Published in final edited form as:

Biometrics. 2008 June ; 64(2): 490–500. doi:10.1111/j.1541-0420.2007.00892.x.

Aberrant Crypt Foci and Semiparametric Modeling of Correlated Binary Data

Tatiana V. Apanasovich¹, David Ruppert¹, Joanne R. Lupton², Natasa Popovic², Nancy D. Turner², Robert S. Chapkin², and Raymond J. Carroll^{3,*}

¹*School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853, U.S.A.*

²*Department of Nutrition and Food Science, Texas A&M University, College Station, Texas 77843-2253, U.S.A.*

³*Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.*

Summary

Motivated by the spatial modeling of aberrant crypt foci (ACF) in colon carcinogenesis, we consider binary data with probabilities modeled as the sum of a nonparametric mean plus a latent Gaussian spatial process that accounts for short-range dependencies. The mean is modeled in a general way using regression splines. The mean function can be viewed as a fixed effect and is estimated with a penalty for regularization. With the latent process viewed as another random effect, the model becomes a generalized linear mixed model. In our motivating data set and other applications, the sample size is too large to easily accommodate maximum likelihood or restricted maximum likelihood estimation (REML), so pairwise likelihood, a special case of composite likelihood, is used instead. We develop an asymptotic theory for models that are sufficiently general to be used in a wide variety of applications, including, but not limited to, the problem that motivated this work. The splines have penalty parameters that must converge to zero asymptotically: we derive theory for this along with a data-driven method for selecting the penalty parameter, a method that is shown in simulations to improve greatly upon standard devices, such as likelihood crossvalidation. Finally, we apply the methods to the data from our experiment ACF. We discover an unexpected location for peak formation of ACF.

Keywords

Aberrant crypt foci; Colon carcinogenesis; Composite likelihood; Generalized linear mixed models; Longitudinal data; Pairwise likelihood; Partially linear model; Semiparametric regression; Single index models; Spatial statistics

1. Introduction

This article develops methods for semiparametric regression with possibly nonstationary correlated binary data. We are motivated by an example in colon carcinogenesis where the number of observations per subject is too large to use exact likelihood or restricted likelihood (Sherman, Apanasovich, and Carroll, 2004), so we use composite likelihood. This article extends Heagerty and Lele's (1998) work on composite likelihood to nonparametric specifications of the mean. We develop a new smoothing parameter selector, because there

*email: carroll@stat.tamu.edu.

appears to be no previous work on smoothing parameter selection for composite likelihood. In the Web Appendix, the asymptotic theory in Heagerty and Lele (1998) is extended to smoothing parameter estimation. The asymptotic theory we provide is new and should be of general interest. We use increasing domain asymptotics, in which the minimum distance between the sampling points is bounded away from zero and thus the spatial domain of observation is unbounded. This type of asymptotics has been successfully considered in many applications (Cressie, 1993; Heagerty and Lele, 1998; Fuentes, 2002).

Our model has fixed effects for the mean, representing systematic effects, and random effects giving a rather general form of dependence conditional upon the mean. The fixed-effects structure includes partially linear models, single index models, and additive models and uses fixed-knot splines with penalties (Ruppert, Wand, and Carroll, 2003). This structure is new and should be of interest elsewhere.

The problem is motivated by a study in colon carcinogenesis concerning aberrant crypt foci (ACF), colonic crypts that have been changed morphologically by a carcinogen and/or radiation. These ACF are precursors to colon cancer (Bird, 1995; Bird and Good, 2000), and hence of major biological significance. The mean function models the intensity of ACF formation as a function of location in the crypt, whereas the spatial dependency can address the issue of whether and how cells signal to each other. Although suggested by this application, our methodology is sufficiently general and it will apply to a wide variety of problems.

The ACF study has a large number (about 800) of repeated measures per subject and a small number of subjects (about 14). Many papers on longitudinal data address data sets of the opposite type where there are many subjects but relatively few repeated measures per subject. To analyze data such as ours, we provide a theoretical framework for a large number of outcomes on a single subject. This methodology can be applied separately to each subject, as we do in the ACF study.

Our binary mixed model is a type of generalized linear mixed model (GLMM) called a spatial GLMM. Spatial GLMMs (Diggle, Tawn, and Moeed, 1998) are flexible models for a variety of applications where we have observations of spatially dependent and non-Gaussian random variables. As in a standard GLMM, given the random effects, which they model by a Gaussian random field, the observations are conditionally independent and follow a generalized linear model. Both Bayesian and frequentist methods have been developed for inference and forecasting in spatial GLMMs. Diggle et al. (1998) and Christensen and Waagepetersen (2002) used a Bayesian Markov chain Monte Carlo framework. The computational burden for those methods is high because the number of correlated random effects to be simulated is equal to the number of observations. Maximum likelihood estimation (MLE) in spatial GLMMs generally involve numerical integration of a high-dimensional integral. The integral may be computed by Monte Carlo integration. McCulloch (1997) reviews several Monte Carlo techniques for MLE within GLMMs. Booth and Hobert (1999) described a Monte Carlo expectation-maximization (EM) algorithm for the spatial probit model, whereas Zhang (2002) used a maximum likelihood approach together with Monte Carlo EM algorithm to estimate parameters of a general spatial GLMM. Alternatively the integral may be computed by Laplace's method, which uses a Gaussian approximation of the integrand. In GLMM inference, this method has been used by Breslow and Clayton (1993), who call it penalized quasilielihood (PQL) and has been implemented in SAS's GLIMMIX (SAS 9.1 from SAS Institute, Inc.). However, PQL is known to be biased (Breslow and Lin, 1995; Lin and Breslow, 1996; McCulloch, 1997; Neuhaus and Segal, 1997). In two examples where the exact MLE can be computed, Neuhaus and Segal find that the exact MLE and the PQL estimator can differ by more than one standard error of the PQL and even Breslow and Lin's (1995) bias correction does not lead to accurate approximation of the exact MLE. Neuhaus and Segal's conclusion is

that “the biases exhibited by the Taylor approximation and PQL approaches indicate that these approaches require further development and modification before they should be used in practice.”

