Chapter 4

Modeling and Analysis of Spatially Correlated Data [*]

Yi Li [†]

Abstract

The last decade has seen a resurgence of statistical methods for the analysis of spatial data arising from various scientific fields. This chapter reviews these methodologies, mainly within the geostatistical framework. We consider data measured at a finite set of locations and draw inference about the underlying spatial process, based on the partial realization over this subset of locations. Methodologically, we employ linear mixed models and generalized linear mixed models that enable likelihood inference for fully observable spatial data. For spatial data subject to censoring, we review a class of semiparametric normal transformation models for spatial survival data. A key feature of this model is that it provides a rich class of models, where regression coefficients have a population-level interpretation and the spatial dependence of survival times is conveniently modeled using flexible normal random fields. This would be appealing to practitioners, especially given that there are virtually no spatial failure time distributions that are convenient to work with.

Keywords: Geostatistical data; areal data; semivariogram; kriging; spatial (generalized) linear mixed models; EM algorithm; PQL approximation; spatial survival data; semiparametric normal transformation; conditional Martingale covariance rate function.

1 Introduction

Spatial data are commonly acquired to serve diverse scientific purposes. For example, meterologists are interested in the amount of precipitation over adjacent tropical forests. Mining engineers are keen to predict the conserve of a new oil field based on the product of nearing fields. Political geologists link election results to spatial regions so as to map political power on geographical space. Economists study economic activities on macro spatial scales, as factors like access to the sea and to the raw materials tend to impact economic activities at a country level. Epidemiologists investigate the incidence of a disease and its variations across regions for a better understanding of the etiology. Environmentalists monitor the

[*] The author is partly supported by an NIH R01 grant.

[†] Harvard University, Boston, MA, E-mail: yili@jimmy.harvard.edu

level of PM2.5 (a particulate matter that can travel into a person's lungs) in air-pollution monitoring sites and characterize its spatial distributions. The shared features of data arising from these studies fall into the following two categories: (a) geostatistical (or point-referenced) type, where the outcomes are random variables indexed by locations that vary continuously over a subset of a Euclidean space. (b) areal type, where the locations are finite number of areal units with well-defined boundaries and the data are typically summary statistics over these units. Some spatial data sets even feature both geostatistical and areal types. For example, in the NCI Surveillance Epidemiology and End Results (SEER) data, the outcome for each individual was measured, while the location information was only available at the county (areal unit) level. As individual-level or point-referenced data have become increasingly common in public health studies with the use of the Geographic Information System (GIS) technology and geocoding of individua addresses, this chapter focuses on the statistical analysis of geostatistical data, while necessary modifications for analysis of the areal data will be discussed.

Spatial data analysis is challenged by the presence of spatial dependence among observations. We give a simple example to illustrate the effect of correlation on analysis. First consider independent samples $Y_1, \cdots, Y_n$ from a normal distribution with mean μ and known variance σ^2. The most efficient unbiased estimator of μ is the sample average $\bar{Y} = \sum_i Y_i/n$, which follows a normal distribution with mean μ and variance σ^2/n, yielding a two-sided 95% confidence interval for μ

$$(\bar{Y} - 1.96\sigma/\sqrt{n}, \bar{Y} + 1.96\sigma/\sqrt{n}).$$

Now instead of independent data, suppose that the data exhibit a spatial correlation in R^1 that decreases exponentially as the separation between data points increases

$$\operatorname{cov}(Y_i, Y_j) = \sigma\rho^{|i-j|}. \tag{1.1}$$

Under (1.1), $\bar{Y}$ will still follow a normal distribution with mean μ, but with variance

$$\operatorname{var}(\bar{Y}) = \sigma^2/n[1 + 2\{\rho/(1-\rho)\}(1-1/n) - 2\{\rho/(1-\rho)\}^2(1-\rho^{n-1})/n]. \tag{1.2}$$

For $n = 10$ and $\rho = 0.26$, $\operatorname{var}(\bar{Y}) = \sigma^2/10 \times 1.608$, resulting in a two-sided 95% confidence interval for μ

$$(\bar{Y} - 2.485\sigma/\sqrt{10}, \bar{Y} + 2.485\sigma/\sqrt{10});$$

see Cressie (1993). Thus, failure to account for the underlying correlations among data tends to narrow the confidence intervals. More intuitive explanation of the impact of spatial correlation can be obtained from (1.2), based on which the *effective sample size* can be computed as

$$n^* = n[1 + 2\{\rho/(1-\rho)\}(1-1/n) - 2\{\rho/(1-\rho)\}^2(1-\rho^{n-1})/n]^{-1}.$$

For large n, it follows that $n^*/n = (1-\rho)/(1+\rho)$, hinting that the effect of correlation is palpable even for large samples.

Hence, it is important to take into account the spatial correlation for correct inference. Mixed effects models have provided a convenient means of modeling spatial correlations by using random effects, with common spatial correlation structures including, for example, random fields (for geostatistical data) and autoregressive (CAR) structure (for areal data) (Yasui and Lele, 1997; Waller, et al., 1997). Over the past two decades, spatial statistical methods have been well established for normally distributed data (Cressie, 1993; Haining, et al., 1989) and discrete data (Journel, 1983; Cressie, 1993; Carlin and Louis, 1996; Diggle et al., 1998). Statistical models for such data are often fully parameterized, and inference procedures are based on maximum likelihood (Clayton and Kaldor, 1987; Cressie, 1993), penalized maximum likelihood (Breslow and Clayton, 1993) and Markov chain Monte Carlo (Besag, York, Mollie, 1991; Waller et al., 1997).

Further complicate the analysis of spatial data is the presence of censoring for outcomes as reflected in many epidemiological and social behavioral studies. For example, in the East Boston Asthma Study on childhood asthma, subjects were enrolled at community health clinics in the east Boston area, and questionnaire data, documenting ages at onset of childhood asthma and other environmental factors, were collected during regularly scheduled visits. Apart from the basic demographic data, residential addresses were geocoded for each study subject so that the latitudes and longitudes were available. Residents of East Boston are mainly relatively low income working families. Children residing in this area have similar social economical backgrounds and are often exposed to similar physical and social environments. These environmental factors are important triggers of asthma but are often difficult to measure in practice. Ages at onset of asthma of the children in this study were hence likely to be subject to spatial correlation. The statistical challenge is to identify significant risk factors associated with age at onset of childhood asthma while taking the possible spatial correlation into account. Li and Ryan (2002) proposed semi-parametric frailty models for spatially correlated survival data, where the spatial units are prespecified geographic regions (e.g. census tract). Their approach is to extend the ordinary frailty models to accommodate spatial correlations and exploit a robust rank likelihood-based inferential procedure. However, frailty survival models typically do not possess regression coefficients with population-level interpretations, less appealing to population scientists.

In this chapter, we review the methodologies for the analysis of spatial data, mainly within the geostatistical framework. That is, the data consist of the measurements at a finite set of locations and the statistical problem is to draw inference about the spatial process, based on the partial realization over this subset of locations. The rest of this chapter is organized as follows. We introduce the building blocks of geostatistical modeling, including stationarity, isotropy and (semi-)variograms. We next discuss statistical models, including linear mixed models and generalized linear models, that enable likelihood inference for fully observable spatial data. For spatial data that are subject to censoring, we review a semiparametric normal transformation model that was recently developed by Li and Lin (2006). A key feature of this model is that it provides a rich class of models where regression coefficients have a population-level interpretation and the spatial dependence of survival times is conveniently modeled using flexible normal random fields. We conclude this chapter with further topics and some open questions.

2 Basic concepts of spatial process

We present the essential elements of geostatistical spatial models, starting with the fundamental underlying concept of a stochastic spatial process $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$. For example, $Y(\mathbf{s})$ represents the level of PM2.5 at monitor site $\mathbf{s}$, and $\mathcal{D}$ is a fixed subset of Euclidean space R^r, containing all the pollution monitoring sites. In the spatial context, The dimension of the Euclidean space, r, is often 2 (latitude and longitude) or 3 (latitude, longitude and altitude above sea level). Assuming the existence of the first two moments of $Y(\mathbf{s})$ for every $\mathbf{s} \in \mathcal{D}$, the first moment $E\{Y(\mathbf{s})\} = \mu(\mathbf{s})$ is often termed *trend* or *drift*, while the existence of the second moment allows the definition of (weak) stationarity.

To be more specific, a spatial process is *weakly stationary* if $\mu(\mathbf{s}) \equiv \mu$ (i.e. the process has a constant mean) and

$$\operatorname{cov}\{Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})\} = C(\mathbf{h})$$

for all $\mathbf{h} \in R^r$ such that $\mathbf{s}, \mathbf{s}+\mathbf{h} \in \mathcal{D}$, where $C(\mathbf{h})$ is termed the covariance function. In contrast, a spatial process is termed *strongly stationary* if, for any given $n \geqslant 1$ and any given $\mathbf{s}_1, \cdots, \mathbf{s}_n$ and any $\mathbf{h} \in R^r$ (as long as $\mathbf{s}_i + \mathbf{h} \in \mathcal{D}$), $(Y(\mathbf{s}_1), \cdots, Y(\mathbf{s}_n))$ has the same joint distribution as $(Y(\mathbf{s}_1+\mathbf{h}), \cdots, Y(\mathbf{s}_n+\mathbf{h}))$. Of course, strong stationarity implies weak stationarity, but not vice versa.

Apart from these two stationary types, there indeed exists a third type of stationarity called *intrinsic stationarity.* Here, we assume $E(Y(\mathbf{s})) \equiv \mu$ and define intrinsic stationarity if $E(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s}))^2$ depends only on $\mathbf{h}$. If that is the case, we write $2\gamma(\mathbf{h}) = E(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s}))^2$ and call $2\gamma(\mathbf{h})$ the variogram and $\gamma(\mathbf{h})$ the semivariogram. In the following, we use $|\cdot|$ to denote the Euclidean norm for a vector.

The behavior of $\gamma(\mathbf{h})$ near $|\mathbf{h}| = 0$ is informative about the continuity properties of $Y(\cdot)$. Specifically, (i) if $\gamma(\mathbf{h}) \to 0$ as $|\mathbf{h}| \to 0$, then $Y(\cdot)$ is L_2 continuous, namely, $|Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})|_{L_2} \to 0$ for any $\mathbf{s}$ as $|\mathbf{h}| \to 0$. For example, a Brownian motion in R^1 is L_2 continuous; (ii) if $\gamma(\mathbf{h})$ does not approach 0 as $|\mathbf{h}| \to 0$, then $Y(\cdot)$ is not L_2 continuous and termed irregular. The discontinuity of $\gamma(\mathbf{h})$ at 0 is called the nugget effect reflecting microscale variation, which will be discussed later; (iii) if $\gamma(\mathbf{h})$ is a positive constant, then $Y(\mathbf{s}_1)$ and $Y(\mathbf{s}_2)$ are uncorrelated for any $\mathbf{s}_1 \neq \mathbf{s}_2$, regardless of their proximity.

