)$.

Specifically, let $\mathbf{K}_n$ be the $n \times n$ kernel matrix and $\mathbf{Q}_n = \mathbf{I}_n - \mathbf{J}_n/n$, 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 [15], $P_{\boldsymbol{\Omega}}$ is given by $(\mathbf{w}_1, \ldots, \mathbf{w}_d)$, where $\mathbf{w}_j$ is the j-th 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 $\mathbf{k}_n(\boldsymbol{X}(\boldsymbol{s})) = \Big\{K(\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s}_i)) - \sum_{j=1}^{n} K(\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s}_j))$ $/n, i = 1, \ldots, n\Big\}$. We use $d = n/4$ for our simulations and the real data example.

4. SIMULATION STUDY

We conducted a simulation study to compare WPSVM with other dimension reduction methods, specifically, the SIR, SAVE and DR methods. These methods are briefly described in the Appendix. We use data generated from spatial point models similar to those used in [13].

We first simulated a stationary multivariate Gaussian random field $\{\boldsymbol{X}(\boldsymbol{s})\}$ over a 2×2 window to serve as the covariates, where $\boldsymbol{X}(\boldsymbol{s}) = \{X_1(\boldsymbol{s}), \ldots, X_p(\boldsymbol{s})\}^\top \in \mathbb{R}^p$. In our simulations, we considered $p = 5, 10$ and 20. For each $1 \leq j \leq p$, $\{X_j(\boldsymbol{s})\}$ 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

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

We used $\gamma = 0.1$ and 0.2 to vary the range of correlation.

Another quantity $\{\epsilon(\boldsymbol{s})\}$ is also independently simulated as a stationary univariate Gaussian random field having the same mean and covariance structure as the X's. Conditional on $\{X_j(\boldsymbol{s})\}$ and $\{\epsilon(\boldsymbol{s})\}$ we constructed three first-order intensity functions:

$$\begin{aligned}
&\text{(I)} \quad \lambda_1(\boldsymbol{s}) = \alpha \exp\left\{X_1(\boldsymbol{s}) + X_2(\boldsymbol{s}) + 0.4\,\epsilon(\boldsymbol{s})\right\},\\
&\text{(II)} \quad \lambda_2(\boldsymbol{s}) = \alpha \exp\left\{\tfrac{X_1^2(\boldsymbol{s})}{4} + 0.4\,\epsilon(\boldsymbol{s})\right\},\\
&\text{(III)} \quad \lambda_3(\boldsymbol{s}) = \alpha \exp\left\{\tfrac{X_1(\boldsymbol{s})}{0.5 + \{1.5 + X_2(\boldsymbol{s})\}^2} + 0.4\,\epsilon(\boldsymbol{s})\right\}.
\end{aligned}$$

With these intensity functions, we generated spatial point patterns from the inhomogeneous Poisson process model. In each case, we chose values of the constant $\alpha > 0$ such that the expected number of events produced is 200 for a 1×1 window and 800 for the 2×2 window.

We also generated inhomogeneous point patterns from a modified Thomas model ([12, 31]). The modified Thomas model is a Neyman–Scott model where the data points are the locations of offspring points from parent locations. The (unobserved) parent point process is homogeneous Poisson with intensity κ, and offspring points are scattered around their parent point according to an isotropic bivariate normal density with standard deviation σ. In the inhomogeneous version of the Thomas model, the offspring points are thinned or retained based on a retension probability that depends on an intensity function λ. These point patterns can be generated using the `rThomas` function in the `spatstat` R package. For this study, we used $\kappa = 20$ and 50, and $\sigma = 0.05$ and 0.1. We used only $\gamma = 0.1$.

For each model and parameter set, we generated 500 realizations. We then applied linear WPSVM, as well as the SIR, SAVE, and DR methods. We also estimated the structure dimensionality, k using the CVBIC procedure. For the non-linear Model II, we applied KWPSVM as well, and compared it to SAVE, DR and kernel sliced inverse regression (KSIR; [35]), a non-linear version of SIR.

4.1 Comparision with linear WPSVM (LWPSVM)

Our measurement of performance is based on an estimation error used in [17], [36] and [13]. For an estimated $\widetilde{\mathbf{B}}$ and true $\mathbf{B}_0$, the measurement error is given by

$$\Delta\left(\mathbf{B}_0, \widetilde{\mathbf{B}}\right) = \left\|\mathbf{B}_0\left(\mathbf{B}_0^\top \mathbf{B}_0\right)^{-1}\mathbf{B}_0^\top - \widetilde{\mathbf{B}}\left(\widetilde{\mathbf{B}}^\top \widetilde{\mathbf{B}}\right)^{-1}\widetilde{\mathbf{B}}^\top\right\|_{\max}, \tag{17}$$

where $\|\mathbf{A}\|_{\max}$ is the maximum absolute singular value of an arbitrary matrix $\mathbf{A}$, and $0 \leq \|\mathbf{A}\|_{\max} \leq 1$. Smaller $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$ signify better performance.

For each of 500 simulated realizations corresponding to a spatial point model, we ran linear WPSVM, SIR, SAVE and DR, obtained the estimate $\widetilde{\mathbf{B}}$ and calculated $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$. Tables 1, 2, and 3 show the mean and standard deviation of $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$ for the Poisson process based on Models I, II and III respectively.

Model I is a linear model with $\mathbf{B}_0 = (1, 1, 0, \cdots, 0)^\top \in \mathbb{R}^p$, with $\lambda_1(\boldsymbol{s})$ monotonic in $\mathbf{B}_0^\top \boldsymbol{X}(\boldsymbol{s})$. Hence, SIR is expected to perform well. DR, which includes SIR in its formulation, also performs reasonably well, while SAVE performs most poorly. In this case LWPSVM outperforms all three SDR methods, with the smallest values of Δ.

For Model II, the true structural dimension is still $k = 1$, with $\mathbf{B}_0 = (1, 0, \cdots, 0)^\top \in \mathbb{R}^p$. The intensity function is symmetric and the performance of SIR is poor. As SAVE and DR are both sensitive to symmetric directions, they perform reasonably well. LWPSVM performs poorly due to the symmetry, but is comparable with SIR. In Section 4.2

Table 1. Mean (standard deviation) of $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$ for Poisson Model I

γ	p	Win	Results for:			
			SIR	SAVE	DR	LWPSVM
.1	5	1×1	.23 (.03)	.67 (.12)	.24 (.03)	.18 (.07)
		2×2	.30 (.02)	.52 (.20)	.32 (.02)	.08 (.03)
	10	1×1	.34 (.03)	.94 (.09)	.37 (.03)	.31 (.07)
		2×2	.30 (.02)	1.00 (.00)	.32 (.02)	.20 (.05)
	20	1×1	.45 (.03)	.99 (.01)	.51 (.03)	.43 (.07)
		2×2	.34 (.02)	1.00 (.00)	.37 (.02)	.26 (.04)
.2	5	1×1	.25 (.07)	.82 (.19)	.30 (.08)	.21 (.11)
		2×2	.29 (.02)	.81 (.09)	.32 (.02)	.16 (.05)
	10	1×1	.35 (.05)	.90 (.03)	.45 (.04)	.27 (.07)
		2×2	.25 (.03)	1.00 (.00)	.28 (.03)	.17 (.04)
	20	1×1	.42 (.05)	.98 (.02)	.47 (.05)	.60 (.07)
		2×2	.40 (.03)	.99 (.00)	.47 (.03)	.29 (.04)

Table 2. Mean (standard deviation) of $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$ for Poisson Model II