Maximum likelihood inference, whether by EM or PQL, requires high-dimensional matrices to be inverted repeatedly. Because matrix inversion is an order $O(n^3)$ operation, computational effort increases rapidly with n . Here n is the number of random effects, which in our context is the number of observations. This is the motivation for pairwise likelihood (Heagerty and Lele, 1998; Curriero and Lele, 1999; Nott and Rydén, 1999; Kuk and Nott, 2000; Renard, Molenberghs, and Geys, 2004). Even for ordinary kriging where the spatial process is directly observable and there is no need to integrate latent variables, Curriero and Lele (1999) state that ML or restricted maximum likelihood estimation (REML) “involve inversion of large matrices which can be computationally prohibitive” and use alternative methods. In a simulation study of multilevel probit models, Renard et al. (2004) find that the time to compute the MLE increases rapidly with the number of the random effects per cluster and the time to compute the MLE is much greater than for pairwise likelihood when this number exceeds 3; see their Table 3. Obviously, the crossover point where the pairwise likelihood estimator is less intensive than the MLE is problem specific, but it is noteworthy that the crossover point is very low in this study.

A possible solution to this computational problem is to reduce the number of random effects, for example, with low-rank splines as in Ruppert et al. (2003, Section 13.8). However, this technique is applicable only when the observations are conditionally independent given the random effects determining the mean. In our model, the observations are conditionally dependent, given the mean. The dependencies in the ACF example are short-range and cannot be modeled with only a small number of random effects.

In summary, both Bayesian and maximum likelihood approaches are extremely difficult to implement for large sets of correlated binary data. To gain computational efficiency, we consider composite likelihood (Lindsay, 1988), which is the product of likelihoods for subsets of data, and estimate parameters by maximizing this product. This approach has been used in many problems for correlated binary response data, for example, in spatial models (Heagerty and Lele, 1998). Note that the integrals require the use of low-dimensional numerical integration.

We know of no theoretical results comparing the efficiency of composite to full maximum likelihood, but in an example of Kuk and Nott (2000) pairwise likelihood for clustered data has nearly full asymptotic efficiency. In another simulation study, Renard et al. (2004) found that the efficiency losses of pairwise likelihood relative to the MLE averaged about 10% and ranged from 5% to 18% across three parameters (two fixed effects and a variance component) in four simulation studies. Cox and Reid (2004) studied asymptotic relative efficiency for estimators based on pairwise likelihood in the case of dichotomized normals. They reported a relatively small loss of efficiency. Varin, Høst, and Skare (2005) compared the maximum pairwise likelihood estimator with the approximate MLE based on Laplace's approximation for integration and reported similar values for mean squared errors. Zhao and Joe (2005) showed a relatively small loss of efficiency for models they considered.

Also, in a related context of autologistic models, Sherman et al. (2004) found the maximum pseudolikelihood and generalized pseudolikelihood estimators comparable to the MLE approximated by MCMC.

In Section 2, we describe our estimation methodology and give a brief discussion of asymptotic theory and estimation of standard errors. A very general asymptotic theory, including technical conditions and proofs, is given in Web Appendix A. Section 3 presents our smoothing

parameter selectors. A small simulation study is presented in Section 4, followed by the analysis of the ACF experiment in Section 5. Discussion and extensions are given in Section 6.

2. Models

2.1 Binary Mixed Model and General Fixed Effects Structure

Our models are semiparametric GLMMs, with a binary response variable D_i and predictors (Z_i, X_i) , all measured within a spatial domain. In our case, we will model the effects of Z_i parametrically and the results of X_i semiparametrically, and as in our example we focus on the case that X is scalar.

To complete the model specification, let ϵ_i be independent $\text{Normal}(0, 1)$ random variables. Let λ_i denote random effects responsible for possible spatial dependence. For a parameter ρ and a correlation matrix $\Omega(\rho)$, the $\{\lambda_i\}$ are assumed to be normally distributed with mean 0 and covariance matrix $\sigma_\lambda^2 \Omega(\rho)$ and independent of the ϵ_i . Let μ_i be the mean function incorporating systematic effects. Then the model is defined as $D_i = I(\mu_i + \lambda_i + \epsilon_i > 0)$, so that

$$\text{pr}(D_i=1|\lambda_i, \mu_i) = \Phi(\mu_i + \lambda_i), \quad (1)$$

where $\Phi(\bullet)$ is the standard normal cumulative distribution function (CDF).

In the remainder of this section, we describe the structure of the μ_i (Section 2.3), the fixed-knot splines used to model μ_i (Section 2.4), and the composite likelihood method (Section 2.5) to estimate parameters and functions.

2.2 Random Effects Structure

We will consider one approach to modeling the covariance function of the spatial process $\{\lambda_i\}$, although our methods and theoretical techniques apply more generally.

Let $d(i, j)$ be the Euclidian distance between sites i and j . The simplest type of correlation structure is stationary, for example, the Matérn family (Handcock and Stein, 1993; Handcock and Wallis, 1993; Stein, 1999). Thus, the (i, j) element of $\Omega(\rho)$ is $\Omega_{ij}(\rho) = M\{d(i, j), \rho\}$, where $M(\bullet)$ is a known function with unknown parameter ρ .

There are several choices available for the correlation function, see Stein (1999) for an extensive overview. In this article we work with a parametric family of autocorrelation functions, the Matérn family (Handcock and Stein, 1993; Handcock and Wallis, 1999). Let $\mathcal{K}_\nu(\bullet)$ be the modified Bessel function of order ν . Then the Matérn correlation at distance d can be written as either $K_\nu(d/\rho)(d/\rho)^\nu 2^{1-\nu}/\Gamma(\nu)$ or its more stable version $2^{1-\nu}(2d\nu^{1/2}/\rho)^\nu K_\nu(2d\nu^{1/2}/\rho)/\Gamma(\nu)$. For example, when $\nu = 5/2$, the correlation is $\exp(-d/\rho) \{1 + d/\rho + (1/3)(d/\rho)^2\}$.

2.3 General Fixed Effects Structure

We focus here on the partially linear model for the fixed effects, namely, that for some unknown parameter ζ_0 , and an unknown function $\Lambda(\cdot)$, the fixed effects structure is

$$\mu_i = Z_i^T \zeta_0 + \Lambda(X_i). \quad (2)$$

Our results will be phrased in terms of the partially linear model (2), but they actually apply to far more general situations, most notably the single-index model $\mathcal{G}(Z_i^T \zeta_0) + \Lambda(X_i)$, where $\mathcal{G}$

$(\cdot)$ is an unknown function modeled by splines with its own smoothing parameter, and for identifiability $|\zeta_0| = 1$. However, the partially linear model is important in its own right, applicable to the ACF example, and the results for it are the most transparent.