If the semivariogram $\gamma(\mathbf{h})$ depends on vector $\mathbf{h}$ only though its length, we call the underlying process *isotropic*, reflecting that the pairwise correlations among subjects depend only on their distances; otherwise, it is called *anisotropic.* Isotrop ic processes are popular in spatial data analysis for their interpretability and availability of parametric functions for the semivariogram, which can be simply written as $\gamma(|\mathbf{h}|)$. The Matèrn model has recently emerged as a powerful model for $\gamma(|\mathbf{h}|)$

in practice, which is given by

$$\gamma(|\mathbf{h}|) = \begin{cases} \tau^2 + \sigma^2 - m(\sigma^2, \zeta, \nu, |\mathbf{h}|) \; , if \; |\mathbf{h}| > 0, \\ 0, \qquad\qquad otherwise, \end{cases} \tag{2.1}$$

where

$$m(\sigma^2, \zeta, \nu, d) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(2\zeta\sqrt{\nu}d)^{\nu}\mathcal{K}_{\nu}(2\zeta\sqrt{\nu}d) \tag{2.2}$$

is the Matèrn function. Here, ζ measures the correlation decay with the distance and ν is a smoothness parameter, $\Gamma(\cdot)$ is the conventional Gamma function, $\mathcal{K}_{\nu}(\cdot)$ is the modified Bessel function of the second kind of order ν (see, e.g. Abramowitz and Stegun, 1965). While $\gamma(0) = 0$ by definition, $\lim_{|\mathbf{h}|\to 0+} = \tau^2$, termed *nugget*, which characterizes local variations; in addition, $\lim_{|\mathbf{h}|\to\infty} = \tau^2 + \sigma^2$ is called *sill*; finally, sill minus nugget is termed partial sill, which is σ^2 in this case. Model (2.2) is rather general, special cases including the exponential function $\sigma^2 \exp(-d)$ when the smoothness parameter $\nu = 0.5$ and the "decay parameter" $\zeta = 1$, and the "Gaussian" correlation function $\sigma^2 \exp(-d^2)$ corresponding to $\nu \to \infty$ and $\zeta = 1$. In cases like these, ζ characterizes the effective range, the distance at which there is practically no lingering spatial correlation. Other common choices of semivariogram functions can be found in Cressie (1993) and Banerjee et al. (2003).

Given a variety of choices of semivariogram function, a natural question, of course, is to decide which best fits a given data or whether the data can distinguish them. It is customary to empirically estimate the semivariogram, with the goal of comparing it to the theoretical shapes. Under the constant mean assumption, a straight-forward estimator, due to Matheron (1962), would be

$$2\hat{\gamma}(|\mathbf{h}|) = \frac{1}{|N(\mathbf{h})|}\sum_{N(\mathbf{h})}(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2, \tag{2.3}$$

where $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : |\mathbf{s}_i - \mathbf{s}_j| = |\mathbf{h}|\}$ denotes the collection of pairs that are distanced by $|\mathbf{h}|$ and $|N(\mathbf{h})|$ is the number of distinct pairs in $N(\mathbf{h})$. In practice, however, Matheron's estimator is of limited value unless the observations fall on a regular grid. Instead, we would partition the half-line into distance bins $I_1 = (0, h_1), I_2 = [h_2, h_3)$ and up to $I_K = [h_{K-1}, h_K)$ for some prespecified $0 < h_1 < \cdots < h_K$. Then we can alter the definition of $N(h)$ by

$$N(h_k) = \{(\mathbf{s}_i, \mathbf{s}_j) : |\mathbf{s}_i - \mathbf{s}_j| \in I_k\}, k = 1, \cdots, K.$$

For the choice of h_k and K, see Journel and Huijbregts (1979). Moreover, as (2.3) is sensitive to outliers, a more robust fourth-root-of-squared-difference estimator can be constructed; see Cressie and Hawkins (1980).

We close this section by noting the relationship between the semivariogram $\gamma(\mathbf{h})$ and the covariance function $C(\mathbf{h})$. Apparently,

$$\gamma(\mathbf{h}) = C(\boldsymbol{o}) - C(\mathbf{h}). \tag{2.4}$$

Hence, given C, we can recover γ. But how about recovering C from γ? It turns out that we need to add more condition, i.e. $C(\mathbf{h}) \to 0$ when $|\mathbf{h}| \to 0$. This condition is sensible as it regulates that the covariance between two observations diminishes as they become farther apart. Taking the limit of both sides of (2.4), we have $C(\boldsymbol{o}) = \lim_{|\mathbf{h}|\to\infty} \gamma(\mathbf{h})$. Thus, we have that

$$C(\mathbf{h}) = \lim_{|\mathbf{h}|\to\infty} \gamma(\mathbf{h}) - \gamma(\mathbf{h}). \tag{2.5}$$

In general, such limit may not exist; but if it does, the process is weakly stationary with covariance function $C(\mathbf{h})$.

2.1 Spatial regression models for normal data

In many spatial studies, also observed along with the outcomes are exploratory variables, measuring for example the characteristics of each individual at given locations. Within the framework of spatial process, we denote such covariate process by $(\mathbf{X}(\mathbf{s}), \mathbf{s} \in \mathcal{D})$. The spatial linear mixed model of $Y(\mathbf{s})$ given $\mathbf{X}(\mathbf{s})$ can be written as

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + Z(\mathbf{s}), \tag{2.6}$$

where $\mathbf{X}(s)$ is a $q \times 1$ covariate vector, $\boldsymbol{\beta}$ is a vector of regression coefficients in some open subset of R^q, say, $\mathcal{B}$, and $Z(\mathbf{s})$ is a mean zero (weakly stationary) Gaussian process with spatial covariance function

$$\text{cov}\{Z(\mathbf{s}), Z(\mathbf{s}')\} = C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) \tag{2.7}$$

where $\boldsymbol{\theta}$ is a $k \times 1$ vector of spatial dependence parameters in some open subset of R^k, say, $\mathcal{A}$. The parameter $\boldsymbol{\beta}$ characterizes the deterministic part of the spatial data and is sometimes called the *trend* parameter, while $\boldsymbol{\theta}$ characterizes the variability of the underlying spatial field through the spatial covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$. A common practice is to further assume $Z(\mathbf{s})$ is isotropic so that the covariance function $C(\cdot)$ depends on $\mathbf{s}$ and $\mathbf{s}'$ only through their distance $|\mathbf{s} - \mathbf{s}'|$. In this case, we write $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$ as $C(|\mathbf{s} - \mathbf{s}'|; \boldsymbol{\theta})$, which can be easily specified from the semivariogram γ, e.g. the Matern model (2.1), via relationship (2.5).

We are in a position to draw inference based on model (2.6). Given a finite number of locations $\mathbf{s}_1, \cdots, \mathbf{s}_n$, then the $n \times n$ matrix $\boldsymbol{\Sigma}_n = [C(|\mathbf{s}_i - \mathbf{s}_j|; \boldsymbol{\theta})]$ is positive-definite. Let $Y_n = (Y(\mathbf{s}_1), \cdots, Y(\mathbf{s}_n))$ denote the outcome data and $\mathbf{X}_n = (\mathbf{X}(\mathbf{s}_1), \cdots, \mathbf{X}(\mathbf{s}_n))'$ denote the design matrix. Let $\boldsymbol{\Theta} = (\boldsymbol{\beta}', \boldsymbol{\theta}')'$ denote the $(q+k) \times 1$ parameter vector. Then the log likelihood is

$$\begin{aligned} L_n(\boldsymbol{\Theta}) = {} & -n/2\log(2\pi) - 1/2\log(|\boldsymbol{\Sigma}_n|) \\ & -1/2(\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta})'\boldsymbol{\Sigma}_n^{-1}(\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta}) \end{aligned} \tag{2.8}$$

The maximum likelihood estimator $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\theta}}')'$ satisfies

$$L_n(\hat{\boldsymbol{\Theta}}) = \sup\{L_n(\boldsymbol{\Theta}) : \boldsymbol{\Theta} \in \mathcal{B} \times \mathcal{A}\}.$$

Computational details are given in Cressie (1993). However, the MLE of $\boldsymbol{\theta}$ can be seriously biased, especially when the sample size is small. A common remedy is the *restricted maximum likelihood estimation* (REML) approach that filters the data so that the joint distribution of the filtered data is free of $\boldsymbol{\beta}$. To proceed with the REML, we need to introduce a new concept, *error contrast.* Specifically, we call a linear combination of outcome, say, $\mathbf{a}'\mathbf{Y}_n$, where $\mathbf{a} \in R^n$, an error contrast if $E(\mathbf{a}'\mathbf{Y}_n) = 0$ for all $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$; hence, $\mathbf{a}'\mathbf{Y}_n$ is an error contrast if and only if $\mathbf{a}'\mathbf{X}_n = \boldsymbol{o}_q'$, where $\boldsymbol{o}_q$ is a $q \times 1$ zero vector.

We assume that the design matrix $\mathbf{X}_n$ is full rank, i.e. $\text{rank}(\mathbf{X}_n) = q$. Hence, the kernel space of $\mathbf{X}_n$ has dimension $n - q$. That is, we can find an $n \times (n - q)$ full rank matrix $\mathbf{A}$ (i.e. $\text{rank}(\mathbf{A}) = n - q$)such that $\mathbf{A}'\mathbf{X}_n = 0$, yielding a vector of $n - q$ linearly independent error contrast $\mathbf{A}'\mathbf{Y}_n \overset{\text{def}}{=} \mathbf{W}_n$. Under the normality assumption, $\mathbf{W}_n \sim MVN(\boldsymbol{o}_n, \mathbf{A}'\boldsymbol{\Sigma}_n(\boldsymbol{\theta})\mathbf{A})$, which does not depend on $\boldsymbol{\beta}$. Here, $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

The choice of $\mathbf{A}$ is not unique; however, Harville (1974) showed that the log likelihood function would differ each other only by an additive constant for various $\mathbf{A}$'s that form the $n - q$ linearly independent contrasts. Indeed, for the $\mathbf{A}$ that satisfies $\mathbf{A}\mathbf{A}' = \mathbf{I} - \mathbf{X}_n(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'$ and $\mathbf{A}\mathbf{A}' = \mathbf{I}$, the log likelihood function for $\mathbf{W}_n = \mathbf{A}'\mathbf{Y}_n$ is

$$\begin{aligned} L_R(\boldsymbol{\theta}) = &-(n-q)/2\log(2\pi) - 1/2\log|\mathbf{X}_n'\mathbf{X}_n| + 1/2\log|\boldsymbol{\Sigma}_n(\boldsymbol{\theta})| \\ &+1/2\log|\mathbf{X}_n'\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\theta})\mathbf{X}_n| + 1/2\mathbf{Y}_n'\Pi(\boldsymbol{\theta})\mathbf{Y}_n \end{aligned} \tag{2.9}$$

where $\Pi(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\theta})\mathbf{X}_n(\mathbf{X}_n'\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\theta})\mathbf{X}_n)^{-1}\mathbf{X}_n'\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\theta})$. A REML estimator $\hat{\theta}_{REML}$ would satisfy

$$L_R(\hat{\theta}_{\text{REML}}) = \sup\{L_R(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{A}\}.$$

A Basyesian justification has been provided by Harville (1974), which showed the marginal posterior density for $\boldsymbol{\theta}$ is proportional to (2.9) with a noninformative prior on $\boldsymbol{\beta}$.

Estimation is also feasible based on (2.6) without imposing normality assumption on the spatial component $Z(\mathbf{s})$. Instead, we assume the existence of the first two components of $Z(\mathbf{s})$ such that $E(Z(\mathbf{s})) = 0$ and $Z(\mathbf{s})$ has spatial covariance function $C(\cdot; \boldsymbol{\theta})$ as defined in (2.7). Assuming for now that $\boldsymbol{\theta}$ is known, we can obtain the best linear unbiased estimator of $\boldsymbol{\beta}$ by minimizing the quadratic "loss" function"

$$(\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta})'\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\theta})(\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta})$$

where $\boldsymbol{\Sigma}_n(\boldsymbol{\theta}) = [C_{ij}(\boldsymbol{\theta})]$ and $C_{ij}(\boldsymbol{\theta}) = C(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta}), i, j = 1, \cdots, n$, leading to

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}_n'\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\theta})\mathbf{X}_n]^{-1}\mathbf{X}_n'\boldsymbol{\Sigma}_n(\boldsymbol{\theta})\mathbf{Y}_n. \tag{2.10}$$

In reality, $\boldsymbol{\theta}$ is most likely unknown, necessitating an iterative reweighted least squares estimation:

$$\hat{\boldsymbol{\beta}}^{(k+1)} = [\mathbf{X}_n'\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\theta}^{(k)})\mathbf{X}_n]^{-1}\mathbf{X}_n'\boldsymbol{\Sigma}_n(\boldsymbol{\theta}^{(k)})\mathbf{Y}_n,$$

where $\boldsymbol{\theta}^{(k)}$ is a variogram-based estimate (see Cressie, 1993) based on the current estimate of $\hat{\boldsymbol{\beta}}^{(k)}$. Write $\tilde{\boldsymbol{\beta}} = \lim_{k\to\infty} \hat{\boldsymbol{\beta}}^{(k)}$ and $\tilde{\boldsymbol{\theta}} = \lim_{k\to\infty} \theta^{(k)}$. It can be shown that this procedure yields an asymptotically efficient and consistent estimate for $\boldsymbol{\beta}$, whose variance matrix is subsequently

$$\mathrm{cov}(\tilde{\boldsymbol{\beta}}) = [\mathbf{X}_n' \boldsymbol{\Sigma}_n^{-1}(\tilde{\boldsymbol{\theta}})\mathbf{X}_n]^{-1}.$$

Further details of statistical properties can be found in del Pino (1989).

