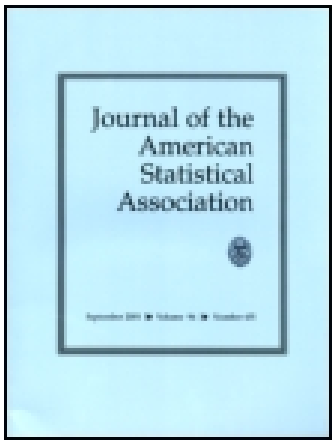
This article was downloaded by: [141.211.66.220]
On: 02 October 2014, At: 11:33
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Functional Principal Component Analysis of Spatiotemporal Point Processes With Applications in Disease Surveillance

Yehua Li & Yongtao Guan

Accepted author version posted online: 18 Feb 2014.Published online: 02 Oct 2014.

To cite this article: Yehua Li & Yongtao Guan (2014) Functional Principal Component Analysis of Spatiotemporal Point Processes With Applications in Disease Surveillance, Journal of the American Statistical Association, 109:507, 1205-1215, DOI: 10.1080/01621459.2014.885434

To link to this article: http://dx.doi.org/10.1080/01621459.2014.885434

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Functional Principal Component Analysis of Spatiotemporal Point Processes With Applications in Disease Surveillance

Yehua LI and Yongtao GUAN

In disease surveillance applications, the disease events are modeled by spatiotemporal point processes. We propose a new class of semiparametric generalized linear mixed model for such data, where the event rate is related to some known risk factors and some unknown latent random effects. We model the latent spatiotemporal process as spatially correlated functional data, and propose Poisson maximum likelihood and composite likelihood methods based on spline approximations to estimate the mean and covariance functions of the latent process. By performing functional principal component analysis to the latent process, we can better understand the correlation structure in the point process. We also propose an empirical Bayes method to predict the latent spatial random effects, which can help highlight hot areas with unusually high event rates. Under an increasing domain and increasing knots asymptotic framework, we establish the asymptotic distribution for the parametric components in the model and the asymptotic convergence rates for the functional principal component estimators. We illustrate the methodology through a simulation study and an application to the Connecticut Tumor Registry data. Supplementary materials for this article are available online.

KEY WORDS: Composite likelihood; Functional data; Latent process; Semiparametric methods; Spatiotemporal data; Splines; Strong mixing.

Downloaded by [141.211.66.220] at 11:33 02 October 2014

1. INTRODUCTION

Spatiotemporal point patterns commonly arise from many fields including ecology, epidemiology, and seismology (e.g., Brix and Møller 2001; Schoenberg 2003; Diggle 2006). The log-Gaussian Cox processes (LGCPs), first introduced by Møller, Syversveen, and Waagepetersen (1998) in the spatial case and later on extended to the spatiotemporal setting by Brix and Møller (2001) and Brix and Diggle (2001), provide a wide class of useful models for modeling such kind of data. For a typical spatiotemporal LGCP, its intensity function is assumed to be a log-linear model of some latent spatiotemporal Gaussian process, where the mean of the process may depend on some observed covariates. Borrowing ideas from recent developments in functional data analysis (Ramsay and Silverman 2005), we model the latent temporal process at any fixed spatial location as a functional process with a standard functional principal component expansion. We allow the functional principal component scores at different locations to be spatially correlated. The proposed model can accommodate both nonparametric temporal trend and spatiotemporal correlations in the point process.

In functional data analysis (FDA), the data considered are collections of curves, which are usually modeled as independent realizations of a stochastic process. Some recent papers on this topic include Yao, Müller, and Wang (2005a,b), Hall and Hosseini-Nasab (2006), Hall, Müller, and Wang (2006), and Li and Hsing (2010a,b). Di et al. (2009) and Zhou et al. (2010) studied multilevel functional data, where functional data at the lower level of the hierarchy are allowed to be temporally correlated. All the aforementioned papers considered only Gaussian type of functional data. Recently, Hall, Müller, and Yao (2008) studied generalized longitudinal data, where the non-Gaussian longitudinal trajectories are linked to some Gaussian latent processes through a nonlinear link function and these latent random processes are modeled as functional data. For such non-Gaussian longitudinal data, Hall et al. proposed a nonparametric estimation procedure based on a delta method, which is an approximation by ignoring the higher order influence of the latent processes. There has also been some recent work on functional data modeling of point processes, including Bouzas et al. (2006), Illian et al. (2006), and Wu, Müller, and Zhang (2013). These authors considered data with independent replicates of the point process and modeled a summary measure of the point process (e.g., the intensity function or the L-function) as functional data.

To develop FDA tools for spatiotemporal point processes, we encounter many new challenges and our proposed method is hence different from those in the literature in a number of ways. First, in most FDA papers in the literature, the data consist of n independent units (subjects). In our settings, however, there is only one realization of the spatiotemporal process, and the data are correlated both spatially and temporally. Second, unlike the scenarios considered in the classic FDA literature where the functional trajectories can be directly observed, the functional data in our setting are latent processes that determine the rate of events. To estimate the covariance structure of the process, we propose a novel method based on composite likelihood and spline approximation. We develop asymptotic properties of our

Yehua Li is Associate Professor, Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA 50011 (E-mail: yehuali@iastate.edu). Yongtao Guan is Professor, Department of Management Science, University of Miami, Coral Gables, FL 33124 (E-mail: yguan@bus.miami.edu). This research has been partially supported by NIH grant 1R01CA169043 and NSF grants DMS-0845368, DMS-1105634, and DMS 1317118. The Connecticut Tumor Registry is supported by Contract No. HHSN2612013000191 between the National Cancer Institute and State of Connecticut Department of Public Health. This study was approved by the Connecticut Department of Public Health (CDPH). Data used in this article were obtained from the CDPH. The authors assume full responsibility for analysis and interpretation of these data.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jasa.

© 2014 American Statistical Association

DOI: 10.1080/01621459.2014.885434

Downloaded by [141.211.66.220] at 11:33 02 October 2014

estimators under an increasing domain asymptotic framework. Third, we perform spatial prediction of the latent principal component scores using an empirical Bayes method. These predicted spatial random effects can be put into maps to highlight hot areas with unusually high event rates or increasing trends in event rates. Such information can be valuable to government agencies when making public health polices.

Our work is motivated by cancer surveillance data collected by the Connecticut Tumor Registry (CTR). The CTR is a population-based resource for examining cancer patterns in Connecticut, and its computerized database includes all reported cancer cases diagnosed in Connecticut residents from 1935 to the present. Our primary interest here is to study the spatiotemporal pattern of pancreatic cancer incidences based on 8230 pancreatic cancer cases in the CTR database from 1992 to 2009. The residential addresses and time of diagnosis are both available and are assumed to be generated by a spatiotemporal point process.

The rest of the article is organized in the following way. We introduce the model assumptions in Section 2 and propose our estimation procedures in Section 3. Then, we study the asymptotic properties of the proposed estimators in Section 4. The proposed methods are tested by a simulation study in Section 5 and are applied to the CTR data in Section 6. Assumptions for our asymptotic theory are collected in the appendix. All technical proofs and implementation details, including variance estimation, model selection, and model diagnostic, are provided in the online supplementary material.

2. MODEL ASSUMPTIONS

Let N denote a spatiotemporal point process that is observed on $W = D \otimes T$, where $D \subset \mathbb{R}^2$ is a spatial domain and T is a time domain. Let X be an L^2 Gaussian random field on W. We assume that conditional on X, N is a Poisson process with an intensity function $\lambda(\boldsymbol{s}, t)$ given by

$$g\{\lambda(\boldsymbol{s}, t)\} = \boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\beta} + X(\boldsymbol{s}, t), \tag{1}$$

where g is a known link function such that $g^{-1}(\cdot)$ is nonnegative, $\boldsymbol{Z}(\boldsymbol{s}, t)$ is a d-dimensional covariate vector, and X represents spatiotemporal random effects that cannot be explained by $\boldsymbol{Z}$.

In this article, we will focus on the log-link function, that is, $g(\cdot) = \log(\cdot)$. The model given in (1) then becomes an LGCP model. In point process literature, the effect of the covariate $\boldsymbol{Z}(\boldsymbol{s}, t)$ is often assumed to be parametric (Møller and Waagepetersen 2007), although nonparametric approaches have also been recently proposed (Guan 2008a; Guan and Wang 2010). Similarly, a parametric model is generally used for the covariance structure of the latent process $X(\boldsymbol{s}, t)$. For example, Brix and Diggle (2001) assumed a covariance structure from a class of Ornstein–Uhlenbeck processes, while Diggle, Rowlingson, and Su (2005) used a parametric covariance model that is stationary both in space and time. However, we are not aware of any existing literature that models the latent process nonparametrically in a spatiotemporal log-Gaussian Cox process as what we will do next.

For a fixed location $\boldsymbol{s}$, $X(\boldsymbol{s}, t)$ can be considered as an L^2 Gaussian process on T, and hence by the standard Karhunen–Loève expansion (Ash and Gardner 1975),

$$X(\boldsymbol{s}, t) = \mu(t) + \sum_{j=1}^{p} \xi_j(\boldsymbol{s})\psi_j(t), \tag{2}$$

where $\mu(t) = \mathrm{E}\{X(\boldsymbol{s}, t)\}$ with the expectation taken over all locations, $\psi_j(\cdot)$'s are orthonormal functions, and $\xi_j(\cdot)$'s are independent spatial Gaussian random fields. We assume that $\xi_j(\boldsymbol{s})$ is a zero-mean random field with variance ω_j and covariance function, $C_j(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathrm{cov}\{\xi_j(\boldsymbol{s}_1), \xi_j(\boldsymbol{s}_2)\}$, for $j = 1, 2, \ldots, p$. The functions $\psi_j(\cdot)$'s are called the eigenfunctions of process X. We assume that $\psi_j(\cdot)$'s are kept in a descending order of ω_j's, that is, $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_p > 0$. The number of principal components p can be ∞ in theory, but is often assumed to be finite for practical considerations.

The general covariance function of $X(\boldsymbol{s}, t)$ is

$$\begin{aligned} R(\boldsymbol{s}_1, \boldsymbol{s}_2, t_1, t_2) &= \mathrm{cov}\{X(\boldsymbol{s}_1, t_1), X(\boldsymbol{s}_2, t_2)\} \\ &= \sum_{j=1}^{p} C_j(\boldsymbol{s}_1, \boldsymbol{s}_2)\psi_j(t_1)\psi_j(t_2), \end{aligned} \tag{3}$$

which implies that $X(\boldsymbol{s}, t)$ is not necessarily stationary in t. Note that the aforementioned model coincides with the spatial coregionalization model commonly used for multivariate Gaussian random fields (Gelfand, Sherman, and Calvin 2004) and is not separable when $p > 1$. To connect with the FDA literature, it is helpful to consider the covariance function of the latent process $X(\boldsymbol{s}, t)$ at the same location $\boldsymbol{s}$. By setting $\boldsymbol{s}_1 = \boldsymbol{s}_2$, (3) is simplified to

$$\begin{aligned} R_T(t_1, t_2) &= \mathrm{cov}\{X(\boldsymbol{s}, t_1), X(\boldsymbol{s}, t_2)\} \\ &= \sum_{j=1}^{p} \omega_j\psi_j(t_1)\psi_j(t_2), \quad \boldsymbol{s} \in D, \quad t_1, t_2 \in T, \end{aligned} \tag{4}$$

where ω_j and $\psi_j(\cdot)$'s are the eigenvalues and eigenfunctions of $R_T(\cdot, \cdot)$. If a consistent estimator $\widehat{R}_T(\cdot, \cdot)$ exists, one can then estimate $\{\omega_j, \psi_j(\cdot)\}$ by an eigenvalue decomposition of $\widehat{R}_T$ using a standard functional data analysis approach (Ramsay and Silverman 2005).

We estimate the proposed model through the use of the first- and second-order intensity functions of N. Let $N(d\boldsymbol{s}, dt)$ denote the number of events in an infinitesimal window $(d\boldsymbol{s}, dt)$, and let $|d\boldsymbol{s}|$ and $|dt|$ denote the volumes of $d\boldsymbol{s}$ and dt, respectively. The marginal first-order intensity function, which characterizes the probability to observe an event at a given location and time, is defined as

$$\begin{aligned} \lambda(\boldsymbol{s}, t) = \lim_{|d\boldsymbol{s}|, |dt| \to 0} \frac{\mathrm{E}\{N(d\boldsymbol{s}, dt)\}}{|d\boldsymbol{s}||dt|} &= \mathrm{E}[\exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\beta} \\ &\quad + X(\boldsymbol{s}, t)\}] = \exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\beta} + \gamma(t)\}, \end{aligned} \tag{5}$$

where $\gamma(t) = \mu(t) + (1/2)\sum_{j=1}^{p} \omega_j\psi_j^2(t)$. In derivation of (5), we use the fact that $\mathrm{E}\{\exp(Y)\} = \exp(\mu + \sigma^2/2)$ for a $Y \sim \mathrm{Normal}(\mu, \sigma^2)$ and the covariance of $X(\boldsymbol{s}, t)$ in (4). Nonstationarity in the first-order intensity function can be modeled by including proper spatiotemporal covariates $\boldsymbol{Z}(\boldsymbol{s}, t)$. For example, in our disease surveillance application, nonstationarity in the cancer rate caused by spatially varying population level is accommodated by including population density as a covariate.

Downloaded by [141.211.66.220] at 11:33 02 October 2014

The second-order intensity function, which characterizes the correlation within the process, is defined as

$$\begin{aligned}\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2) &= \lim_{\substack{|d\boldsymbol{s}_1|,|d\boldsymbol{s}_2|,\\ |dt_1|,|dt_2|\to 0}} \frac{\mathrm{E}\{N(d\boldsymbol{s}_1,dt_1)N(d\boldsymbol{s}_2,dt_2)\}}{|d\boldsymbol{s}_1||dt_1||d\boldsymbol{s}_2||dt_2|}\\ &= \mathrm{E}[\exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s}_1,t_1)\boldsymbol{\beta}+\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s}_2,t_2)\boldsymbol{\beta}\\ &\quad + X(\boldsymbol{s}_1,t_1)+X(\boldsymbol{s}_2,t_2)\}]\\ &= \lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2)\exp\left\{\sum_{j=1}^{p} C_j(\boldsymbol{s}_1,\boldsymbol{s}_2)\right.\\ &\quad \left.\times\,\psi_j(t_1)\psi_j(t_2)\right\}, \end{aligned} \tag{6}$$

for $(\boldsymbol{s}_1,t_1)\neq(\boldsymbol{s}_2,t_2)$, where the last equality is a result of the Gaussian assumption for the principal component random processes $\xi_j(\boldsymbol{s})$. The Gaussian assumption is also commonly made in other settings such as generalized linear mixed models and spatial hierarchical models (Banerjee, Gelfand, and Carlin 2003).

Given the first- and second-order intensity functions, the pair correlation function (e.g., Møller and Waagepetersen 2004) for the point process is

$$\begin{aligned}\mathcal{G}_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2) &= \frac{\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)}{\lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2)}\\ &= \exp\left\{\sum_{j=1}^{p} C_j(\boldsymbol{s}_1,\boldsymbol{s}_2)\psi_j(t_1)\psi_j(t_2)\right\}. \end{aligned} \tag{7}$$

If $C_j(\boldsymbol{s}_1,\boldsymbol{s}_2)$ is stationary, that is, it only depends on the spatial lag $\boldsymbol{s}_1-\boldsymbol{s}_2$, then $\mathcal{G}_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)$ is a function of $\boldsymbol{s}_1-\boldsymbol{s}_2$ and the time points (t_1,t_2). Hence, the point process is second-order intensity reweighted stationary in space (Braddley et al. 2000).

3. ESTIMATION PROCEDURE

3.1 Estimation of the Mean Components

The Poisson maximum likelihood (Schoenberg 2005) method is a general approach to fit parametric models for the intensity function of a point process, where the point process can be purely spatial, temporal, or spatiotemporal. Asymptotic properties of the resulting estimator such as consistency and asymptotic normality were considered in Guan and Loh (2007). In the spatiotemporal case, let $\lambda(\mathbf{s},t;\boldsymbol{\theta})$ be such a model under consideration where $\boldsymbol{\theta}$ is some unknown parameter. Then, $\boldsymbol{\theta}$ can be estimated by maximizing

$$\ell(\boldsymbol{\theta}) = \sum_{(\boldsymbol{s},t)\in N\cap W} \log\{\lambda(\boldsymbol{s},t;\boldsymbol{\theta})\} - \int_D\int_T \lambda(\boldsymbol{s},t;\boldsymbol{\theta})dtd\boldsymbol{s}. \tag{8}$$

In our setting, we will apply the aforesaid method to estimate $\boldsymbol{\beta}$ and $\gamma(t)$. Observe that $\lambda(\boldsymbol{s},t;\boldsymbol{\theta})=\exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}+\gamma(t)\}$. We then modify (8) as

$$\begin{aligned}\ell(\boldsymbol{\beta},\gamma) &= \sum_{(\boldsymbol{s},t)\in N\cap W}\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}+\gamma(t)\}\\ &\quad - \int_D\int_T \exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}+\gamma(t)\}dtd\boldsymbol{s}. \end{aligned} \tag{9}$$

To further parameterize $\gamma(t)$, we propose to approximate it by regression splines (Zhou et al. 1998; Zhu, Fung, and He 2008). For simplicity, we assume the time domain to be $T=[0,1]$. Let $\kappa_j=j/(J_1+1)$, $j=0,\ldots,J_1+1$ be equally spaced knots on T, then we can define $K_1=J_1+r_1$ normalized B-spline basis functions of order r_1, which form the basis of a functional space $\mathcal{S}^{r_1}_{K_1}$. The B-spline basis functions are

$$\begin{aligned}B_j(t) &= (\kappa_j-\kappa_{j-r_1})[\kappa_{j-r_1},\ldots,\kappa_j](\kappa-t)_+^{r_1-1},\\ &\quad j=1,\ldots,K_1,\end{aligned}$$

where $[\kappa_{j-r_1},\ldots,\kappa_j]\phi(\kappa)$ denotes the r_1th order divided difference of the function $\phi(\kappa)$ on r_1+1 points $\kappa_{j-r_1},\ldots,\kappa_j$, $\kappa_j=\kappa_{\min\{\max(j,0),J_1+1\}}$ for $j=1-r_1,\ldots,K_1$, and $(x)_+=\max(x,0)$.

