The Canadian Journal of Statistics
Vol. xx, No. yy, 20??, Pages 1–25
La revue canadienne de statistique

Variable Selection for Inhomogeneous Spatial Point Process Models

Yu (Ryan) Yue[1]* and Ji Meng Loh[2]

[1]*Zicklin School of Business, Baruch College, The City University of New York*
[2]*Department of Mathematical Sciences, New Jersey Institute of Technology*

Key words and phrases: Berman-Turner approximation; intensity function; maximum pseudo-likelihood estimator ; spatial point processes; variable selection via regularization ; weighted Poisson likelihood.

MSC 2010: Primary 62M30; secondary 62J07

Abstract: In this work, we consider variable selection when modeling the intensity and clustering of inhomogeneous spatial point processes, integrating well-known procedures in the respective fields of variable selection and spatial point process modeling to introduce a simple procedure for variable selection in spatial point process modeling. Specifically, we consider modeling spatial point data with Poisson, pairwise interaction and Neyman-Scott cluster models, and incorporate LASSO, adaptive LASSO and elastic net regularization methods into the generalized linear model framework for fitting these point models. We perform simulation studies to explore the effectiveness of using each of the three regularization methods in our procedure. We then use the procedure in two applications, modeling the intensity and clustering of rainforest trees with soil and geographical covariates using a Neyman-Scott model, and of fast food restaurant locations in New York City with Census variables and school locations using a pairwise interaction model.
The Canadian Journal of Statistics xx: 1–25; 20?? © 20?? Statistical Society of Canada

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–25; 20?? © 20?? Société statistique du Canada

1. INTRODUCTION

Spatial point data occur in a wide range of applications such as ecology, astronomy and epidemiology and consist of random locations representing the observations of the objects (e.g. trees, galaxies, disease cases) of interest. Different statistical models for spatial point processes have been developed and studied, see e.g. Møller and Waagepetersen (2004). For stationary point patterns, the interest is often the study of the clustering properties. With the more recent interest in inhomogeneous point patterns, one focus is the modeling of the spatially varying intensity of the process in terms of covariates.

Broadly speaking, spatial point processes can be divided into three classes according to how the points are clustered or dispersed: complete spatial randomness, clustering or regularity. Complete spatial randomness is exhibited by the Poisson spatial point process, where there is independence between the points of the process. This process has been well-studied due to its theoretical and computational tractability. Despite the strong assumption of independence, Schoenberg (2005) showed that the inhomogeneous Poisson process model with a linear expression for the log-intensity can yield consistent estimates for non-Poisson point data under certain regularity conditions on the process generating the data.

* *Author to whom correspondence may be addressed.*
E-mail: yu.yue@baruch.cuny.edu

Point patterns that exhibit clustering are often modeled using cluster point processes. They produce clustering patterns through "offspring" points that are distributed around the locations of (unobserved) underlying "parent" points. The parent point process is often taken to be Poisson, although Yau and Loh (2010) considered a regular parent process. Waagepetersen (2007) and Waagepetersen and Guan (2009) exploited the consistency property of Schoenberg (2005) to introduce a two-step procedure for modeling the intensity and an inhomogeneous K function that is often used to describe the clustering properties of a point process.

A pairwise interaction process uses a non-negative *interaction function* to describe the interactions between pairs of points of the process. More general interaction processes involve higher-order interaction functions but are less often used in applications. Pairwise interaction processes involve analytically intractable normalizing constants hence likelihood-based methods for fitting such models are computationally intensive. These models are also more suited for regular point processes (Stoyan and Stoyan, 1994). We describe these three spatial point processes in more detail in Section 2 and in particular, describe how these models are fit to data.

Variable selection via regularization became a very active research area with the introduction of a penalized likelihood procedure by Tibshirani (1996), where a least absolute shrinkage and selection operator (LASSO) penalty is added to the likelihood function and used to shrink small coefficients to zeros while retaining large coefficients in the model. Minimizing the negative log likelihood plus the LASSO penalty simultaneously performs variable selection and parameter estimation. Since then, many regularization methods have been developed, e.g. SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005) and adaptive LASSO (Zou, 2006). In particular, the elastic net uses both the LASSO and the ridge-regression penalties to achieve a sparse solution as well as shrink correlated coefficients to each other. The adaptive LASSO and SCAD have so-called *oracle* properties, with asymptotic performance that is as good as if the true model were known.

More recently, there has been work on implementing variable selection for spatial point models in order to reduce variance inflation from overfitting and bias from underfitting. Renner and Warton (2013) used the LASSO penalty with a Poisson point process to introduce a maximum entropy approach for modeling the spatial distribution of a species in ecology. Thurman and Zhu (2014) employed an adaptive LASSO penalty to select variables for the Poisson point process model. Note that in both cases, only the Poisson point process was considered. In this work, we propose a unified yet flexible approach to perform variable selection when modeling spatial point processes. The approach allows the use of a variety of shrinkage methods with a wide range of inhomogeneous spatial point processes, essentially giving the investigator more control over the process of variable selection and spatial modeling. To introduce the procedure, we consider the use of the LASSO, adaptive LASSO and elastic net shrinkage methods to select variables for the Poisson, the pairwise interaction as well as the cluster point process models. The procedure is computationally efficient and can be easily implemented in `R` (R Core Team, 2013). We study the proposed method using a comprehensive simulation study, comparing the use of these three shrinkage methods together with the three types of point models. We also apply our method to two real data sets: the Barro Colorado Island (BCI) forest dataset and a New York City (NYC) fast food restaurant and school dataset.

1.1. Data sets

The BCI dataset used here is part of a series of censuses conducted on a 1000m by 500m plot of tropical moist forest in the Barro Colorado Island in central Panama. The censuses were conducted over a period of 25 years during which over 350,000 trees consisting of as many as 3000 tree species were inspected. For this paper, we focus on the census conducted in 1995. Spatial covariates are available, including elevation and slope of the land, soil pH and concentrations of

minerals such as nitrogen (N), potassium (K) and phosphorus (P) in the soil.

The NYC dataset consists of locations of fast food restaurants (FFRs) in NYC, NYC public elementary school locations with demographic information, as well as city demographic and infrastructure information at the census block group level. The FFR locations were obtained from a 2005 online directory of restaurant inspections by the NYC Department of Health and Mental Hygiene. There are various ways to characterize a fast food restaurant. The dataset we use is the one analyzed by Kwate et al. (2009) and the paper describes the criteria used to select the FFRs. There were a total of 818 FFRs consisting of both national chains and local stores. Data on the NYC public elementary schools were obtained from the Department of Education. Besides the actual locations of 913 elementary schools, there was school demographic and socio-economic data such as racial composition and percent of students eligible for free lunch. Area income, median age, racial composition and population density at the block group level were obtained from the 2000 US Census Summary Files 1 and 3 (SF-1 and SF-3). Average household expenditures for food (lunch, dinner and snacks) away from home for 2006 were derived from a Consumer Expenditure Survey by the US Bureau of Labor Statistic and supplied by a commercial GIS firm. Finally, information on zoning was obtained from NYC tax lot base map files.