Though our models, along with the inferential procedures, are framed within the geostatistical context, they can easily accommodate areal data with a proper $\boldsymbol{\Sigma}_n$. A popular choice of $\boldsymbol{\Sigma}_n$ for the areal data is the conditional auto-regressive (CAR) structure, possessing both appealing theoretical properties and attractive interpretation (Cressie, 1993). The CAR structure assumes that the full conditional distribution of the $Z(\mathbf{s}_i)$ [or $Y(\mathbf{s}_i)$] on the rest of data depends only on its "neighbors". In particular, for the normal data, the CAR model states that

$$Z(\mathbf{s}_i)|Z(\mathbf{s}_j), j \neq i \sim N\left(\theta \sum q_{ij} Z(\mathbf{s}_j)/q_{i+}, \tau^2/q_{i+}\right),$$

where $q_{ii} = 0$, $q_{i+} = \sum_j q_{ij}$ and the nonnegative q_{ij} controls the strength of connection between areas i and j, and often takes value 0 when areas i,j are not neighbors, $-1 \leqslant \theta \leqslant 1$ is the spatial dependence parameter controlling the amount of information in an area provided by its neighbors, and τ^2 is the nugget, measuring local variations. When areas i and j are neighbors, a common choice of q_{ij} is $q_{ij} = 1$, reflecting equal weights from neighboring areas. Brook's lemma (Brook, 1964) suggests a multivariate normal distribution for $Z(\mathbf{s}_1), \cdots, Z(\mathbf{s}_n)$ with mean 0 and variance matrix (Yasui and Lele, 1997)

$$\boldsymbol{\Sigma}_n = \tau^2(\mathbf{M}^{-1} - \theta\mathbf{Q})^{-1}, \tag{2.11}$$

where $\mathbf{Q} = \{q_{ij}\}$ is an $n \times n$ symmetric matrix; $\mathbf{M}$ is an $n \times n$ diagonal matrix with diagonal elements $1/q_{i+}$. It is worth noting that the flexibility of the CAR structure allows for a more general neighborhood concept than a mere geographical proximity (Cressie, 1993).

2.2 Spatial prediction (Kriging)

Having estimated the trend parameter $\boldsymbol{\beta}$ and the spatial correlation parameter $\boldsymbol{\theta}$, we are ready to discuss geostatistical techniques to predict the value of a random field at a given location from nearby observations. Such techniques are termed *kriging*, developed by a French mathematician Georges Matheron and named after Daniel Gerhardus Krige, a mining engineer who developed a distance-weighted average method for determining gold grades based on nearby samples.

Specifically, kriging is about interpolating the value $Y(\mathbf{s}_0)$ of a random field $Y(\mathbf{s})$ at a prespecified location $\mathbf{s}_0$ from observations $Y_i = Y(\mathbf{s}_i)$, $i = 1, \cdots, n$. In essence, it is a minimum-mean-squared-error method of prediction that depends on the second-order properties of $Y(\cdot)$, via computing the best linear unbiased estimator $\hat{Y}(\mathbf{s}_0)$ for $Y(\mathbf{s}_0)$, given by

$$\hat{Y}_K(\mathbf{s}_0) = c_0 + \sum_{i=1}^{n} w_i Y(\mathbf{s}_i) = c_0 + \mathbf{w}'\mathbf{Y}_n.$$

The weights $\mathbf{w} \overset{\text{def}}{=} (w_1, \cdots, w_n)'$ are chosen in such a way that the prediction error variance (also called kriging variance or kriging error)

$$\begin{aligned}\sigma_p^2(x_0) &= \text{var}\left(\hat{Y}_k(\mathbf{s}_0) - Y(\mathbf{s}_0)\right)\\ &= \text{var}(Y(\mathbf{s}_0)) - 2\mathbf{w}'\mathbf{k}(\mathbf{s}_0) + \mathbf{w}'\mathbf{\Sigma}_n\mathbf{w},\end{aligned} \tag{2.12}$$

where $\mathbf{k}(s_0) = (C(\mathbf{s}_0 - \mathbf{s}_1; \boldsymbol{\theta}), \cdots, C(\mathbf{s}_0 - \mathbf{s}_n; \boldsymbol{\theta}))'$ and $\mathbf{\Sigma}_n = [C(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta})]_{n\times n}$, is minimized subject to the unbiasedness condition:

$$E[\hat{Y}_K(\mathbf{s}_0) - Y(\mathbf{s}_0)] = 0. \tag{2.13}$$

This optimization problem can be solved by using the Lagrange multipliers, yielding a closed-form kriging estimator

$$\hat{Y}_K(\mathbf{s}_0) = \hat{\boldsymbol{\beta}}'\mathbf{X}(\mathbf{s}_0) + \mathbf{k}(\mathbf{s}_0)'\mathbf{\Sigma}_n^{-1}(\mathbf{Y}_n - \mathbf{X}_n\hat{\boldsymbol{\beta}}), \tag{2.14}$$

where $\hat{\boldsymbol{\beta}}$ is obtained from (2.10), with prediction error variance given by

$$\begin{aligned}\sigma_p^2(\mathbf{s}_0) = {} & [\text{var}(Y(\mathbf{s}_0)) - \mathbf{k}(\mathbf{s}_0)'\mathbf{\Sigma}_n^{-1}\mathbf{k}(\mathbf{s}_0)]\\ & + (\mathbf{X}(\mathbf{s}_0) - \mathbf{X}_n'\mathbf{\Sigma}_n^{-1}\mathbf{k}(\mathbf{s}_0))'(\mathbf{X}_n'\mathbf{\Sigma}_n^{-1}\mathbf{X}_n)^{-1}(\mathbf{X}(\mathbf{s}_0) - \mathbf{X}_n'\mathbf{\Sigma}_n^{-1}\mathbf{k}(\mathbf{s}_0)).\end{aligned}$$

Refer to Ripley (1981) for detailed derivations.

Several issues merit attention. First, $\mathbf{\Sigma}_n$ is an $n \times n$ matrix, presenting much difficulty for inverting it, especially when n is large. An attractive solution is to use a subset of $(\mathbf{s}_1, \cdots, \mathbf{s}_n)$, say, $(\kappa_1, \cdots, \kappa_K)$, where $K << n$, as representative knots and perform low rank kriging as proposed by Kammann and Wand (2003). This subset can be obtained while an efficient space filling algorithm (e.g. Nychka and Saltzman, 1998).

Secondly, the development so far needs the covariance function $C(\cdot; \boldsymbol{\theta})$ or the spatial variance component $\boldsymbol{\theta}$ is known. When $\boldsymbol{\theta}$ is unknown (which is often the case), one needs to replace $\boldsymbol{\theta}$ that involves in (2.14) by its consistent estimate $\hat{\boldsymbol{\theta}}$ (discussed in the previous section), but it remains an open problem to derive the prediction error variance that also accounts for the variability of $\hat{\boldsymbol{\theta}}$. A kriging method that blends prediction and nonparametric estimation of the covariance function $C(\cdot)$ was given by Opsomer et al. (1999), though the estimate of the prediction error that fully accounts variations from all sources, e.g. owing to estimation of β and variance function, is still elusive.

Thirdly, equation (2.14) reveals that the classical kriging uses the linear combination of the observed values to approximate that of a new location, with larger weights assigned to more nearby locations. However, in many situations, e.g. for non-normal data, such linearity assumption is too strict and may not be plausible. We therefore consider an optimal predictor that minimizes the following conditional-mean-squared-error function

$$E[\{p(\mathbf{Y}_n;\mathbf{s}_0) - Y(\mathbf{s}_0)\}^2 | \mathbf{Y}_n], \tag{2.15}$$

where $p(\mathbf{Y}_n;\mathbf{s}_0)$ is a predictor at $\mathbf{s}_0$ based on the observed $\mathbf{Y}_n$. In view of

$$E[\{p(\mathbf{Y}_n;\mathbf{s}_0) - Y(\mathbf{s}_0)\}^2 | \mathbf{Y}_n] = \text{var}\{Y(\mathbf{s}_0)|\mathbf{Y}_n\} + [p(\mathbf{Y}_n;\mathbf{s}_0) - E\{Y(\mathbf{s}_0)|\mathbf{Y}_n\}]^2,$$

it is obvious that the optimal predictor is

$$\hat{Y}_o(\mathbf{s}_0) = E\{Y(\mathbf{s}_0)|\mathbf{Y}_n\}.$$

When $Y(\cdot)$ is a Gaussian process, the optimal predictor $\hat{Y}_o(\mathbf{s}_0)$ coincides with the classical kriging $\hat{Y}_k(\mathbf{s}_0)$ in (2.14). For non-normal data, we will consider a generalized linear mixed model, based on which the optimal predictor will be non-linear and often requires numerical approximations.

Finally, as we have been confined to use a fully parametric form to model the *trend* surface, i.e. the deterministic part of the spatial process, as in (2.6), one possible alternative is to estimate the trend surface by nonparametric regression. This would produce enough flexibility to absorb the signal almost completely into the trend, effectively diminishing spatial dependence. Mueller (2000) considered the following model in the spirit of Hastie and Tibshirani (1993),

$$Y(\mathbf{s}) = \eta(\mathbf{X}(\mathbf{s}), \boldsymbol{\beta}(\mathbf{s})) + Z(\mathbf{s})$$

where η is assumed to be a smooth function and $Z(\mathbf{s})$ is a white noise with $\text{cov}(Z(\mathbf{s}), Z(\mathbf{s}')) = 0$ if $\mathbf{s} \neq \mathbf{s}'$ and $\text{var}(Z(\mathbf{s})) = \sigma^2(\mathbf{s})$. By allowing $\boldsymbol{\beta}$ to change smoothly over the location $\mathbf{s}$, one would be able to recover the trend surface of the underlying spatial process. A unified approach that encompasses both regression and kriging based on this model is given by Host (1999).

3 Spatial models for non-normal/discrete data

While by far we have focused on the normal outcome data, non-normal data do frequently arise from spatial studies. Sometimes it is possible to transform the data so that they feature more as realizations from the Gaussian process, but in most cases, especially for discrete data, such transformation is not possible. Examples include the forest defoliation study reported in Heagerty and Lele (1998), with binary outcomes indicating the presence or absence of Gypsy moth egg masses at given locations, and the sudden infant death (SID) study (Cressie and Read, 1989), which studied the counts of SIDs cross 100 counties of North Carolina. Statistical models for independent non-normal data, can be traced back as early as 1934, when Bliss (1934) proposed the first probit regression model for binary data. It was not, however, until four decades later did Nelder and Wedderburn (1972) and McCullagh and Nelder (1983 1st ed., 1989 2nd ed.) propose Generalized Linear

Models (GLMs) to unify the models and modeling techniques for analyzing more general data (e.g. count data and polytomous data). Several authors (Laird and Ware, 1982; Stiratelli et al., 1984; Schall, 1991, among others) considered a natural generalization of the GLMs to accommodate correlated non-normal data by incorporating random terms into the linear predictor parts. The resulting models are termed generalized linear mixed models (GLMMs), providing a convenient and flexible way to model multivariate non-normal data. In particular, GLMMs constitute a unified framework for modeling geostatistical non-normal data, using mixed terms to model the underlying spatial process. We call the special application of GLMMs to the geostatistical data as spatial generalized linear mixed models (see e.g. Diggle et al., 1998; Zhang, 2002), which is to be discussed in the next section.