γ	p	Win	Results for:			
			SIR	SAVE	DR	LWPSVM
.1	5	1×1	.46 (.16)	.33 (.07)	.33 (.06)	.64 (.17)
		2×2	.90 (.13)	.32 (.03)	.32 (.03)	.88 (.12)
	10	1×1	.65 (.16)	.43 (.07)	.41 (.06)	.86 (.09)
		2×2	.83 (.11)	.40 (.05)	.40 (.05)	.91 (.09)
	20	1×1	.87 (.07)	.52 (.06)	.56 (.06)	.95 (.03)
		2×2	.98 (.03)	.55 (.04)	.55 (.04)	.99 (.02)
.2	5	1×1	.90 (.12)	.56 (.21)	.57 (.21)	.88 (.11)
		2×2	.52 (.10)	.44 (.04)	.41 (.04)	.65 (.08)
	10	1×1	.86 (.13)	.80 (.14)	.78 (.14)	.84 (.12)
		2×2	.95 (.06)	.49 (.09)	.50 (.09)	.94 (.06)
	20	1×1	.89 (.09)	.88 (.09)	.85 (.10)	.91 (.07)
		2×2	.94 (.04)	.55 (.08)	.56 (.08)	.93 (.05)

Table 3. Mean (Standard Deviation) of $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$ for Poisson Model III

γ	p	Win	Results for:			
			SIR	SAVE	DR	LWPSVM
.1	5	1×1	1.00 (.00)	.83 (.16)	.80 (.17)	.69 (.20)
		2×2	1.00 (.00)	.44 (.20)	.34 (.09)	.40 (.18)
	10	1×1	1.00 (.00)	.92 (.08)	.91 (.09)	.81 (.08)
		2×2	1.00 (.00)	.62 (.14)	.58 (.13)	.82 (.08)
	20	1×1	1.00 (.00)	.97 (.04)	.97 (.04)	.82 (.08)
		2×2	1.00 (.00)	.90 (.11)	.86 (.13)	.78 (.11)
.2	5	1×1	1.00 (.00)	.77 (.18)	.73 (.19)	.63 (.21)
		2×2	1.00 (.00)	.58 (.09)	.57 (.11)	.77 (.14)
	10	1×1	1.00 (.00)	.82 (.09)	.83 (.10)	.53 (.12)
		2×2	1.00 (.00)	.61 (.09)	.59 (.10)	.69 (.15)
	20	1×1	1.00 (.00)	.99 (.02)	.98 (.02)	.84 (.07)
		2×2	1.00 (.00)	.91 (.10)	.93 (.06)	.69 (.09)

Table 4. Mean (standard deviation) of $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$ for Thomas model I

p, κ, σ	Win	Results for:			
		SIR	SAVE	DR	LWPSVM
$5, 50, .1$	1×1	.28 (.06)	.91 (.16)	.32 (.06)	.17 (.07)
	2×2	.20 (.05)	.92 (.11)	.22 (.05)	.14 (.06)
$5, 50, .05$	1×1	.31 (.08)	.90 (.14)	.36 (.08)	.22 (.09)
	2×2	.23 (.07)	.86 (.19)	.26 (.07)	.16 (.05)
$5, 25, .1$	1×1	.31 (.06)	.92 (.13)	.35 (.06)	.22 (.08)
	2×2	.20 (.06)	.91 (.12)	.23 (.06)	.14 (.05)
$5, 25, .05$	1×1	.35 (.10)	.88 (.15)	.40 (.11)	.27 (.12)
	2×2	.26 (.08)	.85 (.20)	.29 (.09)	.18 (.07)
$10, 50, .1$	1×1	.35 (.08)	.98 (.01)	.44 (.09)	.26 (.07)
	2×2	.33 (.05)	.99 (.01)	.36 (.06)	.22 (.06)
$10, 50, .05$	1×1	.41 (.09)	.98 (.02)	.48 (.11)	.33 (.07)
	2×2	.36 (.07)	.97 (.07)	.40 (.08)	.24 (.07)
$10, 25, .1$	1×1	.39 (.09)	.98 (.01)	.47 (.11)	.31 (.07)
	2×2	.34 (.07)	.98 (.02)	.38 (.07)	.24 (.06)
$10, 25, .05$	1×1	.46 (.11)	.98 (.02)	.52 (.12)	.41 (.10)
	2×2	.38 (.08)	.98 (.02)	.42 (.08)	.28 (.08)
$20, 50, .1$	1×1	.47 (.07)	.98 (.01)	.61 (.09)	.40 (.08)
	2×2	.38 (.06)	1.00 (.01)	.44 (.08)	.33 (.06)
$20, 50, .05$	1×1	.53 (.08)	.98 (.01)	.65 (.08)	.46 (.08)
	2×2	.40 (.06)	.99 (.01)	.46 (.08)	.36 (.07)
$20, 25, .1$	1×1	.50 (.07)	.98 (.01)	.63 (.08)	.43 (.08)
	2×2	.40 (.07)	.99 (.01)	.47 (.09)	.35 (.06)
$20, 25, .05$	1×1	.58 (.08)	.98 (.01)	.68 (.08)	.55 (.09)
	2×2	.43 (.07)	.99 (.01)	.50 (.09)	.42 (.06)

we show that kernel WPSVM performs much better for this model.

Model III is two-dimensional with $\mathbf{B}_0 = \{(1, 0, \cdots, 0)^\top, (0, 1, \cdots, 0)^\top\} \in \mathbb{R}^{p \times 2}$. Since, SIR can extract only one direction it performs poorly. SAVE and DR both perform better than SIR. LWPSVM also performs better than SIR. Relative to SAVE and DR, the performance of LPSVM is mixed, outperforming them in some cases but not in others. In Section 4.2 we see that kernel WPSVM also performs much better for this model.

Tables 4, 5, and 6 show the mean and standard deviation of $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$ for the Thomas model based on Models I, II and III respectively. The actual performance of LWPSVM for the Thomas model is slightly worse when compared to

Table 5. Mean (standard deviation) of $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$ for Thomas model II

p, κ, σ	Win	Results for:			
		SIR	SAVE	DR	LWPSVM
$5, 50, .1$	1×1	.84 (.15)	.46 (.14)	.46 (.14)	.85 (.17)
	2×2	.57 (.12)	.45 (.04)	.45 (.03)	.69 (.21)
$5, 50, .05$	1×1	.83 (.16)	.52 (.18)	.52 (.17)	.86 (.15)
	2×2	.62 (.17)	.47 (.05)	.47 (.05)	.76 (.20)
$5, 25, .1$	1×1	.85 (.16)	.50 (.18)	.50 (.17)	.88 (.13)
	2×2	.63 (.15)	.44 (.05)	.45 (.05)	.74 (.21)
$5, 25, .05$	1×1	.85 (.15)	.60 (.21)	.61 (.19)	.87 (.14)
	2×2	.71 (.19)	.46 (.08)	.46 (.08)	.80 (.21)
$10, 50, .1$	1×1	.90 (.10)	.55 (.11)	.54 (.11)	.94 (.08)
	2×2	.77 (.12)	.49 (.04)	.50 (.04)	.94 (.09)
$10, 50, .05$	1×1	.91 (.11)	.60 (.12)	.61 (.12)	.94 (.09)
	2×2	.81 (.13)	.51 (.05)	.52 (.06)	.95 (.09)
$10, 25, .1$	1×1	.93 (.09)	.58 (.11)	.59 (.10)	.95 (.08)
	2×2	.79 (.12)	.51 (.04)	.52 (.05)	.95 (.07)
$10, 25, .05$	1×1	.92 (.09)	.67 (.14)	.68 (.13)	.94 (.08)
	2×2	.84 (.13)	.54 (.07)	.55 (.07)	.94 (.09)
$20, 50, .1$	1×1	.94 (.06)	.77 (.13)	.77 (.13)	.97 (.03)
	2×2	.88 (.07)	.62 (.06)	.64 (.07)	.98 (.02)
$20, 50, .05$	1×1	.94 (.05)	.86 (.10)	.86 (.10)	.97 (.03)
	2×2	.92 (.08)	.62 (.07)	.63 (.08)	.98 (.03)
$20, 25, .1$	1×1	.94 (.06)	.82 (.12)	.82 (.12)	.97 (.05)
	2×2	.89 (.08)	.61 (.09)	.63 (.09)	.98 (.03)
$20, 25, .05$	1×1	.96 (.04)	.86 (.11)	.86 (.11)	.97 (.04)
	2×2	.93 (.08)	.63 (.07)	.64 (.07)	.98 (.03)