Denote the estimators of $\boldsymbol{\beta}$ and $\gamma(t)$ as

$$\{\widehat{\boldsymbol{\beta}},\widehat{\gamma}(t)\} = \mathrm{argmax}_{(\boldsymbol{\beta},\gamma)\in\mathbb{R}^d\otimes\mathcal{S}^{r_1}_{K_1}}\,\ell(\boldsymbol{\beta},\gamma). \tag{10}$$

Let $\boldsymbol{B}(t)=\{B_1(t),\ldots,B_{K_1}(t)\}^{\mathrm{T}}$ be the vector of spline basis, and write $\gamma(t)=\boldsymbol{B}^{\mathrm{T}}(t)\boldsymbol{\nu}$. For convenience of developing asymptotic theory, we denote $\widetilde{\boldsymbol{B}}(t)=\sqrt{K_1}\boldsymbol{B}(t)$, $\mathbb{X}(\boldsymbol{s},t)=\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t),\widetilde{\boldsymbol{B}}^{\mathrm{T}}(t)\}^{\mathrm{T}}$, and $\boldsymbol{\theta}=(\boldsymbol{\beta}^{\mathrm{T}},K_1^{-1/2}\boldsymbol{\nu}^{\mathrm{T}})^{\mathrm{T}}$. Then, $\widehat{\boldsymbol{\theta}}$ is the solution of the estimating equation

$$\begin{aligned}\mathbf{0} = \frac{\partial\ell}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}) &= \sum_{(\boldsymbol{s},t)\in N\cap W}\mathbb{X}(s,t)\\ &\quad - \int_D\int_T \mathbb{X}(\boldsymbol{s},t)\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\theta}\}dtd\boldsymbol{s}. \end{aligned} \tag{11}$$

The estimating equation can be solved numerically by a Newton–Raphson algorithm, where the integral in the equation is evaluated numerically. Asymptotic properties of these estimators are studied in Section 4. The number of spline basis functions K_1 is often deemed as a tuning parameter in spline smoothing. Selection of this tuning parameter is discussed in Section W.5 of the supplementary material.

3.2 Estimation of the Eigenvalues and Eigenfunctions

As mentioned before, the eigenvalues and eigenfunctions, $\{\omega_j,\psi_j(\cdot)\}$, can be estimated by an eigendecomposition of the covariance function R_T. By (6), the second-order intensity of the point process across time at a given spatial location is

$$\lambda_{2,\boldsymbol{s}}(t_1,t_2)=\lambda_2(\boldsymbol{s},\boldsymbol{s},t_1,t_2)=\lambda(\boldsymbol{s},t_1)\lambda(\boldsymbol{s},t_2)\exp\{R_T(t_1,t_2)\}. \tag{12}$$

We will approximate R_T by tensor product splines. Let $\{B_j(t);j=1,\ldots,K_2\}$ be B-spline functions with order r_2 defined on J_2 equally spaced knots on $[0,1]$. The tensor product spline basis functions are given by $B_{jj'}(t_1,t_2)=B_j(t_1)B_{j'}(t_2)$, $j,j'=1,\ldots,K_2$. Denote $\boldsymbol{B}_{[2]}(t_1,t_2)=(B_{11},B_{12},\ldots,B_{1K_2},B_{21},\ldots,B_{K_2K_2})^{\mathrm{T}}(t_1,t_2)$, and the functional space spanned by $\boldsymbol{B}_{[2]}$ as $\mathcal{S}^{r_2}_{[2],K_2}$. Then, the spline approximation for the covariance function is $R_T(t_1,t_2)\approx\boldsymbol{B}^{\mathrm{T}}_{[2]}(t_1,t_2)\boldsymbol{b}$.

We estimate $\boldsymbol{b}$ and hence R_T by generalizing the composite likelihood approach of Guan (2006) and Waagepetersen (2007). Let $\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2;\boldsymbol{\eta})$ be a parametric model for the second-order intensity function of a point process depending on some

Downloaded by [141.211.66.220] at 11:33 02 October 2014

parameter vector $\boldsymbol{\eta}$. Then, $\boldsymbol{\eta}$ can be estimated by maximizing

$$\ell_c(\boldsymbol{\eta}) = \sum_{(\boldsymbol{s}_1,t_1)\in N\cap W} \sum_{\substack{(\boldsymbol{s}_2,t_2)\in \\ N\cap\{D\otimes T-(\boldsymbol{s}_1,t_1)\}}} \times w(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)\log\{\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2;\boldsymbol{\eta})\} - \int_D\int_D\int_T\int_T w(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2;\boldsymbol{\eta}) \times dt_1dt_2d\boldsymbol{s}_1d\boldsymbol{s}_2, \quad (13)$$

where $D\otimes T-(\boldsymbol{s}_1,t_1)=\{(\boldsymbol{s}_2,t_2):(\boldsymbol{s}_2,t_2)\in D\otimes T,$ and $(\boldsymbol{s}_2,t_2)\neq(\boldsymbol{s}_1,t_1)\}$ and $w(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)$ is a prespecified weight function. For our purpose, it is sufficient to estimate $\lambda_{2,\boldsymbol{s}}(t_1,t_2)$ as a result of (12). Thus, we may want to consider pairs of events that occurred at the same location. However, for an orderly spatial point process, the probability of observing two events at the same location is zero. Instead, we define a small spatial neighborhood for every event location $\boldsymbol{s}$, denoted as $D_{\boldsymbol{s},\delta}=\{\boldsymbol{u}\in D,\|\boldsymbol{u}-\boldsymbol{s}\|<\delta\}$, and consider pairs of the given event and any other events within the neighborhood. This can be achieved by defining the weight function as

$$w(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)=I(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|<\delta),$$

where $I(\cdot)$ is an indicator function. We assume that the spatial covariance functions $C_j(\cdot)$'s are continuous at 0, and we choose δ to be small so that $\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)\approx\lambda_{2,\boldsymbol{s}_1}(t_1,t_2)$ for any $\boldsymbol{s}_2\in D_{\boldsymbol{s}_1,\delta}$. The role of δ in our estimation procedure will be discussed in Section 4, after developing the asymptotic theory of the proposed covariance estimator, and a practical criterion to choose δ is provided in the online supplementary material.

With the aforementioned modifications, the composite likelihood criterion in (13) becomes

$$\ell_c(\boldsymbol{\beta},\gamma,R_T)=\sum_{(\boldsymbol{s}_1,t_1)\in N\cap W}\sum_{\substack{(\boldsymbol{s}_2,t_2)\in \\ N\cap\{D_{\boldsymbol{s}_1,\delta}\otimes T-(\boldsymbol{s}_1,t_1)\}}} \times\log\{\lambda_{2,\boldsymbol{s}_1}(t_1,t_2)\} - \int_D\int_{D_{\boldsymbol{s}_1,\delta}}\int_T\int_T \lambda_{2,\boldsymbol{s}_1}(t_1,t_2)dt_1dt_2d\boldsymbol{s}_2d\boldsymbol{s}_1, \quad (14)$$

where $D_{\boldsymbol{s}_1,\delta}\otimes T-(\boldsymbol{s}_1,t_1)=\{(\boldsymbol{s}_2,t_2):\|\boldsymbol{s}_2-\boldsymbol{s}_1\|\le\delta,$ and $(\boldsymbol{s}_2,t_2)\neq(\boldsymbol{s}_1,t_1)\}$. We propose to estimate the covariance function as the maximizer of the composite likelihood restricted in the spline space, that is,

$$\widehat{R}_T=\operatorname{argmax}_{R_T\in\mathcal{S}^{r_2}_{[2],K_2}}\ell_c(\widehat{\boldsymbol{\beta}},\widehat{\gamma},R_T), \quad (15)$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$ are the estimators defined in (10).

The covariance estimator can be rewritten as $\widehat{R}_T(t_1,t_2)=\boldsymbol{B}^{\mathrm{T}}_{[2]}(t_1,t_2)\widehat{\boldsymbol{b}}$, where $\widehat{\boldsymbol{b}}$ is the solution of

$$\mathbf{0}=\frac{\partial\ell_c}{\partial\boldsymbol{b}}(\widehat{\boldsymbol{\beta}},\widehat{\gamma},\boldsymbol{b})=\sum_{(\boldsymbol{s}_1,t_1)\in N}\sum_{\substack{(\boldsymbol{s}_2,t_2)\in \\ N\cap\{D_{\boldsymbol{s}_1,\delta}\otimes T-(\boldsymbol{s}_1,t_1)\}}}\boldsymbol{B}_{[2]}(t_1,t_2) - \int_D\int_{D_{\boldsymbol{s}_1,\delta}}\int_T\int_T \boldsymbol{B}_{[2]}(t_1,t_2)\widehat{\lambda}(\boldsymbol{s}_1,t_1) \times\widehat{\lambda}(\boldsymbol{s}_2,t_2)\exp\left\{\boldsymbol{B}^{\mathrm{T}}_{[2]}(t_1,t_2)\boldsymbol{b}\right\}dt_1dt_2d\boldsymbol{s}_2d\boldsymbol{s}_1, \quad (16)$$

where $\widehat{\lambda}(\boldsymbol{s},t)=\exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\widehat{\boldsymbol{\beta}}+\widehat{\gamma}(t)\}$. When the neighborhood $D_{\boldsymbol{s},\delta}$ is sufficiently small and the number of knots of the spline basis is sufficiently large, (16) is an approximately unbiased estimating equation.

The estimates of the eigenvalues and eigenfunctions are obtained by solving the eigendecomposition problems

$$\int_T\widehat{R}_T(t_1,t_2)\widehat{\psi}_j(t_1)dt_1=\widehat{\omega}_j\widehat{\psi}_j(t_2),\quad j=1,\ldots,p. \quad (17)$$

Since our estimator $\widehat{R}_T(\cdot,\cdot)$ is constrained in a functional subspace spanned by tensor products of a spline basis $\boldsymbol{B}(\cdot)$, the estimated eigenfunction function is spanned by the same spline basis. Hence, the functional eigendecomposition problem in (17) can be translated into a multivariate problem. Notice that our estimator $\widehat{R}_T$ is inherently symmetric because the same pairs of events contribute equally in estimating $R_T(t_1,t_2)$ and $R_T(t_2,t_1)$. We can arrange the coefficient vector $\widehat{\boldsymbol{b}}$ into a symmetric matrix $\widehat{\boldsymbol{G}}$, so that $\widehat{R}_T(t_1,t_2)=\boldsymbol{B}^{\mathrm{T}}(t_1)\widehat{\boldsymbol{G}}\boldsymbol{B}(t_2)$. Define an inner product matrix $\mathcal{J}=\int_T\boldsymbol{B}(t)\boldsymbol{B}^{\mathrm{T}}(t)dt$, then the eigendecomposition problem in (17) is equivalent to the multivariate generalized eigenvalue decomposition

$$\widehat{\boldsymbol{\phi}}^{\mathrm{T}}_j\mathcal{J}\widehat{\boldsymbol{G}}\mathcal{J}\widehat{\boldsymbol{\phi}}_j=\widehat{\omega}_j,\quad\text{subject to }\widehat{\boldsymbol{\phi}}^{\mathrm{T}}_j\mathcal{J}\widehat{\boldsymbol{\phi}}_{j'}=I(j=j'), \quad (18)$$

where $I(\cdot)$ is an indicator function. Then, $\widehat{\psi}_j(t)=\boldsymbol{B}^{\mathrm{T}}(t)\widehat{\boldsymbol{\phi}}_j$, $j=1,\ldots,p$.

In the procedures described earlier, selecting the tuning parameters K_2 and δ as well as selecting the number of principal components p are important issues, which are addressed in Section W.5 of the supplementary material.

3.3 Estimation of the Spatial Correlation

In the previous section, we estimate the eigenfunctions ψ_j's and eigenvalues ω_j's using pairs of events that occurred within a close distance to avoid the complications of spatial correlation. With ψ_j's and ω_j's consistently estimated, we now estimate the spatial correlation functions using another composite likelihood that includes pairs of events further apart. Suppose the spatial covariance functions are of a parametric form $C_j(\boldsymbol{s}_1,\boldsymbol{s}_2;\boldsymbol{\vartheta}_j)$, where $\boldsymbol{\vartheta}_j$'s are unknown parameters. We will focus on stationary covariance models such as the flexible class of Matérn covariance models (Stein 1999). Stationarity in space is commonly assumed in spatial statistics including spatiotemporal log-Gaussian Cox processes (e.g., Brix and Møller 2001; Diggle, Rowlingson, and Su 2005). In what follows, we use $C_j(\boldsymbol{s}_1-\boldsymbol{s}_2;\boldsymbol{\vartheta}_j)$ instead to reflect the assumption of stationarity.

Define $\boldsymbol{\vartheta}=(\boldsymbol{\vartheta}^{\mathrm{T}}_1,\ldots,\boldsymbol{\vartheta}^{\mathrm{T}}_p)^{\mathrm{T}}$. To estimate $\boldsymbol{\vartheta}$, we again modify the composite likelihood (13) through the use of a proper weight function w. Specifically, we use

$$w(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)=\lambda^{-1}(\boldsymbol{s}_1,t_1)\lambda^{-1}(\boldsymbol{s}_2,t_2)I(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\le\varrho),$$

where ϱ is a prespecified spatial distance. By (6),

$$w(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2) = \exp\left\{\sum_{j=1}^p C_j(\boldsymbol{s}_1-\boldsymbol{s}_2;\vartheta_j)\psi_j(t_1)\psi_j(t)\right\}I(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\le\varrho).$$

Thus, we avoid integrating the covariate process $\boldsymbol{Z}(\boldsymbol{s},t)$ over the entire spatial temporal domain while evaluating the

Downloaded by [141.211.66.220] at 11:33 02 October 2014

integral in (13). Let $\ell_{c,\mathrm{spat}}(\boldsymbol{\vartheta}, \boldsymbol{\beta}, \gamma, \boldsymbol{\omega}, \boldsymbol{\psi})$ be the resulting modified composite likelihood, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^{\mathrm{T}}$ and $\boldsymbol{\psi}(t) = (\psi_1, \ldots, \psi_p)^{\mathrm{T}}(t)$. We substitute $(\boldsymbol{\beta}, \gamma, \boldsymbol{\omega}, \boldsymbol{\psi})$ with their estimators described in Section 3.2 and therefore define the estimator of $\boldsymbol{\vartheta}$ as

$$\widehat{\boldsymbol{\vartheta}} = \mathrm{argmax}_{\boldsymbol{\vartheta}} \ell_{c,\mathrm{spat}}(\boldsymbol{\vartheta}, \widehat{\boldsymbol{\beta}}, \widehat{\gamma}, \widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\psi}}). \tag{19}$$

The proposed weight function w excludes the pairs of events with a distance greater than ϱ, since the spatial correlation tends to diminish as the spatial lag increases and including events that are too far away may provide little information about the correlation function. The parameter ϱ therein can be considered as a tuning parameter. A reasonable choice of ϱ is the range of the spatial correlation, which can be estimated by fitting a pilot parametric model to the data or by checking the empirical pair correlation function of the point pattern (e.g., Guan 2008b).

3.4 Prediction of the Spatial Random Effects

To predict the random fields $\xi_k(\boldsymbol{s})$, we use a maximum a posteriori (MAP) predictor as in Møller, Syversveen, and Waagepetersen (1998). For ease of presentation, we assume that the spatial domain D is a rectangle $[0, L]^2$. We partition D into smaller rectangles, $D_{ij} = [(i-1)/M, i/M) \times [(j-1)/M, j/M)$, $i, j = 1, \ldots, ML$. We take each D_{ij} sufficiently small so that $\xi_k(\boldsymbol{s})$ is approximately a constant for $\boldsymbol{s} \in D_{ij}$, and denote this value as $\xi_{ij,k}$, for $k = 1, \ldots, p$. Given $\boldsymbol{\xi}_{ij} = (\xi_{ij,1}, \ldots, \xi_{ij,p})^{\mathrm{T}}$, the conditional log-likelihood for the events in $D_{ij} \otimes T$ is

$$\ell_{ij} = \sum_{\substack{(\boldsymbol{s},t) \in \\ N \cap \{D_{ij} \otimes T\}}} \left\{ \boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\beta} + \mu(t) + \sum_{k=1}^{p} \xi_{ij,k}\psi_k(t) \right\} - \int_{D_{ij}} \int_T \exp\left\{ \boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\beta} + \mu(t) + \sum_{k=1}^{p} \xi_{ij,k}\psi_k(t) \right\} dt d\boldsymbol{s}. \tag{20}$$

By the model, $\boldsymbol{\xi}_k$ has a prior distribution Normal$(\mathbf{0}, \Sigma_k)$, where Σ_k is the covariance matrix for the kth principal component by interpolating $C_k(\boldsymbol{s}_1 - \boldsymbol{s}_2; \vartheta_k)$ on the discrete grid points. Collect the grid point values of the kth principal component score into $\boldsymbol{\xi}_k = \{\xi_{ij,k}; i, j = 1, \ldots, M\}$, then the log posterior density of $\boldsymbol{\xi} = \{\boldsymbol{\xi}_k; k = 1, \ldots, p\}$ is

$$p(\boldsymbol{\xi}) \propto \sum_{i=1}^{M} \sum_{j=1}^{M} \ell_{ij} - \frac{1}{2} \sum_{k=1}^{p} \boldsymbol{\xi}_k^{\mathrm{T}} \Sigma_k^{-1} \boldsymbol{\xi}_k. \tag{21}$$

We substitute $\boldsymbol{\beta}$, $\psi_k(t)$, and ω_k with their estimators defined earlier, and $\mu(t)$ with $\widehat{\mu}(t) = \widehat{\gamma}(t) - 1/2\widehat{R}_T(t, t)$. The empirical Bayes estimator $\widehat{\boldsymbol{\xi}}$ is then obtained by maximizing the posterior (21). We can also draw samples from the posterior (21) using the Metropolis-adjusted Langevin algorithm (MALA) described in Møller, Syversveen, and Waagepetersen (1998), and estimate the prediction error by the posterior variance.