The remainder of this paper is organized as follows. The Poisson, pairwise interaction and cluster process models for spatial point data are described in Section 2, specifically describing how estimates of model parameters are obtained. We then describe the implementation of LASSO, adaptive LASSO and elastic net selection techniques to these spatial point process models in Section 3. Simulation studies are presented in Section 4. The results of the applications to the BCI forest data and NYC fast food restaurant data are described in Section 5. We conclude the paper with a short discussion in Section 6.

2. STATISTICAL MODELS FOR SPATIAL POINT PROCESSES

Let X be a spatial point process on a domain $D \subset \mathbb{R}^2$ with first- and second-order intensity functions defined by

$$\rho(s) = \lim_{|ds|\to 0} \left(\frac{\mathrm{E}[X(ds)]}{|ds|} \right) \text{ and } \rho^{(2)}(s_1, s_2) = \lim_{|ds_1|,|ds_2|\to 0} \left(\frac{\mathrm{E}[X(ds_1)X(ds_2)]}{|ds_1||ds_2|} \right), \quad (1)$$

where ds represents an infinitesimal region around location s, $|ds|$ its area and $X(ds)$ the number of points in ds. The quantity $\rho(s)|ds|$ can be thought of as the probability of observing one point in ds and $\rho^{(2)}(s_1, s_2)|ds_1||ds_2|$ the probability of observing one point in each of ds_1 and ds_2. The first-order intensity function is often referred to as the *intensity function* and, for an inhomogeneous point process, is often assumed to be a parametric function ρ_θ that depends on some spatial covariates through parameters $\boldsymbol{\theta}$. Moreover, we assume that the process X is second-order intensity reweighted stationary (Baddeley and Turner, 2000), such that $\rho^{(2)}(s_1, s_2) = \rho(s_1)\rho(s_2)g(s_1 - s_2)$, where $g(\cdot)$ is the *pair correlation function*. Under this assumption, the so-called K*-function* is well-defined and given in terms of an integral involving g. A parametric model g_ψ is often assumed for the pair correlation function and hypotheses regarding clustering may be formulated in terms of ψ. See for example Waagepetersen and Guan (2009). Yue and Loh (2011, 2013), however, estimated both the intensity function and the pair correlation function non-parametrically. Diggle (2003), Møller and Waagepetersen (2004) and Illian et al. (2008) provide comprehensive introductions to spatial point processes.

2.1. Poisson point processes

Suppose $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ denotes a realization of a spatial point process X observed within a bounded region D, where n is random and $x_i, i = 1, \ldots, n$ represent the locations of the ob-

served points. If X is a Poisson point process, then the likelihood of X is given by

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \rho_\theta(x_i) \exp\left(-\int_D \rho_\theta(s)\, ds\right), \tag{2}$$

where $\rho_\theta(s)$ is the intensity function with vector parameter $\boldsymbol{\theta}$. To estimate $\boldsymbol{\theta}$ one may maximize $L(\boldsymbol{\theta})$ using the Berman-Turner device (Berman and Turner, 1992). More explicitly, we approximate the integral in (2) by

$$\int_D \rho_\theta(s)\, ds \approx \sum_{j=1}^{N} w_j \rho_\theta(s_j), \quad j = 1, \ldots, N,$$

where the set $\mathcal{S} = \{s_j\}_{j=1}^N$ of N points in D consist of the n data points and $N - n$ dummy points. The so-called quadrature weights $w_j > 0$ are such that $\sum_{j=1}^n w_j = |D|$. This yields an approximation to the weighted log-likelihood,

$$\begin{aligned} \log L(\boldsymbol{\theta}) &\approx \sum_{i=1}^{n} \log \rho_\theta(x_i) - \sum_{j=1}^{N} w_j \rho_\theta(s_j) \\ &= \sum_{j=1}^{N} \Big[y_j \log \rho_\theta(s_j) - \rho_\theta(s_j) \Big] w_j, \end{aligned} \tag{3}$$

where $y_j = 1/w_j$ if s_j is a data point and $y_j = 0$ if s_j is a dummy point. It is easy to see that the right-hand side of (3) is equivalent to the log-likelihood of independent Poisson variables Y_j with mean $\rho_\theta(s_j)$ and weights w_j. This use of this procedure allows for any choice of dummy points and quadrature weights. Waagepetersen (2008) suggested two ways of obtaining the dummy points, using stratified dummy points combined with grid-type weights or binomial dummy points with the Dirichlet-type weights. In general, a large number of dummy points are required in order to obtain accurate parameter estimates.

To study the spatial heterogeneity determined by specified covariates, we assume that the intensity function has a parametric form

$$\rho_\theta(s) = \exp\{\beta_0 + z_1(s)\beta_1 + \cdots + z_p(s)\beta_p\}, \quad s \in D, \tag{4}$$

where $z_k(s)$, $k = 1, \ldots, p$, are p covariates measured at location s, and $\boldsymbol{\theta} \equiv \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the vector of corresponding regression coefficients. To estimate $\boldsymbol{\beta}$ one can maximize the approximate log-likelihood (3) using standard software for fitting generalized linear models (Baddeley and Turner, 2005). This procedure has also been applied to spatial point processes that are not Poisson. The resulting estimates are referred to as Poisson estimates. Schoenberg (2005) showed that estimates of $\boldsymbol{\beta}$ obtained from maximizing the Poisson likelihood (2) are consistent under certain regularity conditions on the point process. More specifically, the regularity conditions (A1)-(A3) of Schoenberg (2005) ensure that the parameter space contains the Poisson likelihood maxima, restrict the variability of the point process, and ensure that $\rho_{\hat{\theta}}$ is sufficiently different from ρ_θ for θ outside a neighborhood of $\hat{\theta}$. See Schoenberg (2005) for further details.

However, since the procedure ignores any dependence between the points in the process, the estimates may have poor efficiency for finite spatial point patterns with strong interaction.

2.2. Pairwise interaction processes

A pairwise interaction process on D with trend/activity function b_β and interaction function h_γ has likelihood function

$$L(\boldsymbol{\theta}) = \alpha_\theta \prod_{i=1}^{n} b_\beta(x_i) \prod_{i<j} h_\gamma(x_i, x_j), \tag{5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ and α_θ is a normalizing constant (Ripley, 1991). The function b_β influences the intensity of points while h_γ controls the interaction (or dependence) among the points in the pattern. If the interaction function $h_\gamma \equiv 1$, the process becomes inhomogeneous Poisson with intensity function b_β. A simple pair interaction process is the Strauss process (Strauss, 1975), where the interaction function is constant with value $0 \leq \gamma < 1$ for point pairs closer than distance r apart, and equal to 1 otherwise. Hence there is repulsion up to a range of r, with smaller values of γ representing stronger repulsion.