2.4 Regression Splines and Penalization

In our work, we model an unknown function $\Lambda(\bullet)$ as a fixed-knot regression spline, which has the representation $\Lambda(x) = \tilde{B}^T(x)\eta_0$, where $\tilde{B}^T(x) = \{B_1(x), \dots, B_k(x)\}$. For example, $\tilde{B}(x)$ might be the B-spline basis functions or the truncated power series basis (Ruppert et al., 2003) of

order q with K knots $x_1, \dots, x_K$ given as $\tilde{B}^T(x) = (1, x, \dots, x^q, (x - x_1)^q_+, \dots, (x - x_K)^q_+)$, where the subscripted plus sign is the positive part function.

Regression splines with a fixed number of knots have become an increasingly popular means of semiparametric inference (see Ruppert et al. [2003] and references therein). Generally, not many knots are required to capture most fixed effects structures (Ruppert, 2002), and for binary data in particular, capturing very complex structure that cannot be modeled by a low-order basis representation is unlikely to be practical. Of course some sort of smoothing is required. This is generally done either by knot selection devices to greatly lower the dimensionality, or by penalization to achieve smoothness. In this article, we use the latter device, see Section 3 for details. Our general model is a generalization of that of Yu and Ruppert (2002) who also use fixed-knot splines and found that this technique is a numerically stable approach to single-index model.

2.5 Penalized Composite Likelihood Estimation

In composite likelihood estimation, the n observations are split into many subsets of a computationally convenient size k . We use $k = 2$ in our numerical work and, for ease of notation, will keep k fixed at 2 in the main body of the article. The extension to general k is discussed in the Appendix and proofs in the Web Appendix are for general k . For each subset of observations, the log-likelihood function is computed: the idea of using a subset of observations is to make this calculation easy. Then the various log likelihoods are summed across all the selected subsets, and maximized to get the composite likelihood estimator. Inference for composite likelihood is somewhat tricky because the subsets are not independent sets of data.

In this section, we first set up the formal calculation of a composite likelihood function. We then introduce the penalization that is necessary when using regression splines. Section 3 introduces a method for data-based estimation of the smoothing parameters.

2.5.1 Composite likelihood calculation—Let $\Phi_2(\mu_1, \mu_2; \Delta)$ be the bivariate standard normal probability of being below μ_1, μ_2 when the correlation matrix is Δ . If

$\mu_i^* = \mu_i / (1 + \sigma_\lambda^2)^{1/2}$, then marginal probabilities can be expressed as

$$\begin{aligned} & \text{pr}(D_{i_1} = 1, D_{i_2} = 1 | \mu_{i_1}^*, \mu_{i_2}^*, \sigma_\lambda^2, \rho) \\ &= \Phi_2 \left[\mu_{i_1}^*, \mu_{i_2}^*; \frac{\sigma_\lambda^2}{(1 + \sigma_\lambda^2)} \left\{ \Omega_{i_j}(\rho) \right\}_{i,j \in \{i_1, i_2\}} \right]. \end{aligned} \quad (3)$$

The composite likelihood is defined as a sum of log likelihoods based on the marginal distribution of the responses, see below for more details. At least in principle, both the correlation function and the variance of the random effects can be estimated from the marginal component likelihoods.

Because μ_i and μ_i^* are multiples of one another, for notational simplicity, we will use only one of them. In what follows, we will use only the function μ_i^* . Also, we can now drop the subscript “*” and henceforth μ_i^* will be denoted by μ_i .

2.5.2 Penalization—Recall from (2) that the fixed effects depend on ζ_0 and the regression spline is written as $\Lambda(x) \tilde{B}^T(x) \eta_0$. Let β be the collection of the parameters (ζ, η) , with true value β_0 , and write $\mu_i = \mu_i(\beta)$. We need to penalize this function to account for the nonparametric regression, which we do with a smoothing parameter k and a penalty matrix $\mathcal{G}$. The penalty matrix depends on the basis functions used. For example (Ruppert et al., 2003), for the truncated polynomial series basis of order q and K knots defined in Section 2.4, $\mathcal{G} = \text{diag}(0I_q, I_K)$, where I_q is the identity matrix of size q .

In order to estimate the correlation function, we must use at least pairs of observations. Let Θ denote the unknown parameters (σ_λ^2, ρ) and let θ_0 denote its true value. Define Θ as the collection of all parameters to be estimated, and let θ_0 be its true value.

If we use sets of size 2, then we need to keep track of the values of the sets and their probabilities. To do this, we define

$$\begin{aligned} Z_{i_1, i_2}^{(j_1, j_2)} &= I(D_{i_1} = j_1, D_{i_2} = j_2); \\ \pi_{i_1, i_2}^{(j_1, j_2)} &= E(Z_{i_1, i_2}^{(j_1, j_2)} | \mu_{i_1}^*, \mu_{i_2}^*, \sigma_\lambda^2, \rho). \end{aligned} \quad (4)$$

Using the marginal probabilities (3), we can write the likelihood at distinct locations (i_1, i_2) as

$$\mathcal{L}_{i_1, i_2}(\Theta) = \sum_{j_1, j_2=0}^1 Z_{i_1, i_2}^{(j_1, j_2)} \log \{ \pi_{i_1, i_2}^{(j_1, j_2)}(\beta, \theta) \}.$$

It is useful to note that $\mathcal{L}_{i_1, i_2}(\Theta)$ is an actual likelihood function, albeit for a reduced set of data.

2.5.3 A penalized composite likelihood—We propose to maximize a weighted penalized composite likelihood. Let w_{i_1, i_2} be weights, which take nonzero values only if the maximum distance between two locations from a set of locations (i_1, i_2) is less than a specified value apart and 0 otherwise, a choice that is useful to cut down on the size of the summations. It is assumed that w_{i_1, i_2} does not depend upon (D_{i_1}, D_{i_2}) or any parameters, because otherwise, if the weights depend upon parameters, then our methods will be inconsistent, a well-known fact in generalized least squares that is made explicit by Müller (1999). Here we just used indicators, whereas Guan (2006) proposed to use a data-driven choice of the maximum distance used to select sets and the weight function. However, the choice should be driven by efficiency considerations, and goodness of fit discussed by Guan (2006) is not directly applicable in our case.