Choosing the partition in spatial prediction is a compromise between prediction bias and computation feasibility. The latent processes are defined in a continuous space, hence a finer spatial grid leads to smaller bias. On the other hand, using a finer grid increases the dimension of the latent random vector $\boldsymbol{\xi}_k$ and makes it harder to simulate from the posterior distribution in (21). Specifically, a higher dimension of $\boldsymbol{\xi}_k$ make it harder for the Markov chain to mix and, as a result, the Markov chain takes a longer time to converge. In many real applications such as the CTR data considered in this article, there are natural choices for the partition of the spatial domain, for example we used the census tracts to partition the state of Connecticut.

4. ASYMPTOTIC PROPERTIES

To distinguish from other possible values in the parameter space, we denote the true parameters (functions) as $\boldsymbol{\beta}_0$, γ_0, R_{T0}, ω_{j0}, and ψ_{j0}. We study the asymptotic properties of the proposed estimators under an increasing domain asymptotic framework as in Guan (2006). We consider a sequence of spatial domains D_n with expanding areas, but the time domain T remains fixed. Specifically, we assume

$$C_1 n^2 \le |D_n| \le C_2 n^2 \quad \text{and } C_1 n \le |\partial D_n| \le C_2 n \quad \text{for some } 0 < C_1 < C_2 < \infty, \tag{22}$$

where $|\partial D_n|$ denotes the perimeter of D_n. Condition (22) is satisfied by many commonly encountered shape sequences. For example, let $D \subset (0, 1] \times (0, 1]$ be the interior of a simple closed curve with nonempty interior. If we multiply D by n to obtain D_n, then D_n satisfies (22). This formulation allows for a wide variety of shapes as the sequence of observation windows, including both rectangular and elliptical regions.

For any function $f(\boldsymbol{x})$ defined on a compact set $\mathcal{I}$, where $\mathcal{I} \subset \mathbb{R}$ or $\mathbb{R}^2$, define the supremum and L^2 norm of g to be $\|f\|_\infty = \sup_{\boldsymbol{x} \in \mathcal{I}} |f(\boldsymbol{x})|$ and $\|f\| = \{\int_{\mathcal{I}} f^2(\boldsymbol{x}) d\boldsymbol{x}\}^{1/2}$, respectively. For any m dimensional vector $\boldsymbol{a}$, define its L^2 norm $\|\boldsymbol{a}\| = (\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a})^{1/2}$, and its L^∞ norm $\|\boldsymbol{a}\|_\infty = \max_{j=1}^{m} |a_j|$. For any real valued $m_1 \times m_2$ matrix $\boldsymbol{A} = (a_{ij})$, define its L^2 norm as $\|\boldsymbol{A}\| = \sup_{\boldsymbol{x} \in \mathbb{R}^{m_2}} \|\boldsymbol{A}\boldsymbol{x}\|/\|\boldsymbol{x}\|$, its L^∞ norm as $\|\boldsymbol{A}\|_\infty = \max_{i=1}^{m_1} \sum_{j=1}^{m_2} |a_{ij}|$, and its Frobenius norm as $\|\boldsymbol{A}\|_F = \{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})\}^{1/2}$.

Theorem 1. Let T be a fixed time domain, D_n satisfies condition (22), then under Assumptions 1–5 in Appendix A,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \|\widehat{\gamma} - \gamma_0\| = O_p\{K_1^{-r_1+1/2} + (K_1/|D_n|)^{1/2}\}.$$

The convergence rate in Theorem 1 is not optimal. A more detailed study in Theorem 2 below reveals that $\widehat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}_0$ in a parametric convergence rate and is asymptotically normal, and $\widehat{\gamma}(t)$ converges to $\gamma_0(t)$ with the usual nonparametric asymptotic convergence rate. To facilitate this result, we first define the residual process in the spatiotemporal pattern (Baddeley et al. 2005) as

$$\mathcal{E}(d\boldsymbol{s}, dt) = N(d\boldsymbol{s}, dt) - \exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\beta}_0 + \gamma_0(t)\} d\boldsymbol{s} dt. \tag{23}$$

We also define

$$\boldsymbol{\mu}_Z(t; \boldsymbol{\beta}, \gamma) = \mathrm{E}[\boldsymbol{Z}(\boldsymbol{s}, t) \exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\beta} + \gamma(t)\}]/q(t; \boldsymbol{\beta}, \gamma) \tag{24}$$

Downloaded by [141.211.66.220] at 11:33 02 October 2014

where $q(t;\boldsymbol{\beta},\gamma)$ is defined in Assumption 3, and

$$\Sigma_Z(\boldsymbol{\beta},\gamma) = \mathrm{E}\left[\int_T \{\boldsymbol{Z}(\boldsymbol{s},t)-\boldsymbol{\mu}_Z(t;\boldsymbol{\beta},\gamma)\}^{\otimes 2}\exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\beta+\gamma(t)\}dt\right], \tag{25}$$

where $\boldsymbol{x}^{\otimes 2}=\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$ for any vector $\boldsymbol{x}$.

Theorem 2. Under the same conditions as in Theorem 1, we have the following weak convergence result

$$|D_n|^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\to \mathrm{Normal}\{\mathbf{0},\,\Sigma_{Z,0}^{-1}\Omega\Sigma_{Z,0}^{-1}\},$$

where $\Sigma_{Z,0}$ is a shorthand for $\Sigma_Z(\boldsymbol{\beta}_0,\gamma_0)$ and

$$\Omega=\lim_{n\to\infty}\frac{1}{|D_n|}\mathrm{var}\left[\int_T\int_{D_n}\{\boldsymbol{Z}(\boldsymbol{s},t)-\boldsymbol{\mu}_Z(t;\boldsymbol{\beta}_0,\gamma_0)\}\mathcal{E}(d\boldsymbol{s},dt)\right].$$

A tighter asymptotic convergence rate for $\widehat{\gamma}$ is $\|\widehat{\gamma}-\gamma_0\|=O_p(K_1^{1/2}|D_n|^{-1/2}+K_1^{-r_1})$.

For statistical inference, we need to estimate the covariance matrix of $\widehat{\boldsymbol{\beta}}$. We follow Heinrich and Prokešová (2010) to derive a consistent moment estimator for $\mathrm{cov}(\widehat{\boldsymbol{\beta}})$. Details of the derivations are given in the Web Appendix B. We also outline a strategy on how to estimate the variance of $\widehat{\gamma}(\cdot)$, in light of the fact that both $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}(\cdot)$ are obtained by solving the estimating equation (11).

Next, we study the asymptotic properties of the estimated covariance function and those of the estimated eigenvalues and eigenfunctions. The radius of the local neighborhood in the composite likelihood (14) should depend on n. However, we will continue to use δ for ease of exposition.

Theorem 3. Assume that condition (22) and Assumptions 1–9 in Appendix A are true. Then, $\|\widehat{R}_T-R_{T0}\|=O_p\{K_2(|D_n||D_\delta|)^{-1/2}+\delta+K_2^{-r_2}+(K_1/|D_n|)^{1/2}+K_1^{-r_1}\}$.

Theorem 3 implies that the radius parameter δ plays a similar role to the bandwidth used in nonparametric regressions. As such, there is a trade-off between bias and variance when choosing the optimal δ. Specifically, increasing δ will include more data into the estimation and hence reduce the variance of $\widehat{R}_T$, but it will increase the bias due to the use of pairs of events that are much further apart; the opposite can be said when decreasing δ. A practical method to select δ is proposed in Section W.5.2..

Following Theorem 1 in Hall and Hosseini-Nasab (2006), the following asymptotic properties for the functional principal component estimators in (17) and (18) are immediate.

Corollary 1. Letting $\Delta_n=\|\widehat{R}_T-R_{T0}\|$, then $\sup_j|\widehat{\omega}_j-\omega_j|\le\Delta_n$. If p is finite, define $\omega_{p+1}=0$. Put $\tau_j=\min_{1\le k\le j}(\omega_k-\omega_{k+1})$, $J=\inf\{j\ge 1:\omega_j-\omega_{j+1}\le 2\Delta_n\}$, then $\|\widehat{\psi}_j-\psi_j\|\le C\Delta_n/\tau_j$, for $1\le j\le J-1$.

Theorems 2 and 3 and Corollary 1 show that our estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\gamma}(\cdot)$ and $\{\widehat{\omega}_j,\widehat{\psi}_j(\cdot);j=1,\ldots,p\}$ are consistent, and hence by plugging in these consistent estimators the method described in Section 3.3 also provides a consistent estimator for the spatial correlation parameter ϑ. Using the theory in Guan (2006), we have the following corollary.

Corollary 2. Under condition (22), the assumptions in Appendix A and the regularity conditions in Theorem 1 of Guan (2006), $\widehat{\vartheta}$ defined by (19) is consistent to the true correlation parameter ϑ.

5. SIMULATION STUDY

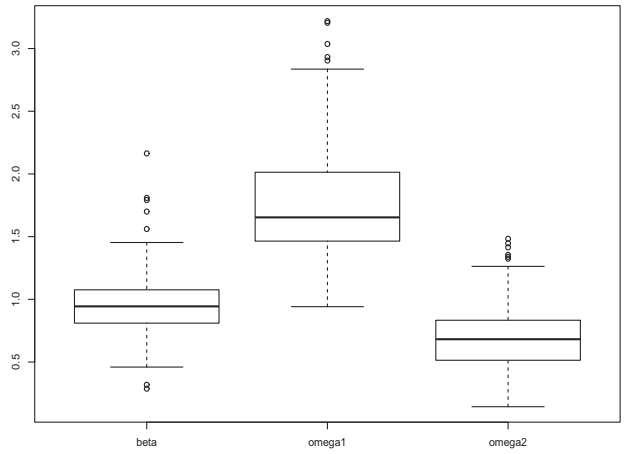
Let the spatial region be $D=[0,2]^{\otimes 2}$, and the time window be $T=[0,1]$. We assume that $Z(\boldsymbol{s})$ is a one-dimensional covariate, which is generated as an isotropic Gaussian random field on D with an exponential covariance structure. In particular, $\mathrm{cov}\{Z(\boldsymbol{s}_1),Z(\boldsymbol{s}_2)\}=\exp(-\|\boldsymbol{s}_1-\boldsymbol{s}_2\|/\rho)$, and we set the scale parameter to be $\rho=0.2$. The random field $X(\boldsymbol{s},t)$ is generated with $\mu(t)=3+2t^2$ and $p=2$ principal components, where $(\omega_1,\omega_2)=(2,1)$, $\psi_1(t)=1$, and $\psi_2(t)=\sqrt{2}\cos(2\pi t)$. Both principal component scores, $\xi_j(\boldsymbol{s})$, $j=1,2$, are generated from Gaussian random fields with isotropic exponential covariance structures $C_j(\boldsymbol{s}_1-\boldsymbol{s}_2;\vartheta_j)=\omega_j\exp(-\|\boldsymbol{s}_1-\boldsymbol{s}_2\|/\vartheta_j)$, and the scale parameters ϑ_j are also set to be 0.2. Both $Z(\boldsymbol{s})$ and $\xi_j(\boldsymbol{s})$ are simulated on a regular grid with increments 0.01, using the *RandomFields* package in R. The events are generated using rejection sampling.

In this setting, the covariance function is $R_T(t_1,t_2)=2+2\cos(2\pi t_1)\cos(2\pi t_2)$. The two principal components have clear interpretations. The first principal component is a random intercept. If $\xi_1(\boldsymbol{s})$ is high in a location $\boldsymbol{s}$, the event intensity is high at that location. The second principal component can be interpreted as a periodic random effect. On average, there are 1661 events in the defined spatiotemporal domain.

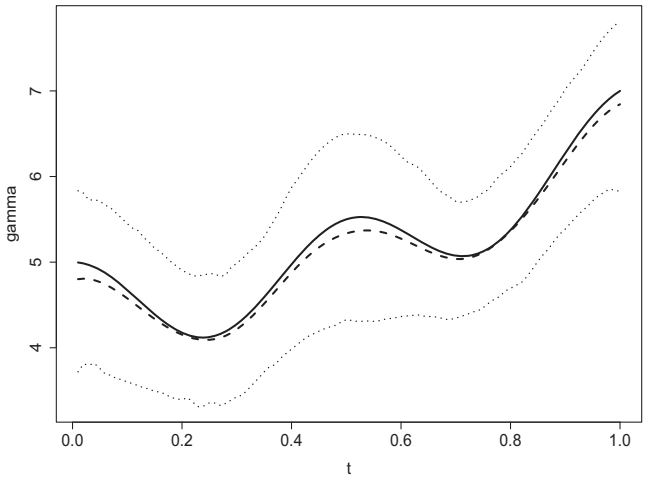
The simulation is repeated 200 times, and the proposed model selection and estimation procedures are applied to each simulated dataset. We first choose the tuning parameters as described in Section W.5 of the supplementary material. The AIC (W.9) picks $K_1=10$ for most of the simulated datasets, and the cross-validation procedure in Section W.5.2 chooses $K_2=7$ and $\delta=0.01$ most frequently. Therefore, we fix the value of these tuning parameters for further estimation. Under our choice of δ, we include, on average, one neighboring event for every event in the composite likelihood (14). Next, we apply our second AIC (W.11) to choose the number of principal components, and it chooses the correct number, $p=2$, of principal components 57% of the time. We find that AIC tends to choose an overfitted model, and 88% of the time, it chooses the number of principal components to be between 2 and 4. Such an overfitting tendency is consistent with what has been discovered in the literature. Since underfitting is usually a more serious problem than overfitting, the performance of AIC seems satisfactory.

The estimation results are summarized in Figure 1. In panel (a) of Figure 1, we show the boxplots of $\widehat{\beta}$, $\widehat{\omega}_1$, and $\widehat{\omega}_2$. As we can see, $\widehat{\beta}$ is almost unbiased to the true value $\beta_0=1$, which is consistent with our asymptotic theory. The estimated eigenvalues are slightly biased but nevertheless close to the truth. Although these estimators are consistent, some bias is often reported in FDA literature in a finite-sample setting, see Li and Hsing (2010b). This is true even when direct measurements are made on the curves. In our setting, $X(\boldsymbol{s},t)$ are latent processes which make estimation of these parameters considerably harder.

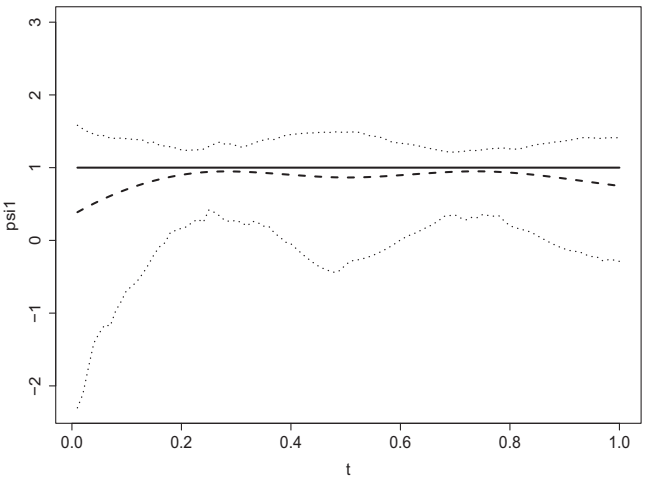
Downloaded by [141.211.66.220] at 11:33 02 October 2014



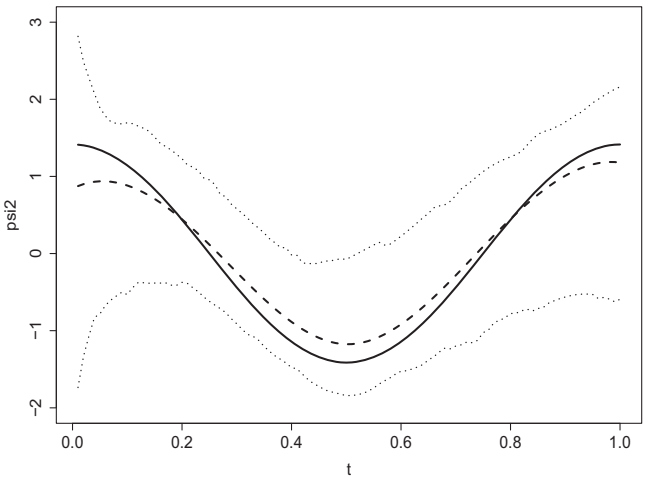
(a)



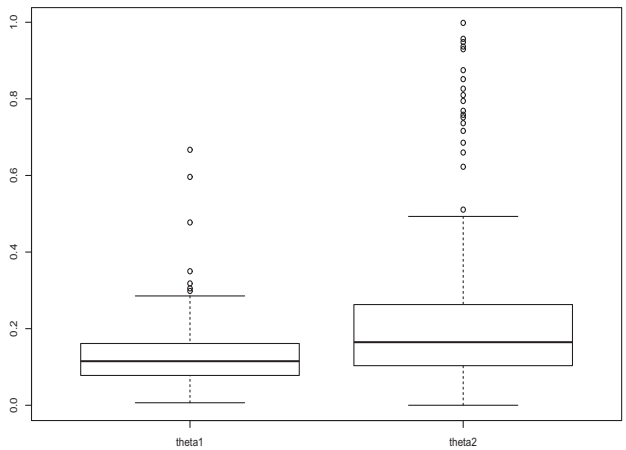
(b)



(c)



(d)



(e)

Figure 1. Estimation results in the simulation study. Panel (a) shows the boxplots of $\widehat{\beta}$, $\widehat{\omega}_1$, and $\widehat{\omega}_2$. Panels (b)–(d) show the estimation results for $\widehat{\gamma}(t)$, $\widehat{\psi}_1(t)$, and $\widehat{\psi}_2(t)$, respectively, where the solid curve in each panel is the true curve, the dashed curve is the mean of the estimator, and the two dotted curves are the pointwise 5% and 95% percentiles of the estimator. Panel (e) shows boxplots of the estimated spatial correlation parameters ϑ_1 and ϑ_2 for the two principal components.