3.1 Spatial generalized linear mixed models (SGLMMs)

Consider a simple illustration of a spatial logistic regression for binary data:

$$\begin{aligned} Y(\mathbf{s})|\mathbf{X}(\mathbf{s}), Z(\mathbf{s}) &\overset{ind}{\sim} Bernoulli(\mu_{\mathbf{s}}); \\ logit(\mu_{\mathbf{s}}) &= \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + Z(\mathbf{s}) \end{aligned} \tag{3.1}$$

where $Y(\mathbf{s})$ denotes the binary outcome (e.g. 1 corresponds to the presence of Gypsy moth egg masses and 0 to the absence) at location $\mathbf{s}$, $\mathbf{X}(\mathbf{s})$ is a vector of additional individual-level covariates of interest and $Z(\mathbf{s})$ are unobserved spatially correlated random effects indexed by $\mathbf{s}$. In practice, we posit a random field structure on $Z(\mathbf{s})$, with covariance structure specified as in Section 2. This class of logistic regression model was originally designed for prospective studies, but is also applicable to case-control studies. Inference based on (3.1) has been detailed in Paciorek (2007).

It is straightforward to generalize (3.1) to accommodate more general data beyond binary outcomes. Specifically, conditional on unobserved spatial random variables $Z(\mathbf{s})$, the $Y(\mathbf{s})$ are assumed to be independent and follow a distribution of the exponential family:

$$Y(\mathbf{s})|\mathbf{X}(\mathbf{s}), Z(\mathbf{s}) \overset{ind}{\sim} f(Y(\mathbf{s})|\mathbf{X}(\mathbf{s}), Z(\mathbf{s})), \tag{3.2}$$

$$f(Y(\mathbf{s})|\mathbf{X}(\mathbf{s}), Z(\mathbf{s})) = \exp\{[Y(\mathbf{s})\alpha_{\mathbf{s}} - h(\alpha_{\mathbf{s}})]/\tau^2 - c(Y(\mathbf{s}), \tau)\}. \tag{3.3}$$

The conditional mean of $Y(\mathbf{s})|\mathbf{X}(\mathbf{s}), Z(\mathbf{s})$, denoted by $\mu_{\mathbf{s}}$, is related to $\alpha_{\mathbf{s}}$ through the identity $\mu_{\mathbf{s}} = \partial h(\alpha_{\mathbf{s}})/\partial\alpha_{\mathbf{s}}$. It is to be modeled, after a proper transformation, as a linear model in both the fixed and spatial random effects:

$$g(\mu_s) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + Z(\mathbf{s}). \tag{3.4}$$

Here, $g(\cdot)$ is coined a *link function*, often chosen as an invertible and continuous function, and $Z(\mathbf{s})$ is assumed to have a random field structure, whose covariance function is characterized by a finite dimensional parameter $\boldsymbol{\theta}$, termed the *spatial variance components*.

Model (3.4) is comprehensive and encompasses a variety of models, including the aforementioned spatial logistic regression model as a special case. Specifically, for binary outcome data, let

$$h(\alpha) = \log\{1 + \exp(\alpha)\},\ \tau \equiv 1,\ c(y, \tau^2) \equiv 0.$$

Choosing $g(\mu) = logit(\mu)$ yields spatial logistic regression model (3.1), while choosing $g(\mu) = \Phi^{-1}(\mu)$, where $\Phi(\cdot)$ is the CDF for a standard normal, gives a probit random effects model. On the other hand, for continuous outcome data, by setting

$$h(\alpha) = \frac{1}{2}\alpha^2, c(y, \tau^2) = \frac{1}{2}y^2/\tau^2 - \frac{1}{2}\log(2\pi\tau^2)$$

and $g(\cdot)$ to be an identity function, model (3.4) reduces to a linear mixed model. For count data, putting

$$h(\alpha) = e^{\alpha},\ \tau \equiv 1,\ c(y, \tau^2) = \log y$$

and choosing $g(\mu) = \log(\mu)$ results in a Poisson regression model.

Given data $(Y(\mathbf{s}_i), i = 1, \cdots, n)$, (3.2) and (3.3) induce the following log likelihood that the inference will be based on:

$$\ell = \log \int \prod_{i=1}^{n} f(Y(\mathbf{s}_i)|\mathbf{X}(\mathbf{s}_i), Z(\mathbf{s}_i); \boldsymbol{\beta}) f(\mathbf{Z}|\mathbf{X}_n; \boldsymbol{\theta}) d\mathbf{Z},$$

where the integration is over the n-dimensional random effect $\mathbf{Z} = (Z(\mathbf{s}_1), \cdots, Z(\mathbf{s}_n))'$.

We can further reformulate model (3.4) in a compact vectorial form. With $\mathbf{Y}_n, \mathbf{X}_n$ defined as in the previous section, we write

$$g\{E(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z})\} = \mathbf{X}_n\boldsymbol{\beta} + \mathbf{A}\mathbf{Z}, \tag{3.5}$$

where $\mathbf{A}$ is a non-random design matrix, compatible with the random effects $\mathbf{Z}$. The associated log likelihood function can be rewritten as

$$\ell(\mathbf{Y}_n|\mathbf{X}_n; \boldsymbol{\beta}, \boldsymbol{\theta}) = \log L(\mathbf{Y}_n|\mathbf{X}_n; \boldsymbol{\beta}, \boldsymbol{\theta}) = \log \int f(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta}) f(\mathbf{Z}|\mathbf{X}_n; \boldsymbol{\theta}) d\mathbf{Z}, \tag{3.6}$$

where $f(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta})$ is the conditional likelihood for $\mathbf{Y}_n$ and $f(\mathbf{Z}|\mathbf{X}_n; \boldsymbol{\theta})$ is the density function for $\mathbf{Z}$, given the observed covariates $\mathbf{X}_n$.

Model (3.5) is not a simple reformat - it accommodates more complex data structure beyond spatial data. For example, with properly defined $\mathbf{A}$ and random effects $\mathbf{Z}$ it encompasses non-normal clustered data and crossed factor data (Breslow and Clayton, 1993). When $\mathbf{A}$ is defined as matrix indicating membership of spatial regions (e.g. counties or census tracts), (3.5) models areal data as well. Model (3.5) accommodates a low-rank kriging spatial model, where the spatial random effects $\mathbf{Z}$ will have a dimension that does not increase with the sample size n and, in practice, is often far less than n. Specifically, consider a subset of locations $(\mathbf{s}_1, \cdots, \mathbf{s}_n)$, say, $(\kappa_1, \cdots, \kappa_K)$, where $K << n$, as representative knots. Let

$$\mathbf{A} = (C(\mathbf{s}_i - \kappa_k; \boldsymbol{\theta}))_{1 \leqslant i \leqslant n, 1 \leqslant k \leqslant K},$$

$$\boldsymbol{\Omega} = (C(\kappa_{k'} - \kappa_k; \boldsymbol{\theta}))_{1 \leqslant k, k' \leqslant K}$$

and $\mathbf{Z}$ be a $K \times 1$ vector with covariance $\boldsymbol{\Omega}^{-1}$. Then (3.5) represents a low-kriging model by taking a linear combination of radial basis functions $C(\mathbf{s} - \kappa_k; \boldsymbol{\theta})), 1 \leqslant k \leqslant K$, centered at the knots $(\kappa_1, \cdots, \kappa_K)$, and can be viewed a generalization of Kammann and Wand's (2003) linear geoadditive model to accommodate non-normal spatial data.

Because of the generality of (3.5), the ensuing inferential procedures in Section 3.2 will be based on (3.5) and (3.6), facilitating the prediction of spatial random effects and, hence, each individual's profile. Two routes can be taken. The best predictor of random effects minimizing the conditional-mean-squared-error (2.15) is $E(\mathbf{Z}|\mathbf{Y}_n)$, not necessarily linear in $\mathbf{Y}_n$. But if we confine our interest to an unbiased linear predictors of the form

$$\hat{\mathbf{Z}} = \mathbf{c} + \mathbf{Q}\mathbf{Y}_n,$$

for some conformable vector $\mathbf{c}$ and matrix $\mathbf{Q}$, minimizing the mean squared error (2.12) subject to constraint (2.13) leads to the best linear unbiased predictor (BLUP)

$$\hat{\mathbf{Z}} = E(\mathbf{Z}) + \mathrm{cov}(\mathbf{Z}, \mathbf{Y}_n)\{\mathrm{var}(\mathbf{Y}_n)\}^{-1}\{\mathbf{Y}_n - E(\mathbf{Y}_n)\}. \tag{3.7}$$

Equation (3.7) holds true without any normality assumptions (McCulloch and Searle, 2001).

For illustration, consider a Dirichlet model for binary spatial outcomes such that

$$Y(\mathbf{s})|Z(\mathbf{s}) \sim \mathit{Bernoulli}(Z(\mathbf{s}))$$

and the random effect $\mathbf{Z} = (Z(\mathbf{s}_1), \cdots, Z(\mathbf{s}_n)) \sim \mathit{Dir}(\alpha_1, \cdots, \alpha_n)$, where $\alpha_i > 0$. Using (3.7), we obtain the best linear predictor for $Z(\mathbf{s}_i)$,

$$\hat{Z}(\mathbf{s}_i) = \frac{\alpha_i}{\alpha_0} + \boldsymbol{\xi}_i' \boldsymbol{\Sigma}_0^{-1}(\mathbf{Y}_n - \boldsymbol{\mu}_0)$$

where $\alpha_0 = \sum_{i=1}^K \alpha_i$, $\boldsymbol{\mu}_0 = (\alpha_1/\alpha_0, \cdots, \alpha_n/\alpha_0)'$, $\boldsymbol{\Sigma}_0 = [c_{ij}]_{n \times n}$, $c_{ij} = -\alpha_i\alpha_j/\alpha_0^2$ $(\alpha_0 + 1)$ for $i \neq j$, $c_{ii} = \alpha_i(\alpha_0 - \alpha_i)/\alpha_0^2(\alpha_0 + 1)$ and $\boldsymbol{\xi}_i$ is the i-th column of $\boldsymbol{\Sigma}_0$. As a simple example, when $n = 2$,

$$\hat{Z}(\mathbf{s}_i) = \frac{\alpha_i + \bar{Y}_2}{\alpha_1 + \alpha_2 + 1},$$

where $\bar{Y}_2 = (Y(\mathbf{s}_1) + Y(\mathbf{s}_2))/2$.

3.2 Computing MLEs for SGLMMs

A common theme in fitting a SGLMM has been the difficulty of computation of likelihood-based inference. Computing the likelihood itself actually is often challenging for SGLMMs, largely due to high dimensional intractable integrals. We present below several useful likelihood-based approaches to estimating the coefficients and variance components, including iterative maximization procedures, such as the Expectation and Maximization (EM) algorithm, and approximation procedures, such as the Penalized Quazi-likelihood method and the Laplace method.

The EM algorithm (Dempster et al., 1977) was originally designed for likeliho od-based inference in the presence of missing observations, and involves an iterative procedure that increases likelihood at each step. The utility of the EM algorithm in a spatial setting lies in treating the unobserved spatial random terms as 'missing' data, and imputing the missing information based on the observed data, with the goal of maximizing the marginal likelihood of the observed data.