Table 6. Mean (standard deviation) of $\Delta(\mathbf{B}_0, \widetilde{\mathbf{B}})$ for Thomas model III

p, κ, σ	Win	Results for:			
		SIR	SAVE	DR	LWPSVM
$5, 50, .1$	1×1	1.00 (.00)	.51 (.16)	.51 (.17)	.46 (.15)
	2×2	1.00 (.00)	.44 (.14)	.39 (.14)	.50 (.19)
$5, 50, .05$	1×1	1.00 (.00)	.63 (.19)	.62 (.20)	.60 (.19)
	2×2	1.00 (.00)	.51 (.17)	.48 (.18)	.52 (.19)
$5, 25, .1$	1×1	1.00 (.00)	.60 (.20)	.61 (.20)	.53 (.17)
	2×2	1.00 (.00)	.67 (.20)	.65 (.20)	.68 (.19)
$5, 25, .05$	1×1	1.00 (.00)	.50 (.14)	.47 (.16)	.60 (.21)
	2×2	1.00 (.00)	.57 (.19)	.54 (.19)	.64 (.22)
$10, 50, .1$	1×1	1.00 (.00)	.69 (.16)	.71 (.18)	.75 (.12)
	2×2	1.00 (.00)	.64 (.17)	.64 (.19)	.70 (.14)
$10, 50, .05$	1×1	1.00 (.00)	.79 (.16)	.78 (.17)	.82 (.12)
	2×2	1.00 (.00)	.68 (.16)	.66 (.19)	.73 (.15)
$10, 25, .1$	1×1	1.00 (.00)	.73 (.16)	.75 (.18)	.79 (.13)
	2×2	1.00 (.00)	.85 (.14)	.85 (.15)	.89 (.10)
$10, 25, .05$	1×1	1.00 (.00)	.70 (.16)	.68 (.17)	.75 (.17)
	2×2	1.00 (.00)	.77 (.17)	.76 (.16)	.82 (.14)
$20, 50, .1$	1×1	1.00 (.00)	.94 (.10)	.94 (.09)	.88 (.07)
	2×2	1.00 (.00)	.80 (.14)	.80 (.15)	.83 (.09)
$20, 50, .05$	1×1	1.00 (.00)	.98 (.05)	.97 (.05)	.93 (.07)
	2×2	1.00 (.00)	.88 (.13)	.87 (.12)	.87 (.08)
$20, 25, .1$	1×1	1.00 (.00)	.94 (.09)	.94 (.08)	.90 (.07)
	2×2	1.00 (.00)	.98 (.04)	.97 (.04)	.96 (.04)
$20, 25, .05$	1×1	1.00 (.00)	.86 (.13)	.86 (.13)	.90 (.07)
	2×2	1.00 (.00)	.94 (.08)	.95 (.07)	.92 (.07)

the Poisson case. This is true for the SIR, SAVE and DR methods also. We find that the performance improves with larger κ and/or larger σ, which is expected, as the spatial correlation is less. We also find that the performance improves with the 2×2 window. With the same correlation range but a larger observation window, the effect of correlation is less, hence the increased performance. Compared with the competing methods, LWPSVM still performs well, with the same relative performance against SIR, SAVE and DR as we saw in the Poisson case: performing the best under Model I, poorly under Model II due to symmetry and with mixed performance under Model III.

Hence we find that WPSVM performs well against SIR, SAVE and DR, and the empirical performance of WPSVM matches what we expect from the theory.

4.1.1 Structural dimensionality

Here we describe the performance of the CVBIC procedure (Section 2.1) for determining the structural dimension of the three spatial point models. Only results for the 1×1 window are shown. Table 7 contains the empirical probabilities (proportion) of correctly estimating the true value of k out of 100 independent simulations.

In general, the procedure seems to perform better for Model I, with $p = 5$ possible covariates, and correlation parameter $\gamma = 0.1$, though the actual proportions are not high. However, we note that the highest empirical proportion consistently occurs at the true value of k. Hence, in practice, we can still use the CVBIC procedure on a real data set, selecting the value of k that has highest empirical probability.

4.2 Kernel WPSVM (KWPSVM)

An advantage of WPSVM is that a kernel technique can be easily used with the procedure, allowing it to deal with non-linear problems with minimal adjustments to the fitting procedure. In Section 4.1, we saw that LWPSVM did not perform well for Model II, which is a non-linear model. Here, we show results from applying kernel WPSVM to the simulated data. We compare KWPSVM with the SAVE, DR and kernel sliced inverse regression (KSIR; [35]) methods. KSIR, as the name suggests, is a nonlinear version of the SIR.

Table 7. Empirical probabilities (proportion) of correctly estimating the true k based on 100 independent simulations for 1×1 window

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

Table 8. Mean (standard deviation) of p-values from Wilcoxon rank sum test based on 100 simulation replications for Poisson Model II

γ	p	Window	Results for:			
			SAVE	DR	KSIR	KWPSVM
.1	5	1×1	.17 (.22)	.17 (.22)	.36 (.30)	.00 (.00)
		2×2	.40 (.28)	.40 (.27)	.50 (.23)	.00 (.00)
	10	1×1	.27 (.27)	.26 (.27)	.41 (.31)	.00 (.00)
		2×2	.45 (.26)	.45 (.26)	.49 (.31)	.00 (.00)
	20	1×1	.18 (.22)	.14 (.19)	.27 (.30)	.00 (.00)
		2×2	.46 (.27)	.46 (.27)	.40 (.29)	.00 (.00)
.2	5	1×1	.56 (.27)	.53 (.28)	.48 (.30)	.00 (.00)
		2×2	.00 (.01)	.00 (.01)	.23 (.29)	.00 (.00)
	10	1×1	.43 (.29)	.34 (.30)	.51 (.29)	.00 (.00)
		2×2	.40 (.28)	.39 (.28)	.52 (.26)	.00 (.00)
	20	1×1	.53 (.27)	.40 (.30)	.48 (.30)	.00 (.00)
		2×2	.21 (.25)	.18 (.25)	.40 (.28)	.00 (.00)

Note that for kernel WPSVM, there is no estimated $\mathbf{B}$, so we cannot use the distance measure Δ (17) to measure the performance. Commonly used techniques include computing the correlation between the response and the estimated sufficient predictor, and the Hotelling T^2 statistic, which [29] used. Using the correlation is inappropriate with binary responses. The same is true for the Hotelling T^2 since the underlying assumptions of normality and independence do not hold. Instead, we use the Wilcoxon rank sum test to measure the performance. Specifically, we test if the estimated sufficient predictors can separate out the two populations of data and dummy points. For each SDR method, we compute the p-values for 100 independent realizations from Model II. The smaller the p-values, the more evidence toward class separation. Tables 8 and 9 summarize the results for the Poisson and Thomas models respectively. We find that the mean p values for KWPSVM are consistently smaller than those from the other methods, both for the Poisson model and for the Thomas model, suggesting that KWPSVM achieves greater outperformance in terms of separation of the classes when the decision curve is non-linear.