Downloaded by [141.211.66.220] at 11:33 02 October 2014

In that sense, the behavior of these estimators is reasonable. In panels (b)–(d) of Figure 1, we summarize the estimation results for $\gamma(t)$ and the two eigenfunctions, where we compare the mean, 5% and 95% pointwise percentiles of the functional estimators with the true functions. The plots suggest that these estimators behave reasonably well. We also provide, in panel (e) of Figure 1, the boxplots of $\widehat{\vartheta}_1$ and $\widehat{\vartheta}_2$, which are the spatial correlation parameters for the two principal components estimated using the composite likelihood method in Section 3.3. The boxplots show that these estimates are reasonably close to the true value 0.2. Since the second principal component is less prominent in the data, its spatial correlation is harder to estimate. Consequently, $\widehat{\vartheta}_2$ is more variable then $\widehat{\vartheta}_1$.

We also perform spatial prediction for the latent processes $\xi_1(\boldsymbol{s})$ and $\xi_2(\boldsymbol{s})$ as described in Section 3.4. Plots of the prediction results in a typical run is provided in Section W.6 of the online supplementary material. These predicted maps can provide useful information on hot spatial regions due to clustering.

6. DATA ANALYSIS

We apply the proposed methodology to historical cancer incidence records collected by the CTR. The CTR is the oldest cancer registry in the United States. Since the Surveillance, Epidemiology, and End Results (SEER) Program was launched by the National Cancer Institute in 1973, it has always been a program-participating SEER site. The CTR has reciprocal reporting agreements with cancer registries in all adjacent states (and Florida, a popular winter destination for retirees) to identify Connecticut residents with cancer diagnosed and/or treated in these states. For each identified CTR case, both the date of diagnosis and residential address at the time of diagnosis were recorded, along with a list of demographic and diagnostic variables. The longitude and latitude of a diagnosis are recorded in the Universal Transverse Mercator (UTM) coordinate system.

Pancreatic cancer is the fourth most common cause of cancer-related deaths in both men and women in the United States. We consider the CTR data of 8230 pancreatic cancer incidences that were diagnosed from 1992 to 2009. Our primary interest is to study the spatiotemporal pattern of pancreatic cancer incidences, after having accounted for heterogeneities in both population density and socioeconomic status (SES) scores at the census block group level. There are 2616 block groups within the state of Connecticut. The SES score is an aggregated measure to reflect poverty level in a neighborhood, where a higher SES score indicates a more deprived neighborhood (Wang et al. 2009).

Similar to model (1), we assume that the conditional intensity for the cancer incidences is

$$\lambda(\boldsymbol{s}, t) = \lambda_0(\boldsymbol{s}) \exp\{Z(\boldsymbol{s})\beta + X(\boldsymbol{s}, t)\},$$

where $\lambda_0(\boldsymbol{s})$ and $Z(\boldsymbol{s})$ are the population density and SES score at $\boldsymbol{s}$, and $X(\boldsymbol{s}, t)$ is a latent process with the same structure as in (2). We assume that $\lambda_0(\boldsymbol{s})$ and $Z(\boldsymbol{s})$ are constants within a block group.

We first estimate the parameters in the first-order intensity. The AIC in (W.9) picks $K_1 = 9$ cubic B-splines to model the function $\gamma(t)$. We apply the proposed method in Section 3.1 to estimate β and $\gamma(t)$, and use the method described in Section W.4 of the supplementary material to estimate the standard error of the estimators. The estimated coefficient for the SES score is $\widehat{\beta} = 1.63 \times 10^{-2}$, with standard error 3.87×10^{-3}. We therefore conclude that the SES score is positively associated with the pancreatic cancer rate in a neighborhood. The estimated temporal trend function $\gamma(t)$ and the 95% confidence bands are presented in Figure 2. The plot suggests that the pancreatic cancer rate was increasing over the years in the study period.

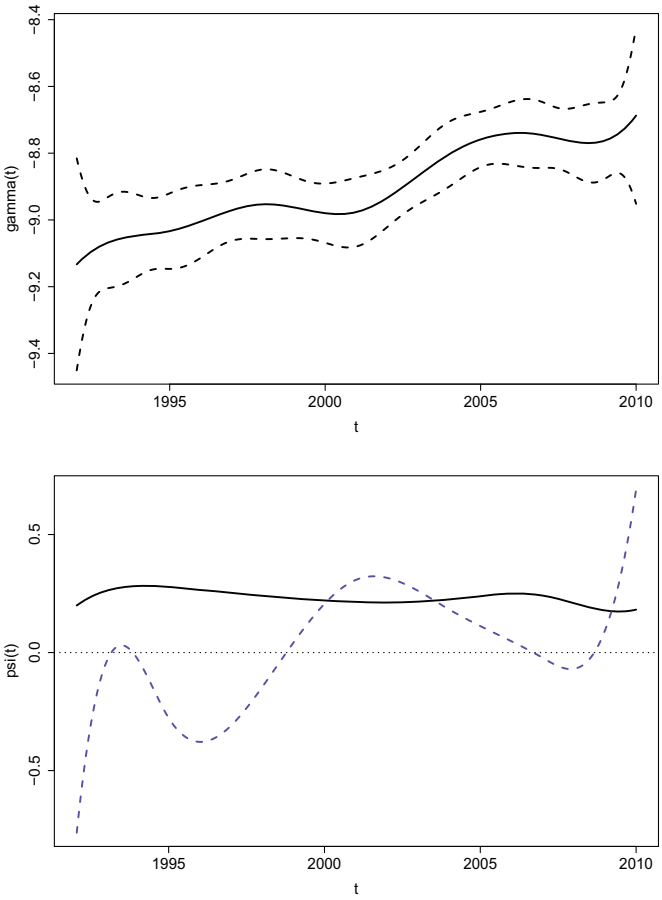
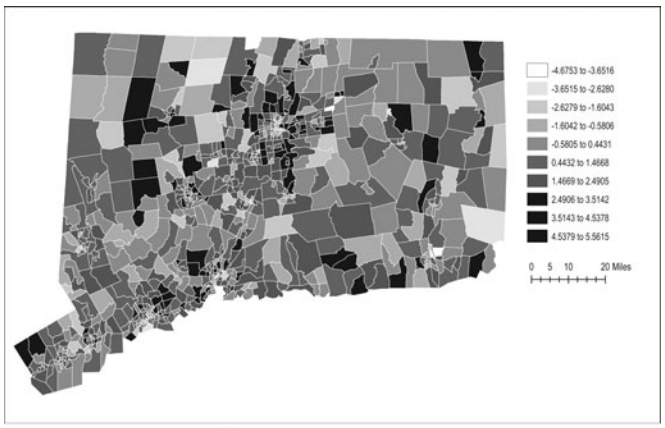


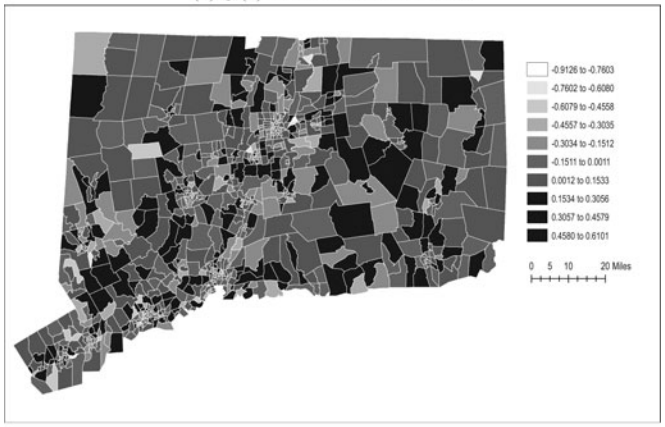
Figure 2. Estimation results for the Connecticut Tumor Registry data. The first plot is the estimated temporal trend $\widehat{\gamma}(t)$ (the dashed curves are the 95% confidence band), the second plot shows the first two estimated eigenfunctions (the solid curve is the $\widehat{\psi}_1(t)$ and the dashed curve is $\widehat{\psi}_2(t)$).

To estimate the covariance function R_T, the cross-validation procedure in Section W.5 picks $K_2 = 9$ and $\delta = 1000$ UTM units. Therefore, block groups with a UTM distance less than 1000 are considered neighbors, the pancreatic cancer incidences in neighboring block groups are considered neighboring events, and all such pairs of neighboring events are used in the composite likelihood (14). Note that 1000 UTM units is about 1 km, which is a small distance in the scale of this application. The AIC defined in (W.11) suggests that there are two principal components in $X(\boldsymbol{s}, \cdot)$. The first two eigenvalues are 8.742 and 0.345 which explain a total of 93% of variation in the covariance function. The first two eigenfunctions are shown in the second plot of Figure 2. As we can see, $\widehat{\psi}_1(t)$ given by the solid curve is almost a constant over time, indicating that the first principal component score $\xi_1(\boldsymbol{s})$ is a spatial random intercept. When $\xi_1(\boldsymbol{s})$

Downloaded by [141.211.66.220] at 11:33 02 October 2014

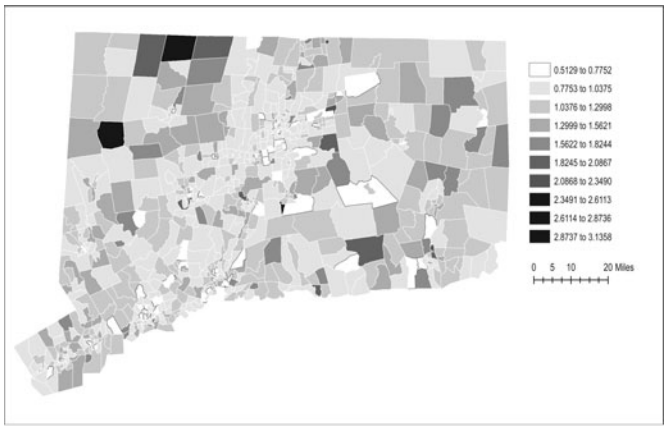


(a) $\widehat{\xi}_1(\boldsymbol{s})$ at the census tract level

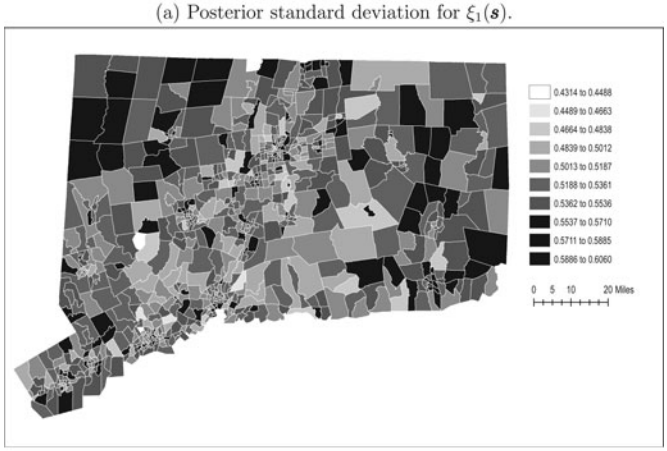


(b) $\widehat{\xi}_2(\boldsymbol{s})$ at the census tract level

Figure 3. Estimated scores for the first two principal components in the Connecticut Tumor Registry data. The principal component scores are estimated at census tract level and highlighted by gray levels on the map of Connecticut.



(a) Posterior standard deviation for $\xi_1(\boldsymbol{s})$.



(b) Posterior standard deviation for $\xi_2(\boldsymbol{s})$.

Figure 4. Estimated prediction errors for the first two principal components in the Connecticut Tumor Registry data. The prediction errors are the square root of the posterior variance of the scores at census tract level.

is higher, the cancer rate at $\boldsymbol{s}$ is also higher than the average rate. On the other hand, $\widehat{\psi}_2(t)$ represents an increasing trend in time, even though it does not increase linearly. Hence, when $\xi_2(\boldsymbol{s})$ is higher, the cancer rate at $\boldsymbol{s}$ increases faster than average.

We also model the spatial correlation in $\xi_1(\boldsymbol{s})$ and $\xi_2(\boldsymbol{s})$ by the exponential correlation function, and estimate correlation range parameters by the composite likelihood method in Section 3.3. The estimated range is about 2400 UTM units for both principal components. We perform spatial prediction for $\xi_1(\boldsymbol{s})$ and $\xi_2(\boldsymbol{s})$ at the census tract level, by simulating samples from the posterior (21). We use the posterior means as the predicted values of the principal component scores, and the posterior standard deviations as the prediction errors.

In the two panels of Figure 3, the predicted values of $\xi_1(\boldsymbol{s})$ and $\xi_2(\boldsymbol{s})$ are highlighted by gray levels in maps of Connecticut, respectively, where black represents the highest positive values and white represents the lowest negative values. By the interpretation of the two principal components described earlier, we believe the dark census tracts in panel (a) of Figure 3 have higher pancreatic cancer rates than average, while the dark tracts in panel (b) have higher increasing rate in pancreatic cancer than others. The posterior standard deviations of the two latent random fields are also given in the two panels of the Figure 4. These maps help us to understand the uncertainty in the spatial prediction. We perform z-tests on the predicted principal component scores and find that about 27% of census tracts have ξ_1 significant different from 0, however none of the predictions for ξ_2 are significant. The latter result is simply because of the relatively large amount of prediction error for ξ_2. We think that the signal in the second principal component is much weaker compared with the first component and the test on a local signal (i.e., at any single census tract) does not have enough power.

APPENDIX A: ASSUMPTIONS FOR THE THEORETICAL RESULTS

A.1 Notation

For any subset $E \subset \mathbb{R}^2$, let $\mathcal{F}(E)$ be the σ-algebra generated by $N \cap (E \otimes T)$ and $\{\boldsymbol{Z}(\boldsymbol{s}, t), \xi_j(\boldsymbol{s}, t), j = 1, \ldots, p : (\boldsymbol{s}, t) \in (E \otimes T)\}$. To quantify the spatial dependence, we introduce the strong mixing coefficient (Rosenblatt 1956),

$$\alpha_{k,l}(h) = \sup\{|P(A_1 \cap A_2) - P(A_1)P(A_2)| : A_1 \in \mathcal{F}(E_1), A_2 \in \mathcal{F}(E_2), |E_1| \le k, |E_2| \le l, d(E_1, E_2) \ge h\}, \quad \text{(A.1)}$$

where $d(E_1, E_2)$ denotes the minimal spatial distance between E_1 and E_2.

Downloaded by [141.211.66.220] at 11:33 02 October 2014

Define

$$\lambda_4(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\boldsymbol{s}_4,t_1,t_2,t_3,t_4) = \lim_{\substack{|d\boldsymbol{s}_j|,|dt_j|\to 0\\ j=1,\ldots,4}} \times \frac{\mathrm{E}\{N(d\boldsymbol{s}_1,dt_1)N(d\boldsymbol{s}_2,dt_2)N(d\boldsymbol{s}_3,dt_3)N(d\boldsymbol{s}_4,dt_4)\}}{|d\boldsymbol{s}_1||dt_1||d\boldsymbol{s}_2||dt_2||d\boldsymbol{s}_3||dt_3||d\boldsymbol{s}_4||dt_4|},$$

$$\mathcal{G}_4(\boldsymbol{u}_1,\boldsymbol{u}_2,\boldsymbol{u}_3,t_1,t_2,t_3,t_4) = \frac{\lambda_4(\boldsymbol{s},\boldsymbol{s}+\boldsymbol{u}_1,\boldsymbol{s}+\boldsymbol{u}_3,\boldsymbol{s}+\boldsymbol{u}_2+\boldsymbol{u}_3,t_1,t_2,t_3,t_4)}{\lambda(\boldsymbol{s},t_1)\lambda(\boldsymbol{s}+\boldsymbol{u}_1,t_2)\lambda(\boldsymbol{s}+\boldsymbol{u}_3,t_3)\lambda(\boldsymbol{s}+\boldsymbol{u}_2+\boldsymbol{u}_3,t_4)}, \quad (A.2)$$

and

$$\mathcal{M}(t_1,t_2) = \frac{1}{|D_n||D_\delta|}\int_{D_n}\int_{D_{\boldsymbol{s}_1,\delta}} \lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2)d\boldsymbol{s}_2 d\boldsymbol{s}_1, \quad (A.3)$$

where $|D_\delta|$ is the common area for all $D_{\boldsymbol{s},\delta}$, for example, $|D_\delta| = \pi\delta^2$ if $D_{\boldsymbol{s},\delta}$ is a disc centered at $\boldsymbol{s}$. Put $\mu_{\mathcal{M}}(t_1,t_2) = \mathrm{E}\{\mathcal{M}(t_1,t_2)\}$.

A.2 Assumptions

We make the following assumptions to derive our asymptotic theory.