Specifically, if the random effects $\mathbf{Z}$ were observed, we would be able to write the 'complete' data as $(\mathbf{Y}_n, \mathbf{Z})$ with a joint log likelihood

$$\ell(\mathbf{Y}_n, \mathbf{Z}|\mathbf{X}_n; \boldsymbol{\beta}, \boldsymbol{\theta}) = \log f(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta}) + \log f(\mathbf{Z}|\mathbf{X}_n; \boldsymbol{\theta}). \tag{3.8}$$

As $\mathbf{Z}$ is unobservable, directly computing (3.8) is not feasible. Rather the EM algorithm adopts a two-step iterative process. The Expectation step ('E' step) computes the expectation of (3.8) conditional on the observed data. That is, calculate

$$\tilde{\ell} = E\{\ell(\mathbf{Y}_n, \mathbf{Z}|\mathbf{X}_n; \boldsymbol{\beta}, \boldsymbol{\theta})|\mathbf{Y}_n, \mathbf{X}_n; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0\},$$

where $\boldsymbol{\beta}_0, \boldsymbol{\theta}_0$ are the current values, followed by a Maximization step (M step), which maximizes $\tilde{\ell}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The E step and M step are iterated until convergence is achieved; however, the former is much costly, as the conditional distribution of $\mathbf{Z}|\mathbf{X}_n, \mathbf{Y}_n$ involves the distribution $f(\mathbf{Y}_n|\mathbf{X}_n)$, a high dimensional intractable integral. A useful remedy is the Metropolis-Hastings algorithm that approximates the conditional distribution of $\mathbf{Z}|\mathbf{X}_n, \mathbf{Y}_n$ by making random draws from $\mathbf{Z}|\mathbf{X}_n, \mathbf{Y}_n$ without calculating the density $f(\mathbf{Y}_n|\mathbf{X}_n)$ (McCulloch, 1997).

Apart from the common EM algorithm that requires a full likelihood analysis, several less costly techniques have proved useful for approximate inference in the SGLMMs and other nonlinear variance component models, among which the Penalized Quasi-likelihood (PQL) method Penalized Quasi-likelihood (PQL) method is most widely used.

The PQL method was initially exploited as an approximate Bayes procedure to estimate regression coefficients for semiparametric models; see Green (1987). Since then, several authors have explored the PQL to draw approximate inferences based on random effects models: Schall (1991) and Breslow and Clayton(1993) developed iterative PQL algorithms, Lee and Nelder (1996) applied the PQL directly to hierarchical models.

We consider the application of the PQL for the SGLMM (3.5). For notational simplicity we write the integrand of the likelihood function

$$f(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta})f(\mathbf{Z}|\mathbf{X}_n; \boldsymbol{\theta}) = \exp\{-K(\mathbf{Y}_n, \mathbf{Z})\}, \tag{3.9}$$

where, for notational simplicity, we do not list $\mathbf{X}_n$ as an argument in function K. Next evaluate the marginal likelihood. Temporarily we assume that $\boldsymbol{\theta}$ is known. For any fixed $\boldsymbol{\beta}$, expanding $K(\mathbf{Y}_n, \mathbf{Z})$ around its mode $\hat{\mathbf{Z}}$ up to the second order term, we have

$$\begin{aligned} L(\mathbf{Y}_n|\mathbf{X}_n; \boldsymbol{\beta}, \boldsymbol{\theta}) &= \int \exp\{-K(\mathbf{Y}_n, \mathbf{Z})\} d\mathbf{Z} \\ &= ||2\pi\{K^{(2)}(\mathbf{Y}_n, \tilde{\mathbf{Z}})\}^{-1}||^{1/2} \exp\{-K(\mathbf{Y}_n, \hat{\mathbf{Z}})\}, \end{aligned}$$

where $K^{(2)}(\mathbf{Y}_n, \mathbf{Z})$ denotes the second derivative of $K(\mathbf{Y}_n, \mathbf{Z})$ with respect to $\mathbf{Z}$, and $\tilde{\mathbf{Z}}$ lies in the segment joining 0 and $\hat{\mathbf{Z}}$. If $K^{(2)}(\mathbf{Y}_n, \mathbf{Z})$ does not vary too much as $\mathbf{Z}$ changes (for instance, $K^{(2)}(\mathbf{Y}_n, \mathbf{Z}) = \textit{constant}$ for normal data), maximizing the marginal likelihood (3.6) is equivalent to maximizing

$$e^{-K(\mathbf{Y}_n, \hat{\mathbf{Z}})} = f(\mathbf{Y}_n|\mathbf{X}_n, \hat{\mathbf{Z}}, \boldsymbol{\beta}) f(\hat{\mathbf{Z}}|\mathbf{X}_n; \boldsymbol{\theta}).$$

This step is also equal to jointly maximizing $f(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta}) f(\mathbf{Z}|\mathbf{X}_n; \boldsymbol{\theta})$ w.r.t $\boldsymbol{\beta}$ and $\mathbf{Z}$ with $\boldsymbol{\theta}$ being held constant. Finally, only $\boldsymbol{\theta}$ is left to be estimated, but it can be estimated by maximizing the approximate profile likelihood of $\boldsymbol{\theta}$,

$$||2\pi\{K^{(2)}(\mathbf{Y}_n, \hat{\mathbf{Z}}(\boldsymbol{\theta}))\}^{-1}||^{1/2} \exp\{-K(\mathbf{Y}_n, \hat{\mathbf{Z}}(\boldsymbol{\theta}))\};$$

refer to Breslow and Clayton (1993).

As no close-form solution is available, the PQL is often performed through an iterative process. In particular, Schall (1991) derived an iterative algorithm when the random effects follow normal distributions. Specifically, with the current estimated values of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\mathbf{Z}$, a working 'response' $\tilde{\mathbf{Y}}_n$ is constructed by the first order Taylor expansion of $g(\mathbf{Y})$ around $\boldsymbol{\mu}^z$, or explicitly,

$$\tilde{\mathbf{Y}}_n = g(\boldsymbol{\mu}^z) + g^{(1)}(\boldsymbol{\mu}^z)(\mathbf{Y} - \boldsymbol{\mu}^z) = \mathbf{X}\boldsymbol{\beta} + \mathbf{A}\mathbf{Z} + g^{(1)}(\boldsymbol{\mu}^z)(\mathbf{Y}_n - \boldsymbol{\mu}^z), \tag{3.10}$$

where $g^{(1)}(\cdot)$ denotes the first derivative and $g(\cdot)$ is defined in (3.4).

When viewing the last term in (3.10) as a random error, (3.10) suggests fitting a linear mixed model on $\tilde{\mathbf{Y}}_n$ to obtain the updated values of $\boldsymbol{\beta}, \mathbf{Z}$ and $\boldsymbol{\theta}$, followed by a recalculation of the working 'responses'. The iteration shall continue until convergence. Computationally, the PQL is easy to implement, only requiring repeatedly invoking existing macros, for example, SAS 'PROC MIXED'. The PQL procedure yields exact MLEs for normally distributed data and for some cases when the conditional distribution of $\mathbf{Y}_n$ and the distribution of $\mathbf{Z}$ are conjugate.

Several variations of the PQL are worth mentioning. First, the PQL is actually applicable in a broader context where only the first two conditional moments of $\mathbf{Y}_n$ given $\mathbf{Z}$ are needed, in lieu of a full likelihood specification. Specifically, $f(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta})$ in (3.9) can be replaced by the quasi-likehood function $\exp\{ql(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta})\}$, where

$$ql(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta}) = \sum_{i=1}^{m} \int_{Y_i}^{\mu_i^z} \frac{Y_i - t}{V(t)} dt.$$

Here $\mu_i^z = E(Y(\mathbf{s}_i)|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta})$ and $V(\mu_i^z) = \mathrm{var}(Y(\mathbf{s}_i)|\mathbf{X}_n, \mathbf{Z}; \boldsymbol{\beta})$.

Secondly, the PQL is tightly related to other approximation approaches, such as the Laplace method and the Solomon-Cox method, which have also received much attention. The Laplace method (see, e.g. Liu and Pierce (1993)) differs from the PQL only in that the former obtains $\hat{\mathbf{Z}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ by maximizing the integrand $e^{-K(\mathbf{Y}_n,\mathbf{Z})}$ with $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ being held fixed, and subsequently estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ by jointly maximizing

$$||2\pi\{K^{(2)}(\mathbf{Y}_n, \hat{\mathbf{Z}})\}^{-1}||^{1/2} \exp\{-K(\mathbf{Y}_n, \hat{\mathbf{Z}})\}.$$

On the other hand, with the assumption of $E(\mathbf{Z}|\mathbf{X}_n) = 0$, the Solomon-Cox technique approximates the integral $\int f(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z}) f(\mathbf{Z}|\mathbf{X}_n) d\mathbf{Z}$ by expanding the integrand $f(\mathbf{Y}_n|\mathbf{X}_n, \mathbf{Z})$ around $\mathbf{Z} = 0$; see Solomon and Cox (1992).

In summary, none of these approximate methods produce consistent estimates, with exception in some special cases, e.g. normal data. Moreover, as these methods are essentially normal approximation-based, they typically do not perform well for sparse data, e.g. for binary data, and when the cluster size is relatively small (Lin and Breslow, 1996). Nevertheless, they provide a much needed alternative, especially given that full likelihood approaches are not always feasible for spatial data.

4 Spatial models for censored outcome data

Biomedical and epidemiological studies have spawned an increasing interest in and practical need for developing statistical methods for modeling time-to-event data that are subject to spatial dependence. Little work has been done in this area. Li and Ryan (2002) proposed a class of spatial frailty survival models. A further extension accommodating time-varying and nonparametric covariate effects, namely geoadditive survival model, was proposed by Hennerfeind et al. (2006). However, the regression coefficients of these frailty models do not have an easy population-level interpretation, less appealing to practitioners. In this section, we focus on a new class of semiparametric likelihood models recently developed by Li and Lin (2006). A key advantage of this model is that observations marginally follow the Cox proportional hazard model and regression coefficients have a population level interpretation and their joint distribution can be specified using a likelihood function that allows for flexible spatial correlation structures.

Consider in a geostatistical setting a total of n subjects, who are followed up to event (e.g. death or onset of asthma) or being censored, whichever comes first. For each individual, we observe a $q \times 1$ vector of covariates $\mathbf{X}$, and an observed event time $\tilde{T} = min(T, U)$ and a non-censoring indicator $\delta = I(T \leqslant U)$, where T and U are underlying true survival time and censoring time respectively, and $I(\cdot)$ is an indicator function. We assume noninformative censoring, i.e., the censoring time U is independent of the survival time T given the observed covariates, and the distribution of U does not involve parameters of the true survival model. The covariates $\mathbf{X}$ are assumed to be a predictable time-dependent (and space-dependent) process. Also documented is each individual's geographic location $\mathbf{s}_i$.

Denote by $\bar{\mathbf{X}}(t) = (\mathbf{X}(s) : 0 \leqslant s \leqslant t)$ the $\mathbf{X}$-covariate path up to time t. We specify that the survival time T marginally follows the Cox model

$$\lambda\{t|\bar{\mathbf{X}}(t)\} = \lambda_0(t)\psi\{\mathbf{X}(t), \boldsymbol{\beta}\} \tag{4.1}$$

where $\psi\{\cdot,\cdot\}$ is a positive function, $\boldsymbol{\beta}$ is a regression coefficient vector and $\lambda_0(t)$ is an unspecified baseline hazard function. A common choice of ψ is the exponential function, in which case, $\psi\{\mathbf{X}(t), \boldsymbol{\beta}\} = \exp\{\boldsymbol{\beta}'\mathbf{X}(t)\}$, corresponding to the Cox proportional hazards model discussed in Li and Lin (2006). This marginal model refers to the assumption that the hazard function (4.1) is with respect to each individual's own filtration, $\mathcal{F}_t = \sigma\{I(\tilde{T} \leqslant s, \delta = 1), I(\tilde{T} \geqslant s), \mathbf{X}(s), 0 \leqslant s \leqslant t\}$, the sigma field generated by the survival and covariate paths up to time t. The regression coefficients $\boldsymbol{\beta}$ hence have a population-level interpretation.