It is generally difficult to maximize the likelihood (5) because the constant α_θ is an intractable function of $\boldsymbol{\theta}$. An alternative to likelihood function is pseudo-likelihood function (Besag, 1975), defined as the product of the conditional likelihoods of each random variable given the other variables. Besag (1977) defined the pseudo-likelihood of a spatial point process over a subset $D \subset W$ to be

$$PL(\boldsymbol{\theta}; \boldsymbol{x}) = \left(\prod_{x_i \in D} \rho_\theta(x_i; \boldsymbol{x}) \right) \exp\left(- \int_D \rho_\theta(s; \boldsymbol{x}) ds \right),$$

where $\rho_\theta(s; \boldsymbol{x})$ is the Papangelou conditional intensity function at location s. It can be shown that the general pairwise interaction process (5) has conditional intensity

$$\rho_\theta(s; \boldsymbol{x}) = b_\beta(s) \prod_{i=1}^{n} h_\gamma(s, x_i) I(x_i \neq s), \tag{6}$$

and therefore its pseudo-likelihood can be written as

$$PL(\boldsymbol{\theta}; \boldsymbol{x}) = \left(\prod_{i=1}^{n} b_\beta(x_i) \prod_{i \neq j} h_\gamma(x_i, x_j) \right) \exp\left(- \int_D b_\beta(s) \prod_{i=1}^{n} h_\gamma(s, x_i) ds \right). \tag{7}$$

Note that if the process is Poisson the pseudo-likelihood (7) coincides with the likelihood (2) up to the factor $\exp(|D|)$, suggesting that the pseudo-likelihood is a useful approximation to the likelihood. When the process has 'weak interactions', the maximum pseudo-likelihood estimator (MPLE) should be efficient. However, it is believed to be inefficient for strong interactions.

To allow for covariate effects and spatial interactions, we assume that $b_\beta(s) = \exp(\beta_0 + \boldsymbol{z}'(s)\boldsymbol{\beta})$ and $h_\gamma(s,t) = \exp(\boldsymbol{H}'(s,t)\boldsymbol{\gamma})$, where $\boldsymbol{z}(s)$ are vectors of covariates and $\boldsymbol{H}(s,t)$ the interaction functions, defined for every $s, t \in D$. Then, the conditional intensity (6) becomes

$$\rho_\theta(s; \boldsymbol{x}) = \exp\left(\beta_0 + \boldsymbol{z}'(s)\boldsymbol{\beta} + \sum_{i=1}^{n} \boldsymbol{H}'(s, x_i)\boldsymbol{\gamma} \right).$$

The pseudo-likelihood (7) can then be maximized to obtain an MPLE for $\boldsymbol{\theta}$ using the Berman-Turner procedure (Baddeley and Turner, 2000). The MPLE is known to be consistent and asymptotically normal, at least for stationary pairwise interaction processes whose interaction functions satisfy suitable regularity conditions.

2.3. Cluster point processes

A cluster point process can be represented as $X = \cup_{c \in C} X_c$, so that X is a superposition of clusters X_c of "offspring" points associated with "parent" points c from a process C. A popular cluster process is the inhomogeneous Neyman-Scott process, introduced by Waagepetersen (2007), where the parent process C is stationary Poisson with intensity $\kappa > 0$ and the clusters X_c are independent Poisson processes with intensity functions

$$\rho_{c,\theta}(s) = \alpha k(s - c; \omega) \exp\left(\boldsymbol{z}'(s)\boldsymbol{\beta}\right),$$

where $\alpha > 0$ is the expected number of offspring for each parent point and $k(\cdot)$ is a probability density determining the distribution of offspring points around the parent points. The intensity is modulated by the covariates $\boldsymbol{z}$ through the expression $\exp\left(\boldsymbol{z}'(s)\boldsymbol{\beta}\right)$. When k is the density of a bivariate normal distribution $N(\mathbf{0}, \omega \boldsymbol{I})$, a so-called modified Thomas process is obtained. For a large κ and a small ω, the Thomas process has many spatially tight clusters whereas a small κ and a large ω produce few and widely dispersed clusters.

The intensity function of the cluster process X is

$$\rho_\theta(s) = \kappa\alpha \exp(\boldsymbol{z}'(s)\boldsymbol{\beta}) = \exp(\beta_0 + \boldsymbol{z}'(s)\boldsymbol{\beta}), \tag{8}$$

where $\beta_0 = \log(\kappa\alpha)$. Although likelihood-based inference can be performed using Markov Chain Monte Carlo methods, this can be very computationally intensive (Møller and Waagepetersen, 2004). Waagepetersen (2007) and Waagepetersen and Guan (2009) suggested a two-step estimation procedure. First, they maximize the Poisson likelihood (2) with intensity function (8) to obtain unbiased regression coefficient estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. This can be easily done using the estimation procedure presented in Section 2.1. Second, they estimate the inhomogeneous K-function for X with the estimated intensity function obtained in the first step, and obtain estimates $\hat{\kappa}$ and $\hat{\omega}$ by minimizing the contrast between the estimated and theoretical inhomogeneous K-functions. Finally, $\hat{\alpha} = \exp(\hat{\beta}_0)/\hat{\kappa}$. These estimates have been shown to be consistent and asymptotically normal (Waagepetersen and Guan, 2009).

3. VARIABLE SELECTION VIA REGULARIZATION METHODS

3.1. Penalized likelihoods for spatial point processes

The three inhomogeneous spatial point process models described above share a common property: their intensity functions are log-linear and the parameters can be estimated by maximizing a weighted Poisson log-likelihood of form

$$\ell(\theta_0, \boldsymbol{\theta}) = \sum_{j=1}^{N} \left[y_j(\theta_0 + \boldsymbol{Z}_j'\boldsymbol{\theta}) - \exp\left(\theta_0 + \boldsymbol{Z}_j'\boldsymbol{\theta}\right) \right] w_j,$$

where θ_0 is the intercept and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)'$ is a vector of q regression coefficients for covariates $\boldsymbol{Z}_j = (Z_{j1}, Z_{j2}, \ldots, Z_{jq})'$ that include spatial covariates and interaction terms. For simplicity the covariates are standardized to have zero mean and unit variance. To simultaneously select and estimate $\boldsymbol{\theta}$, we incorporate regularization into the above log-likelihood function. Specifically, we maximize the penalized likelihood function

$$\frac{1}{N}\ell(\theta_0, \boldsymbol{\theta}) - \lambda P(\boldsymbol{\theta}) \tag{9}$$

with respect to $(\theta_0, \boldsymbol{\theta})$, where $P(\cdot)$ is a penalty function.

We here focus on three specific forms for the penalty function. Zou and Hastie (2005) proposed an elastic net penalty P_E so that the selection and estimation procedure becomes

$$\max_{(\theta_0,\boldsymbol{\theta})\in\mathbf{R}^{q+1}} \left[\frac{1}{N}\ell(\theta_0,\boldsymbol{\theta}) - \lambda P_E(\boldsymbol{\theta})\right], \;\; P_E(\boldsymbol{\theta}) = \sum_{k=1}^{q}\left[\frac{1}{2}(1-a)\theta_k^2 + a|\theta_k|\right], \tag{10}$$

where λ is a tuning (or smoothing) parameter. The function P_E balances between the ridge-regression penalty ($a = 0$) and the LASSO penalty ($a = 1$). Ridge regression shrinks the coefficients of correlated covariates towards each other, while the LASSO shrinks many coefficients to zero, leaving a small subset of nonzero coefficients. Therefore, the elastic-net penalty is particularly useful in the situation where $q \gg N$, or when there are many correlated covariates. Note that with $a = 1 - \epsilon$ for small $\epsilon > 0$ the elastic net performs much like the LASSO, but avoids any degeneracies or unstable behavior caused by extreme correlations.