- *Assumption 1.* Define the class of Hölder continuous functions on $[0,1]$ as $C_1^{r,a}[0,1] = \{f : \sup_{t_1,t_2\in[0,1]} |f^{(r)}(t_1) - f^{(r)}(t_2)|/|t_1 - t_2|^a < \infty\}$ for some nonnegative integer r and some $a > 0$. We assume $\gamma_0 \in C_1^{r_1,a}[0,1]$, where $r_1 \geq 2$ is the order of the spline estimator and $a > 0$.
- *Assumption 2.* We assume that the processes N, $\boldsymbol{Z}$ and ξ_j, $j = 1,\ldots,p$, are strictly stationary in $\boldsymbol{s}$ and satisfy the following mixing condition (Guyon, 1995):

$$\sum_{m\geq 1} m\alpha_{k,l}(m) < \infty \text{ and } \alpha_{k,\infty}(m) = \mathrm{o}(m^{-2}) \quad \text{for some } k > 0 \text{ and } m > 0.$$

  We also assume that $\mathrm{E}[\boldsymbol{Z}(\boldsymbol{s},t)\exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta} + \gamma(t)\}]^C < \infty$ for some $C > 2$, and $\sup_{t_1,t_2}\int_{\mathbb{R}^2} \mid \mathcal{G}_2(\boldsymbol{u},t_1,t_2) - 1 \mid d\boldsymbol{u} < \infty$.
- *Assumption 3.* Define $q(t;\boldsymbol{\beta},\gamma) = \mathrm{E}[\exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta} + \gamma(t)\}]$, which does not depend on $\boldsymbol{s}$ by the stationarity of $\boldsymbol{Z}(\boldsymbol{s},t)$ for any fixed t. Assume that $0 < C_1 \leq \min_t q(t;\boldsymbol{\theta},\gamma) \leq \max_t q(t;\boldsymbol{\theta},\gamma) \leq C_2 < \infty$, for all $(\boldsymbol{\beta},\gamma) \in \mathcal{N}_{C_0} = \{(\boldsymbol{b},g) : \boldsymbol{b} \in \mathbb{R}^d, g \in C_1^{r,a}[0,1], \|\boldsymbol{b} - \boldsymbol{\beta}_0\| + \|g - \gamma_0\|_\infty < C_0\}$.
- *Assumption 4.* Let $\boldsymbol{\mu}_Z(t;\boldsymbol{\beta},\gamma)$ and $\Sigma_Z(\boldsymbol{\beta},\gamma)$ be defined in (24) and (25), and $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ be the functionals to take the maximum and minimum eigenvalues of a matrix. Assume that $\boldsymbol{\mu}_Z(t;\boldsymbol{\beta},\gamma) \in C_1^{r_1,a}[0,1]$ and $0 < C_3 \leq \lambda_{\min}(\Sigma_Z) \leq \lambda_{\max}(\Sigma_Z) \leq C_4 < \infty$, for all $(\boldsymbol{\beta},\gamma) \in \mathcal{N}_{C_0}$. We also assume $\boldsymbol{\mu}_Z$ is continuous in $\boldsymbol{\beta}$ and γ, with $\|\boldsymbol{\mu}_Z(\bullet;\boldsymbol{\beta}_1,\gamma_1) - \boldsymbol{\mu}_Z(\bullet;\boldsymbol{\beta}_2,\gamma_2)\| \leq C(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| + \|\gamma_1 - \gamma_2\|)$, and similar for Σ_Z.
- *Assumption 5.* Let $C > 0$ be a genuine constant. $K_1 = C|D_n|^{v_1}$, $1/(4r_1) < v_1 < 1/2$.
- *Assumption 6.* The spatial covariance functions $C_j(\boldsymbol{u}) = \mathrm{cov}\{\xi_j(\boldsymbol{s}),\xi_j(\boldsymbol{s}+\boldsymbol{u})\}$ are Lipschitz continuous at $\boldsymbol{0}$. There exists a constant M_0 such that $|C_j(\boldsymbol{u}) - \omega_j| \leq M_0\|\boldsymbol{u}\|$, for $j = 1,\ldots,p$.
- *Assumption 7.* Define the class of bivariate Hölder continuous functions on $[0,1]^{\otimes 2}$ as $C_2^{r,a} = \{f : \sup_{\boldsymbol{t}_1,\boldsymbol{t}_2\in[0,1]^{\otimes 2}} |f^{(u_1,u_2)}(\boldsymbol{t}_1) - f^{(u_1,u_2)}(\boldsymbol{t}_2)|/\|\boldsymbol{t}_1 - \boldsymbol{t}_2\|^a < \infty$, for $u_1,u_2 \geq 0, u_1 + u_2 \leq r\}$. We assume that $R_{T0} \in C_2^{r_2,a}$, where $r_2 \geq 2$ is the order of the tensor product spline function and $a > 0$.
- *Assumption 8.* Assume that $0 < C_5 \leq \mu_{\mathcal{M}}(t_1,t_2) \leq C_6 < \infty$ for all $t_1,t_2 \in T$,

$$\sup_{\boldsymbol{u}_1,\boldsymbol{u}_2,t_1,\ldots,t_4}\int_{\mathbb{R}^2} \mid \mathcal{G}_4((\boldsymbol{u}_1,\boldsymbol{u}_2,\boldsymbol{u}_3,t_1,t_2,t_3,t_4) - \mathcal{G}_2(\boldsymbol{u}_1,t_1,t_2)\mathcal{G}_2(\boldsymbol{u}_2,t_3,t_4) \mid d\boldsymbol{u}_3 < \infty.$$

- *Assumption 9.* Assume that $\delta \to 0$, $|D_n||D_\delta| \to \infty$ and $K_2 = C(|D_n||D_\delta|)^{v_2}$, $1/(4r_2) < v_2 < 1/2$.

SUPPLEMENTARY MATERIALS

The online supplementary material provides technical proofs of Theorems 1–3, variance estimation procedure for the mean estimators, model and tuning parameter selection methods, additional plots from the simulation study and model checking results for the CTR data analysis.

[Received August 2013. Revised November 2013.]

REFERENCES

Ash, R. B., and Gardner, M. F. (1975), *Topics in Stochastic Processes*, New York: Academic Press. [1206]

Baddeley, A. J., Møller, J., and Waagepetersen, R. (2000), "Non- and Semi-Parametric Estimation of Interaction in Inhomogeneous Point Patterns," *Statistica Neerlandica*, 54, 329–350. [1207]

Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005), "Residual Analysis for Spatial Point Processes" (with discussion), *Journal of the Royal Statistical Society,* Series B, 67, 617–666. [1209]

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2003), *Hierarchical Modeling and Analysis for Spatial Data*, New York: Chapman & Hall/CRC. [1207]

Brix, A., and Diggle, P. J. (2001), "Spatiotemporal Prediction for Log-Gaussian Cox Processes," *Journal of the Royal Statistical Society,* Series B, 63, 823–841. [1205,1206]

Brix, A., and Møller, J. (2001), "Space-Time Multi Type Log Gaussian Cox Processes With a View to Modelling Weeds," *Scandinavian Journal of Statistics*, 28, 471–488. [1205,1208]

Bouzas, P. R., Valderrama, M., Aguilera, A. M., and Ruiz-Fuentes, N. (2006), "Modeling the Mean of a Doubly Stochastic Poisson Process by Functional Data Analysis," *Computational Statistics & Data Analysis*, 50, 2655–2667. [1205]

Di, C., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009), "Multilevel Functional Principal Component Analysis," *The Annals of Applied Statistics*, 3, 458–488. [1205]

Diggle, P. J. (2006), "Spatio-Temporal Point Processes, Partial Likelihood, Foot and Mouth Disease," *Statistical Methods in Medical Research*, 15, 325–336. [1205]

Diggle, P., Rowlingson, B., and Su, T. (2005), "Point Process Methodology for Online Spatio-Temporal Disease Surveillance," *EnvironMetrics*, 16, 423–434. [1206,1208]

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004), "Nonstationary Multivariate Process Modeling Through Spatially Varying Coregionalization," *TEST*, 263–312. [1206]

Guan, Y. (2006), "A Composite Likelihood Approach in Fitting Spatial Point Process Models," *Journal of the American Statistical Association*, 101, 1502–1512. [1207,1209,1210]

——— (2008a), "On Consistent Nonparametric Intensity Estimation for Inhomogeneous Spatial Point Processes," *Journal of the American Statistical Association*, 103, 1238–1247. [1206]

——— (2008b), "A KPSS Test for Stationarity for Spatial Point Processes," *Biometrics*, 64, 800–806. [1209]

Guan, Y., and Loh, J. M. (2007), "A Thinned Block Bootstrap Variance Estimation Procedure for Inhomogeneous Spatial Point Patterns," *Journal of the American Statistical Association*, 102, 1377–1386. [1207]

Guan, Y., and Wang, H. (2010), "Sufficient Dimension Reduction for Spatial Point Processes Directed by Gaussian Random Fields," *Journal of the Royal Statistical Society,* Series B, 367–387. [1206]

Guyon, X. (1995), *Random Fields on a Network: Modeling, Statistics, and Applications*, New York: Springer-Verlag. [1214]

Hall, P., and Hosseini-Nasab, M. (2006), "On Properties of Functional Principal Components Analysis," *Journal of the Royal Statistical Society*, Series B, 68, 109–126. [1205,1210]

Hall, P., Müller, H. G., and Wang, J. -L. (2006), "Properties of Principal Component Methods for Functional and Longitudinal Data Analysis," *The Annals of Statistics*, 34, 1493–1517. [1205]

Hall, P., Müller, H. G., and Yao, F. (2008), "Modelling Sparse Generalized Longitudinal Observations With Latent Gaussian Processes," *Journal of the Royal Statistical Society*, Series B, 70, 703–723. [1205]

Heinrich, L., and Prokešová, M. (2010), "On Estimating the Asymptotic Variance of Stationary Point Processes," *Methodology and Computing in Applied Probability*, 12, 451–471. [1210]

Illian, J., Benson, E., Crawford, J., and Staines, H. (2006), "Principal Component Analysis for Spatial Point Processes –Assessing the Appropriateness of the Approach in an Ecological Context," in *Case Studies in Spatial Point Process Modeling, Lecture Notes in Statistics* (Vol. 185), New York: Springer, pp. 135–150. [1205]

Li, Y., and Hsing, T. (2010a), "Deciding the Dimension of Effective Dimension Reduction Space for Functional and High-Dimensional Data," *The Annals of Statistics*, 38, 3028–3062. [1205]

——— (2010b), "Uniform Convergence Rates for Nonparametric Regression and Principal Component Analysis in Functional/Longitudinal Data," *The Annals of Statistics*, 38, 3321–3351. [1205,1210]

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998), "Log-Gaussian Cox Processes," *Scandinavian Journal of Statistics*, 25, 451–482. [1205,1209]

Møller, J., and Waagepetersen, R. (2007), "Modern Statistics for Spatial Point Processes," *Scandinavian Journal of Statistics*, 34, 643–684. [1206]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer-Verlag. [1205,1206]

Rosenblatt, M. (1956), "A Central Limit Theorem and a Strong Mixing Condition," *Proceedings of the National Academy of Science*, 42, 43–47. [1213]

Schoenberg, F. P. (2003), "Multidimensional Residual Analysis of Point Process Models for Earthquake Occurrences," *Journal of the American Statistical Association*, 98, 789–795. [1205]

——— (2005), "Consistent Parametric Estimation of the Intensity of a Spatial temporal Point Process," *Journal of Statistical Planning and Inference*, 128, 79–93. [1207]

Stein, M. L. (1999), *Interpolation of Spatial Data*, New York: Springer. [1208]

Waagepetersen, R. P. (2007), "An Estimating Function Approach to Inference for Inhomogeneous Neyman–Scott Processes," *Biometrics*, 63, 252–258. [1207]

Wang, R., Gross, C. P., Halene, S., and Ma, X. (2009), "Neighborhood Socioeconomic Status Influences the Survival of Elderly Patients With Myelodysplastic Syndromes in the United States," *Cancer Causes and Control*, 20(8), 1369–1376. [1212]

Wu, S., Müller, H. G., and Zhang, Z. (2013), "Functional Data Analysis for Point Processes With Rare Events," *Statistica Sinica*, 23, 1–23. [1205]

Yao, F., Müller, H. G., and Wang, J. L. (2005a), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [1205]

——— (2005b), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903. [1205]

Zhou, L., Huang, J., Martinez, J. G., Maity, A., Baladandayuthapani, V., and Carroll, R. J. (2010), "Reduced Rank Mixed Effects Models for Spatially Correlated Hierarchical Functional Data," *Journal of the American Statistical Association*, 105, 390–400. [1205]

Zhou, S., Shen, X., and Wolfe, D. A. (1998), "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, 26, 1760–1782. [1207]

Zhu, Z., Fung, W. K., and He, X. (2008), "On the Asymptotics of Marginal Regression Splines With Longitudinal Data," *Biometrika*, 95, 907–917. [1207]

Downloaded by [141.211.66.220] at 11:33 02 October 2014

Supplementary Material to *Functional Principal Component Analysis of Spatio-Temporal Point Processes with Applications in Disease Surveillance*

Yehua Li
Department of Statistics & Statistical Laboratory, Iowa State University, Ames, IA 50011,
yehuali@iastate.edu

Yongtao Guan
Department of Management Science, University of Miami, Coral Gables, FL 33124,
yguan@bus.miami.edu

Web Appendix W. Technical Proofs

W.1 Lemmas

LEMMA **1** *Under Assumption 1 and 7, there exist* $\gamma^* \in \mathcal{S}^{r_1}_{K_1}$ *and* $R^* \in \mathcal{S}^{r_2}_{[2],K_2}$ *so that* $\|\gamma^* - \gamma\|_\infty = O(K_1^{-r_1})$ *and* $\|R^* - R_T\|_\infty = O(K_2^{-r_2})$.

Proof: The lemma follows directly from Corollary 6.21 and Theorem 12.7 of Schumaker (1981).

LEMMA **2** *Let* $g\{t, \boldsymbol{Z}(\boldsymbol{s},t)\}$ *be a measurable function with* $\sup_t \mathrm{E}|g\{t, \boldsymbol{Z}(\boldsymbol{s},t)\}|^C < \infty$ *and* $\sup_t \mathrm{E}|g\{t, \boldsymbol{Z}(\boldsymbol{s},t)\}\lambda(\boldsymbol{s},t)|^C < \infty$ *for some* $C > 2$. *Under the strong mixing condition in Assumption 3, we have the following weak convergence*

$$\frac{1}{\sqrt{|D_n|}}\Big[\int_T\int_{D_n} g\{t, \boldsymbol{Z}(\boldsymbol{s},t)\}d\boldsymbol{s}dt - \mathrm{E}\int_T\int_{D_n} g\{t, \boldsymbol{Z}(\boldsymbol{s},t)\}d\boldsymbol{s}dt\Big] \to \mathit{Normal}(0, \sigma_g^2),$$
$$\frac{1}{\sqrt{|D_n|}}\Big[\int_T\int_{D_n} g\{t, \boldsymbol{Z}(\boldsymbol{s},t)\}\{N(d\boldsymbol{s},dt) - \lambda(\boldsymbol{s},t)d\boldsymbol{s}dt\}\Big] \to \mathit{Normal}(0, \sigma_{N,g}^2),$$

where

$$\sigma_g^2 = \lim_{n\to\infty} |D_n|^{-1} var\Big[\int_T\int_{D_n} g\{t, \boldsymbol{Z}(\boldsymbol{s},t)\}d\boldsymbol{s}dt\Big]$$

and $$\sigma_{N,g}^2 = \lim_{n\to\infty} |D_n|^{-1} var\Big[\int_T\int_{D_n} g\{t, \boldsymbol{Z}(\boldsymbol{s},t)\}\{N(d\boldsymbol{s},dt) - \lambda(\boldsymbol{s},t)d\boldsymbol{s}dt\}\Big].$$

Proof: The lemma is proved by applying the central limit theorem for strong mixing random processes (Theorem 3.3.1, Guyon, 1995).

W.2 Proof of Theorems 1 and 2

By Lemma 1, there exists a spline function $\gamma^*(t) = \boldsymbol{B}^{\mathrm{T}}(t)\boldsymbol{\nu}^*$ so that $\|\gamma^* - \gamma_0\|_\infty = O(K_1^{-r_1})$. Put $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_0^{\mathrm{T}}, K_1^{-1/2}\boldsymbol{\nu}^*)$.

LEMMA **3** *(Consistency of the estimators) Under the assumptions in Theorem 1,* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = o_p(1)$ *and* $\|\widehat{\gamma} - \gamma_0\|_\infty = o_p(1)$.

Proof: It is easy to see that the Poisson likelihood function (9) is convex function of β and γ, hence there is a unique maximizer $(\widehat{\beta}, \widehat{\gamma})$. We only need to show that they are in a $o(1)$ neighborhood of (β_0, γ_0) with a probability approaching 1.

For any $\boldsymbol{\beta}_n \in \mathbb{R}^d$ and $\gamma_n(t) = \boldsymbol{B}^{\mathrm{T}}(t)\boldsymbol{\nu}_n$ with $\|\boldsymbol{\beta}_n\|^2 + \|\gamma_n(t)\|^2 = O(K_1^{-a})$ such that $1 < a < 1/(2\nu_1)$ where ν_1 is defined in Assumption 5. Define $H(x) = |D_n|^{-1}\ell(\boldsymbol{\beta}_0 + x\boldsymbol{\beta}_n, \gamma^* + x\gamma_n)$ for a scalar x. We will show that for any $x_0 > 0$, $H'(x_0) < 0$ and $H'(-x_0) > 0$ except on an event with probability tending to 0, which implies $\widehat{\beta}$ is between $\boldsymbol{\beta}_0 \pm x_0\boldsymbol{\beta}_n$ and $\widehat{\gamma}$ is between $\gamma^* \pm x_0\gamma_n$.

$$H'(x) = \frac{1}{|D_n|}\int\int \{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}_n + \gamma_n(t)\}\Big[N(d\boldsymbol{s},dt) - \exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)(\boldsymbol{\beta}_0 + x\boldsymbol{\beta}_n) + (\gamma^* + x\gamma_n)(t)\}d\boldsymbol{s}dt\Big].$$

By Lemma 2, when $|D_n|$ is large enough,

$$\begin{aligned}
H'(x) &= \frac{1}{|D_n|}\mathrm{E}\int_T\int_{D_n} \{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}_n + \gamma_n(t)\}\Big[\exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}_0 + \gamma_0(t)\} \\
&\qquad - \exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)(\boldsymbol{\beta}_0 + x\boldsymbol{\beta}_n) + (\gamma^* + x\gamma_n)(t)\}\Big]d\boldsymbol{s}dt + O_p(|D_n|^{-1/2}) \\
&= -\frac{x}{|D_n|}\mathrm{E}\Big(\int_T\int_{D_n}\Big[\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}_n + \gamma_n(t)\}^2 + \{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}_n + \gamma_n(t)\}(\gamma^* - \gamma_0)(t)\Big] \\
&\qquad \times \exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}_0 + \gamma_0(t)\}d\boldsymbol{s}dt\Big) \times \{1 + o(1)\} + O_p(|D_n|^{-1/2}) \\
&\le -xCK_1^{-a} + O(K_1^{-r_1 - a/2}) + O_p(|D_n|^{-1/2}),
\end{aligned}$$

which, by Assumption 5, is negative with a probability tending to 1. Similarly, one can show that $H'(-x) > 0$ with a probability tending to 1. Thus, we have shown that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 =$

$O_p(K_1^{-a/2}) = o_p(1)$ and $\|\widehat{\gamma} - \gamma^*\| = o_p(K_1^{-1/2})$. By Lemma 6.1 of Zhou et al. (1998),

$$\|\widehat{\gamma} - \gamma^*\|^2 \geq cK_1^{-1}\|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\|^2,$$

hence $\|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\| = o_p(1)$. By Lemma 7 in Stone (1986), $\|\widehat{\gamma} - \gamma^*\|_\infty \leq cK_1^{1/2}\|\widehat{\gamma} - \gamma^*\| = o_p(1)$.