For fixed smoothing parameter k , the composite likelihood estimator is denoted by $\hat{\Theta}(k)$ and is obtained by maximizing over Θ :

$$\tilde{\mathcal{L}}_{w'}(\Theta, k) = W^{-1} \sum_{i_1, i_2=1} w_{i_1, i_2} \mathcal{L}_{i_1, i_2}(\Theta) - (1/2) k \eta^T \mathcal{G} \eta, \quad (5)$$

where $\mathcal{W} = \sum_{i_1, i_2} w_{i_1, i_2}$.

2.6 Asymptotic Results and Standard Error Estimation

Web Appendix A gives technical conditions and proofs for models far more general than the partially linear model (2), including single-index models and multiple smoothing parameters. The theory is notable because it describes precisely how the smoothing parameter k must decrease to zero with the sample size. Here we provide a brief description of the main results.

Web Appendix A shows that as long as the smoothing parameter k decreases to zero at a rate faster than $\mathcal{W}^{-1/2}$, then the penalized composite likelihood function $\tilde{\mathcal{L}}_{\mathcal{W}}(\Theta, \kappa)$ given in (5) can be treated as if it were an ordinary composite likelihood function. Thus, if the composite score function is

$$\Psi_{\mathcal{W}}(\Theta, \kappa) = \partial \tilde{\mathcal{L}}_{\mathcal{W}}(\Theta, \kappa) / \partial \Theta,$$

then $\widehat{\Theta}(\kappa)$ is asymptotically normally distributed with mean Θ_0 , the true value, and a covariance matrix given by Godambe's (1960) information matrix, also known as the sandwich formula:

$$\begin{aligned} \text{cov}\{\widehat{\Theta}(\kappa)\} &\approx \left[E \left\{ \frac{\partial \Psi_{\mathcal{W}}(\Theta, \kappa)}{\partial \Theta^T} \right\} \right]^{-1} \text{cov}\{\Psi_{\mathcal{W}}(\Theta, \kappa)\} \\ &\quad \times \left[E \left\{ \frac{\partial \Psi_{\mathcal{W}}(\Theta, \kappa)}{\partial \Theta^T} \right\} \right]^{-1}. \end{aligned} \quad (6)$$

Estimating the covariance matrix given in (6) has a standard and a nonstandard part. The “bread,” that is, the outer matrices, can be estimated simply by plug-in as

$$\left[\partial \Psi_{\mathcal{W}}\{\widehat{\Theta}(\kappa), \kappa\} / \partial \Theta^T \right]^{-1}.$$

The “meat” of the sandwich, $\text{cov}\{\Psi_{\mathcal{W}}(\Theta, \kappa)\}$ in (6), is more complex because of the spatial dependence. However, a consistent estimate of it can be well approximated by a Monte Carlo calculation. Specifically, if Θ_0 is the true parameter, then we have a model for the data generation process. Replacing the unknown Θ_0 by the estimated $\widehat{\Theta}(\kappa)$, we can generate simulated realizations of the data generation process, and hence realizations of $\Psi_{\mathcal{W}}(\Theta, \kappa)$. We then approximate $\text{cov}\{\Psi_{\mathcal{W}}(\Theta, \kappa)\}$ by the sample covariance matrix of the generated realizations of $\Psi_{\mathcal{W}}(\Theta, \kappa)$.

3. Smoothing Parameter Estimation

It is well known that with dependent data, standard approaches to smoothing parameter selection, such as crossvalidation (CV), lead to undersmoothing, see, for example, Opsomer, Wang, and Yang (2001) for discussion and extensive references. The usual device for numerical response data is either to attempt to select the smoothing parameter in such a way as to minimize asymptotic average mean squared error, or to treat the parameters in the mean as random rather than fixed effects so that the smoothing parameter is a ratio of variance components, and to maximize a resulting mixed model likelihood. Given that in our case the correlated responses are binary and that a likelihood estimate with or without an extra variance component is extremely difficult to compute, it seems, at the very least, useful to explore different approaches.

Let Θ denote all the parameters, let Θ_0 be their true value and let $CL(\Theta)$ be a composite log likelihood, as in the first term in (5). Our approach is to adapt the idea of Kullback-Leibler (KL) distance to estimate the smoothing parameters. Define

$$KL(\Theta, \Theta_0) = E_{\Theta_0} \{CL(\Theta_0) - CL(\Theta)\},$$

where the subscript Θ_0 means that the expectation is taken at the distribution induced by Θ_0 . It is technically convenient to work with a symmetrized version of this distance, namely,

$$SKL(\Theta, \Theta_0) = KL(\Theta, \Theta_0) + KL(\Theta_0, \Theta).$$

It is easy to see that $SKL(\Theta, \Theta_0)$ is always nonnegative and equals zero when $\Theta = \Theta_0$. In our problem we estimate parameters by maximizing a composite log-likelihood function that depends on a vector of smoothing parameters k . If we plug an estimated $\widehat{\Theta}(k)$ into this expression we get a random variable, whose expectation we then take to find

$$E_{\Theta_0} [SKL\{\widehat{\Theta}(k), \Theta_0\}]. \quad (7)$$

Our goal is to choose the smoothing parameter so as to minimize (7). More precisely, we will find an asymptotically equivalent version of $\widehat{\Theta}(k)$, substitute it into (7) and minimize.