Table 9. Mean (standard deviation) of p-values from Wilcoxon rank sum test based on 100 simulation replications for Thomas model II

p, κ, σ	Wi	Results for:			
		SAVE	DR	KSIR	KWPSVM
$5, 50, .1$	1×1	.26 (.28)	.25 (.28)	.38 (.29)	.00 (.00)
	2×2	.05 (.12)	.05 (.12)	.16 (.24)	.00 (.00)
$5, 50, .1.05$	1×1	.25 (.29)	.23 (.30)	.30 (.31)	.00 (.00)
	2×2	.13 (.24)	.13 (.24)	.31 (.33)	.00 (.00)
$5, 25, .1.1$	1×1	.27 (.28)	.25 (.27)	.35 (.30)	.00 (.00)
	2×2	.10 (.23)	.10 (.23)	.24 (.35)	.00 (.00)
$5, 25, .1.05$	1×1	.27 (.33)	.23 (.31)	.23 (.28)	.00 (.00)
	2×2	.16 (.29)	.16 (.29)	.20 (.31)	.00 (.00)
$10, 50, .1.1$	1×1	.24 (.31)	.23 (.31)	.38 (.31)	.00 (.00)
	2×2	.11 (.22)	.11 (.22)	.18 (.24)	.00 (.00)
$10, 50, .05$	1×1	.24 (.28)	.22 (.28)	.30 (.29)	.00 (.00)
	2×2	.12 (.25)	.12 (.25)	.16 (.29)	.00 (.00)
$10, 25, .1$	1×1	.28 (.31)	.26 (.30)	.36 (.32)	.00 (.00)
	2×2	.06 (.13)	.06 (.13)	.29 (.31)	.00 (.00)
$10, 25, .05$	1×1	.29 (.33)	.24 (.32)	.23 (.28)	.00 (.00)
	2×2	.15 (.26)	.14 (.25)	.17 (.25)	.00 (.00)
$20, 50, .1$	1×1	.22 (.29)	.15 (.26)	.39 (.31)	.00 (.00)
	2×2	.06 (.14)	.05 (.13)	.25 (.32)	.00 (.00)
$20, 50, .05$	1×1	.30 (.31)	.20 (.29)	.34 (.32)	.00 (.00)
	2×2	.07 (.19)	.06 (.19)	.26 (.31)	.00 (.00)
$20, 25, .1$	1×1	.30 (.31)	.15 (.23)	.36 (.31)	.00 (.00)
	2×2	.15 (.28)	.14 (.27)	.18 (.26)	.00 (.00)
$20, 25, .05$	1×1	.28 (.31)	.13 (.26)	.23 (.28)	.00 (.00)
	2×2	.20 (.29)	.18 (.27)	.21 (.28)	.00 (.00)

5. APPLICATION TO DATA ON BURGLARIES IN LONDON

Here, we briefly describe the use of WPSVM on a data set of burglaries that occurred in London in January 2013. [25] analyzed a larger version of the data and performed a Bayesian analysis of cell counts, and the authors have a Github repository with the raw and processed data available. Besides burglary locations, the processed data provided by [25] include covariate values on a fairly dense grid of points corresponding to the centroids of census regions called lower super-output areas (LSOA). Each LSOA contains between 400 to 1200 households. The covariates can be classified into three categories corresponding to reward, effort and risk that criminology studies have identified as possible factors affecting burglary target selection. Examples of the covariates are number of households, home prices, accessibility, residential turnover and ethnic heterogeneity. See [25] for more detailed information on the data and criminology background.

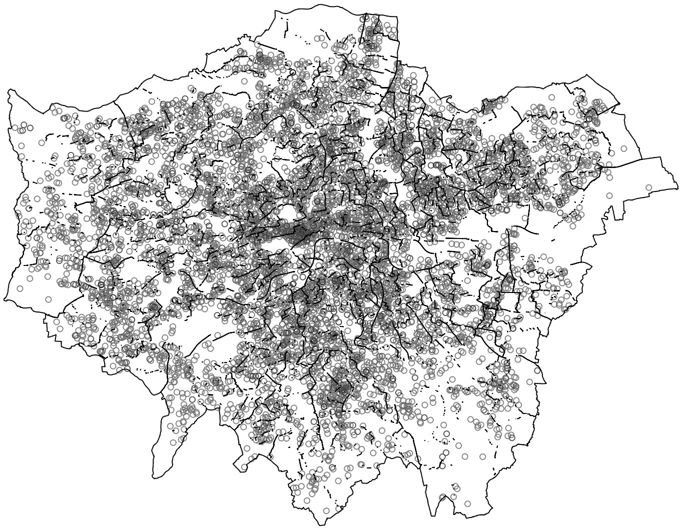

Figure 1. Locations of burglaries in London in January 2013.

Figure 1 shows plots of the January 2013 burglary locations. We generated a set of dummy points, and to each data and dummy point we associate covariate values obtained from the LSOA centroid closest to that point. We then apply linear and kernel WPSVM to the data. For kernel WPSVM we use a Gaussian kernel $K\left(\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s})^\top\right) = \exp(-\|\boldsymbol{X}(\boldsymbol{s}) - \boldsymbol{X}(\boldsymbol{s})^\top\|^2/2\sigma^2)$ with bandwidth parameter σ equal to the median of the pairwise Euclidean distances between the two classes. The same set of values of π and λ were used.

Figure 2 shows 2D histograms of the dummy and data points plotted using the first three estimated bases of the central subspace. Note that we do not expect to achieve complete separation of data and dummy points, since the dummy points are scattered over the whole observation region and there will be many dummy points close to the data points. We find that the data points are slightly more confined to a smaller region than the dummy points. The peaks for the dummy and data points are also at slightly different locations.

If we examine sufficient predictors obtained with LWPSVM, we find that the loadings lean heavily on these covariates: ethnic heterogeneity, number of houses, occupation variation, urban-suburban proportion, household fractions (single parent etc), population turnover, and sports/entertainment points of interest (POIs). This agrees with some findings in criminology studies on factors affecting crime rates – number of dwellings [33], residential turnover [4], rate of single-parent households [28], and ethnic diversity [27, 6] for example.

This also agrees generally with [25]. The model in [25] had several components to account for spatial heterogeneity and these components have slightly different covariate effects. However, in general, they found that number of households, POIs, accessibility, ethnic heterogeneity, occupation variation and residential turnover to be in all three model components. Residential turnover and house prices have opposite effects in their model components. These effects might cancel out and not show up in our SDR method which is applied to the data globally.

Figure 3 shows similar 2D histograms for kernel WPSVM, with greater separation between the data and dummy points compared with LWPSVM. A scatter plot version of Figure 3 is shown in Figure 4. As expected, there is still quite a lot of overlap. However, not only are the peaks for the dummy and data points at different locations, some of the dummy and data points are in separate regions. The better performance of KWPSVM agrees with [25], who found that a linear model was inadequate, and used a 3-component Bayesian model to capture differing local effects.

6. DISCUSSION

As seen from the simulation results, WPSVM may have an advantage over other SDR methods when used with spatial point patterns. This is because we characterize a spatial point pattern into a correlated binary response, and WPSVM is designed for binary classification where the other methods are known to falter. Another advantage of WPSVM is the ease with which a kernel technique can be implemented to deal with non-linear problems, so that the same procedure covers both linear SDR and non-linear SDR. Application to the rainforest data also gave favorable results.

The translation of a spatial point pattern to a binary response involves placing a set of dummy points. It is clear that a dummy point right in the middle of a cluster of point data locations cannot reasonably be separated by any method. The dummy points are meant to represent locations with no point observations. As an extreme example, a set of dummy points where each dummy point is right next to a point process location would be useless. Hence, it might be possible to further improve SDR performance by modifying how the dummy points are generated. A way to do this is by thinning the initial set of dummy points using an ad-hoc estimate of the intensity function of the trees, before applying an SDR procedure. The commonly used kernel estimate ([12]) serves this purpose well.