Use subscript i to flag each individual. A spatial joint likelihood model for $T_1, \cdots, T_n$ is to be developed, which allows T_i to marginally follow the Cox model (4.1) and allows for a flexible spatial correlation structure among the T_i's. Denote by $\Lambda_i(t) = \int_0^t \lambda_i(s|\mathbf{X}_i)ds$ the cumulative hazard and $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ the cumulative baseline hazard. Then $\Lambda_i(T_i)$ marginally follows a unit exponential distribution, and its probit-type transformation

$$T_i^* = \Phi^{-1}\left\{1 - e^{-\Lambda_i(T_i)}\right\} \tag{4.2}$$

follows the standard normal distribution marginally, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. We then conveniently posit a spatial structure on the underlying random fields of $\mathbf{T}^* = \{T_i^*, i = 1, \cdots, n\}$ within the traditional Gaussian geostatistical framework. Hence such a normal transformation of the cumulative hazard provides a general framework to construct a flexible joint likelihood model for spatial survival data by preserving the Cox model for each individual marginally. This also provides a convenient way to generate spatially correlated survival data whose marginal distributions follow the Cox model.

Specifically, we assume $\mathbf{T}^*$ to be a Gaussian random field as specified in Section 2, such that $\mathbf{T}^*$ follows a joint multivariate normal distribution as

$$\mathbf{T}^* = \{T_i^*, i = 1, \cdots, n\} \sim MVN(\boldsymbol{o}_n, \boldsymbol{\Gamma}), \tag{4.3}$$

where $\boldsymbol{\Gamma}$ is a positive definite matrix with diagonal elements being 1. Denote by θ_{ij} the (i,j)th element of $\boldsymbol{\Gamma}$. We assume that the correlation θ_{ij} between a pair of normalized survival times, say T_i^* and T_j^*, depends on their geographic locations $\mathbf{s}_i$ and $\mathbf{s}_j$, i.e.

$$\text{corr}(T_i^*, T_j^*) = \theta_{ij} = \theta_{ij}(\mathbf{s}_i, \mathbf{s}_j) \tag{4.4}$$

for $i \neq j$ $(i, j = 1, \cdots, n)$, where $\theta_{ij} \in (-1, 1)$. Generally a parametric model is assumed for θ_{ij}, which depends on a parameter vector $\boldsymbol{\alpha}$ as $\theta_{ij}(\boldsymbol{\alpha})$.

Since the transformed times $\mathbf{T}^*$ are normally distributed, a rich class of models can be used to model the spatial dependence by specifying a parametric model for θ_{ij}. For instance, $\theta_{ij}(\boldsymbol{\alpha})$ may be parameterized as $\rho(d_{ij}, \boldsymbol{\alpha})$, an isotropic correlation function which decays as the Euclidean distance d_{ij} between two individuals increases. A widely adopted choice for the correlation function is the Matèrn function $m(\sigma^2, \zeta, \nu, d)$ as defined in (2.2). Recall that σ^2 is a scale parameter and corresponds to the 'partial sill', ζ measures the correlation decay with the distance and ν is a smoothness parameter, characterizing the behavior of the correlation function near the origin, but its estimation is difficult as it requires dense space data and may even run into identifiability problems. Stein (1999) has argued that data can not distinguish between $\nu = 2$ and $\nu > 2$. Li and Lin (2006) fixing ν to estimate the other parameters and performing a sensitivity analysis by varying ν for data analysis, in which case the unknown $\boldsymbol{\alpha} = (\sigma^2, \zeta)'$.

4.1 A class of semiparametric estimation equations

As a full likelihood-based inferential procedure, which involves a large dimensional integral, is difficult, we opt for a class of spatial semiparametric estimating equations constructed using the first two moments of individual survival times and the covariance functions of all pairs of survival times.

First derive the Martingale covariance rate function under the semiparametric normal transformation model (4.2)—(4.3). We denote the counting process $N_i(t) = I(\tilde{T}_i \leqslant t, \delta_i = 1)$ and the at-risk process $Y_i(t) = I(\tilde{T}_i \geqslant t)$. Next define a Martingale, adapted to the filtration $\mathcal{F}_{i,t} = \sigma(N_i(s), Y_i(s), \mathbf{X}_i(s), 0 \leqslant\leqslant s < t)$, as

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\psi\{\mathbf{X}_i(s), \boldsymbol{\beta}\}d\Lambda_0(s).$$

To relate the correlation parameters $\boldsymbol{\alpha}$ to the counting processes, one needs to consider the joint counting process of two individuals. Define the conditional Martingale covariance rate function for the joint counting process of two individuals, a multi-dimensional generalization of the conditional hazard function, as (Prentice and Cai, 1992)

$$A_{i,j}(dt_1, dt_2) = E\{M_i(dt_1)M_j(dt_2)|T_i > t_1, T_j > t_2\}.$$

Then we have

$$E\{M_i(t_1)M_j(t_2) - \int_0^{t_1}\int_0^{t_2} Y_i(s_1)Y_j(s_2)A_{i,j}(ds_1, ds_2)\} = 0.$$

Denote by $\tilde{S}_{ij}(v_1, v_2)$ the joint survival function of $\Lambda_i(T_i)$ and $\Lambda_j(T_j)$, the exponential transformations of the original survival times. Then

$$\tilde{S}_{ij}(v_1, v_2; \theta_{ij}) = P\{\Lambda_i(T_i) > v_1, \Lambda_j(T_j) > v_2; \theta_{ij}\}. \tag{4.5}$$

Following Prentice and Cai (1992), one can show that the covariance rate can be written as

$$A_{i,j}(dt_1, dt_2; \theta_{ij}) = A_0\{\Lambda_i(t_1), \Lambda_j(t_2); \theta_{ij}\}\Lambda_i(dt_1)\Lambda_j(dt_2),$$

where

$$A_0(v_1,v_2;\theta)=\left\{\frac{\partial^2}{\partial v_1\partial v_2}\tilde{S}_{ij}(v_1,v_2;\theta)+\tilde{S}_{ij}(v_1,v_2;\theta)\right.$$
$$\left.+\frac{\partial}{\partial v_1}\tilde{S}_{ij}(v_1,v_2;\theta)+\frac{\partial}{\partial v_2}\tilde{S}_{ij}(v_1,v_2;\theta)\right\}\Big/\tilde{S}_{ij}(v_1,v_2;\theta).$$

As a special case, $A_0(v_1,v_2;\theta=0)\equiv 0$. Li and Lin (2006) showed that as $\theta\to 0+$, $A_0(v_1,v_2;\theta)$ converges to 0 uniformly at the same rate as that when (v_1,v_2) lies in a compact set. We simultaneously estimate the regression coefficients $\boldsymbol{\beta}$ (a $q\times 1$ vector) and the correlation parameters $\boldsymbol{\alpha}$ (a $k\times 1$ vector) by considering the first two moments of the Martingale vector $(M_1,\cdots,M_n)$. In particular, for a pre-determined constant $\tau>0$ such that it is within the support of the observed failure time, i.e $P(\tau<U_i\wedge T_i)>0$ (in practice τ is usually the study duration), we consider the following unbiased estimating functions for $\boldsymbol{\Theta}=\{\boldsymbol{\beta},\boldsymbol{\alpha}\}$ for an arbitrary pair of two individuals, indexed by u and v:

- if $u=v$,

$$\mathbf{U}_{u,u}(\boldsymbol{\Theta})=\begin{bmatrix}\int_0^\tau \mathbf{X}_u(s)W_{(u,u)}(s)dM_u(s)\\ \mathbf{v}_{uu}\{M_u^2(\tau)-\int_0^\tau Y_u(s)d\Lambda_u(s)\}\end{bmatrix}$$

  where $W_{(u,u)}(s)$ (a scalar) and $\mathbf{v}_{uu}$ (a length-q vector) are non-random weights.

- if $u\neq v$,

$$\mathbf{U}_{u,v}(\boldsymbol{\Theta})=\begin{bmatrix}\int_0^\tau \mathbf{X}_{u,v}(s)\mathbf{W}_{(u,v)}(s)d\mathbf{M}_{u,v}(s)\\ \mathbf{v}_{uv}\{M_u(\tau)M_v(\tau)-A_{uv}\}\end{bmatrix}$$

  where $\mathbf{X}_{u,v}(s)=\{\mathbf{X}_u(s),\mathbf{X}_v(s)\}$, $d\mathbf{M}_{u,v}(s)=\{dM_u(s),dM_v(s)\}'$, and $\mathbf{W}_{(u,v)}\ (s)=\{w_{ij}^{(u,v)}\}_{2\times 2}$ and $\mathbf{v}_{uv}$ (a length-q vector) are non-random weigh ts and

$$A_{uv}=\int_0^\tau\int_0^\tau Y_u(s)Y_v(t)A_0\{\Lambda_u(s),\Lambda_v(t);\theta_{uv}\}d\Lambda_u(s)d\Lambda_v(t)$$
$$=\int_0^{\Lambda_u(X_u\wedge\tau)}\int_0^{\Lambda_v(X_v\wedge\tau)}A_0\{t_1,t_2;\theta_{uv}\}dt_1dt_2.$$

It can be easily shown that $\mathbf{U}_{u,v}$ is an unbiased estimating function, since $E\{\mathbf{U}_{u,v}(\boldsymbol{\Theta}_0)\}=0$, where the expectation is taken under the true $\boldsymbol{\Theta}_0=(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)$ and the true cumulative hazard function $\Lambda_0(\cdot)$. In fact, the first component of $\mathbf{U}_{u,v}$, which is the estimating equation for $\boldsymbol{\beta}$, is unbiased even when the spatial correlation structure is misspecified. Hence the regression coefficient estimator $\widehat{\boldsymbol{\beta}}$ is robust to misspecification of the spatial correlation structure.

As $\Lambda_0(t)$ in the estimating equations is unknown, a natural alternative is to substitute it with the Breslow-type estimator

$$\hat{\Lambda}_0(t)=\int_0^t\frac{\sum_{i=1}^m dN_i(s)}{\sum_{i=1}^m Y_i(s)\psi\{\mathbf{X}_i(s),\boldsymbol{\beta}\}}.$$

As a result, the parameters of interest $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ are estimated by solving the following estimating equations, constructed by weightedly pooling individual Martingale residuals and weightedly pooling all pairs of Martingale residuals respectively

$$\mathbf{G}_n = n^{-1} \sum_{u \geqslant v} \hat{\mathbf{U}}_{u,v}(\boldsymbol{\Theta}) = 0, \tag{4.6}$$

where $\widehat{\mathbf{U}}(\cdot)$ arises from $\mathbf{U}(\cdot)$ by substituting $\Lambda_0(t)$ by $\hat{\Lambda}_0(t)$.

With the matrix notation, (4.6) can be expressed conveniently as

$$n^{-1} \begin{bmatrix} \int_0^\tau \mathbf{X}(s)\mathbf{W} d\hat{\mathbf{M}}(s) \\ \hat{\mathbf{M}}'(\tau)\mathbf{V}_1\hat{\mathbf{M}}(\tau) - tr(\mathbf{V}_j\hat{\mathbf{A}}) \end{bmatrix} = 0, \tag{4.7}$$

where $j = 1, \cdots, k$, $\mathbf{W}$ and $\mathbf{V}_j$ are weight matrices, $\hat{\mathbf{M}} = (\hat{M}_1, \cdots, \hat{M}_n)'$, $\mathbf{X}(s) = \{\mathbf{X}_1(s), \cdots, \mathbf{X}_n(s)\}'$, $\hat{\mathbf{A}}$ is an $n \times n$ matrix whose uv-th $(u \neq v)$ entry is $\hat{A}_{uv}$ obtained from A_{uv} with $\Lambda_0(t)$ replaced by $\hat{\Lambda}_0(t)$, and $\hat{A}_{uu} = \int_0^\tau Y_u(s) d\hat{\Lambda}_u(s)$.