LEMMA **4** *Let $\mathbb{X}(\boldsymbol{s}, t)$ and $\boldsymbol{\theta}$ be defined as in (11). For $\boldsymbol{\theta} \in \mathbb{R}^{K_1+p}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < C$, define*

$$\mathcal{D}_n(\boldsymbol{\theta}) = \frac{1}{|D_n|}\int\int \mathbb{X}^{\otimes 2}(\boldsymbol{s}, t)\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\theta}\}d\boldsymbol{s}dt, \tag{W.1}$$

then there exist constants $0 < c_1 < c_2 < \infty$, which do not depend on K_1 such that

$$\{c_1 + o_p(1)\} \leq \lambda_{\min}(\mathcal{D}_n) \leq \lambda_{\max}(\mathcal{D}_n) \leq \{c_2 + o_p(1)\}. \tag{W.2}$$

As a result, $\|\mathcal{D}_n^{-1}\| = O_p(1)$.

Proof: We partition $\mathcal{D}_n(\boldsymbol{\theta})$ as

$$\mathcal{D}_n(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{D}_{11} & \mathcal{D}_{12} \\ \mathcal{D}_{21} & \mathcal{D}_{22} \end{pmatrix}, \tag{W.3}$$

where $\mathcal{D}_{11} = |D_n|^{-1}\int\int \boldsymbol{Z}^{\otimes 2}(\boldsymbol{s}, t)\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\theta}\}d\boldsymbol{s}dt$ is the upper left $p \times p$ block of $\mathcal{D}_n$.

We first show that all eigenvalues of $\mathcal{D}_{22}$ are bounded. Define $q_\theta(t) = \mathrm{E}\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\theta}\}$ and $q_{n\theta}(t) = |D_n|^{-1}\int_{D_n}\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s}, t)\boldsymbol{\theta}\}d\boldsymbol{s}$. By Lemma 2, $q_{n\theta}(t) - q_\theta(t) = O_p(|D_n|^{-1/2})$ and hence $\|q_{n\theta} - q_\theta\| = O_p(|D_n|^{-1/2})$. Let $\boldsymbol{a}$ be any K_1 dimensional vector and denote $S(t) = \widetilde{\boldsymbol{B}}^{\mathrm{T}}(t)\boldsymbol{a}$, then

$$\boldsymbol{a}^{\mathrm{T}}\mathcal{D}_{22}\boldsymbol{a} = \int_T S^2(t)q_{n\theta}(t)dt = \Big\{\int_T S^2(t)q_\theta(t)dt\Big\}\{1 + o_p(1)\}.$$

By Assumption 3, $0 < C_1 \leq \min_t q_\theta(t) \leq \max_t q_\theta(t) \leq C_2 < \infty$, hence

$$C_1\Big\{\int S^2(t)dt\Big\}\{1 + o_p(1)\} \leq \boldsymbol{a}^{\mathrm{T}}\mathcal{D}_{22}\boldsymbol{a} \leq C_2\Big\{\int S^2(t)dt\Big\}\{1 + o_p(1)\}.$$

By equation (13) in Zhou et al. (1998), $\widetilde{C}_1\|\boldsymbol{a}\|^2 \leq \int S^2(t)dt \leq \widetilde{C}_2\|\boldsymbol{a}\|^2$, where $\widetilde{C}_1 > 0$ and $\widetilde{C}_2 < \infty$ are constants not relying on K_1. Therefore,

$$C_1\widetilde{C}_1 \times \{1 + o_p(1)\} \leq (\boldsymbol{a}^{\mathrm{T}}\mathcal{D}_{22}\boldsymbol{a})/\|\boldsymbol{a}\|^2 \leq C_2\widetilde{C}_2 \times \{1 + o_p(1)\}, \tag{W.4}$$

and hence all eigenvalues of $\mathcal{D}_{22}$ are bounded by constants.

Define

$$\widetilde{\mathcal{D}}_n = \mathcal{P}^{\mathrm{T}}\mathcal{D}_n(\boldsymbol{\theta})\mathcal{P} := \begin{pmatrix} I & -\mathcal{D}_{12}\mathcal{D}_{22}^{-1} \\ \mathbf{0} & I \end{pmatrix} \mathcal{D}_n(\boldsymbol{\theta}) \begin{pmatrix} I & \mathbf{0} \\ -\mathcal{D}_{22}^{-1}\mathcal{D}_{21} & I \end{pmatrix} = \begin{pmatrix} \widetilde{\mathcal{D}}_{11} & \mathbf{0} \\ \mathbf{0} & \mathcal{D}_{22} \end{pmatrix},$$

where $\widetilde{\mathcal{D}}_{11} = \mathcal{D}_{11} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21}$. Since all eigenvalues of $\mathcal{P}$ are equal to 1, eigenvalues of $\mathcal{D}_n(\boldsymbol{\theta})$ are the same as those of $\widetilde{\mathcal{D}}_n$. To show the lemma, we only need to show that all eigenvalues of $\widetilde{\mathcal{D}}_{11}$ are bounded by nonzero constants.

Let $\boldsymbol{\theta}_2$ be a subvector of $\boldsymbol{\theta}$ containing the last K_1 entries, and $\gamma(t) = \widetilde{\boldsymbol{B}}^{\mathrm{T}}(t)\boldsymbol{\theta}_2$. We want to show that $\widehat{\boldsymbol{\mu}}_Z(t;\boldsymbol{\beta},\gamma) = \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\widetilde{\boldsymbol{B}}(t)$ is a regression spline estimator of $\boldsymbol{\mu}_Z(t;\boldsymbol{\beta},\gamma)$ defined in Assumption 4. By Assumption 4 and Lemma 1, there exists $\boldsymbol{\mu}_Z^*(t;\boldsymbol{\beta},\gamma) = \boldsymbol{V}_Z^*\widetilde{\boldsymbol{B}}(t)$ such that $\|(\mu_{Z,j}^* - \mu_{Z,j})(t;\boldsymbol{\beta},\gamma)\|_\infty = O(K_1^{-r_1})$, where $\mu_{Z,j}(t;\boldsymbol{\beta},\gamma)$ is the jth entry of $\boldsymbol{\mu}_Z(t;\boldsymbol{\beta},\gamma)$ and $\boldsymbol{V}_Z^*$ is a $d \times K_1$ matrix of spline coefficients. We have

$$\begin{aligned} \mathcal{D}_{12}\mathcal{D}_{22}^{-1} &= \Big[\frac{1}{|D_n|}\int\int\{\boldsymbol{Z}(\boldsymbol{s},t) - \boldsymbol{\mu}_Z^*(t;\boldsymbol{\beta},\gamma) + \boldsymbol{V}_Z^*\widetilde{\boldsymbol{B}}(t)\}\widetilde{\boldsymbol{B}}^{\mathrm{T}}(t)\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\theta}\}d\boldsymbol{s}dt\Big]\mathcal{D}_{22}^{-1} \\ &= \boldsymbol{V}_Z^* + (\boldsymbol{G}_1 + \boldsymbol{G}_2)\mathcal{D}_{22}^{-1}, \end{aligned}$$

where $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ are $d \times K_1$ matrices with the (j,ℓ)th entries

$$\begin{aligned} G_{1,j\ell} &= |D_n|^{-1}\int\int\{Z_j(\boldsymbol{s},t) - \mu_{Z,j}(t;\boldsymbol{\beta},\gamma)\}\widetilde{B}_\ell(t)\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\theta}\}d\boldsymbol{s}dt \\ &\quad + \int\{\mu_{Z,j}(t;\boldsymbol{\beta},\gamma) - \mu_{Z,j}^*(t;\boldsymbol{\beta},\gamma)\}\widetilde{B}_\ell(t)\{q_{n\theta}(t) - q_\theta(t)\}dt \\ \text{and} \quad G_{2,j\ell} &= \int\{\mu_{Z,j}(t;\boldsymbol{\beta},\gamma) - \mu_{Z,j}^*(t;\boldsymbol{\beta},\gamma)\}\widetilde{B}_\ell(t)q_\theta(t)dt. \end{aligned}$$

By Lemma 2, $G_{1,j\ell} = O_p(|D_n|^{-1/2})$, and by the definition of $\mu_{Z,j}$ we have $\max_\ell |G_{2,j\ell}| = O(K_1^{-r_1})$. Let $\boldsymbol{G}_{1,j\bullet}^{\mathrm{T}}$ and $\boldsymbol{G}_{2,j\bullet}^{\mathrm{T}}$ be the jth row of $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ respectively.

Therefore

$$\begin{aligned} \|\widehat{\mu}_{Z,j} - \mu_{Z,j}^*\|^2 &= \|(\boldsymbol{G}_{1,j\bullet} + \boldsymbol{G}_{2,j\bullet})^{\mathrm{T}}\mathcal{D}_{22}^{-1}\widetilde{\boldsymbol{B}}\|^2 \le C\|\mathcal{D}_{22}^{-1}(\boldsymbol{G}_{1,j\bullet} + \boldsymbol{G}_{2,j\bullet})\|^2 \\ &\le 2C(\|\mathcal{D}_{22}^{-1}\boldsymbol{G}_{1,j\bullet}\|^2 + \|\mathcal{D}_{22}^{-1}\boldsymbol{G}_{2,j\bullet}\|^2). \end{aligned}$$

It is easy to see that $\|\mathcal{D}_{22}^{-1}\boldsymbol{G}_{1,j\bullet}\|^2 \le \lambda_{\max}^2(\mathcal{D}_{22}^{-1})\|\boldsymbol{G}_{1,j\bullet}\|^2 = O_p(K_1/|D_n|)$. Since the B-splines are compact supported basis functions, it is easy to see that $\mathcal{D}_{22}$ is a band matrix and by Theorem 2.2 of Demko (1977), the off-diagonal entries of $\mathcal{D}_{22}^{-1}$ decay exponentially. As a

result, $\|\mathcal{D}_{22}^{-1}\|_\infty = O(1)$ and $\|\mathcal{D}_{22}^{-1}\boldsymbol{G}_{2,j\bullet}\|^2 \le \|\mathcal{D}_{22}^{-1}\|_\infty^2\|\boldsymbol{G}_{2,j\bullet}\|_\infty^2 = O(K_1^{-2r_1})$. We have just shown that $\|(\widehat{\boldsymbol{\mu}}_Z - \boldsymbol{\mu}_Z)(t;\boldsymbol{\beta},\gamma)\| = O_p\{K_1^{-r_1} + (K_1/|D_n|)^{1/2}\}$.

One can also see that

$$\begin{aligned}\widetilde{\mathcal{D}}_{11} &= \mathcal{D}_{11} - (\mathcal{D}_{12}\mathcal{D}_{22}^{-1})\mathcal{D}_{22}(\mathcal{D}_{22}^{-1}\mathcal{D}_{21}) \\ &= \frac{1}{|D_n|}\int\int\{\boldsymbol{Z}^{\otimes 2}(\boldsymbol{s},t) - \widehat{\boldsymbol{\mu}}_Z^{\otimes 2}(t;\boldsymbol{\beta},\gamma)\}\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\theta}\}d\boldsymbol{s}dt \\ &= \Sigma_Z(\boldsymbol{\beta},\gamma) + O_p\{K_1^{-r_1} + (K_1/|D_n|)^{1/2}\},\end{aligned}$$

and hence by Assumption 4, all eigenvalues of $\widetilde{\mathcal{D}}_{11}$ are bounded. Inequality (W.2) follows immediately. Finally, $\|\mathcal{D}_n^{-1}\| = \lambda_{\min}^{-1}(\mathcal{D}_n) = O_p(1)$.

Proof of Theorem 1: Let $\mathbb{X}$ and $\boldsymbol{\theta}$ be defined as above. The estimator $\widehat{\boldsymbol{\theta}}$ satisfies

$$\mathbf{0} = \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\widehat{\boldsymbol{\theta}}) = \int\int\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\Big[N(d\boldsymbol{s},dt) - \exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\widehat{\boldsymbol{\theta}}\}d\boldsymbol{s}dt\Big].$$

By Lemma 3, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = o_p(1)$, we take a Taylor expansion of the equation around $\boldsymbol{\theta}^*$,

$$\mathbf{0} = \mathcal{C}_n(\boldsymbol{\theta}^*) - \mathcal{D}_n(\bar{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \tag{W.5}$$

where

$$\mathcal{C}_n(\boldsymbol{\theta}^*) = |D_n|^{-1}\int\int\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)[N(d\boldsymbol{s},dt) - \exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\theta}^*\}d\boldsymbol{s}dt], \tag{W.6}$$

$\mathcal{D}_n(\boldsymbol{\theta})$ is defined in Lemma 4, and $\bar{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}^*$ and $\widehat{\boldsymbol{\theta}}$. By Lemma 2 and 4, it is easy to see that

$$\|\mathcal{C}_n(\boldsymbol{\theta}^*)\|^2 = K_1 \times O_p(|D_n|^{-1} + K_1^{-2r_1}), \quad \|\mathcal{D}_n^{-1}(\bar{\boldsymbol{\theta}})\| = O_p(1).$$

Therefore,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = \|\mathcal{D}_n^{-1}(\bar{\boldsymbol{\theta}})\mathcal{C}_n(\boldsymbol{\theta}^*)\| \le \|\mathcal{D}_n^{-1}(\bar{\boldsymbol{\theta}})\|\|\mathcal{C}_n(\boldsymbol{\theta}^*)\| = O_p\{(K_1/|D_n|)^{1/2} + K_1^{-r_1+1/2}\}. \tag{W.7}$$

The convergence rate of $\widehat{\boldsymbol{\beta}}$ follows directly from (W.7). By equation (13) in Zhou et al. (1998), $\|\widehat{\gamma} - \gamma_0\| \le \|\widehat{\gamma} - \gamma^*\| + \|\gamma^* - \gamma_0\| \le cK_1^{-1/2}\|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\| + O(K_1^{-r_1})$, and the convergence rate of $\widehat{\gamma}$ follows.

Proof of Theorem 2: Since $\widehat{\boldsymbol{\theta}}$ satisfies the score equation (W.5), we have $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \mathcal{D}_n(\bar{\boldsymbol{\theta}})^{-1}\mathcal{C}_n(\boldsymbol{\theta}^*)$. Suppose $\mathcal{D}_n(\boldsymbol{\theta})$ has the partition as in Lemma 4, and partition $\mathcal{C}_n(\boldsymbol{\theta})$ accordingly as $(\mathcal{C}_{1n}^{\mathrm{T}}, \mathcal{C}_{2n}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{\theta})$. By matrix algebra (Mardia et al., 1979, page 459),

$$\mathcal{D}_n^{-1} = \begin{pmatrix} \mathcal{D}_n^{11} & \mathcal{D}_n^{12} \\ \mathcal{D}_n^{21} & \mathcal{D}_n^{22} \end{pmatrix},$$

where $\mathcal{D}_n^{11} = (\mathcal{D}_{11} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21})^{-1}$, $\mathcal{D}_n^{12} = -(\mathcal{D}_{11} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21})^{-1}\mathcal{D}_{12}\mathcal{D}_{22}^{-1}$, $\mathcal{D}_n^{21} = -\mathcal{D}_{22}^{-1}\mathcal{D}_{21}$ $(\mathcal{D}_{11} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21})^{-1}$, $\mathcal{D}_n^{22} = \mathcal{D}_{22}^{-1} + \mathcal{D}_{22}^{-1}\mathcal{D}_{21}(\mathcal{D}_{11} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21})^{-1}\mathcal{D}_{12}\mathcal{D}_{22}^{-1}$. Therefore,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &= \mathcal{D}_n^{11}(\bar{\boldsymbol{\theta}})\mathcal{C}_{1n}(\boldsymbol{\theta}^*) + \mathcal{D}_n^{12}(\bar{\boldsymbol{\theta}})\mathcal{C}_{2n}(\boldsymbol{\theta}^*) \\
&= (\mathcal{D}_{11} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21})^{-1} \\
&\quad \times \frac{1}{|D_n|}\int\int \Big\{\boldsymbol{Z}(\boldsymbol{s},t) - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\widetilde{\boldsymbol{B}}(t)\Big\}\Big[N(d\boldsymbol{s},dt) - \exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\theta}^*\}d\boldsymbol{s}dt\Big].
\end{aligned}$$

Using the arguments in Lemma 4, we can show that $\mathcal{D}_{11} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21} = \Sigma_Z(\boldsymbol{\beta}_0, \gamma_0) + o_p(1)$ and $\widehat{\boldsymbol{\mu}}_Z(t;\boldsymbol{\beta}_0,\gamma^*) = \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\widetilde{\boldsymbol{B}}(t) = \boldsymbol{\mu}_Z(t;\boldsymbol{\beta}_0,\gamma^*) + O_p\{K_1^{-r_1} + (K_1/|D_n|)^{1/2}\} = \boldsymbol{\mu}_Z(t;\boldsymbol{\beta}_0,\gamma_0) + O_p\{K_1^{-r_1} + (K_1/|D_n|)^{1/2}\}$.

Denote $\lambda_0(\boldsymbol{s},t) = \exp\{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\beta}_0 + \gamma_0(t)\}$, then

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &= \frac{1}{|D_n|}\Sigma_{Z0}^{-1}\Big[\int\int \Big\{\boldsymbol{Z}(\boldsymbol{s},t) - \widehat{\boldsymbol{\mu}}_Z(t;\boldsymbol{\beta}_0,\gamma^*)\Big\}\mathcal{E}(d\boldsymbol{s},dt) \\
&\qquad + \int\int \Big\{\boldsymbol{Z}(\boldsymbol{s},t) - \widehat{\boldsymbol{\mu}}_Z(t;\boldsymbol{\beta}_0,\gamma^*)\Big\}(\gamma^* - \gamma_0)(t)\lambda_0(\boldsymbol{s},t)d\boldsymbol{s}dt\Big] \times \{1 + o_p(1)\} \\
&= \frac{1}{|D_n|}\Sigma_{Z0}^{-1}\int\int \Big\{\boldsymbol{Z}(\boldsymbol{s},t) - \boldsymbol{\mu}_Z(t;\boldsymbol{\beta}_0,\gamma_0)\Big\}\mathcal{E}(d\boldsymbol{s},dt) + o_p(|D_n|^{-1/2}) \\
&\qquad + O_p[K_1^{-r_1} \times \{K_1^{-r_1} + (K_1/|D_n|)^{1/2}\}].
\end{aligned}$$

By Assumption 4, $K_1^{-r_1} \times \{K_1^{-r_1} + (K_1/|D_n|)^{1/2}\} = o_p(|D_n|^{-1/2})$, then the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ in Theorem 2 follows directly from the central limit theorem in Lemma 2.