Specifically, we generate the dummy points as before, and apply independent thinning on the set of dummy points with retention probability $1 - \widehat{\lambda}(s)/\max_s \widehat{\lambda}(s)$, where $\widehat{\lambda}$ is a kernel estimate of the intensity function. We tried thinning in our simulations, and there was a slight performance boost to kernel WPSVM, though not so much for linear WPSVM or the other competing SDR methods. Future work includes fine-tuning the use of thinning to achieve greater performance.

A criticism of WPSVM is a loss of interpretability. This is true of SDR methods in general. The role of sufficient dimension reduction is complementary to other techniques, such

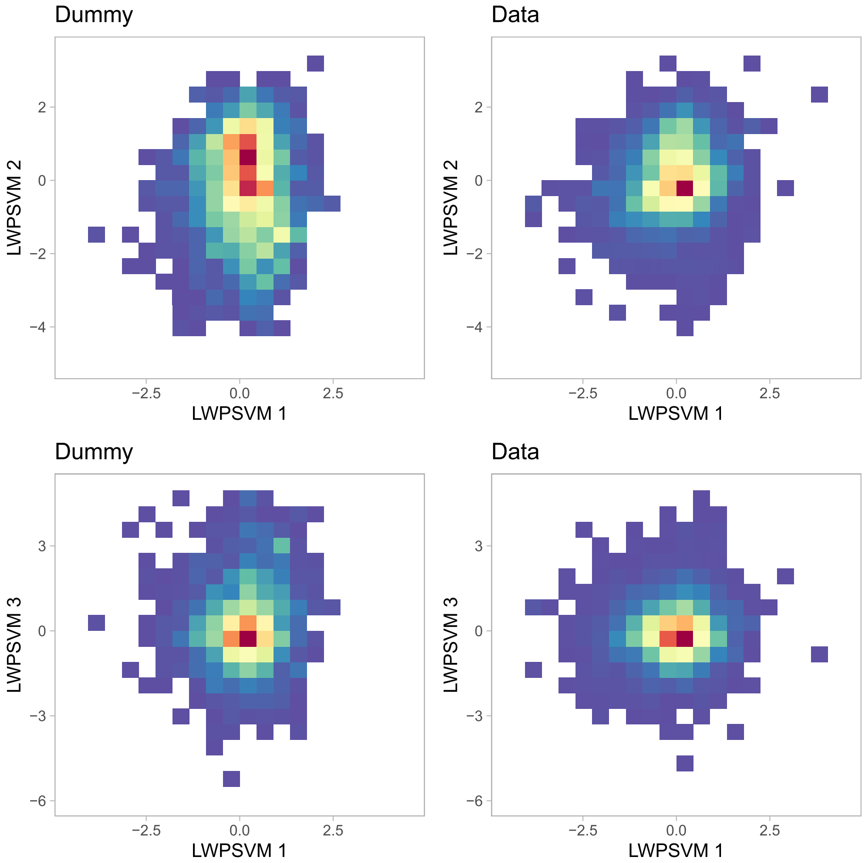

Figure 2. 2D histograms of data (right) and dummy (left) points using the first and second LWPSVM bases (top row) and using the first and third LWPSVM bases (bottom row).

as variable selection methods. Also, there is work that employ sparse SDR to improve interpretability (e.g. [16, 19]). Some of these ideas can be extended to WPSVM.

WPSVM may fail to work when the number of predictors, p is larger than the number of observations, n. One way to tackle this is to use a penalty term for $\boldsymbol{\beta}$. However, this adds an extra level of complexity to an already complex problem. Another way is to use joint screening, which seeks to eliminate uninformative features to reduce dimensionality before further analysis. 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 is part of future work. Additional future work includes SDR methods for the second- and higher-order central intensity subspaces.

APPENDIX A

Central subspace (CS) and Central intensity subspace (CIS)

We denote the linear subspace spanned by the column vectors of a matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$ by $\mathcal{S}(\mathbf{B})$. In the context of (1), $\mathcal{S}(\mathbf{B})$ is called a *sufficient dimension reduction subspace*, and the intersection of all such dimension reduction subspaces is denoted the *central subspace*, $\mathcal{S}_{Y|\mathbf{X}}$. [8, 9] provided mild conditions under which $\mathcal{S}_{Y|\mathbf{X}}$ uniquely exists and has the lowest dimension among all dimension reduction subspaces.

For a spatial point process, [13] suggested that a sufficient dimension reduction subspace $\mathcal{S}(\mathbf{B})$ would be such that, for any positive integer k, and any bounded Borel sets $B_1, \dots B_k \subseteq \mathbb{R}^2$

$$\{\mathcal{N}(\mathcal{B}_1)\dots\mathcal{N}(\mathcal{B}_k)\} \perp \{\mathbb{X}(\mathcal{B}_1),\dots,\mathbb{X}(\mathcal{B}_k)\} \text{ given } \left\{\mathbf{B}^\top\mathbb{X}(\mathcal{B}_1),\dots,\mathbf{B}^\top\mathbb{X}(\mathcal{B}_k)\right\}, \tag{18}$$

where $\mathbb{X}(\mathcal{B}) = \{\boldsymbol{X}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{B}\}$. The *central subspace* $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$ is then the intersection of all dimension reduction subspaces satisfying (18). Like in [13], 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 $\mathcal{S}_{\mathcal{N}|\mathbb{X}}$.

Since the probability distribution of $\mathcal{N}$ can be uniquely determined by the moment measures (2) (see [38]), and each k-th order moment measure can be expressed in terms of intensity functions up to order k, the probability distribution of $\mathcal{N}$ can be specified using the full set of k-th order intensity

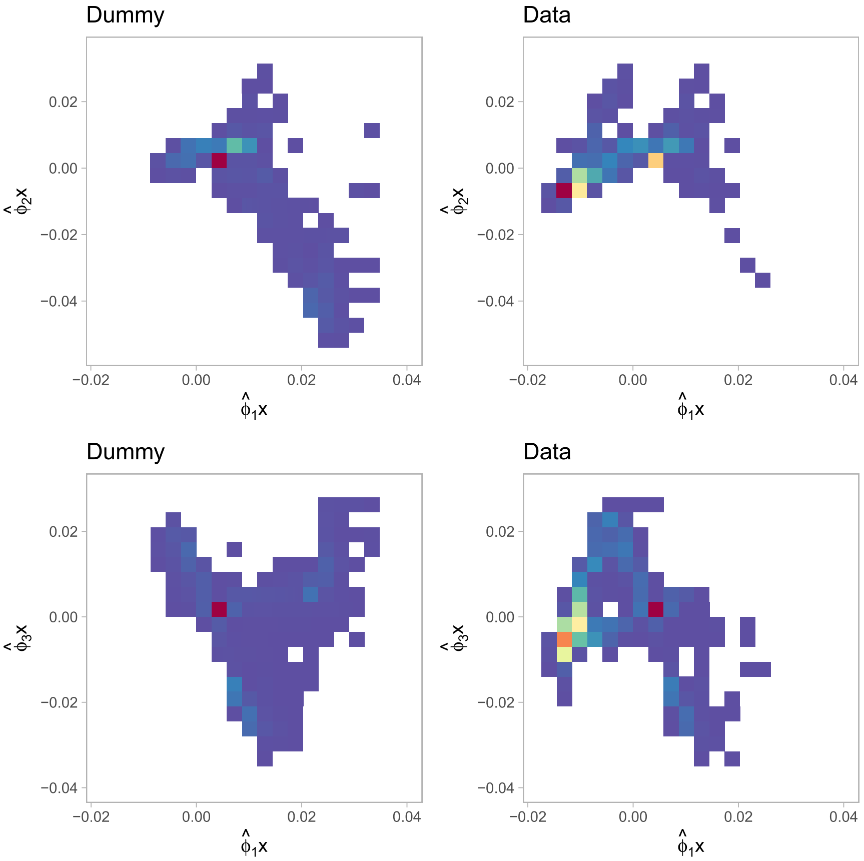

Figure 3. 2D histograms of data (right) and dummy (left) points using the first and second kernel bases (top row) and using the first and third kernel bases (bottom row).