The weight matrices $\mathbf{W}$ and $\mathbf{V}_1, \cdots, \mathbf{V}_k$ are meant to improve efficiency and convergence of the estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. In particular, following Cai and Prentice (1997) $\mathbf{W}$ can be specified as $(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2})^{-1}$, the inverse of the correlation matrix of the Martingale vector $\mathbf{M}(\tau)$, where $\mathbf{D} = diag(A_{11}, \cdots, A_{nn})$. In the absence of spatial dependence, $\mathbf{W}$ is an identity matrix and hence the first set of equations of (4.7) is reduced to the ordinary partial likelihood score equation for regression coefficients $\boldsymbol{\beta}$. To specify $\mathbf{V}_j$ $(j = 1, \cdots, q)$, one could assume $\mathbf{V}_j = \mathbf{A}^{-1}(\partial\mathbf{A}/\partial\alpha_j)\mathbf{A}^{-1}$. Under this specification, the second set of estimating equations in (4.7) resembles the score equations of the variance components $\boldsymbol{\alpha}$ if the 'response' $\hat{\mathbf{M}}$ followed a multivariate normal distribution $MVN(\boldsymbol{o}_n, \mathbf{A})$ (Cressie, 1993, p483).

To ensure numerical stability, we consider a modification of the spatial estimating equation (4.7) by adding a penalty term,

$$\mathbf{G}_n^*(\boldsymbol{\Theta}) = \mathbf{G}_n(\boldsymbol{\Theta}) - \frac{1}{n}\boldsymbol{\Omega}\boldsymbol{\Theta}$$

where $\boldsymbol{\Omega}$ is a $(q+k) \times (q+k)$ positive definite matrix, acting like a penalty term. This penalized version of the spatial estimating equation (4.7) can be motivated from the perspective of ridge regression or from Bayesian perspectives by putting a Gaussian prior $MVN(\boldsymbol{o}_{q+k}, \boldsymbol{\Omega}^{-1})$ on $\boldsymbol{\Theta}$, and results in stabilized variance component estimates of $\boldsymbol{\alpha}$ for example, for moderate sample sizes, and is likely to force the resulting estimates to lie in the interior of the parameter space (Heagerty and Lele, 1998). Therefore in practice, especially when the sample size is not large, we consider using a small penalty, $\boldsymbol{\Omega} = \omega\mathbf{I}$, where $0 < \omega < 1$, for numerical stability. As the sample size n goes to ∞, we have $\frac{1}{n}\boldsymbol{\Omega}\boldsymbol{\Theta} \to 0$. Therefore $\mathbf{G}_n(\boldsymbol{\Theta})$ and $\mathbf{G}_n^*(\boldsymbol{\Theta})$ are asymptotically equivalent, and therefore the large sample results of the original and penalized estimating equations are equivalent.

4.2 Asymptotic Properties and Variance Estimation

The large sample properties of the estimators can be established, facilitating drawing inference based on the semiparametric normal transformation model. Under the regularity conditions listed in Li and Lin (2006), the estimators obtained by solving $\mathbf{G}_m(\boldsymbol{\Theta}) = 0$ exist and are consistent for the true values of $\boldsymbol{\Theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$ and that $n^{1/2}\{\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\}$ is asymptotic normal with mean zero and a covariance matrix that can be easily estimated using a sandwich estimator. The results are formally stated in the following Proposition and can be proved along the line of Li and Lin (2006), which focused on the proportional hazards models.

Proposition 4.1. *Assume the true $\boldsymbol{\Theta}_0$ is an interior point of an compact set, say, $\mathcal{B} \times \mathcal{A} \in R^{q+k}$, where q is the dimension of $\boldsymbol{\beta}$ and k is the dimension of $\boldsymbol{\alpha}$. When n is sufficiently large, the estimating equation $\mathbf{G}_n(\boldsymbol{\Theta}) = 0$ has a unique solution in a neighborhood of $\boldsymbol{\Theta}_0$ with probability tending to 1 and the resulting estimator $\hat{\boldsymbol{\Theta}}$ is consistent for $\boldsymbol{\Theta}_0$. Furthermore, $\sqrt{n}\{\boldsymbol{\Sigma}^{(2)}\}^{-1/2}\boldsymbol{\Sigma}\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})' - (\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)'\} \xrightarrow{d} MVN\{\boldsymbol{o}_{q+k}, \mathbf{I}\}$, where $\mathbf{I}$ is an identity matrix whose dimension is equal to that of $\boldsymbol{\Theta}_0$, and*

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{u \geqslant v} E\left\{\frac{\partial}{\partial \boldsymbol{\Theta}}\mathbf{U}_{u,v}(\boldsymbol{\Theta})\right\}$$

$$\boldsymbol{\Sigma}^{(2)} = \frac{1}{n^2}\sum_{u_1 \geqslant v_1}\sum_{u_2 \geqslant v_2} E\{\mathbf{U}_{u_1,v_1}(\boldsymbol{\Theta}_0)\mathbf{U}_{u_2,v_2}(\boldsymbol{\Theta}_0)\}.$$

It follows that the covariance of $\hat{\boldsymbol{\Theta}}$ can be estimated in finite samples by

$$\mathbf{I}_n^{-1} = \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Sigma}}^{(2)}\left\{\widehat{\boldsymbol{\Sigma}}^{-1}\right\}' \tag{4.8}$$

where $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}^{(2)}$ are estimated by replacing $\mathbf{U}_{uv}(\cdot)$ by $\hat{\mathbf{U}}_{uv}(\cdot)$ and evaluated at $\widehat{\boldsymbol{\Theta}}_0$.

Although each $E\left\{\hat{\mathbf{U}}_{u_1,v_2}(\boldsymbol{\Theta}_0)\hat{\mathbf{U}}'_{u_2,v_2}(\boldsymbol{\Theta}_0)\right\}$ could be evaluated numerically, the total number of these calculations would be prohibitive, especially when the sample size m is large. To numerically approximate $\widehat{\boldsymbol{\Sigma}}^{(2)}$, one can explore the resampling techniques of Carlstein (1986) and Sherman (1996). Specifically, under the assumption of

$$n \times E\left\{\mathbf{G}_n\mathbf{G}_n'\right\} \to \boldsymbol{\Sigma}_\infty,$$

$\boldsymbol{\Sigma}_\infty$ can be estimated by averaging K randomly chosen subsets of size n_j $(j = 1, \cdots, K)$ from the n subjects as

$$\widehat{\boldsymbol{\Sigma}}_\infty = K^{-1}\sum_{j=1}^{K} n_j\left\{\widehat{\mathbf{G}}_{n_j}\widehat{\mathbf{G}}_{n_j}'\right\},$$

where $\widehat{\mathbf{G}}_{n_j}$ is obtained by substituting $\boldsymbol{\Theta}$ with $\widehat{\boldsymbol{\Theta}}$ in $\mathbf{G}_{n_j}$. The n_j is often chosen to be proportional to n so as to capture the spatial covariance structure. For practical utility, Li and Lin recommended to choose n_j to be roughly 1/5 of the total population. Given the estimates $\widehat{\mathbf{\Sigma}}_\infty$ and $\widehat{\mathbf{\Sigma}}$, the covariance of $\widehat{\mathbf{\Theta}}$ can be estimated by $\widehat{\mathbf{\Sigma}}^{-1}[1/n \times \widehat{\mathbf{\Sigma}}_\infty](\widehat{\mathbf{\Sigma}}^{-1})'$.

To estimate the covariance matrix of the estimates arising from the penalized estimator obtained by solving $\mathbf{G}_n^*(\Theta) = 0$, $\widehat{\mathbf{\Sigma}}$ is replaced by $\widehat{\mathbf{\Sigma}} - \frac{1}{n}\mathbf{\Omega}$. A similar procedure was adopted by Heagerty and Lele (1998) for spatial logistic regression.

4.3 A data example: east boston asthma study

Li and Lin reported the application of the proposed method to analyze the East Boston Asthma study, focusing on assessing how the familial history of asthma may have attributed to disparity in disease burden. In particular, this study was to establish the relationship between the Low Respiratory Index (LRI) in the first year of life, ranging from 0 to 16, with high values indicating worse respiratory functioning, and age at onset of childhood asthma, controlling for maternal asthma status (MEVAST), coded as 1=ever had asthma and 0=never had asthma, and log-transformed maternal cotinine levels (LOGMCOT). This investigation would help better understand the natural history of asthma and its associated risk factors and to develop future intervention programs.

Subjects were enrolled at community health clinics throughout the east Bost on area, with questionnaire data collected during regularly scheduled well-baby visits. The ages at onset of asthma were identified through the questionnaires. Residential addresses were recorded and geocoded, with geographic distance measured in the unit of kilometer. A total of 606 subjects with complete information on latitude and longitude were included in the analysis, with 74 events observed at the end of the study. The median followup was 5 years. East Boston is a residential area of relatively low income working families. Participants in this study were largely white and hispanic children, aging from infancy to 6 years old. Asthma is a disease strongly affected environmental triggers. Since the children living in adjacent locations might have had similar backgrounds and living environments and, therefore, were exposed with similar unmeasured similar physical and social environments, their ages at onset of asthma were likely to be subject to spatial correlation.

The age at onset of asthma was assumed to marginally follow a Cox model

$$\lambda(t) = \lambda_0(t)\exp\{\beta_L \times \text{LRI} + \beta_M \times \text{MEVAST} + \beta_C \times \text{LOGMCOT}\}, \qquad (4.9)$$

while the Matèrn model (2.1) was assumed for the spatial dependence. Evidently, $beta_L, \beta_M$ and β_C measured the impact of main covariates and have population-level interpretations. The regression coefficients and the correlation parameters were estimated using the spatial semiparametric estimating equation approach, and the associated standard error estimates were computed using (4.8). To check the robustness of the method, Li and Lin varied the smoothness parameter ν in (2.1) to be 0.5, 1 and 1.5.

As the East Boston Asthma Study was conducted in a fixed region, to examine the performance of the variance estimator in (4.8), which was developed under the increasing-domain-asymptotic, Li and Lin calculated the variance using a 'delete-a-block' jackknife method (see, e.g. Kott (1998)). Specifically, they divided the samples into B nonoverlapping blocks based on their geographic proximity and then formed B jackknife replicates, where each replicate was formed by deleting one of the blocks from the entire sample. For each replicate, the estimates based on the semiparametric estimating equations were computed, and the jackknife variance was formulated as

$$\mathrm{var}_{jackknife} = \frac{B-1}{B}\sum_{j=1}^{B}(\hat{\mathbf{\Theta}}_j - \hat{\mathbf{\Theta}})(\hat{\mathbf{\Theta}}_j - \hat{\mathbf{\Theta}})' \tag{4.10}$$

where $\hat{\mathbf{\Theta}}_j$ was the estimate produced from the jackknife replicate with the j-th 'group' deleted and $\hat{\mathbf{\Theta}}$ was the estimate based on the entire population. In their calculation, B was chosen to be 40, which appeared large enough to render a reasonably good measure of variability. This jackknife scheme, in a similar spirit of Carlstein (1986, 1988), treated each block approximately independent and seemed plausible for this data set, especially in the presence of weak spatial dependence. Loh and Stein (2004) termed this scheme as the *splitting method* and found it work even better than more complicated block-bootstrapping methods (e.g. Kunsch, 1989; Liu and Singh, 1992; Politis and Romano, 1992; Bulhmann and Kunsch, 1995). Other advanced resampling schemes for spatial data are also available, e.g double-subsampling method (Lahiri et al., 1999; Zhu and Morgan, 2004) and linear estimating equation Jackknifing (Lele, 1991), but are subject to much more computational burden compared with the simple jackknife scheme we used.

Their results are summarized in the following table, with the large sample standard errors (SE_a) computed using the method described in Section 4.3 and the Jackknife standard errors (SE_j) computed using (4.10).