By solving the second set of equations in (W.5), we have

$$\begin{aligned}
K_1^{-1/2}(\widehat{\nu} - \nu^*) &= \mathcal{D}_n^{21}(\bar{\boldsymbol{\theta}})\mathcal{C}_{1n}(\boldsymbol{\theta}^*) + \mathcal{D}_n^{22}(\bar{\boldsymbol{\theta}})\mathcal{C}_{2n}(\boldsymbol{\theta}^*) \\
&= \mathcal{D}_{22}^{-1}\mathcal{C}_{2n} + \mathcal{D}_{22}^{-1}\mathcal{D}_{21}(\mathcal{D}_{11} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21})^{-1}(\mathcal{C}_{1n} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{C}_{2n}) \\
&= \mathcal{D}_{22}^{-1}\mathcal{C}_{2n} + \mathcal{D}_{22}^{-1}\mathcal{D}_{21}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).
\end{aligned}$$

Using the argument as above, we can show that $\mathcal{D}_{22}^{-1}\mathcal{D}_{21}$ is $K_1^{-1/2}$ times the spline coefficients of $\widehat{\boldsymbol{\mu}}_Z(t;\bar{\boldsymbol{\beta}},\bar{\gamma})$, and hence $\|\mathcal{D}_{22}^{-1}\mathcal{D}_{21}\|^2 \le C\|\widehat{\boldsymbol{\mu}}_Z(t;\bar{\boldsymbol{\beta}},\bar{\gamma})\|^2 = O_p(1)$. Denote the jth entry of $\mathcal{C}_{2n}$

as $\mathcal{C}_{2n,j}$, then

$$\begin{aligned}\mathcal{C}_{2n,j}(\boldsymbol{\theta}^*) &= \frac{K_1^{1/2}}{|D_n|}\int\int B_j(t)\Big[N(d\boldsymbol{s},dt)-\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\boldsymbol{\theta}^*\}d\boldsymbol{s}dt\Big]\\ &= \frac{K_1^{1/2}}{|D_n|}\int\int B_j(t)\mathcal{E}(d\boldsymbol{s},dt)+\frac{K_1^{1/2}}{|D_n|}\int\int B_j(t)\lambda_0(\boldsymbol{s},t)(\gamma^*-\gamma_0)(t)d\boldsymbol{s}dt\times\{1+o_p(1)\}\\ &= O_p(|D_n|^{-1/2}+K_1^{-r_1-1/2}).\end{aligned}$$

Therefore, $\|\mathcal{C}_{2n}\|^2=O_p(K_1/|D_n|+K_1^{-2r_1})$. By (W.4), $\|\mathcal{D}_{22}^{-1}\|=O_p(1)$. Hence,

$$\begin{aligned}&K_1^{-1/2}\|\widehat{\nu}-\nu^*\|\le\|\mathcal{D}_{22}^{-1}\|\|\mathcal{C}_{2n}\|+\|\mathcal{D}_{22}^{-1}\mathcal{D}_{21}\|\|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|=O_p(K_1^{1/2}|D_n|^{-1/2}+K_1^{-r_1}),\\ &\|\widehat{\gamma}-\gamma^*\|^2\le CK_1^{-1}\|\widehat{\nu}-\nu^*\|^2=O_p(K_1/|D_n|+K_1^{-2r_1}).\end{aligned}$$

By the definition of γ^*, we have $\|\widehat{\gamma}-\gamma_0\|=O_p(K_1^{1/2}|D_n|^{-1/2}+K_1^{-r_1})$.

W.3 Proof for Theorem 3

Define the process

$$N_\delta(d\boldsymbol{s},dt_1,dt_2)=N(d\boldsymbol{s},dt_1)\int_{D_{\boldsymbol{s},\delta}}I\{(\boldsymbol{u},t_2)\neq(\boldsymbol{s},t_1)\}N(d\boldsymbol{u},dt_2),$$

with the intensity function

$$\lambda_\delta(\boldsymbol{s},t_1,t_2)=\lim_{|d\boldsymbol{s}|,|dt_1|,|dt_2|\to0}\frac{\mathrm{E}\{N_\delta(d\boldsymbol{s},dt_1,dt_2)\}}{|d\boldsymbol{s}||dt_1||dt_2|}=\int_{D_{\boldsymbol{s},\delta}}\lambda_2(\boldsymbol{s},t_1,d\boldsymbol{u},t_2)=O_p(|D_{\boldsymbol{s},\delta}|).$$

Letting $\mathcal{E}_\delta(d\boldsymbol{s},dt_1,dt_2)=|D_{\boldsymbol{s},\delta}|^{-1/2}\{N_\delta(d\boldsymbol{s},dt_1,dt_2)-\lambda_\delta(\boldsymbol{s},t_1,t_2)d\boldsymbol{s}dt_1dt_2\}$, one can show that

$$\lim_{|d\boldsymbol{s}|,|dt_1|,|dt_2|\to0}\frac{\mathrm{E}\{\mathcal{E}_\delta(d\boldsymbol{s},dt_1,dt_2)\}^2}{|d\boldsymbol{s}|^2|dt_1|^2|dt_2|^2}=O(1).$$

It is easy to verify that $\mathcal{E}_\delta$ satisfies the strong mixing condition in $\boldsymbol{s}$ as in Assumption 2.

By Lemma 1, there exists a $R^*(t_1,t_2)=\boldsymbol{B}_{[2]}^{\mathrm{T}}(t_1,t_2)\boldsymbol{b}^*$ such that $\|R^*-R_T\|_\infty=O(K_2^{-r_2})$. It is easy to see that composite likelihood function (14) is convex and the estimating equation (16) has an unique solution. By taking an Taylor's expansion at $(\boldsymbol{\beta}_0,\gamma_0,\boldsymbol{b}^*)$, (16) can be re-written as

$$\mathbf{0}\approx\mathcal{A}-\mathcal{B}(\widehat{\boldsymbol{b}}-\boldsymbol{b}^*),$$

where $\mathcal{A} = \mathcal{A}_1 - \mathcal{A}_2 - (\mathcal{A}_3 + \mathcal{A}_4) \times \{1 + o_p(1)\}$,

$$\mathcal{A}_1 = \frac{1}{|D_n||D_\delta|^{1/2}} \int_{D_n} \int_T \int_T \boldsymbol{B}_{[2]}(t_1, t_2) \mathcal{E}_\delta(d\boldsymbol{s}, dt_1, dt_2),$$

$$\mathcal{A}_2 = \frac{1}{|D_n||D_\delta|} \int_T \int_T \int_{D_n} \int_{D_{\boldsymbol{s}_1,\delta}} \boldsymbol{B}_{[2]}(t_1, t_2)[\lambda_2(\boldsymbol{s}_1, t_1, \boldsymbol{s}_2, t_2) - \lambda(\boldsymbol{s}_1, t_1)\lambda(\boldsymbol{s}_2, t_2) \exp\{R^*(t_1, t_2)\}] d\boldsymbol{s}_2 d\boldsymbol{s}_1 dt_2 dt_1,$$

$$\mathcal{A}_3 = \Big[\frac{1}{|D_n||D_\delta|} \int_T \int_T \int_{D_n} \int_{D_{\boldsymbol{s}_1,\delta}} \boldsymbol{B}_{[2]}(t_1, t_2)\{\boldsymbol{Z}(\boldsymbol{s}_1, t_1) + \boldsymbol{Z}(\boldsymbol{s}_2, t_2)\}^{\mathrm{T}} \lambda(\boldsymbol{s}_1, t_1)\lambda(\boldsymbol{s}_2, t_2) \times \exp\{R^*(t_1, t_2)\} d\boldsymbol{s}_2 d\boldsymbol{s}_1 dt_2 dt_1\Big](\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

$$\mathcal{A}_4 = \Big[\frac{1}{|D_n||D_\delta|} \int_T \int_T \int_{D_n} \int_{D_{\boldsymbol{s}_1,\delta}} \boldsymbol{B}_{[2]}(t_1, t_2)\{\widehat{\gamma}(t_1) - \gamma_0(t_1) + \widehat{\gamma}(t_2) - \gamma_0(t_2)\} \times \lambda(\boldsymbol{s}_1, t_1)\lambda(\boldsymbol{s}_2, t_2) \exp\{R^*(t_1, t_2)\} d\boldsymbol{s}_2 d\boldsymbol{s}_1 dt_2 dt_1,$$

$$\mathcal{B} = \frac{1}{|D_n||D_\delta|} \int_T \int_T \int_{D_n} \int_{D_{\boldsymbol{s}_1,\delta}} \boldsymbol{B}_{[2]}^{\otimes 2}(t_1, t_2) \lambda(\boldsymbol{s}_1, t_1)\lambda(\boldsymbol{s}_2, t_2) \exp\{R^*(t_1, t_2)\} d\boldsymbol{s}_2 d\boldsymbol{s}_1 dt_2 dt_1.$$

Note that $\mathcal{A}_1 - \mathcal{A}_4$ are K_2^2-dim vectors, and denote the entries corresponding to the spline basis function $B_{jj'}(t_1, t_2)$ as $A_{i,jj'}$, $i = 1, \ldots, 4$, $j, j' = 1, \ldots, K_2$. Since $B_{jj'}(\cdot, \cdot)$ are bounded functions compact supported on $O(K_2^{-2})$ rectangular region, using similar derivations as in Theorem 1, we can show that

$$A_{1,jj'} = O_p\{(|D_n||D_\delta|K_2^2)^{-1/2}\}, \quad A_{2,jj'} = O_p\{(|D_n||D_\delta|K_2^2)^{-1/2} + K_2^{-2}\} \times O_p(K_2^{-r_2} + \delta),$$
$$A_{3,jj'} = O_p\{(|D_n||D_\delta|K_2^2)^{-1/2} + K_2^{-2}\} \times O_p(|D_n|^{-1/2}),$$
$$A_{4,jj'} = O_p\{(|D_n||D_\delta|K_2^2)^{-1/2} + K_2^{-2}\} \times O_p\{(K_1^{1/2}|D_n|^{-1/2} + K_1^{-r_1})\}.$$

Therefore $\|\mathcal{A}\|^2 = O_p[(|D_n||D_\delta|)^{-1} + K_2^{-2r_2-2} + K_2^{-2}\delta^2 + K_2^{-2}(K_1/|D_n| + K_1^{-2r_1})]$.

Letting $\mathcal{M}(t_1, t_2)$ be as defined in (A.3), we can see that $\mathcal{B} = \int_T \int_T \boldsymbol{B}_{[2]}^{\otimes 2}(t_1, t_2) \exp\{R^*(t_1, t_2)\} \mathcal{M}(t_1, t_2) dt_1 dt_2$. Using arguments as above we can show that $\mathcal{M}(t_1, t_2) = \mu_{\mathcal{M}}(t_1, t_2) + O_p\{(|D_n||D_\delta|)^{-1/2}\}$. For any vector $\boldsymbol{b} \in \mathbb{R}^{K_2^2}$, let $\mathcal{S}(t_1, t_2) = \boldsymbol{B}_{[2]}^{\mathrm{T}}(t_1, t_2)\boldsymbol{b}$. By Assumptions 7 and 8, we can see that with a probability tending to one

$$C_1 \int\int \mathcal{S}^2(t_1, t_2) dt_1 dt_2 \;\le\; \boldsymbol{b}^{\mathrm{T}} \mathcal{B} \boldsymbol{b} = \int_T \int_T \mathcal{S}^2(t_1, t_2) \exp\{R^*(t_1, t_2)\} \mathcal{M}(t_1, t_2) dt_1 dt_2 \le C_2 \int\int \mathcal{S}^2(t_1, t_2) dt_1 dt_2,$$

for some $0 < C_1 < C_2 < \infty$. Using the arguments similar to Lemma 3.10 of Stone (1994),

we can show

$$C_3 K_2^{-2}\|\boldsymbol{b}\|^2 \le \int\int \mathcal{S}^2(t_1,t_2)dt_1dt_2 \le C_4 K_2^{-2}\|\boldsymbol{b}\|^2.$$

for some $0 < C_3 < C_4 < \infty$. Therefore, $C_1C_3K_2^{-2} \le \lambda_{\min}(\mathcal{B}) \le \boldsymbol{b}^{\mathrm{T}}\mathcal{B}\boldsymbol{b}/\|\boldsymbol{b}\|^2 \le \lambda_{\max}(\mathcal{B}) \le C_2C_4K_2^{-2}$, and

$$\|\widehat{\boldsymbol{b}} - \boldsymbol{b}^*\| = \|\mathcal{B}^{-1}\mathcal{A}\| \le \lambda_{\min}^{-1}(\mathcal{B})\|\mathcal{A}\| = O_p(K_2^2\|\mathcal{A}\|).$$

Finally,

$$\begin{aligned}\|\widehat{R}_T - R^*\|^2 &= O_p(K_2^{-2}\|\widehat{\boldsymbol{b}} - \boldsymbol{b}^*\|^2) = O_p(K_2^2\|\mathcal{A}\|^2) \\ &= O_p[K_2^2(|D_n||D_\delta|)^{-1} + K_2^{-2r_2} + \delta^2 + K_1/|D_n| + K_1^{-2r_1}].\end{aligned}$$

Theorem 3 follows directly from the fact that $\|R^* - R_{T0}\|_\infty = O(K_2^{-r_2})$.

W.4 Variance estimation for the mean estimators

The most important inference problem that we need to address is about the standard errors in $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}(t)$. Since the coefficient vector $\boldsymbol{\theta}$ is estimated by solving the estimating equation (11), the covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by a sandwich formula. By the theoretical derivations in Theorem 2, $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx \mathcal{D}_n^{-1}(\boldsymbol{\theta}^*)\mathcal{C}_n(\boldsymbol{\theta}^*)$, where $\mathcal{D}_n$ and $\mathcal{C}_n$ are defined in (W.1) and (W.6). It is easy to see that the asymptotic covariance matrix of $|D_n|^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is

$$\Omega_\theta = \mathcal{D}_n^{-1}(\boldsymbol{\theta}^*)\Omega_{\mathcal{C}}\mathcal{D}_n^{-1}(\boldsymbol{\theta}^*), \quad \text{where } \Omega_{\mathcal{C}} = |D_n| \times \mathrm{cov}\{\mathcal{C}_n(\boldsymbol{\theta}^*)\}.$$

By straightforward calculations,

$$\begin{aligned}|D_n| \times \Omega_{\mathcal{C}} &= \int_{D_n}\int_T \mathbb{X}^{\otimes 2}(\boldsymbol{s},t)\lambda(\boldsymbol{s},t)dtd\boldsymbol{s} + \int_{D_n}\int_T\int_{D_n}\int_T \mathbb{X}(\boldsymbol{s}_1,t_1)\mathbb{X}^{\mathrm{T}}(\boldsymbol{s}_2,t_2) \\ &\quad \times\{\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2) - \lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2)\}dt_2d\boldsymbol{s}_2dt_1d\boldsymbol{s}_1,\end{aligned}$$

where $\mathbb{X}(\boldsymbol{s},t)$ is defined in (11). Suppose that there exists a distance d^* so that

$$\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2) = \lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2) \text{ when } \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| > d^*. \tag{W.8}$$

Then, the covariance matrix above can be re-written as

$$\begin{aligned}|D_n| \times \Omega_{\mathcal{C}} &= \int_{D_n}\int_T \mathbb{X}^{\otimes 2}(\boldsymbol{s},t)\lambda(\boldsymbol{s},t)dtd\boldsymbol{s} + \int_{D_n}\int_T\int_{\|\boldsymbol{s}_2-\boldsymbol{s}_1\|\le d^*}\int_T \mathbb{X}(\boldsymbol{s}_1,t_1)\mathbb{X}^{\mathrm{T}}(\boldsymbol{s}_2,t_2) \\ &\quad \times\{\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2) - \lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2)\}dt_2d\boldsymbol{s}_2dt_1d\boldsymbol{s}_1,\end{aligned}$$

We can estimate $\mathcal{D}_n$ and $\Omega_{\mathcal{C}}$ consistently by

$$\begin{aligned}
\widehat{\mathcal{D}}_n &= \frac{1}{|D_n|}\int_T\int_{D_n} \mathbb{X}^{\otimes 2}(\boldsymbol{s},t)\exp\{\mathbb{X}^{\mathrm{T}}(\boldsymbol{s},t)\widehat{\boldsymbol{\theta}}\}d\boldsymbol{s}dt,\\
\widehat{\Omega}_{\mathcal{C}} &= \frac{1}{|D_n|}\Big[\int_T\int_T\int_{D_n}\int_{D_n} \mathbb{X}(\boldsymbol{s}_1,t_1)\mathbb{X}^{\mathrm{T}}(\boldsymbol{s}_2,t_2)I(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\le d^*)\\
&\qquad\qquad \times\{N_2(d\boldsymbol{s}_1,d\boldsymbol{s}_2,dt_1,dt_2)-\lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2)d\boldsymbol{s}_1d\boldsymbol{s}_2dt_1dt_2\}\\
&\qquad + \int_T\int_{D_n}\mathbb{X}^{\otimes 2}(\boldsymbol{s},t)\widehat{\lambda}(\boldsymbol{s},t)d\boldsymbol{s}dt\Big],
\end{aligned}$$

where $N_2(d\boldsymbol{s}_1,d\boldsymbol{s}_2,dt_1,dt_2) = I\{(\boldsymbol{s}_1,t_1)\neq(\boldsymbol{s}_2,t_2)\}N(d\boldsymbol{s}_1,dt_1)N(d\boldsymbol{s}_2,dt_2)$. Heinrich and Prokešová (2010) established consistency of $\widehat{\Omega}_{\mathcal{C}}$ when (W.8) is only approximately true by letting $d^*\to\infty$ as n increases in the stationary case. Their argument can be generalized to the current setting. In practice, one can check plot of the empirical pair correlation function to get an estimate of d^*.