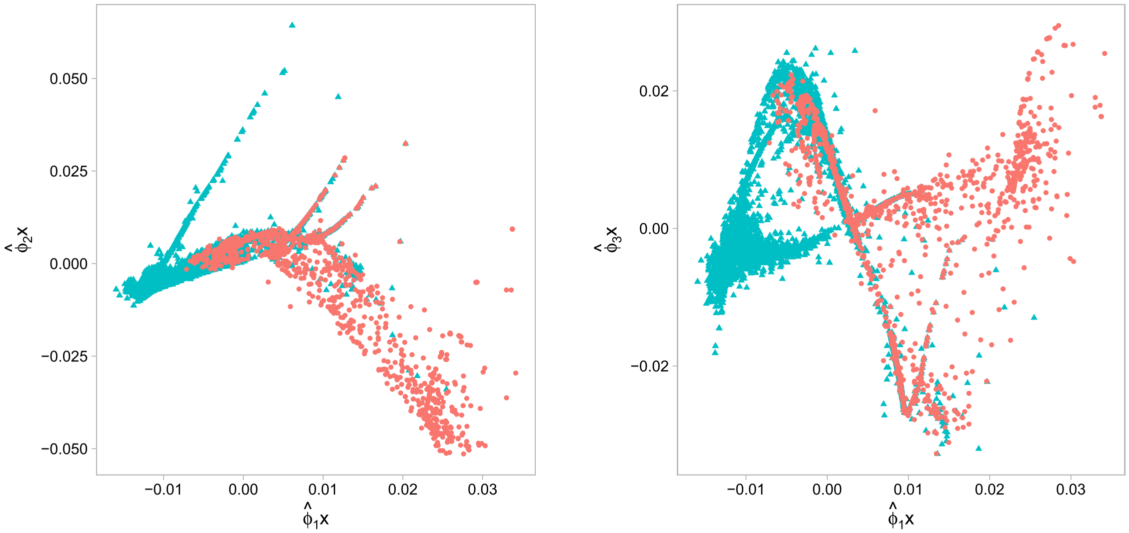

Figure 4. Scatter plots of data (blue-green) and dummy (orange-red) points using the first and second kernel bases (left) and using the first and third kernel bases (right).

functions. If for some function $f_k(.)$ and matrix $\mathbf{B}$,

$$(19) \qquad \lambda_k(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k) = f_k\left\{\mathbf{B}^\top \boldsymbol{X}(\boldsymbol{s}_1), \ldots, \mathbf{B}^\top \boldsymbol{X}(\boldsymbol{s}_k)\right\},$$

the span of $\mathbf{B}$, $\mathcal{S}(\mathbf{B})$, is the k-th order *sufficient intensity dimension reduction subspace* of $\mathcal{N}$ ([13]). If $\mathcal{S}_k$, the intersection of all dimension reduction subspaces satisfying (19), is itself a sufficient intensity dimension reduction subspace, then it should be the smallest and [13] called it the k-th order central intensity subspace (CIS). For the purposes of this paper, we assume that the CIS exists.

Since by definition, the CS contains all information of $\mathbb{X}$ about $\mathcal{N}$, it contains all information of $\mathbb{X}$ on any summary function of $\mathcal{N}$, including the k-th order intensity functions $\{\lambda_k(.) : k \geq 1\}$. Clearly, $\mathcal{S}_k \subseteq \mathcal{S}_{\mathcal{N}|\mathbb{X}}$, for any $k \geq 1$, so that $\bigcup_{k\geq 1} \mathcal{S}_k \subseteq \mathcal{S}_{\mathcal{N}|\mathbb{X}}$. [13] also argued that the reverse relationship holds too, i.e., $\mathcal{S}_{\mathcal{N}|\mathbb{X}} \subseteq \bigcup_{k\geq 1} \mathcal{S}_k$, so that $\mathcal{S}_{\mathcal{N}|\mathbb{X}} = \bigcup_{k\geq 1} \mathcal{S}_k$.

Most analyses of spatial point processes involve only the first- and second-order intensity functions (see e.g. [12, 22]). Like [13], we assume $\mathcal{S}_{\mathcal{N}|\mathbb{X}} = \mathcal{S}_1 \cup \mathcal{S}_2$, known as the *coverage condition*. In fact, they use $\mathcal{S}_{\mathcal{N}|\mathbb{X}} = \mathcal{S}_1$. We do the same here, and focus only on estimating $\mathcal{S}_1$. This holds for a spatial point process whose k-th order intensity function is such that

$$(20) \qquad \lambda_k(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k) = \lambda_1(\boldsymbol{s}_1) \ldots \lambda_1(\boldsymbol{s}_k) g_k(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k),$$

where $g_k(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)$ is free of any covariates for all $k \geq 2$. This means that only the first order intensity function depends on the covariates $\mathbf{X}$.

SIR, SAVE and DR methods

We describe here briefly the SIR, SAVE and DR methods for SDR as applied to spatial point processes by [13], and which we compare WPSVM to. For more details, we refer the reader to [13] as well as the original papers introducing these methods ([18, 10, 16])

For SIR, the spatial point process is treated as a binary random field and $\mathrm{E}(\mathbf{X}(s)|s \in \mathcal{N})$ is estimated by $\widehat{\mathbf{B}}_{\text{SIR}} = \frac{1}{\mathcal{N}(W)} \sum_{s\in\mathcal{N}\cap W} \mathbf{X}(s)$. The estimate $\widehat{\mathbf{B}}_{\text{SIR}}$ converges to $\mathbf{B}_{\text{SIR}}$, and under condition 19 and a p-dimensional random field $\mathbb{X}$, it can be shown that $\mathcal{S}(\mathbf{B}_{\text{SIR}}) \subseteq \mathcal{S}_1$.

The SAVE method uses the conditional covariance of the predictors given the response, estimating $\mathbf{B}_{\text{SAVE}} = [\text{cov}(\mathbf{X}(s)|s \in \mathcal{N}) - \mathbf{I}_p]^2$ by $\widehat{\mathbf{B}}_{\text{SAVE}} = [\widehat{\Sigma} - \mathbf{I}_p]^2$ where $\widehat{\Sigma} = \frac{1}{\mathcal{N}(W)} \sum_{s\in\mathcal{N}\cap W} \mathbf{X}(s)\mathbf{X}(s)^\top - \widehat{\mathbf{B}}_{\text{SIR}}\widehat{\mathbf{B}}_{\text{SIR}}^\top$. [13] describes conditions under which $\widehat{\mathbf{B}}_{\text{SAVE}}$ converges to $\mathbf{B}_{\text{SAVE}}$ in probability, and $\mathcal{S}(\mathbf{B}_{\text{SAVE}}) = \mathcal{S}_1$. Finally, the DR method combines SIR and SAVE, and estimates

$$\begin{aligned}
\mathbf{B}_{\text{DR}} \quad = \quad & 2\mathrm{E}(\mathrm{E}^2[\mathbf{X}(s)\mathbf{X}(s)^\top - \mathbf{I}_p|s \in \mathcal{N}]) \\
& + 2\mathrm{E}^2(\mathrm{E}[\mathbf{X}(s)|s \in \mathcal{N}]\mathrm{E}[\mathbf{X}(s)^\top|s \in \mathcal{N}]) \\
& + 2\mathrm{E}(\mathrm{E}[\mathbf{X}(s)^\top|s \in \mathcal{N}]\mathrm{E}[\mathbf{X}(s)|s \in \mathcal{N}]) \times \\
& \qquad \mathrm{E}(\mathrm{E}[\mathbf{X}(s)|s \in \mathcal{N}]\mathrm{E}[\mathbf{X}(s)^\top|s \in \mathcal{N}]),
\end{aligned}$$

where, like for SAVE, $\mathcal{S}(\mathbf{B}_{\text{DR}}) = \mathcal{S}_1$.