It is known that the LASSO penalty procedure does not have the so-called *oracle* property of asymptotically performing as well as if the true underlying model were known (Fan and Li, 2001). It may also produce biased estimates of the large coefficients and suffer from a conflict between optimal prediction and consistent variable selection. To address these issues, Zou (2006) proposed the *adaptive LASSO* penalty

$$P_A(\boldsymbol{\theta}) = \sum_{k=1}^{q} |\theta_k|/|\hat{\theta}_k|^{\gamma}, \tag{11}$$

where $\hat{\theta}_k$ are the ordinary least square estimates and $\gamma > 0$ is a fixed number. Zou (2006) proved that the adaptive LASSO exhibits the oracle property, even in generalized linear models, under mild regularity conditions. Similar to the LASSO, it yields a near minimax-optimal estimator. It is worth noting that the oracle property does not necessarily yield optimal prediction performance, and the LASSO can still be advantageous in some difficult prediction problems.

3.2. Cyclical coordinate descent algorithm

To model spatial point processes, a large value of N is often needed for the Berman-Turner device to work properly. As a result, fitting such generalized linear models with regularization penalties can be computationally intensive. Therefore we adopt cyclical coordinate descent methods (Friedman et al., 2007, 2010), which can work remarkably efficiently on very large datasets and can take advantage of sparsity in the set of covariates. The algorithm for fitting a Poisson linear regression model with the elastic net penalty is presented below.

Since $\ell(\theta_0,\boldsymbol{\theta})$ is a concave function of the parameters, the Newton-Raphson algorithm for maximizing (10) amounts to the iteratively reweighted least squares (IRLS) method. Letting $(\tilde{\theta}_0,\tilde{\boldsymbol{\theta}})$ be current estimates of the parameters, we construct a quadratic approximation to the log-likelihood $\ell(\theta_0,\boldsymbol{\theta})$ using Taylor's expansion:

$$\ell_Q(\theta_0,\boldsymbol{\theta}) = -\frac{1}{2N}\sum_{j=1}^{N} v_j\left(y_j^* - \theta_0 - \boldsymbol{Z}_j'\boldsymbol{\theta}\right)^2 + C(\tilde{\theta}_0,\tilde{\boldsymbol{\theta}}),$$

where $C(\tilde{\theta}_0,\tilde{\boldsymbol{\theta}})$ is a constant, y_j^* are the working response values and v_j the weights updated in the IRLS procedure,

$$y_j^* = \tilde{\theta}_0 + \boldsymbol{Z}_j'\tilde{\boldsymbol{\theta}} + y_j/v_j - 1$$
$$v_j = w_j \exp\left(\tilde{\theta}_0 + \boldsymbol{Z}_j'\tilde{\boldsymbol{\theta}}\right).$$

For each value of the tuning parameter λ, an outer loop is created to compute ℓ_Q, and then coordinate descent is used to solve the penalized weighted least-squares problem in (10) with $\ell(\theta_0, \boldsymbol{\theta})$ replaced by $\ell_Q(\theta_0, \boldsymbol{\theta})$. More specifically, suppose we wish to partially optimize (10) with respect to θ_k and have estimates $\tilde{\theta}_0$ and $\tilde{\theta}_\ell$ for $\ell \neq k$. Donoho and Johnstone (1995) show that the coordinate-wise update is

$$\tilde{\theta}_k \leftarrow \frac{S\left(\sum_{j=1}^N v_j Z_{jk}(y_j - \tilde{y}_j^{(k)}), \lambda a\right)}{\sum_{j=1}^N v_j Z_{jk}^2 + \lambda(1-a)}, \tag{12}$$

where $\tilde{y}_j^{(k)} = \tilde{\theta}_0 + \sum_{\ell \neq k} Z_{j\ell}\tilde{\beta}_\ell$ is the fitted value excluding the contribution from Z_{jk}, and $S(z, \lambda)$ is the soft-thresholding operator with value

$$\text{sign}(z)(|z| - \lambda)_+ = \begin{cases} z - \lambda \text{ if } z > 0 \text{ and } \lambda < |z| \\ z + \lambda \text{ if } z < 0 \text{ and } \lambda < |z| \\ 0 \qquad \text{if } \lambda \geq |z|. \end{cases}$$

The details of this derivation can be found in Friedman et al. (2007). The update (12) is repeated for $k = 1, 2, \ldots, q$ until convergence. Since the intercept is not regularized, this means that $\hat{\theta}_0$ is the mean of the y_j's for all values of λ and a. Starting with the smallest value of the tuning parameter λ for which the entire vector $\hat{\boldsymbol{\theta}} = \mathbf{0}$, the solutions for a decreasing sequence of values of λ are obtained. With such a path of solutions in λ, the user can select a particular value of λ that gives the best prediction performance measured by, for instance, cross-validation. To implement adaptive LASSO shrinkage, we can simply let $a = 1$ and replace λ by $\lambda_k = \lambda/|\hat{\theta}_k|^\gamma$ in the procedure.

In summary, cyclical coordinate descent methods are a natural approach for solving convex problems. Each coordinate-descent step is fast, with an explicit formula for each coordinate-wise optimization. The method also exploits the sparsity of the model, and its computational speed both for large N and q are quite remarkable (Friedman et al., 2010).

4. SIMULATION STUDIES

We performed a simulation study to compare the three regularization methods: LASSO, adaptive LASSO and elastic net, when they were applied to modeling spatial point data. We generated twenty independent Gaussian random fields with exponential covariance functions to use as covariates. We chose for the intensity function a log-linear form, $\rho_\theta(s) = \exp\{\beta_0 + \boldsymbol{z}'(s)\boldsymbol{\beta}\}$ with $\beta_0 = 0$ and $\beta = (1, 2, 3, 4, 5, 0, \ldots, 0)'$. Thus only the first five covariates are in the true model. With this intensity function we simulated 200 spatial point patterns each from a Poisson process, a Strauss process and a Thomas process, using the `spatstat R` package (Baddeley and Turner, 2005). The Strauss process is a pairwise interaction process, while the Thomas process is a special case of the Neyman-Scott process. The Poisson point patterns were generated using `rpoispp` function in a $[0, 1] \times [0, 1]$ window and had about 6,000 points each. In a $[0, 500] \times [0, 1000]$ window, we used the `rStrauss` and `rThomas` functions to generate the Strauss and Thomas point patterns respectively. For the Strauss process we set $\gamma = 0.5$ for the interaction parameter and $r = 5$ for the range, which yields about 2,500 points in each realization. For the Thomas process we chose parameters $\kappa = 5 \times 10^{-4}$, $\alpha = 20$ and $\omega = 10$, corresponding to about 20 offspring points distributed around each of an average of 250 parent points, so that each realization has about 5,000 points.