We estimate Ω_θ by $\widehat{\Omega}_\theta = \widehat{\mathcal{D}}_n^{-1}\widehat{\Omega}_{\mathcal{C}}\widehat{\mathcal{D}}_n^{-1}$, and partition the covariance matrix into 2×2 blocks according to the partition of $\boldsymbol{\theta}$. Then

$$\begin{aligned}
\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}) &= |D_n|^{-1}\times\widehat{\Omega}_{\theta,11},\\
\widehat{\mathrm{var}}\{\widehat{\gamma}(t)\} &= |D_n|^{-1}\times\widetilde{\boldsymbol{B}}^{\mathrm{T}}(t)\widehat{\Omega}_{\theta,22}\widetilde{\boldsymbol{B}}(t),
\end{aligned}$$

where $\widehat{\Omega}_{\theta,11}$ and $\widehat{\Omega}_{\theta,22}$ are the blocks in $\widehat{\Omega}_\theta$ corresponding to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively.

W.5 Model and tuning parameter selection

To apply the proposed model, there are a number of important model selection issues that need to be addressed, such as choosing the numbers of knots in $\widehat{\gamma}(t)$ and $\widehat{R}_T(t_1,t_2)$, the tuning parameter δ for the composite likelihood in Section 3.2, as well as choosing the number of principal components, p.

W.5.1 Selecting K_1

We first address the tuning parameter selection problem in estimating $\gamma(t)$. In the spline smoothing literature, the number of knots are usually chosen by minimizing some information criterion. Some theoretical justification for the use of Bayesian information criterion (BIC) was provided by Huang and Yang (2004). However, we propose to choose K_1 by an Akaike

information criterion (AIC) criterion for two reasons. First, the BIC uses the logarithm of the sample size, which is hard to decide in our situation because the data are correlated. Second, since the estimator $\widehat{\gamma}$ is used for estimation of $R_T(\cdot,\cdot)$, undersmoothing is usually recommended in this situation to avoid introducing too much bias into $\widehat{R}_T$, and AIC in general tends to choose a larger number of knots than BIC. The proposed AIC is

$$\mathrm{AIC}_\gamma(K_1) = -2\ell\{\widehat{\beta}, \boldsymbol{B}^{\mathrm{T}}(t)\widehat{\boldsymbol{\nu}}\} + 2K_1, \tag{W.9}$$

where $\ell(\cdot)$ is the Poisson maximum likelihood function (9)

W.5.2 Selecting δ and K_2 for the composite likelihood

To choose δ, it is natural to consider cross-validation using the composite likelihood (14) as the loss function. However, as δ becomes smaller, we consider smaller neighborhoods, fewer number of pairs of events, and thus smaller composite likelihood values. To make the composite likelihood function comparable under different δ, we need to consider a standard version of it.

By similar arguments used in our proof of Theorem 3, we can show that there exist a limiting function $\mathcal{L}_c(\boldsymbol{\beta}, \gamma, R_T)$ so that

$$\mathcal{L}_c(\boldsymbol{\beta}, \gamma, R_T) = \lim_{n\to\infty}\lim_{\delta\to 0} \ell_c(\boldsymbol{\beta}, \gamma, R_T)/(|D_n||D_\delta|),$$

where $|D_\delta|$ is the area of a local neighborhood with radius δ. Here, $\mathcal{L}_c$ is the expected composite likelihood functions that we are trying to maximize. Therefore, we propose a m-fold cross-validation procedure using an empirical version of $\mathcal{L}_c$ as the loss function.

We first partition the spatial domain D into m non-overlapping regions with similar size and shape, denoted as $\{D_1, \ldots, D_m\}$. Let $\ell_{c,k}$ be the composite likelihood based on pairs of events within D_k, $\widetilde{\mathcal{L}}_{c,k} = \ell_{c,k}/(|D_k||D_\delta|)$, and let $\widehat{R}^{(-k)}_{T,\delta,K_2}$ be the spline covariance estimator with tuning parameters (δ, K_2) using data in $D\backslash D_k$ only. Define

$$CV(\delta, K_2) = \frac{1}{m}\sum_{k=1}^{m} \widetilde{\mathcal{L}}_{c,k}(\widehat{\boldsymbol{\beta}}, \widehat{\gamma}, \widehat{R}^{(-k)}_{T,\delta K_2}),$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$ are the mean estimators using the Poisson likelihood. We choose δ and K_2 as the minimizer of the cross-validation function, which can be obtained by a two dimensional grid search.

W.5.3 Choosing the numbers of principal components

In FPCA literature, the number of principal components are chosen by various information criteria. See Yao et al. (2005) and Hall et al. (2008) for comprehensive accounts of these methodologies. Under the model with p principal components, we can reconstruct the covariance function $\widehat{R}_T$ as

$$\widetilde{R}_{[p]}(t_1, t_2) = \sum_{j=1}^{p} \widehat{\omega}_j \widehat{\psi}_j(t_1) \widehat{\psi}_j(t_2), \tag{W.10}$$

where $\widehat{\omega}_j$ and $\widehat{\psi}_j$ are defined in (17) and (18). Under this reconstruction, $R_T(\cdot, \cdot)$ is modeled by $p(K_2 + 1)$ parameters, which include the p eigenvalues and K_2 spline coefficients for each eigenfunction. However, the eigenfunctions need to have unit norms and be orthogonal to each other. After adjusting for these constraints, the total number of parameters is $p(2K_2 - p + 1)/2$. We propose the second AIC criterion as

$$\mathrm{AIC}_R(p) = -2\ell_c(\widehat{\beta}, \widehat{\gamma}, \widetilde{R}_{[p]}) + p(2K_2 - p + 1), \tag{W.11}$$

where $\widetilde{R}_{[p]}$ is the adjusted covariance function defined in (W.10) and ℓ_c is the composite likelihood function in (14).

Note that $\widehat{R}_T(\cdot, \cdot)$ is not guaranteed to be positive definite. Let $\widetilde{p}_{\max}$ be the maximum order j such that $\widehat{\omega}_j > 0$. We choose p by minimizing AIC_R in the range $[0, \widetilde{p}_{\max}]$, where $p = 0$ means $R_T \equiv 0$ and hence $X(\boldsymbol{s}, t) = \mu(t) = \gamma(t)$ for any $\boldsymbol{s}$.

W.6 Additional plots

The results of spatial prediction in a typical simulation is presented in Figure W.1. The data were simulated as in Section 5 of the paper. We partition the spatial region into $25 \times 25 = 625$ regular cells. On average, there are about 3 events in each cell. The plots in the left column of Figure W.1 are the image of the true latent processes, where the color illustrates the value of the random fields. The plots in the right column of Figure W.1 are the predicted random fields. Because the empirical Bayes predictor has a shrinkage effect on the estimated values, the predicted random fields have less variation than the original ones. Nevertheless, most hot and cold spots in $\xi_1(\boldsymbol{s})$ and $\xi_2(\boldsymbol{s})$ are recovered in the predictions.

W.7 Model checking for the CTR data

To check the model fitting for the CTR pancreatic cancer data, we first run a residual analysis. We partition the whole spatial region W into small non-overlapping blocks D_k, $k = 1, \ldots, M$, so that $W = \cup_{k=1}^{M} D_k$. Natural choices for this partition are the census tracts or the census block groups. We now define the residuals as

$$R_k = \sum_{(\boldsymbol{s},t)\in N\cap D_K\otimes T} 1 - \int_{D_k}\int_T \lambda(\boldsymbol{s},t)d\boldsymbol{s}dt,$$

where $\lambda(\boldsymbol{s},t)$ is the intensity function from the proposed model (5). When the blocks are small, the variance of the residuals can be approximated by

$$\text{var}(R_k) \approx \int_{D_k}\int_T \lambda(\boldsymbol{s},t)d\boldsymbol{s}dt + \int_{D_k}\int_{D_k}\int_T\int_T \{\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2) - \lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2)\}dt_1dt_2d\boldsymbol{s}_1d\boldsymbol{s}_2.$$

We estimate the standard deviation of the residuals using the formula above with estimated intensity functions λ and λ_2. In panel (a) of Figure W.2, we plot the standardized residuals verses SES score at the block group level. As we can see, the residuals scatter about the 0 horizontal line without any clear pattern. Among the 2616 block groups, only about 2% of the standardized residuals have absolute values greater than 2. This plot indicates that the proposed model for the first order intensity (5) fits the data quite well.

We also check the model fitting using the K function. The estimated K function is defined as

$$\widehat{K}(r) = \sum_{(\boldsymbol{s}_1,t_1)\in N\cap W}\ \sum_{(\boldsymbol{s}_2,t_2)\in N\cap W} \frac{I(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\le r)}{\lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2)}, \tag{W.12}$$

while the theoretical K function under the model (6) is defined as

$$K(r) = \int_D\int_D\int_T\int_T \frac{\lambda_2(\boldsymbol{s}_1,\boldsymbol{s}_2,t_1,t_2)I(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\le r)}{\lambda(\boldsymbol{s}_1,t_1)\lambda(\boldsymbol{s}_2,t_2)}dt_1dt_2d\boldsymbol{s}_1d\boldsymbol{s}_2. \tag{W.13}$$

In panel (b) of Figure W.2, the solid curve is $K(r)$ under the proposed model with estimated parameters, and the dash curve is $\widehat{K}(r)$ which does not rely on the assumptions on the latent process $X(\boldsymbol{s},t)$. The two curves being almost identical indicates that out model on the second order intensity also fits the data well.

The Associate Editor raised a good point that if the data are non-stationary, the eigenfunctions might depend on the spatial location. To check our second-order intensity

reweighted stationary assumption, we divide the state of Connecticut into the southern and northern halves, and compare the estimates obtained from different halves of the state. The reason for this partition is that the southern half of Connecticut is along the coast while the northern half is not. If there is any spatial non-stationarity in the data, the southern half is more likely to be different from the northern half. Our AIC criterion picks two principal components for both halves of the state, indicating that there is no evidence of non-stationarity in terms of the number of principal components. In panels (c) and (d) of Figure W.2, we compare the estimated eigenfunctions using the whole state (the solid curves), southern half of the state (dash curves) and northern half (dotted curves). As we can see, the three estimates for $\psi_1(t)$ are almost identical. The three estimates for $\psi_2(t)$ are slightly different, but the overall increasing trend in these functions is consistent. Since the variance of the second principal component ω_2 is much smaller than that of the first component, the signals in $\xi_2(\cdot)$ is much weaker than $\xi_1(\cdot)$, it is expected that the estimate for $\psi_2(t)$ is more variable. These plots, together with the K function plot, indicate that there is no clear violation of the stationary assumptions and our model fits the data quite well.

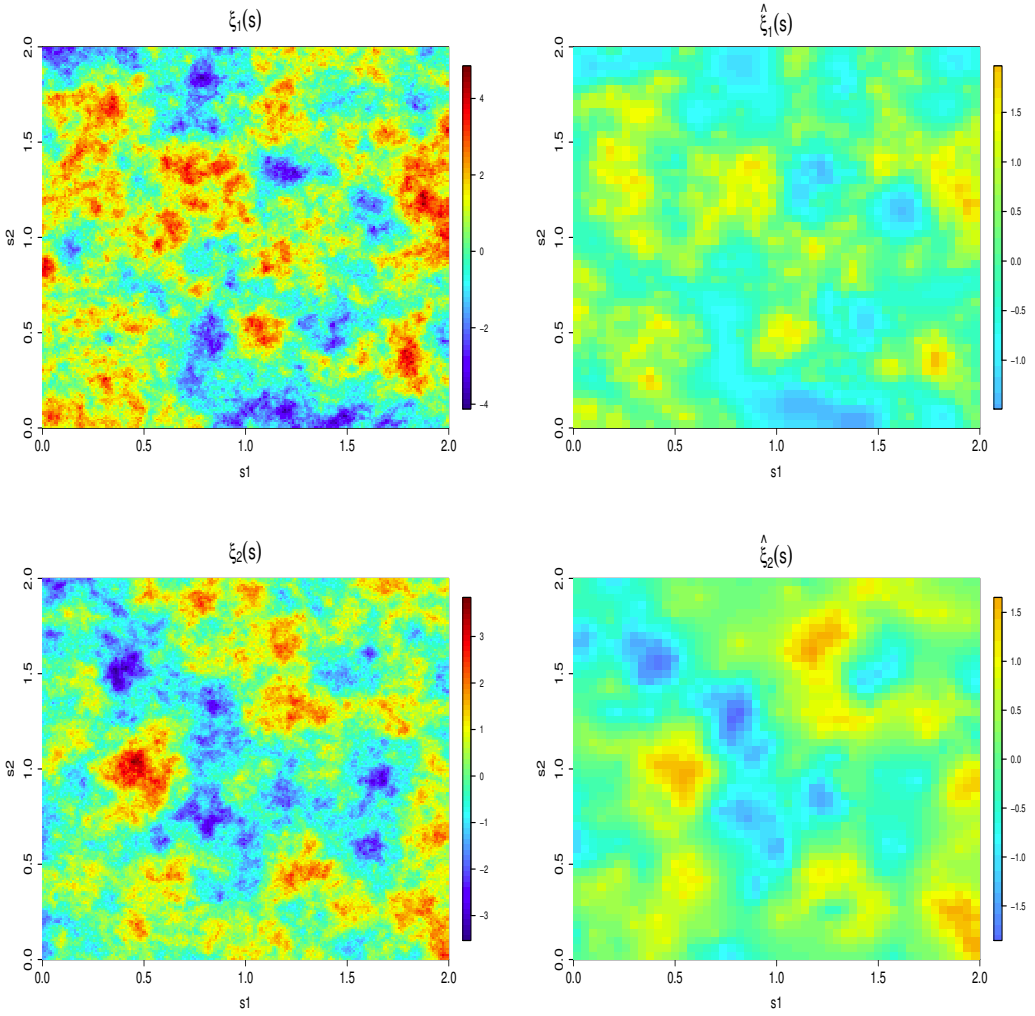
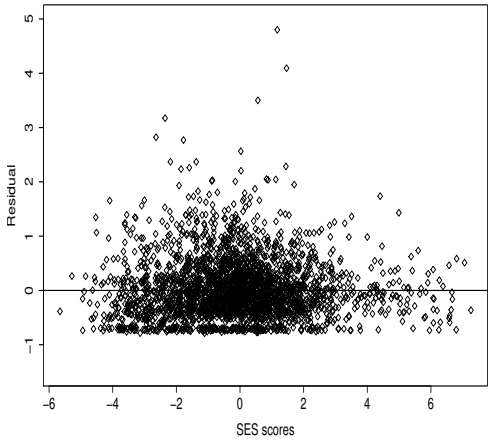
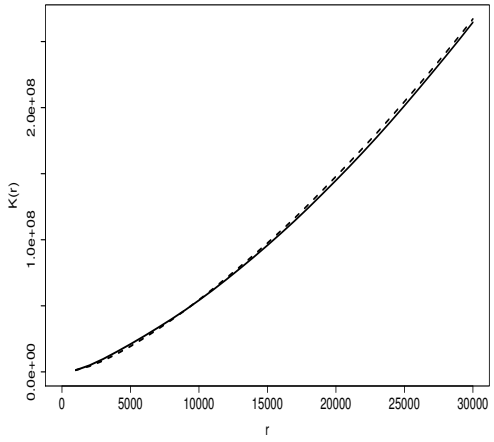


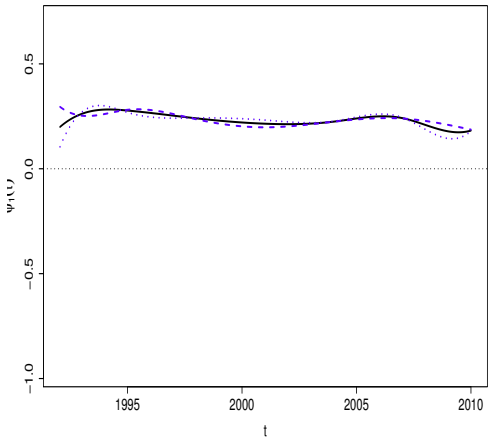
Figure W.1: Spatial prediction in a typical run of the simulation study. The plots in the first column are heat maps of the latent random fields $\xi_1(\boldsymbol{s})$ and $\xi_2(\boldsymbol{s})$ respectively. The plots in the second column are the predicted random fields $\widehat{\xi_1}(\boldsymbol{s})$ and $\widehat{\xi_2}(\boldsymbol{s})$ respectively.



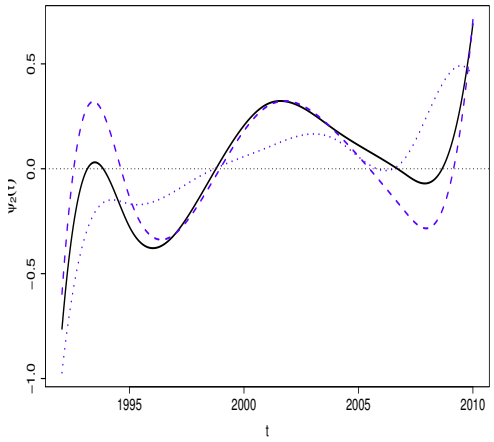
(a) Residual plot.



(b) Estimated and theoretical K functions.



(c) $\widehat{\psi}_1(t)$ from different spatial regions.



(d) $\widehat{\psi}_2(t)$ from different spatial regions.

Figure W.2: Model diagnostic for the CTR pancreatic cancer data. Panel (a) is the scatterplot of the standardized residual verses SES score at the blog group level. In panel (b), the solid curve is the theoretical K function under the proposed model with estimated parameters (functions), and the dash curve is the estimated K function without a model assumption. In panels (c) and (d), the solid curve is the estimated eigenfunction using the whole spatial region, the dash and dotted curves are the estimates using the data from southern and northern part of the state respectively.