Definition of $\tilde{\rho}$-mixing and a strong law for $\tilde{\rho}$-mixing random variables

In this work, we assume that the spatial point processes are $\tilde{\rho}$ mixing. 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 $\mathcal{F}_U = \sigma\{\xi_k : k \in U\}$. Let

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

for σ-fields $\mathcal{F}, \mathcal{G} \subset \mathcal{M}$. [5] defined the coefficient of dependence, $\tilde{\rho}(n) = \sup\{\rho(\mathcal{F}_U, \mathcal{F}_V)\}$ for $n \geq 0$, with supremum 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 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, $0 \leq \tilde{\rho}(n) \leq \tilde{\rho}(n-1) \leq \cdots \leq \tilde{\rho}(1) \leq 1$, this is equivalent to $\tilde{\rho}(n_0) < 1$, for some $n_0 \geq 1$.

To show that (5) is a sample version of the SVM objective function (4), we make use of a theorem in [21]. First, let $f(x), g(x)$ be real positive functions defined on the same domain $[h, +\infty)$, $0 \leq h \leq 1$ and $\psi(x) = f(x)g(x)$. Note that $f(x)$ or $g(x)$ may not be well defined at h, but if so, $\lim_{x\to h^+} f(x)g(x)$ exists, and we can let $\psi(h)$ be equal to this limit.

Theorem A.1 ([21])**.** *Let $f(x), g(x), \psi(x)$ be functions as described above and satisfy the following conditions:*

(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 $t \in \mathbb{R}$, $t^2 \int_{\psi^{-1}(|t|)}^{+\infty} \frac{\mathrm{d}x}{\psi^2(x)} \leq a\psi^{-1}(|t|) + b$.

Let $\{\xi_n, n \in \mathbb{N}\}$ be a sequence of $\tilde{\rho}$-mixing identically distributed random variables. Set

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

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

Corollary 2. *If $f(x) = x^{1/p}$, with $0 < p < 2$, $g(x) = 1$, $\psi(x) = f(x)g(x) = x^{1/p}$, $x \in [0, +\infty)$, 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:*

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

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

Using the above corollary, with $\xi_k = g_\pi(Y(\boldsymbol{s}_i))[1 - Y(s_i)(\alpha + \boldsymbol{\beta}^T \boldsymbol{X}(\boldsymbol{s}_i)))]_+$, we see that (5) is then a sample version of (4).

The proofs for Theorems 2–5 follow those in [29] very closely. We include them here for completeness.

Proof of Theorem 2.2: Consistency

The proof makes use of following theorem and lemma by [23] and [24] respectively:

Theorem A.2 ([23]). *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 $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.

Lemma 2 ([24]). *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 $\mathscr{S}$.*

Looking at the objective function (4), we observe that the first quadratic term of $\Lambda_\pi(\boldsymbol{\theta})$ is strictly convex, since $\boldsymbol{\Sigma}$ is positive definite and $(a+b)_+ \le a_+ + b_+$, $\forall a, b \in \mathbb{R}$. 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 (5) is also convex by the same logic. Since, $\widehat{\boldsymbol{\Sigma}}_n \xrightarrow{P} \boldsymbol{\Sigma}$ and using Theorem A.1 we have that $\widehat{\Lambda}_{n,\pi}(\boldsymbol{\theta})$ converges to $\Lambda_\pi(\boldsymbol{\theta})$ pointwise. By the above lemma, pointwise convergence $\implies$ uniform convergence. Furthermore, since all four conditions of Theorem A.2 hold, $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.

Proof of Theorem 2.3: Bahadur representation

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 (4) 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.

(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

$$|m_\pi(\boldsymbol{\theta}_1, \boldsymbol{Z}) - m_\pi(\boldsymbol{\theta}_2, \boldsymbol{Z})| \le Q(\boldsymbol{Z}) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \tag{21}$$

Note that the first term $\boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta}$ of $m_\pi(\boldsymbol{\theta}, \boldsymbol{Z})$ 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

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

Also, $\mathbb{E}[\pi(Y)(1 + \|\boldsymbol{X}\|^2)^{\frac{1}{2}}] \le \mathbb{E}[(1 + \|\boldsymbol{X}\|^2)^{\frac{1}{2}}] \le (1 + \mathbb{E}\|\boldsymbol{X}\|^2)^{\frac{1}{2}} < \infty$ 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 (9).
To show this, we use the following lemmas from [15]:

Lemma 3 (Lemma 2 of [15]). *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 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, $|m(\boldsymbol{\theta}_2, \boldsymbol{z}) - m(\boldsymbol{\theta}_1, \boldsymbol{z})| \le c(\boldsymbol{z})\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|$.*

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 4 (Lemma 3 of [15]). *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 $c_i(v) \ge 0$ such that $|h_i(u, v)| f_{U|V}(u|v) \le c_i(v)$, 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

$$\begin{aligned}
D_{\epsilon=0}\mathbb{E}[\boldsymbol{h}(U, V)\mathbb{1}(U + \epsilon V < a + \epsilon\eta)] &= \\
f_U(a)\mathbb{E}[(\eta - V)\boldsymbol{h}(U, V)|U = a]&.
\end{aligned} \tag{22}$$

Lemma 5 (Lemma 4 of [15]). *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 (22)

Already having established (a) and (b) above we apply Lemma 3 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

$$\begin{aligned}&\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Lambda_\pi(\boldsymbol{\theta})\\ &= \frac{\partial}{\partial\boldsymbol{\theta}}\left(2\tilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \lambda\mathbb{E}[\pi(Y)\tilde{\boldsymbol{X}}Y\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}}Y < 1\}]\right)\\ &= 2\tilde{\boldsymbol{\Sigma}} - \lambda\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}[\pi(Y)\tilde{\boldsymbol{X}}Y\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}}Y < 1\}]\\ &= 2\tilde{\boldsymbol{\Sigma}} - \lambda\sum_{y=-1,1}P(Y = y)\pi(y)\times\\ &\qquad\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}[\tilde{\boldsymbol{X}}y\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}}y < 1\}|Y = y] \qquad (23)\end{aligned}$$

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 $Y = +1$,

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}A_{+1}(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}[\tilde{\boldsymbol{X}}\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}} < 1\}]\\ &= -f_{\boldsymbol{\beta}^\top\boldsymbol{X}|Y}(1 - \alpha|1)\mathbb{E}[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top|\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}} = 1] \qquad (24)\end{aligned}$$

by applying Lemmas 4 and 5 and under the assumptions (A2)–(A5). Similarly, for $Y = -1$,

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\theta}}A_{-1}(\boldsymbol{\theta})\\ &= \frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}[\tilde{\boldsymbol{X}}\mathbb{1}\{-\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}} < 1\}]\\ &= -f_{\boldsymbol{\beta}^\top\boldsymbol{X}|Y}(-1 - \alpha| - 1)\mathbb{E}[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top|\boldsymbol{\theta}^\top\tilde{\boldsymbol{X}} = -1] \qquad (25)\end{aligned}$$

We plug (24) and (25) into (23) and get the second derivative of $\Lambda_\pi(\boldsymbol{\theta})$ denoted by $\mathbf{H}_{\boldsymbol{\theta}}$ in (9).

Under the consistency established in Theorem 2.2, (7) is a consequence of Theorem 5.23 of [34], given (a)–(c) are true.