The key to the analysis is that in our problem formulation, $SKL(\Theta, \Theta_0)$ can be computed analytically. We cannot of course compute (7) analytically, but we do show how to compute an asymptotically equivalent version of it, and this allows us to estimate the smoothing parameters. With a slight abuse of notation, for composite likelihood, we have that

$$\begin{aligned} SKL\{\widehat{\Theta}(k)\} &= W^{-1} \sum_{i_1, i_2=1}^n \sum_{j_1, j_2=0}^1 w_{i_1, i_2} \\ &\times \left[\pi_{i_1, i_2}^{(j_1, j_2)}(\Theta_0) - \pi_{i_1, i_2}^{(j_1, j_2)}\{\widehat{\Theta}(k)\} \right] \\ &\times \left(\log \left\{ \pi_{i_1, i_2}^{(j_1, j_2)}(\Theta_0) \right\} - \log \left[\pi_{i_1, i_2}^{(j_1, j_2)}\{\widehat{\Theta}(k)\} \right] \right), \end{aligned}$$

where $\pi_{i_1, i_2}^{(j_1, j_2)}(\cdot)$ are defined in (4). In the Appendix, we sketch an argument indicating that with this definition of $SKL(\cdot)$, the minimizer of (7) can be derived as follows. Let

$\Sigma_{W,0} = E\{\partial \Psi_W(\Theta, k) \partial \Theta^T\}$ and let $\Sigma_{W,0} + \Sigma_{W,c}$ be the interior covariance matrix in the sandwich formula (6). We discussed in Section 2.6 how to estimate these components.

Then, as described in the Appendix, there is a matrix $\mathcal{H}(\tilde{\kappa}, \Theta_0, \Sigma_{W,0}, \Sigma_{W,c})$ and a vector $\mathcal{J}(\tilde{\kappa}, \Theta_0, \Sigma_{W,0}, \Sigma_{W,c})$, both defined in the Appendix, such that

$$\kappa = W^{-1} \mathcal{H}^{-1}(\kappa, \Theta, \Sigma_{W,0}, \Sigma_{W,c}) \mathcal{J}(\kappa, \Theta_0, \Sigma_{W,0}, \Sigma_{W,c}).$$

Therefore we estimate κ iteratively such that

$$\kappa^{(l)} = \mathcal{W}^{-1} \mathcal{H}^{-1} \left\{ \kappa^{(l-1)}, \widehat{\Theta}(\kappa^{(l-1)}), \widehat{\Sigma}_{W,0}(\kappa^{(l-1)}), \widehat{\Sigma}_{W,c}(\kappa^{(l-1)}) \right\} \\ \times \mathcal{J} \left\{ \kappa^{(l-1)}, \widehat{\Theta}(\kappa^{(l-1)}), \widehat{\Sigma}_{W,0}(\kappa^{(l-1)}), \widehat{\Sigma}_{W,c}(\kappa^{(l-1)}) \right\}.$$

We conjecture that the KL criterion, rather than its symmetrized version SKL, will yield an asymptotically equivalent smoothing parameter selector.

4. Simulation Study

We performed a small simulation study to understand in part the properties of our methods, patterning the study after the analysis of the ACF data that will be presented in Section 5. There the covariate Z modeled parametrically is the vertical location of a slice and takes on eight equally spaced values between 0 and 1. The covariate, X_i , is the horizontal distance from the origin, normalized to the unit interval. We assume that the latent spatial process is stationary with a Matérn correlation function of index 5/2 (Stein, 1999), so that

$\Omega_{ij}(\rho) = \tilde{\Omega}(d_{ij}, \rho) = \exp(-d_{ij}/\rho) \left\{ 1 + d_{ij}/\rho + (d_{ij}/\rho)^2 \right\} / 3$, where d_{ij} is the Euclidian distance between sites i and j . Mimicking the ACF data, we set $n = 800$, reflecting a 100×8 grid. The function chosen was $\Lambda(X) = 0.5 [\sin\{2\pi(X_i - 0.5)\} - 1]$. For each correlation structure, performed 1000 simulations. In this example we used 14-knot cubic splines with a truncated polynomial series basis. In keeping with the ACF example, where there were seven animals in a treatment group, we replicated the processes seven times. This requires only a few minor changes in notation.

There is one smoothing parameter here, κ . We used two sets of correlation parameters, chosen so that $\sigma_\lambda^2 = 1$, $\tilde{\Omega}(1, \rho) = 0.5$ and $\sigma_\lambda^2 = 4$, $\tilde{\Omega}(1, \rho) = 0.6$, reflecting the moderate dependence typical of ACF data and stronger dependence, respectively.

We studied the performance of the algorithms and compared two methods of smoothing parameter selection: (a) the proposed methods based on SKL criterion (Section 3); and (b) the traditional one based on CV (Ruppert et al., 2003). The integrated mean squared errors and squared bias of the probability function for both methods are given in Table 1. Effectively, we see that for estimating the function, both methods are roughly unbiased, but our method has much smaller mean squared errors, with mean squared error efficiencies being roughly 200%. The reason for this is that CV undersmooths the function, leading to increased variability, which shows up in individual data sets.

Of course, our algorithm allows us to estimate the correlation along with the mean function (see Table 2). Included in Table 2 are the mean estimates of $\psi = \sigma_\lambda^2 / (1 + \sigma_\lambda^2)$ and $\Omega(\rho)$, along with the 2.5th and 97.5th percentiles over the simulated data sets. The estimators appear to be working satisfactorily.

We also addressed the concern whether the asymptotic theory works with the size of the data set we have. We note that asymptotic formulas for variances of the nonparametric function estimates agree with Monte Carlo variances in our simulations. The overall coverage probability averaged across all X values was 95.2% and 94.4% for the case where

$\sigma_\lambda^2 = 1$, $\tilde{\Omega}(1, \rho) = 0.5$ and $\sigma_\lambda^2 = 4$, $\tilde{\Omega}(1, \rho) = 0.6$, respectively, for the selected level of 95%. Let $\hat{\mu}_{0.5}^{(i)}$ be an estimated mean at $X = 0.5$ obtained based on the results from i th simulation run, $i =$

1, ..., 1000. Figure 1 shows Q-Q plots for $\left\{ \hat{\mu}_{0.5}^{(i)} \right\}_{i=1}^{1000}$ for both simulation scenarios. The p-

values for Shapiro-Wilks test for normality were 0.7403 and 0.4201 for both simulations scenarios, respectively. One can clearly see that estimators have a nearly normal distribution.

Separately, we also studied the case of very strong correlation ($\sigma_\lambda^2=1, \tilde{\Omega}(1,\rho)=0.9$), which comes closer to an infill asymptotic framework. There was relatively small bias in both the estimated mean function and in the product $\{\sigma_\lambda^2/(1+\sigma_\lambda^2)\} \tilde{\Omega}(1,\rho)$, although there was, as expected, somewhat greater bias in the component parts of this product.