	$\nu = 0.5$			$\nu = 1$			$\nu = 1.5$		
Parameters	Estimate	SE_a	SE_j	Estimate	SE_a	SE_j	Estimate	SE_a	SE_j
β_L	0.3121	0.0440	0.0357	0.3118	0.0430	0.0369	0.3124	0.0432	0.0349
β_M	0.2662	0.3314	0.3222	0.2644	0.3289	0.3309	0.2676	0.3283	0.3340
β_C	0.0294	0.1394	0.1235	0.02521	0.1270	0.1063	0.0277	0.1288	0.1083
σ^2	1.68E-3	9.8E-3	0.0127	0.74E-3	5.0E-3	7.1E-3	0.72E-3	5.5E-3	4.8E-3
ζ	2.2977	4.974	3.708	2.1917	4.7945	4.1988	1.8886	6.5005	5.01617

The estimates of the regression coefficients and their standard errors were almost constant with various choices of the smoothness parameter ν and indicated that the regression coefficient estimates were not sensitive to the choice of ν in this data set. The standard errors obtained from the large sample approximation and the Jackknife method were reasonably similar. Low respiratory index was highly significantly associated with the age at onset of asthma, e.g. $\widehat{\beta}_L = 0.3121$ ($SE_a = 0.0440, SE_j = 0.0357$) when $\nu = 0.5$; $\widehat{\beta}_L = 0.3118$ ($SE_a = 0.0430, SE_j = 0.0369$) when $\nu = 1.0$; $\widehat{\beta}_L = 0.3124$ ($SE_a = 0.0432, SE_j = 0.0349$) when $\nu = 1.5$, indicating that a child with a poor respiratory functioning was more likely to develop asthma, after controlling for maternal asthma, maternal cotinine levels and accounting for the spatial variation. No significant association was found between ages at onset of asthma and maternal asthma and cotinine levels. The estimates of the spatial dependence parameters, σ^2 and ζ varied slightly with the choices of ν. The scale parameter σ^2 corresponds to the partial sill (Waller and Gotway, 2004, p.279) and measures the correlation between subjects in close geographic proximity. Thi analysis showed that such a correlation is relatively small. The parameter ζ measures global spatial decay of dependence with the spatial distance (measured in kilometers). For example, when $\nu = 0.5$, i.e., under the exponential model, $\zeta = 2.2977$ means the correlation decays by $1 - \exp(-2.2977 \times 1) \doteq 90\%$ for every one kilometer increase in distance.

5 Concluding remarks

This chapter has reviewed the methodologies for the analysis of spatial data within the geostatistical framework. We have dealt with data that consist of the measurements at a finite set of locations, where the statistical problem is to draw inference about the spatial process, based on the partial realization over this subset of locations. Specifically, we have considered using linear mixed models and generalized linear models that enable likelihood inference for fully observable spatial data. The fitting of such models by using maximum likelihood continues to be complicated owing to intractable integrals in the likelihood. In addition to the methods discussed in this chapter, there has been much research on the topic since the last decade, including Wolfinger and O'Connell (1993), Zeger and Karim (1993), Diggle et al. (1998), Booth and Hobert (1999).

We have also reviewed a new class of semiparametric normal transformation model for spatial survival data that was recently developed by Li and Lin (2006). A key feature of this model is that it provides a rich class of models where regression coefficients have a population-level interpretation and the spatial dependence of survival times is conveniently modeled using flexible normal random fields, which is advantageous given that there are virtually none spatial failure time distributions that are convenient to work with. Several open problems, however, remain to be investigated for this new model, including model diagnostics (e.g. examine the spatial correlation structure for censored data), prediction (e.g. predict survival outcome for new locations) and computation (e.g. develop fast convergent algorithms for inference).

Lastly, as this chapter tackles geostatistical data mainly from the frequentist points of view, we have by-passed the Bayesian treatments, which have been, indeed, much active in the past 20 years. Interested readers can refer to the book of Banerjee et al. (2004) for a comprehensive review of Bayesian methods.

References

[1] Abramowitz M., and Stegun I. A. (Editors)(1965), Handbook of Mathematical Functions, Dover Publications, New York.

[2] C. A. Brebbia, J. C. F. Telles and LC. Wrobel(1984), Boundary Element Techniques: Theory and Applications in Engineering, Springer-Verlag, Berlin.

[3] Banerjee, S. and Carlin, B.P (2003), Semiparametric Spatio-temporal Frailty Modeling, *Environmetrics*, 14 (5), 523-535.

[4] Banerjee, S., Carlin, B.P. and Gelfand, A. E (2004), Hierarchical Modeling and Analysis for Spatial Data, Chapman and Hall/CRC Press, Boca Raton.

[5] Besag, J., York, J. and Mollie, A. (1991), Bayesian Image Restoration, With Two Applications in Spatial Statistics, Annals of the Institute of Statistical Mathematics, 43, 1-20.

[6] Breslow, N. E. and Clayton, D. G. (1993), Approximate Inference in Generalized Linear Mixed Models, Journal of the American Statististical Association, 88, 9-25.

[7] Brook, D. (1964), On the Distinction Between the Conditional Probability and the Joint Probability Approaches in the Specification of Nearest-Neighbour Systems, Biometrika, 51, 481-483.

[8] Bulhmann, P. and Kunsch, H. (1995), The Blockwise Bootstrap for General Parameters of a Stationary Time Series, Scandinavian Journal of Statistics, 22, 35-54.

[9] Carlin, B. P. and Louis, T.A. (1996), Bayes and empirical Bayes methods for data analysis, Chapman and Hall Ltd, London.

[10] Carlstein, E. (1986), The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence, em The Annals of Statististics, 14, 1171-1179.

[11] Carlstein, E. (1988), Law of Large Numbers for the Subseries Values of a Statistic from a Stationary sequence, Statistics, 19, 295-299.

[12] Clayton, D. and Kaldor, J. (1987), Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping, Biometrics, 43, 671-681.

[13] Cressie, N. (1993), Statistics for Spatial Data, Wiley, New York.

[14] del Pino, G. (1989), The unifying role of iterative generalized least squares in statistical algorithms (C/R: p403-408) Statistical Science, 4, 394-403.

[15] A.P. Dempster, N.M. Laird, D.B. Rubin(1977), Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. B, 39 1-22

[16] Kamman, E. E. and Wand, M. P. (2003), Geoadditive models. Journal of the Royal Statistical Society, Series C (Applied Statistics), 52, 1-18.

[17] P. J. Green(1987), Penalized likelihood for general semi-parametric regression models, International Statistical Review, 55 (1987) 245-259

[18] Haining, R., Griffith, D. and Bennett, R. (1989), Maximum likelihood estimation with missing spatial data and with an application to remotely sensed data, Communications in Statistics: Theory and Methods, 1875-1894.

[19] Harville, D. A. (1974), Bayesian inference for variance components using only error contrasts, Biometrika, **61**, 383-385

[20] Heagerty, P.J. and Lele, S.R. (1998), A Composite Likelihood Approach to Binary Spatial Data, Journal of the American Statististical Association, 93, 1099-1111.

[21] Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2006), Geoadditive Survival Models, Journal of the American Statistical Association, Vol. 101, 1065-1075.

[22] Host, G. (1999), Kriging by local polynomials, Computational Statistics & Data Analysis, 29, 295-312.

[23] Hougaard, P. (2000), Analysis of Multivariate Survival Data, Springer, New York.

[24] Journel, A. G. (1983), Geostatistics, Encyclopedia of Statistical Sciences (9 vols. plus Supplement), 3, 424-431.

[25] Kott, P. S. (1998), Using the Delete-a-group Jackknife Variance Estimator in Practice, ASA Proceedings of the Section on Survey Research Methods, 763-768.

[26] Lahiri, S. N., Kaiser, M. S., Cressie, N. and Hsu, N. (1999), Prediction of Spatial Cumulative Distribution Functions Using Subsampling (C/R: p97-110), Journal of the American Statististical Association, 94, 86-97.

[27] N. M. Laird, J. H. Ware(1982), Random-effects models for longitudinal data, Biometrics, 38, 963-974

[28] Y. Lee, J. A. Nelder(1996), Hierarchical Generalized Linear Models, *J.R. Statist. Soc.* B, 58, 619-678

[29] Lele, S. (1991), Jackknifing Linear Estimating Equations: Asymptotic Theory and Applications in Stochastic Processes, Journal of the Royal Statistical Society, Series B, 53, 253-267.

[30] Li, Y. and Ryan, L. (2002), Modeling spatial survival data using semi-parametric frailty models. Biometrics 58, 287-297.

[31] Li, Y. and Lin, X. (2006), Semiparametric Normal Transformation Models for Spatially Correlated Survival Data, Journal of the American Statistical Association, 101, 591-603.

[32] X. Lin, N. E. Breslow(1995), Bias Correction in generalized linear mixed models with multiple components of dispersion, Journal of the American Statistical Association, 91, 1007-1016

[33] Lindsay, B. G. (1988), Composite Likelihood Methods, Statistical Inference from Stochastic Processes, 221-239, Prabhu, N. U. (ed.) American Mathematical Society (Providence).

[34] Q. Liu, D. A. Pierce(1993), Heterogeneity in Mantel-Haenszel-type models Biometrika. 80, 543-556.

[35] Liu, R. and Singh, K. (1992), Moving Blocks Jackknife and Bootstrap Capture Weak Dependence, Exploring the Limits of Bootstrap. LePage, Raoul (ed.) and Billard, Lynne (ed.) 225-248.

[36] Loh, J. M. and Stein, M. L. (2004), Bootstrapping a Spatial Point Process, Statistica Sinica, 14, 69-101.

[37] P. McCullagh, J. A. Nelder(1989), Generalized Linear Models, 2nd deition. (Chapman and Hall, London.)

[38] J. A. Nelder, R. W. Wedderburn(1972), Generalized linear models, J. R. Statist. Soc. A, 135, 370-384.

[39] D. Nychka and N. Saltzman (1998), Design of Air-Quality Monitoring Networks. Case Studies in Environmental Statistics, Lecture Notes in Statistics ed. Nychka, D., Cox, L. and Piegorsch, W. , Springer Verlag, New York.

[40] Opsomer, J. D., Hossjer, O., Hoessjer, O., Ruppert, D., Wand, M. P., Holst, U. and Hvssjer, O. (1999), Kriging with nonparametric variance function estimation. Biometrics, 55, 704-710

[41] Paciorek, C.J. (2007), Computational techniques for spatial logistic regression with large datasets. Computational Statistics and Data Analysis 51, 3631-3653.

[42] Politis, D. N. and Romano, J. P. (1992), A General Resampling Scheme for Triangular Arrays of α-mixing Random Variables with Application to the Problem of Spectral Density Estimation, The Annals of Statististics, 20, 1985-2007.

[43] Prentice, R. L. and Cai, J. (1992), Covariance and Survivor Function Estimation Using Censored Multivariate Failure Time Data, Biometrika, 79, 495-512.

[44] Sherman, M. (1996), Variance estimation for statistics computed from spatial lattice data, Journal of the Royal statistical Society,Series B, 58, 509-523.

[45] Stein, M.L. (1999), Interpolation of Spatial Data: Some Theory of Kriging, Springer, New York.

[46] R. Schall(1991), Estimation in generalized linear models with random effects, Biometrika. 78, 719-727.

[47] P. J. Solomon, D. R. Cox(1992), Nonlinear component of variance models, Biometrika. 79, 1-11.

[48] Waller, L.A., Carlin, B.P., Xia, H. and Gelfand, A. E. (1997), Hierarchical Spatio-temporal Mapping of Disease Rates, Journal of the American Statisstistical Association, 92, 607-617.

[49] diggle P.J. Diggle, R.A. Moyeed and J.A. Tawn(1992), Model-based Geostatistics. Applied Statistics, 47, 299–350.

[50] J.D. Opsomer, D. Ruppert, M.P. Wand, U. Holst and O. Hossjer(1999), Kriging with nonparametric variance function estimation. Biometrics, 55, 704–710.

[51] zhanghao H. Zhang (2002), On Estimation and Prediction for Spatial Generalized Linear Mixed Models Biometrics, 58, 129–136.

[52] Zhu, J. and Morgan, G. D. (2004), Comparison of Spatial Variables Over Subregions Using a Block Bootstrap, Journal of Agricultural, Biological, and Environmental Statistics, 9, 91-104