Proof of Theorem 2.4: Asymptotic normality

Let $\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}) = n^{-1}\sum_{i=1}^n \mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z}_i)$, the sample mean of $\mathbf{S}(\boldsymbol{\theta}_{0,h}, \boldsymbol{Z})$. From (10) we have,

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

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,

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

and the result follows from the Central Limit Theorem.

Proof of Theorem 2.5: Structural dimensionality

We have $\widehat{k} = \underset{k\in\{1,\ldots,p\}}{\text{argmax}}\; G_n\left(k; \eta, \widehat{\mathbf{M}}_n\right)$, where $\widehat{\mathbf{M}}_n$ is the candidate matrix of the linear WPSVM as defined in (6). Now,

$$\begin{aligned}&G_n\left(\widehat{k}; \eta, \widehat{\mathbf{M}}_n\right) - G_n\left(k; \eta, \widehat{\mathbf{M}}_n\right)\\ &= \sum_{j=1}^{\widehat{k}} \widehat{\nu}_j - \sum_{j=1}^{k} \widehat{\nu}_j - \eta\frac{\widehat{k}\log n}{\sqrt{n}}\nu_1 + \eta\frac{k\log n}{\sqrt{n}}\nu_1\\ &= \sum_{j=1}^{\widehat{k}} \nu_j - \sum_{j=1}^{k} \nu_j - \eta\frac{(\widehat{k} - k)\log n}{\sqrt{n}}\nu_1 + O_p\left(n^{-1/2}\right), \qquad (26)\end{aligned}$$

where ν_i and $\widehat{\nu}_i$ are the j-th leading eigenvalues of $\mathbf{M}_0$ and $\widehat{\mathbf{M}}_n$ respectively. The last part of (26) is due to the fact that

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

which can be derived as a consequence of Theorem 2.4 and the continuous mapping theorem.

Suppose that $\widehat{k} \neq k$. Thus, we have the following two cases:

Case 1: $\widehat{k} < k$: With increase in sample size, we can see that (26) converges to a negative value, since $rank(\mathbf{M}_0) = k$ and $\sum_{j=1}^{\widehat{k}} \nu_j - \sum_{j=1}^{k} \nu_j < 0$. This leads to a contradiction.

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

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

which leads to a contradiction.

The desired result follows.

Received 5 April 2021

REFERENCES

[1] Baddeley, A., Berman, M., Fisher, N. I., Hardegen, A., Milne, R. K., Schuhmacher, D., & Turner, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. *Electronic Journal of Statistics*, **4**, 1151–1201. MR2735883

[2] Baddeley, A., Coeurjolly, J. F., Rubak, E., & Waagepetersen, R. (2014). Logistic regression for spatial Gibbs point processes. *Biometrika*, **101**, 377–392. MR3215354

[3] Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., & Davies, T. M. (2021). Analysing point patterns on networks – a review. *Spatial Statistics*, **42**, 100435. MR4233256

[4] Bernasco, W., & Luykx, E. (2003). Effects of attractiveness, opportunity, and accessibility to burglars on residential burglary rates of urban neighbourhoods. *Criminology*, **41**, 981–1002.

[5] Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, **2**, 107–144. MR2178042

[6] Clare, J., Fernandez, J., & Morgan, F. (2009). Formal evaluation of the impact of barriers and connectors on residential burglars' macro-level offending location choices. *Australian and New Zealand Journal of Criminology*, **42**, 139–158.

[7] Conway, J. (1990). *A Course in Functional Analysis.* (2nd ed.). New York, NY: Springer.

[8] Cook, R. (1996). Graphics for regressions with a binary response. *J. Am. Stat. Assoc.*, **91**, 983–992. MR1424601

[9] Cook, R. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics.* New York, NY: Wiley. MR1645673

[10] Cook, R., & Weisberg, S. (1991). Discussion of "sliced inverse regression for dimension reduction". *J. Am. Stat. Assoc.*, **86**, 28–33. MR1137117

[11] Cressie, N. (1993). *Statistics for Spatial Data.* (2nd ed.). New York, NY: Wiley. MR1127423

[12] Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns.* New York, NY: Oxford University Press. MR0743593

[13] Guan, Y., & Wang, H. (2010). Sufficient dimension reduction for spatial point processes directed by Gaussian random fields. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, **72**, 367–387. MR2758117

[14] Kolak, M., Bradley, M., Block, D., Pool, L., G. G., Toman, G. K., Boatright, K., Lipiszko, D., Koschinsky, J., Kershaw, K., Carnethod, M., Isakova, T., & Wolk, M. (2018). Chicago supermarket data and food access analytics in census tract shapefiles for 2007–2014. *Data in Brief* (pp. 2482–2488).

[15] Li, B., Artemiou, A., & Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann. Stat.*, **39**, 3182–3210. MR3012405

[16] Li, B., & Wang, S. (2007). On directional regression for dimension reduction. *J. Am. Stat. Assoc.*, **102**, 997–1008. MR2354409

[17] Li, B., Zha, H., & Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Stat.*, **33**, 1580–1616. MR2166556

[18] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.*, **86**, 316–327. MR1137117

[19] Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, **94**, 603–613. MR2410011

[20] Luo, R., Wang, H., & Tsai, C. L. (2009). Contour projected dimension reduction. *Annals of Statistics*, **37**, 3743–3778. MR2572442

[21] Meng, Y.-J., & Lin, Z.-Y. (2010). Strong laws of large numbers for $\tilde{\rho}$-mixing random variables. *J. Math. Anal. Appl.*, **365**, 711–717. MR2587074

[22] Møller, J., & Waagepetersen, R. (2007). Modern statistics for spatial point processes. *Scand. J. Stat.*, **4**, 643–684. MR2392447

[23] Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics IV* (pp. 2113–2245). MR1315971

[24] Pollard, D. (1991). Aymptotics for least absolute deviation regression estimator. *Econom. Theory*, **7**, 186–199. MR1128411

[25] Povala, J., Virtanen, S., & Girolami, M. (2020). Burglary in london: insights from statistical heterogeneous spatial point processes. *Applied Statistics*, **69**, 1067–1090. MR4166857

[26] Roberts, D., Bahn, V., Ciuti, S., Boyce, M., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J., Schroder, B., Thuillier, W., Warton, D., Wintle, B., Hartig, F., & Dormann, C. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, **40**, 913–929.

[27] Sampson, R. J., & Groves, W. B. (1989). Community structure and crime: teting social-disorganization theory. *American Journal of Sociology*, **94**, 774–802.

[28] Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science*, **277**, 918–924.

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

[30] Stoyan, D., Kendall, W. S., & Mecke, J. (1995). *Stochastic Geometry and Its Applications.* New York, NY: Wiley. MR0895588

[31] Thomas, M. (1949). A generalization of Poisson's binomial limit for use in ecology. *Biometrika*, **36**, 18–25. MR0033999

[32] Thurman, A. L., & Zhu, J. (2014). Variable selection for spatial Poisson point processes via a regularization method. *Statistical Methodology*, **17**, 113–125. MR3133589

[33] Townsley, M., Birks, D., Ruiter, S., Bernasco, W., & White, G. (2016). Target selection models with preference variation between offenders. *Journal of Quantitative Criminology*, **32**, 283–304.

[34] van der Vaart, A. W. (1998). *Asymptotic Statistics.* New York, NY: Cambridge University Press. MR1652247

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

[36] Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Stat.*, **35**, 2654–2690. MR2382662

[37] Yue, Y., & Loh, J. M. (2015). Variable selection for inhomogeneous spatial point process models. *Canadian Journal of Statistics*, **43**, 288–305. MR3353384

[38] Zessin, H. (1983). The method of moments for random measures. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **62**, 395–409. MR0688646

Subha Datta
New Jersey Institute of Technology
New Jersey, USA
E-mail address: std8@njit.edu

Ji Meng Loh
New Jersey Institute of Technology
New Jersey, USA
E-mail address: loh@njit.edu