We fitted the point process models to the simulated point patterns using all twenty covariates, without regularization and also with each of the three regularization methods. The ordinary

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
PPM	.0742	.0441	.0427	.0351	.0183	.0718	.0524	.0264	.0246	.0044
Lasso	.1181	.0552	.1106	.0695	.0315	.0290	.0159	.0066	.0079	.0017
Adaptive Lasso	.0806	.0464	.0485	.0372	.0180	.0247	.0131	.0042	.0042	.0006
Elastic Net	.1115	.0452	.1082	.0658	.0401	.0357	.0234	.0101	.0114	.0024
Method	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
PPM	.0868	.0482	.0163	.0227	.0035	.0665	.0482	.0311	.0251	.0054
Lasso	.0277	.0152	.0046	.0084	.0011	.0177	.0147	.0093	.0122	.0021
Adaptive Lasso	.0297	.0099	.0013	.0052	.0001	.0166	.0127	.0041	.0061	.0005
Elastic Net	.0423	.0237	.0066	.0118	.0015	.0256	.0221	.0159	.0170	.0029

TABLE 1: Simulation results of Poisson process: MSE of $\hat{\beta}_k$ for $k = 1, 2, \ldots, 20$, using ppm, LASSO, adaptive LASSO and elastic net.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Lasso	1.00	1.00	1.00	1.00	1.00	0.420	0.335	0.190	0.215	0.130
Adaptive Lasso	0.99	1.00	1.00	1.00	1.00	0.195	0.135	0.060	0.090	0.005
Elastic Net	1.00	1.00	1.00	1.00	1.00	0.455	0.450	0.270	0.290	0.175
Method	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
Lasso	0.095	0.180	0.115	0.285	0.250	0.310	0.245	0.365	0.495	0.090
Adaptive Lasso	0.025	0.055	0.020	0.075	0.020	0.145	0.065	0.090	0.115	0.025
Elastic Net	0.130	0.190	0.180	0.345	0.355	0.400	0.325	0.550	0.640	0.150

TABLE 2: Simulation results of Poisson process: Proportions of samples where $\hat{\beta}_k \neq 0$ for $k = 1, 2, \ldots, 20$, using LASSO, adaptive LASSO and elastic net.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
PPM	4.019	4.437	4.279	3.818	3.787	3.943	3.728	3.538	3.768	3.717
Lasso	1.295	3.024	6.459	7.086	7.628	.6191	.6246	.8532	.7138	.5884
Adaptive Lasso	3.053	3.967	4.526	3.945	3.823	2.700	2.457	2.561	2.506	2.509
Elastic Net	1.296	2.911	6.246	6.672	7.120	.7244	.6951	.8258	.8209	.6283
Method	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
PPM	3.977	4.346	4.472	3.651	3.503	4.484	3.076	4.514	4.033	3.994
Lasso	.6821	.8412	.9163	.6454	.4752	.8784	.4877	.8083	.6025	.7264
Adaptive Lasso	2.703	3.063	3.149	2.346	2.287	3.021	1.863	3.166	2.763	2.734
Elastic Net	.7318	.9633	1.000	.7071	.5271	.8411	.4861	.7953	.7056	.7254

TABLE 3: Simulation results of Strauss process: MSE of $\hat{\beta}_k$ for $k = 1, 2, \ldots, 20$, using ppm, LASSO, adaptive LASSO and elastic net.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Lasso	0.320	0.430	0.455	0.745	0.890	0.310	0.295	0.265	0.290	0.265
Adaptive Lasso	0.695	0.725	0.815	0.945	0.975	0.650	0.610	0.565	0.650	0.580
Elastic Net	0.375	0.480	0.495	0.785	0.915	0.340	0.310	0.295	0.330	0.305
Method	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
Lasso	0.270	0.300	0.280	0.245	0.305	0.340	0.260	0.310	0.235	0.265
Adaptive Lasso	0.620	0.640	0.605	0.570	0.610	0.645	0.645	0.605	0.580	0.580
Elastic Net	0.300	0.335	0.315	0.280	0.315	0.350	0.270	0.325	0.240	0.280

TABLE 4: Simulation results of Strauss process: Proportions of samples where $\hat{\beta}_k \neq 0$ for $k = 1, 2, \ldots, 20$, using LASSO, adaptive LASSO and elastic net.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
KPPM	1.714	2.398	1.757	2.336	1.955	1.870	1.632	2.429	1.821	1.819
Lasso	0.951	1.919	3.704	2.619	5.774	0.280	0.200	0.469	0.275	0.300
Adaptive Lasso	1.045	2.429	3.421	2.189	3.620	0.480	0.267	0.796	0.375	0.446
Elastic Net	0.966	1.839	3.575	2.471	5.780	0.271	0.210	0.497	0.314	0.321
Method	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
KPPM	1.533	1.926	1.992	1.692	2.188	2.629	1.973	2.136	1.899	1.809
Lasso	0.147	0.284	0.400	0.278	0.383	0.499	0.397	0.451	0.354	0.359
Adaptive Lasso	0.235	0.431	0.584	0.373	0.678	0.823	0.562	0.673	0.514	0.492
Elastic Net	0.193	0.327	0.435	0.295	0.424	0.507	0.429	0.479	0.423	0.383

TABLE 5: Simulation results of Thomas process: MSE of $\hat{\beta}_k$ for $k = 1, 2, \ldots, 20$, using kppm, LASSO, adaptive LASSO and elastic net.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Lasso	0.310	0.750	0.800	0.975	0.980	0.175	0.110	0.035	0.175	0.145
Adaptive Lasso	0.285	0.725	0.795	0.980	0.985	0.145	0.105	0.025	0.095	0.115
Elastic Net	0.335	0.765	0.835	0.980	0.985	0.185	0.110	0.050	0.150	0.150
Method	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
Lasso	0.135	0.145	0.165	0.130	0.050	0.315	0.185	0.135	0.095	0.160
Adaptive Lasso	0.105	0.125	0.150	0.125	0.040	0.280	0.130	0.110	0.100	0.095
Elastic Net	0.150	0.165	0.180	0.155	0.065	0.360	0.195	0.160	0.120	0.170

TABLE 6: Simulation results of Thomas process: Proportions of samples where $\hat{\beta}_k \neq 0$ for $k = 1, 2, \ldots, 20$, using LASSO, adaptive LASSO and elastic net.

regression models were fitted using `ppm` function for the Poisson and Strauss point patterns and `kppm` function for the Thomas point patterns. The regularized models were fitted using modified internal functions in `spatstat` and `glmnet` (Friedman et al., 2010). We evaluated the performance of these methods using the mean squared error (MSE) of the estimated coefficients as well as their selection probabilities, i.e. the proportions of the fitted models where the coefficient

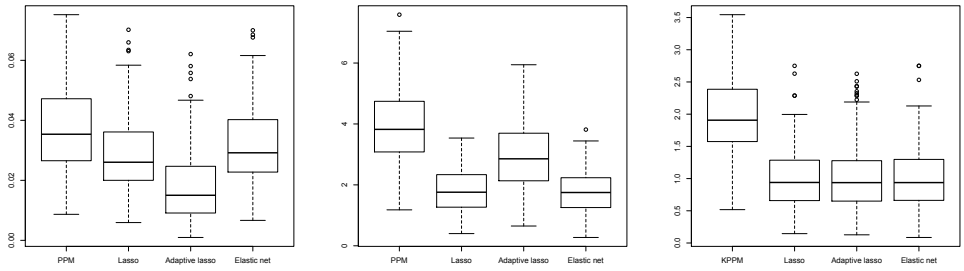

FIGURE 1: The distributions of overall MSEs for Poisson (left), Strauss (middle) and Thomas (right) processes.

estimates are nonzero.