5. Analysis of the ACF Experiment

In the ACF experiments, animals are exposed to a carcinogen, with half of them also exposed to radiation. They are then sacrificed, and images of the colon are obtained by various staining devices. A typical image is given in Figure 2: a color version of this is available at Web Figure A. It is not feasible in practice to measure the exact locations of ACF, so we formed a rectangular grid of locations and recorded (by hand) the existence of an ACF within each location (see Figure 3). Thus the data available to us are the grid of locations along with the indicator of an ACF. Further details on data collection can be found in Apanasovich et al. (2003).

In Figure 2 we see two major types of structures. The small white dots are normal colonic crypts, whose function is to produce cells that line the colon. The larger dark and distended regions are ACF. The response D at a given location is whether there is an ACF or not.

The data are naturally two-dimensional, although in our application, the sections are much longer (X) than wide (Z) by a factor greater than 10. Because the colon strips are not very wide, we modeled the width parametrically and the length nonparametrically, via the partially linear model (2).

A score test for longitudinal/spatial dependence developed by Apanasovich et al. (2003) indicated the presence of spatial dependencies in these data. This finding is interesting because it suggests that damage to the colon is localized regionally, and thus that there may be areas in which greater levels of damage in response to an insult could lead to focused areas of inflammatory responses, or an alteration in the release of signaling molecules that could then affect the regulation of homeostatic mechanisms in colonocytes in adjacent crypts. This localization may help explain why tumors develop from a particular ACF, but not from all ACF formed in response to a carcinogen insult.

Having shown dependence, two questions immediately arise: the extent of the dependence and the nature of the rate of ACF formation depending on the location within the colon. Our hope in this experiment is to identify regions of high ACF formation, and then to see whether regions of high ACF formation are also the regions of high tumor formation predicted by biologists.

In this study we used 14 rats, 7 irradiated and 7 nonirradiated animals that were sacrificed at 6 weeks. Here X_i is the horizontal distance from the distal part of the colon normalized to the unit interval, and Z is the more compact vertical distance. The correlation is modeled with respect to the nominal distance between grid cell, such that the distance between the closest horizontal or vertical neighbors is 1.

We fit a separate partially linear model in each rat, but with the correlation parameters and the smoothing parameter fixed between rats in a treatment group. We formed the composite likelihood using all pairs of sites less than three units apart. For the fixed effects, we used 10-knot cubic splines with the truncated polynomial basis, and the rats were allowed their own smoothing parameters.

In practice, spatial data are observed at a finite number of points and it is not clear which asymptotic framework to use, infill or increasing domain. Physically, we think increasing domain asymptotics makes sense in our example, and of course they have been used in many other contexts to which our methods would apply if splines were used. As the horizontal location we call X increases, we are actually going further out from the proximal (front) part of the colon to the distal (end) of the colon. Thus, neighbors at the proximal part of the colon are 50 times closer in physical distance than they are from locations at the distal region. We also believe that once one takes into account background rate for given rats, there is no reason to think that ACF formation at the front of the colon is highly predictive of ACF formation at the end of the colon. Under increasing domain asymptotics one would expect the correlation to decrease quickly. To test this we ran score tests similar to the one described in Apanasovich et al. (2003) and found that the correlation is not statistically significant at the distance greater than 3 units apart. Moreover, simulations suggested that under increasing domain asymptotics, the distributions of our estimators approximate the finite-sample distributions of those estimators well. Therefore we have reasonably good evidence to support the use of increasing domain asymptotic framework for our data set, and of course more generally this type of asymptotics is often used.

5.1 Results and Model Checks

As described above, Apanasovich et al. (2003) note that a score test indicates spatial correlation. To quantify this, we fit the stationary Matérn correlation model with index $5/2$ (see Section 2.2), obtaining $\widehat{\sigma}_\lambda^2=0.48$, $\widetilde{\Omega}(1,\widehat{\rho})=0.28$ for irradiated animals and $\widehat{\sigma}_\lambda^2=2.62$, $\widetilde{\Omega}(1,\widehat{\rho})=0.20$ for nonirradiated animals. Additional tests based upon our asymptotic theory showed that the estimates for ψ and $\widetilde{\Omega}(1,\widehat{\rho})$ for irradiated and nonirradiated animals are significantly different from 0.

The probability of ACF formation for each group was calculated as an average of individual probabilities. Figure 4 shows the estimated probability of ACF formation as a function of normalized distance from the distal part of the colon for each of the two groups. The nonsolid lines correspond to confidence limits that we constructed the following way. First, we built pointwise confidence intervals for each rat, then we averaged confidence intervals for rats belonging to the same group. This figure suggests three important possible conclusions. First, the shapes are complex, and certainly neither constant nor linear. Second, overall the irradiated and nonirradiated groups are different in their ACF formation, with the former having higher ACF formation overall.

We did two model checks. First, our model fits indicate that the irradiated animals have more ACF. We used a simple t -test to compare the total number of ACFs between two groups, with p -value approximately 0, thus confirming this finding.

Second, and of more potential importance, is the finding that ACF formation is not uniform across the colon, and indeed seems greatest near the middle part of the colon for both the nonirradiated and irradiated rats. This was initially surprising to us, because we expected that in rats who were allowed to live longer, most tumors would be found in the distal region, rather than roughly in the middle.

We used t -tests to show that the data support the claim that there is a higher response in the middle as follows. For each rat we computed the mean of ACF indicators for three regions in the colon: low, middle, and upper, which correspond to the normalized horizontal distance from the proximal part of the colon to be equal to $[0.0, 0.3]$, $(0.3, 0.8]$, and $(0.8, 1.0]$, respectively. The rats are independent by design, therefore we combined the results across the rats. Then we ran paired t -tests. For the rats in the irradiated group, the middle region versus

the low region had a t -statistic of 6.13, whereas the middle region versus the upper region had a t -statistic of 1.18. For the rats from the nonirradiated group, the middle region versus the low region had a t -statistic of 6.06, whereas the middle region versus the upper region had a t -statistic of 3.99. Therefore, there is empirical evidence that the probability of ACF formation is higher in the middle.

5.2 Selection of Pairs for Composite Likelihood

In our example, in computing composite likelihood we have selected pairs that were less than L units apart with $L = 3$. This choice was made on the basis of computational feasibility, as well as the fact that even for a relatively high correlation of 0.50 between nearest neighbors, the correlation between sites 3 or more units apart for the chosen correlation structure is less than 0.02.