The results are shown in Tables 1 to 6, two tables each for the Poisson, Strauss and Thomas models, showing the MSEs of the estimated regression coefficients and the selection probabilities of nonzero coefficients. We find, from Table 1, that the non-regularized method yields slightly smaller MSEs for the nonzero coefficients but much bigger MSEs on the zero coefficients, compared with the regularized methods. Among the three regularization methods, adaptive LASSO provides the best estimation results with smallest MSEs on both the nonzero and zero coefficients. The left plot in Figure 1 shows the distributions of overall MSEs of each method (i.e. the sum of the MSEs of the twenty covariates), where we can clearly see that adaptive LASSO outperforms the other methods in terms of the MSE. With respect to selection accuracy for the three regularization methods, Table 2 shows that the LASSO and elastic net select all nonzero coefficients 100% of the time while adaptive LASSO selects the smallest nonzero coefficient β_1 99% of the time and the other nonzero coefficients 100% of time. For the zero coefficients, the adaptive LASSO has significantly smaller selection probabilities than the other two methods. These simulation results suggest that adaptive LASSO has the best performance for the Poisson point patterns.

The results for the Strauss process are similar, although the individual MSEs become larger overall. Comparatively speaking, the MSEs are larger for the nonzero coefficients with regularization than without regularization, but smaller for the zero coefficients (see Table 3). The adaptive LASSO seems to perform worse than the LASSO and elastic net methods. This can also be seen from the middle plot in Figure 1, where we find that overall MSEs for LASSO and elastic net are similar and quite a bit smaller than both the adaptive LASSO and the non-regularized method (with adaptive LASSO yielding slightly smaller overall MSEs of the two). In terms of the selection probabilities (Table 4), the smaller nonzero coefficients are poorly selected, with e.g. β_1 selected about 30% of the time, not much higher than for the zero coefficients. The regularization methods fair better for the larger coefficients, selecting them from 70% to 90% of the time. Adaptive LASSO seems to perform differently and worse than LASSO and elastic net: although it selects nonzero coefficients with the highest probabilities of the three methods, it also selects zero coefficients about 60% of the time.

For the Thomas point patterns, we find that, compared to the non-regularized method, the three regularization methods yield smaller MSEs on the small and zero coefficients, but bigger MSEs on the relatively large coefficients (see Table 5). The reason may be due to the poor estimation of zero coefficients by `kppm`, whose estimates are included in the penalty function of adaptive LASSO (see (11)). The overall MSEs in Figure 1 show that the three regularization methods perform equally well and outperform the non-regularized method. Regarding selection accuracy, adaptive LASSO consistently selects the zero coefficients less often than LASSO and elastic net do. However, the three methods all poorly select the smaller nonzero coefficients: for

example, they select $\beta_1 = 1$ only about 30% of the time. With respect to the estimates of κ and ω, their MSEs (not shown) are small and quite close across all four methods. It appears that no method has an oracle property when the point patterns are clustered, but the regularization methods clearly outperform the non-regularized method.

As suggested by the referees, we also performed a couple of variations to the above simulation study. The detailed results are not shown, but we report a summary here. First, we re-did simulations for the Thomas process using realizations with about 1000 points, i.e. with smaller number of points than described above. We find that the relative performance between the methods still hold, with the regularized methods yielding slightly larger MSE's for the nonzero coefficients, but much smaller MSE's for the zero coefficients. The individual MSEs were larger in absolute values than with the larger sample sizes, as expected.

We also did a simulation study where the covariates are less sparse, i.e. with fewer number of zero coefficients. Specifically, we kept the 5 non-zero coefficients but included only 5 zero coefficients instead of the 15 zero coefficients previously used. The results are similar. The regularization models outperform the non-regularized model in estimating the zero and small coefficients but underperform in estimating relatively big coefficients. We note that the outperformance of the regularization methods in estimating zero coefficients becomes less when the covariates are less sparse. This is not surprising, since the need for regularization is less when there is less sparsity.

5. APPLICATIONS TO DATA

5.1. Barro Colorado Island (BCI) trees

For this analysis we used the 1995 BCI census of trees and considered in particular the positions of live specimens of the three species: *Acalypha diversifolia* (536 trees), *Lonchocarpus heptaphyllus* (836 trees) and *Capparis frondosa* (3300 trees). These three species are known to have different seed dispersals: for *Acalypha* the seeds are dispersed by exploding capsules, for *Lonchocarpus* by wind and for *Capparis* by birds and mammals. It is hypothesized that the modes of seed dispersal are reflected in the spatial patterns of tree locations with tight clusters for exploding capsules, loose clusters for bird and mammal dispersal and intermediate clustering for species with wind dispersal (Seidler and Plotkin, 2006). Following Waagepetersen and Guan (2009), we considered an inhomogeneous Thomas process for each species with soil minerals and topological attributes as covariates. The two-step estimation procedure in Waagepetersen and Guan (2009) was implemented using `spatstat R` package. More specifically, we used `kppm` function to obtain regression estimates $\hat{\boldsymbol{\beta}}$ and clustering estimates $(\hat{\kappa}, \hat{\omega})$. The asymptotic standard errors of $\hat{\boldsymbol{\beta}}$ were computed by `vcov.kppm` function and those of $(\hat{\kappa}, \hat{\omega})$ were obtained using a parametric bootstrap method. We used the three regularization methods, LASSO, adaptive LASSO and elastic net in turn to select non-zero covariates and estimated their coefficients. In each case, the tuning parameter was chosen by cross-validation. The results are shown in Table 7 and 8.

We found that although `kppm` provides estimates for all the covariates, many of these are not significant at the 0.05 level based on the asymptotic standard errors. In particular, none of the covariates are significant for *Acalypha diversifolia*, while three and one covariates are signficant for *Lonchocarpus heptaphyllus* (P, Zn and altitude) and *Capparis frondosa* (K) respectively.

The three variable selection methods give generally very similar results. and roughly matches what is found using `kppm`. For *Acalypha diversifolia*, all three variable selection methods returned 0 for all the coefficients. For the other two species, the adaptive LASSO was slightly more sparse than the LASSO and elastic net methods, yielding one or two fewer non-zero coefficients. Using the results of adaptive LASSO, we find that *Lonchocarpus* trees occurred more in

	Acalypha diversifolia				*Lonchocarpus heptaphyllus*				*Capparis frondosa*			
	kppm	Lasso	A. Lasso	E. net	kppm	Lasso	A. Lasso	E. net	kppm	Lasso	A. Lasso	E. net
Al	-.04 (.17)	-	-	-	.01 (.14)	-	-	-	.03 (.09)	-	-	-
B	-.22 (.275)	-	-	-	-.08(.215)	-	-	-	.11 (.143)	-	-	-
Ca	.36 (.355)	-	-	-	.36 (.285)	-	-	-	-.06 (.143)	-	-	-
Cu	-.06 (.21)	-	-	-	.12 (.175)	-	-	-	-.00 (.115)	-	-	-
Fe	.15 (.158)	-	-	-	.13 (.128)	-	-	-	-.03 (.088)	-	-	-
K	.21 (.25)	-	-	-	-.24 (.208)	-	-	-	*.27 (.133)	.039	.159	.044
Mg	-.18 (.265)	-	-	-	.32 (.218)	-	-	-	-.14 (.14)	-	-	-
Mn	.00 (.17)	-	-	-	-.05 (.14)	-	-	.008	-.09 (.095)	-	-	-
P	.05 (.15)	-	-	-	*-.29 (.145)	-.102	-.077	-.111	.01 (.085)	-	-	-
Zn	-.16 (.228)	-	-	-	*-.51 (.253)	-.042	-.253	-.068	.01 (.12)	-	-	-
N	.12 (.135)	-	-	-	-.01 (.118)	-	-	-	.04 (.08)	-	-	-
N.min	.08 (.168)	-	-	-	-.18 (.138)	-.169	-	-.155	.17 (.095)	.052	.018	.054
pH	.06 (.153)	-	-	-	-.18 (.125)	-.060	-	-.066	.04 (.085)	.053	-	.054
Altitude	.14 (.158)	-	-	-	*.29 (.143)	-	-	-	.15 (.09)	.077	.117	.080
Slope	.11 (.115)	-	-	-	.08 (.09)	-	-	-	-.06 (.063)	-	-	-

TABLE 7: Barro Colorado Island data analysis: Point estimates of regression coefficients and their standard errors (in parenthesis) for Acalypha, Lonchocarpus and Capparis trees; the estimates labeled with '$*$' are statistically significant at 5% level; for Lasso, adaptive Lasso (A. Lasso) and elastic net (E. Net), the estimates shown are for variables selected by the procedure.

Species	Method	$\kappa \times 10^4$	ω
Acaldi	KPPM	7.612 (5.924, 10.95)	2.843 (2.176, 3.391)
	Lasso	4.253 (2.905, 5.973)	5.812 (4.808, 7.666)
	A. Lasso	5.646 (3.981, 7.149)	4.216 (3.495, 5.249)
	E. Net	4.253 (2.807, 5.943)	5.812 (4.691, 7.418)
Loncla	KPPM	5.311 (3.787, 8.612)	7.430 (5.436, 8.791)
	Lasso	4.082 (2.740, 6.300)	9.487 (7.737, 12.26)
	A. Lasso	4.177 (2.951, 6.843)	9.225 (7.132, 11.92)
	E. Net	4.007 (2.601, 6.433)	9.571 (7.501, 12.24)
Cappfr	KPPM	9.484 (6.226, 13.69)	11.76 (9.866, 13.96)
	Lasso	6.271 (4.034, 9.813)	14.45 (11.77, 18.56)
	A. Lasso	6.821 (4.043, 9.426)	15.63 (12.40, 19.04)
	E. Net	6.709 (4.347, 10.06)	14.67 (11.52, 17.36)

TABLE 8: Barro Colorado Island data analysis: Point estimates and 95% confidence intervals (in parenthesis) of clustering parameters for Acalypha (Acaldi), Lonchocarpus (Loncla) and Capparis (Cappfr) trees, given by KPPM, Lasso, adaptive Lasso (A. Lasso) and elastic net (E. net) methods.

areas with lower levels of Phosphorus (P) and Zinc (Zn). while the presence of *Capparis* trees is positively correlated with Potassium (K), minimum Nitrogen (N.min) and altitude.

The three regularizaton methods yield similar estimates of ω, slightly larger than those obtained with `kppm`. This quantity ω is the variable of interest in terms of capturing the degree of clustering of the different species of trees. The different seed dispersal mechanisms employed by the trees should affect the clustering of these trees which in turn should be reflected in the relative sizes of the ω estimates. We find that this is indeed the case, with *Acalypha diversifolia* trees, which disperses seeds using exploding capsules, having the smallest estimate of ω, followed by *Lonchocarpus heptaphyllus*, dispersed by wind and *Capparis frondosa*, dispersed by animals.

The full BCI data consist of censuses taken over 25 years. The regularized procedures described in this paper should be able to accommodate the covariate selection for the full data set, given an appropriate space-time model for ρ. As an alternative, the covariate selection can be performed for each year separately and the sets of selected covariates are examined to see how informative covariates change over time. However, these topics are beyond the scope of this paper.

5.2. NYC fast food restaurants (FFR)

Since the locations of FFRs are expected to be affected by the locations of other nearby FFRs, we fit a pairwise interaction point process model to the data set. To describe the interaction between points x_i and x_j, we use a piecewise constant potential function $h(x_i, x_j)$ that is defined as

$$h(x_i, x_j) = \sum_k h_k I\{\|x_i - x_j\| \in I_k\},$$

where there is an interaction of h_k between points x_i and x_j whose distance separation lies in $I_k = [r_{k-1}, r_k)$ with r_0 being zero. Here we used two values of r, $r_1 = 400$m and $r_2 = 800$m, which are the two distance cutoffs that we believe are appropriate for the interactions between FFR restaurants. The use of piecewise constant interaction terms was first suggested by Takacs (1986). Then, together with a parameter b controlling the intensity of the points, the pseudo-

Covariates	PPM	Lasso	A. Lasso	E. Net
AGE	.088 (-.006, .237)	-	-	-
MHI	-.007 (-.261, .247)	-.031	-	-.038
POP	*-.159 (-.287, -.032)	.008	-	.012
LUNCH	4.23 (-17.2, 25.7	-	-	-
DINNER	*-16.3 (-22.6, -10.1)	-	-14.67	-
SNACKS	11.9 (-6.9, 30.7)	-	14.50	-
BLK	.002 (-.136, .140)	-	-	-
WHT	*-.230 (-.417, -.042)	-.169	-.176	-.155
NATIONAL	*.235 (.185, .285)	.258	.271	.259
LOCALS	-.037 (-.086, .011)	-	-	-
DIST	-.137 (-.280, .006)	-.046	-	-.042
COUNT	.012 (-.099, .123)	.103	-	.104
ZONE: Residential Medium	.026 (-.104, .15)	.016	-	.015
ZONE: Residential High	.113 (-.0006, .226)	.059	-	.057
ZONE: Commercial	*.152 (.056, .247)	.100	-	.096
I(400)	.086 (-.005, .177)	.171	.065	.169
I(800)	*.073 (.036, .111)	.092	.015	.090

TABLE 9: NYC fast restaurant data analysis: Point estimates of regression coefficients and interaction parameters and their 95% confidence intervals; For the PPM estimates, those labeled with '$*$' are statistically significant at 5% level. For Lasso, adaptive Lasso (A. Lasso) and elastic net (E. Net), the estimates shown are for variables selected by the procedure.

likelihood of this pairwise interaction point process is given by formula in (7).

We study how the intensity of FFRs is related to Census variables based on the census block group that the FFRs fall in. In particular, we use the median age (AGE), median household income (MHI), population density (POP), average household expenditure on food away from home (LUNCH, DINNER, SNACKS), percent black (BLK), percent white (WHT). In addition, we also have a categorical variable for the zone type as classified by the NYC Department of City Planning (ZONE), as well as the number of National-chain and local FFR stores in the block group (NATIONAL and LOCAL). The ZONE variable corresponds to residential classifications of low, medium and high density as well as a category comprising manufacturing and commercial areas. In order to study the inter-relationship between FFR and school locations, we also include two variables associated with proximity of schools: DIST which is the distance to the nearest school, and COUNT which is the number of schools within a 400m radius (roughly walking distance).