Interestingly, there is evidence in the literature that using composite likelihood pairs that are much less correlated than other pairs may have a deleterious effect on the performance of composite likelihood. Varin and Vidoni (2007) in their Figure 1 describe an AR(1) process in time series with an autocorrelation of 0.95. They find that the efficiency of composite likelihood reaches its maximum when pairs up to 6 lags apart: these pairs have a correlation ≥ 0.73 . However, when lags up to 20 were considered, that is, a correlation of 0.36 at lag 20, the mean squared error efficiency for estimating the correlation decreased by approximately 25%.

Varin and Vidoni's calculations indicate that essentially un-correlated pairs should not be used when trying to estimate correlation parameters. In our case, this means pairs more than 3 units apart.

To check whether the choice $L = 3$ was appropriate, we computed estimators using $L = 2$ and 4. The estimated probabilities of ACF formation for all three cases ($L = 2, 3$, and 4) are almost the same. However, confidence intervals using $L = 2$ are 1.23 times wider compared to $L = 3$ and confidence intervals for $L = 4$ are only slightly narrower (by the factor 0.98) than for $L = 3$. The number of terms for $L = 4$ is almost 1.6 times larger than for $L = 3$, which increases computational effort, especially when calculating the standard errors. Therefore, we believe that our choice $L = 3$ is appropriate for the data considered.

6. Discussion and Extensions

Summarizing briefly, we have proposed models for binary data where a fixed effects structure is a combination of parametric and nonparametric models based upon fixed-knot penalized splines. We have focused on the partially linear model, although extensions to single-index models are incorporated into the general theory in Web Appendix A.

The penalized splines have penalty parameters that must converge to zero asymptotically: we derived rates for these parameters that do not lead to an asymptotic bias. We also used the KL distance to develop a data-driven method for selecting a proper amount of smoothing and derived the optimal rate of convergence for penalty parameters by selecting this way. Simulation evidence was positive: even for modeling the mean function our methods did much better than those based on CV. The method is a novel way of performing effective penalization of regression splines in spatial data situations.

The nonparametric method using ideas from Gallant (1987, p. 533) can be implemented for estimating the “meat” of the sandwich. This method is more general and can be applied for problems other than our binary case. This method will be reported elsewhere, but we found that it gave roughly the same answers and small-sample behavior as the exact calculation.

We applied the methods to the data from an experiment on the genesis of colon cancer. We identified regions of high ACF formation that were initially surprising to us but that upon examination of other data actually correspond to regions of high tumor formation. Biologically, this provides a quantification of the localization of ACF formation as precursors to tumors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This article is based on the Ph.D. dissertation of the first author at Texas A&M University. Our research was supported by grants for the National Cancer Institute (CA57030, CA104620, CA61750, CA82907, CA59034), NSF (DMS 04-538), NSBRI (NASA NCC 9-58), and by the Texas A&M Center for Environmental and Rural Health through a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

7. Supplementary Materials

A Web Appendix containing general asymptotic theory for this problem, as well as a color version of Figure 2, are both available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

APPENDIX

Composite Likelihood with Sets of Size k

Equation (3) is replaced by

$$\begin{aligned} \text{pr} \left(D_{i_1}=1, \dots, D_{i_k}=1 \mid \mu_{i_1}^*, \dots, \mu_{i_k}^*, \sigma_{\lambda}^2, \rho \right) \\ = \Phi_k \left[\mu_{i_1}^*, \dots, \mu_{i_k}^*, \frac{\sigma_{\lambda}^2}{(1+\sigma_{\lambda}^2)} \left\{ \Omega_{ij}(\rho) \right\}_{i,j \in \{i_1, \dots, i_k\}} \right]. \end{aligned} \quad (\text{A.1})$$

If we use sets of size k , then we need to keep track of the values of the sets and their probabilities. To do this, we define

$$\begin{aligned} Z_{i_1, \dots, i_k}^{(j_1, \dots, j_k)} &= I(D_{i_1}=j_1, \dots, D_{i_k}=j_k); \\ \pi_{i_1, \dots, i_k}^{(j_1, \dots, j_k)} &= E \left(Z_{i_1, \dots, i_k}^{(j_1, \dots, j_k)} \mid \mu_{i_1}^*, \dots, \mu_{i_k}^*, \sigma_{\lambda}^2, \rho \right). \end{aligned} \quad (\text{A.2})$$

Using the marginal probabilities (A.1), we can write the likelihood at distinct locations $(i_1, \dots, i_k)$ as

$$\mathcal{L}_{i_1, \dots, i_k}(\Theta) = \sum_{j_1, \dots, j_k=0}^1 Z_{i_1, \dots, i_k}^{(j_1, \dots, j_k)} \log \left\{ \pi_{i_1, \dots, i_k}^{(j_1, \dots, j_k)}(\beta, \theta) \right\}.$$

It is useful to note that $\mathcal{L}_{i_1, \dots, i_k}(\Theta)$ is an actual likelihood function, albeit for a reduced set of data.

We propose to maximize a weighted penalized composite likelihood. Let $w_{i_1, \dots, i_k}$ be the weights, for example, the indicator that the maximum distance between two locations from a set of locations $(i_1, \dots, i_k)$ is less than a specified value apart, a choice that is useful to cut down on

the size of the summations. It is assumed that $w_{i_1, \dots, i_k}$ does not depend upon $(D_{i_1}, \dots, D_{i_k})$ or any parameters.

For fixed smoothing parameter, κ , the composite likelihood estimator is denoted by $\widehat{\Theta}(\kappa)$ and is obtained by maximizing over Θ

$$\tilde{\mathcal{L}}_{\mathcal{W}}(\Theta, \kappa) = \mathcal{W}^{-1} \sum_{i_1, \dots, i_k=1} w_{i_1, \dots, i_k} \mathcal{L}_{i_1, \dots, i_k}(\Theta) - (1/2) \kappa \eta^T \mathcal{G} \eta, \quad (\text{A.3})$$

where $\mathcal{W} = \sum_{i_1, \dots, i_k} w_{i_1, \dots, i_k}$.