We tried all three regularization methods (LASSO, adaptive LASSO and elastic net) in our selection procedure, with the tuning parameter selected using cross-validation. We also fit a model without regularization, using `ppm` function in `spatstat R` package. The results are shown in Table 9. Focusing on the regularization methods, we find that the elastic net and LASSO agree on all the selected variables and the coefficient estimates are also very similar. The Adaptive LASSO differs from these two methods, dropping the school variables and variables associated with ZONE. It, however, retains the DINNER and SNACKS variables. We note, though, that the variables LUNCH, DINNER and SNACKS are highly positively correlated with correlations greater than 0.99, suggesting that the elastic net may yield more reliable estimates.

We find that COUNT is positively associated with FFR intensity, i.e. there will be more FFRs if there are more schools nearby. Similarly, an increase in the distance from the nearest school (DIST) is associated with a drop in FFR intensity. This corroborates the finding of Kwate and Loh (2010) using K function analysis. Note that for the adaptive LASSO method, neither of these two variables are retained. Neither of these variables is significant in the PPM method as well.

For all regularization and PPM methods, we find that a higher number of National FFR stores in the block group is associated with higher FFR intensity. Similarly, we find that percent white is significant and negatively associated with FFR intensity. This is in agreement with relevant literature on this subject. From both the LASSO and Elastic Net estimates, we find that increasing FFR intensity is associated with higher residential density, with the greatest FFR intensity associated with commercial areas. This last category is also significant for `ppm` as well. Commercial areas tend to be either tourist areas or areas where there are many workers and hence it is not unusual for such areas to have a lot of eating places, including fast food restaurants.

We also note that with the elastic net and LASSO methods, mean household income (MHI) is retained in the model with higher MHI associated with lower FFR intensity. This is again in agreement with the literature. However, this variable is not significant in the PPM method and also not selected by the adaptive LASSO method. Finally, the two parameters used in the piecewise constant interaction function are selected by all the regularization methods, which suggests that there is some repulsion between FFR locations within 400m and, to a lesser degree, within 800m.

6. DISCUSSION

We introduced a simple method for incorporating existing variable selection methods in the procedures for fitting models to spatial point data. By setting up the non-regularized model fitting procedure into a generalized linear model framework, it is straightforward to incorporate in the model the penalty terms of many existing regularization methods. A cyclical coordinate descent algorithm allows for a fast computation. We examined the use of the LASSO, adaptive LASSO and elastic net methods in a simulation study. In our simulations, we found that adaptive LASSO method performed the best for the Poisson process, which agrees with what was found in Thurman and Zhu (2014). For the Strauss and Thomas processes, adaptive LASSO, however, tends to give worse estimation on small regression coefficients than LASSO and elastic net do. As a result, there is not a uniformly best method, and it appears more appropriate to use LASSO or elastic net when the covariates are highly sparse. More investigation is needed to examine if the relative performance we found hold more generally. In our applications, we find that the clustering parameter estimates for the 3 species of trees agree qualitatively with what is expected based on their different seed dispersal methods. For the locations of FFR restaurants in NYC, we find decreased intensity of FFRs with increased percent White in agreement with findings in other studies.

The proposed variable selection method is quite a general framework, under which a variety of spatial point processes and penalty functions could be used. This generality comes from using the pseudo-likelihood together with the Berman-Turner device, which converts the objective function into a form similar to that for fitting generalized linear models. Besides Poisson, pairwise interaction and cluster point processes, the same procedure can be used with more complicated spatial point process models such as area-interaction models, and with marked point processes. Moreover, other regularization methods could be used as long as they can be applied to generalized linear models. Thus, for example, grouped LASSO (Yuan and Lin, 2006) and nonconcave penalties such as SCAD (Fan and Li, 2001) could be used in addition to the ones we considered here.

Future work will consist of investigating the theoretical properties of the proposed procedure as well as studying its performance under additional scenarios. Many theoretical results have been derived for the various shrinkage methods in non-spatial settings. It will be interesting to see which of these results carry over to spatial point modeling.We will first consider tuning parameter selection and how this affects variable selection consistency, i.e. whether the set of selected covariates tends asymptotically to the true set of covariates. Work is also in progress to convert the R code developed for this work into an `R` package for wider use.

Acknowledgment

We thank associate editor and two referees for their great comments on our manuscript, which has significantly improved its quality. Yu Yue's research has been supported by PSC-CUNY research award #66138-00 44.

BIBLIOGRAPHY

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns. *Australian and New Zealand Journal of Statistics* **42**, 283–322.

Baddeley, A. J. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software* **12**, 1–42.

Berman, M. and Turner, T. R. (1992). Approximating point process likelihoods with glim. *Applied Statistics* 31–38.

Besag, J. (1975). Statistical analysis of non-lattice data. *The statistician* 179–195.

Besag, J. (1977). Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute* **47**, 77–92.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. 2nd edition, Arnold, London.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

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

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Chichester.

Kwate, N. O. and Loh, J. M. (2010). Separate and unequal: the influence of neighborhood and school characteristics on spatial proximity between fast food and schools. *Preventive Medicine* **51**, 153–156.

Kwate, N. O., Yau, C. Y., Loh, J. M. and Williams, D. (2009). Inequality in obesigenic environments: Fast food density in New York City. *Health and Place* **15**, 364–373.

Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall, New York.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renner, I. W. and Warton, D. I. (2013). Equivalence of maxent and poisson point process models for species distribution modeling in ecology. *Biometrics* **69**, 274–81.

Ripley, B. D. (1991). *Statistical inference for spatial processes*. Cambridge University Press, Cambridge.

Schoenberg, F. P. (2005). Consistent parametric estimation of the intensity of a spatial-temporal point process. *Journal of Statistical Planning and Inference* **128**, 79–93.

Seidler, T. G. and Plotkin, J. B. (2006). Seed dispersal and spatial pattern in tropical trees. *PLoS Biology* **4**, e344.

Stoyan, D. and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields*. John Wiley, New York.

Strauss, D. J. (1975). A model for clustering. *Biometrika* **63**, 467–475.

Takacs, R. (1986). Estimator for the pair potential of a Gibbsian point process. *Statistics: A Journal of Theoretical and Applied Statistics* **17**, 429–433.

Thurman, A. L. and Zhu, J. (2014). Variable selection for spatial poisson point processes via a regularization method. *Statistical Methodology* **17**, 113–125.

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

Waagepetersen, R. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics* **63**, 252–258.

Waagepetersen, R. and Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Association Series B* **71**, 685–702.

Waagepetersen, R. P. (2008). Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika* **95**, 351–363.

Yau, C. Y. and Loh, J. M. (2010). A generalization of the Neyman-Scott process. *Statistica Sinica* **22**, 1717–1736.

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

Yue, Y. and Loh, J. M. (2011). Bayesian semiparametric intensity estimation for inhomogeneous spatial point processes. *Biometrics* **67**, 937–946.

Yue, Y. R. and Loh, J. M. (2013). Bayesian nonparametric estimation of pair correlation function for inhomogeneous spatial point processes. *Journal of Nonparametric Statistics* **25**, 463–474.

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

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