With a slight abuse of notation, for composite likelihood, we have that

$$\begin{aligned} \text{SKL} \left\{ \widehat{\Theta}(\kappa) \right\} &= \mathcal{W}^{-1} \sum_{i_1, \dots, i_k=1}^n \sum_{j_1, \dots, j_k=0}^1 w_{i_1, \dots, i_k} \\ &\times \left[\pi_{i_1, \dots, i_k}^{(j_1, \dots, j_k)}(\Theta_0) - \pi_{i_1, \dots, i_k}^{(j_1, \dots, j_k)}\left\{ \widehat{\Theta}(\kappa) \right\} \right] \\ &\times \left(\log \left\{ \pi_{i_1, \dots, i_k}^{(j_1, \dots, j_k)}(\Theta_0) \right\} \right. \\ &\quad \left. - \log \left[\pi_{i_1, \dots, i_k}^{(j_1, \dots, j_k)}\left\{ \widehat{\Theta}(\kappa) \right\} \right] \right), \end{aligned}$$

where $\pi_{i_1, \dots, i_k}^{(j_1, \dots, j_k)}(\cdot)$ are defined in (A.2).

Smoothing Parameter Selection

The main purpose of this section is to show that smoothing parameters should be of order $O(\mathcal{W}^{-1})$, and to sketch the basic algebra to obtain these smoothing parameters. Here we need to be more specific about the definition of Θ , which we take by convention to be $\Theta = (\eta^T, \zeta, \theta^T)^T$.

Define the derivative of $\mathcal{L}_{i_1, \dots, i_k}(\Theta)$ as $\mathcal{A}_{i_1, \dots, i_k}(\Theta)$. Define $\mathcal{T}(\kappa, \Theta) = \kappa \mathcal{M} \Theta$, where $\mathcal{M} = \text{diag}(\mathcal{G}, 0, I_{q\zeta-1})$. Set $\kappa = \mathcal{W}^{-\nu} c$ for some constant c . Assume that $\nu \geq 1/2$, then by theorem 1 in Web Appendix A, $\widehat{\Theta}(\kappa) - \Theta_0 = O_p(\mathcal{W}^{-1/2})$.

By Taylor expansion after some algebraic manipulations,

$$\begin{aligned} \text{SKL} &= \mathcal{W}^{-1} \sum_{i_1, \dots, i_k} \left\{ \mathcal{A}_{i_1, \dots, i_k}(\Theta_0) \right\}^T \left(\widehat{\Theta} - \Theta_0 \right) \\ &\times \left(\widehat{\Theta} - \Theta_0 \right)^T \mathcal{A}_{i_1, \dots, i_k}(\Theta_0) + O_p(\mathcal{W}^{-3/2}). \end{aligned}$$

It is easy to see that

$$\begin{aligned} \text{SKL} &= \mathcal{W}^{-1} \sum_{i_1, \dots, i_k} \left\{ \mathcal{A}_{i_1, \dots, i_k}^n(\Theta_0) \right\}^T \left(\widehat{\Theta} * -\Theta_0 \right) \\ &\times \left(\widehat{\Theta} * -\Theta_0 \right)^T \mathcal{A}_{i_1, \dots, i_k}(\Theta_0) + O_p(\mathcal{W}^{-3/2}), \end{aligned}$$

where $\widehat{\Theta}^* = \Theta_0 + \{\Sigma_{W,0} + M(\kappa)\}^{-1} \mathcal{W}^{-1} \Sigma_{i_1, \dots, i_k} \mathcal{A}_{i_1, \dots, i_k}(\Theta_0) - \mathcal{T}(\kappa, \Theta_0)$. Associated with SKL is its expected value in this asymptotically equivalent version of $\widehat{\Theta}$, namely,

$$\begin{aligned} \text{MASKL} = & \mathcal{W}^{-1} \sum_{i_1, \dots, i_k} \{\mathcal{A}_{i_1, \dots, i_k}(\Theta_0)\} E \\ & \times \left\{ \left(\widehat{\Theta}^* - \Theta_0 \right) \left(\widehat{\Theta}^* - \Theta_0 \right)^T \right\} \mathcal{A}_{i_1, \dots, i_k}(\Theta_0). \end{aligned}$$

Recall that $\Sigma_{W,0} = E \left\{ \partial \Psi_{\mathcal{W}}(\Theta, \kappa) \partial \Theta^T \right\}$ is the exterior noninverted part and $\Sigma_{W,0} + \Sigma_{W,c}$ is the interior covariance matrix in the sandwich formula (6). We discussed in Section 2.6 how to estimate these components

Note that $\mathcal{W}^{-1} \Sigma_{i_1, \dots, i_k} \mathcal{A}_{i_1, \dots, i_k}(\Theta_0) \{\mathcal{A}_{i_1, \dots, i_k}(\Theta_0)\}^T = \Sigma_{W,0} + o_p(\mathcal{W}^{-1/2})$. Define

$R(\kappa) = \{\Sigma_{W,0} + M(\kappa)\}^{-1}$, $S(\kappa) = \mathcal{W}^{-1} (\Sigma_{W,0} + \Sigma_{W,c}) + \mathcal{T}(\kappa, \Theta_0) \mathcal{T}^T(\kappa, \Theta_0)$. Then we have that the asymptotically equivalent version is

$$\widetilde{\text{MASKL}}(\kappa) = \text{trace} \left\{ R(\kappa) S(\kappa) R(\kappa) \Sigma_{W,0} \right\}. \quad (\text{A.4})$$

It is evident by inspection that in order to minimize (A.4), we must have $\kappa = O(\mathcal{W}^{-1})$ as claimed.

The minimizer of (A.4) solves $0 = \partial \widetilde{\text{MASKL}}(\kappa) / \partial \kappa$. Then there is a matrix

$\mathcal{H}(\kappa, \Theta_0, \Sigma_{W,0}, \Sigma_{W,c})$ and a vector $\mathcal{J}(\kappa, \Theta_0, \Sigma_{W,0}, \Sigma_{W,c})$, where $\mathcal{H}^{-1}(\cdot)$ and $\mathcal{J}(\cdot)$ are both of order $O(1)$, such that to terms of order $o(\mathcal{W}^{-1})$, the nontrivial equation for κ is

$\kappa = \mathcal{W}^{-1} \mathcal{H}^{-1}(\kappa, \Theta_0, \Sigma_{W,0}, \Sigma_{W,c}) \mathcal{J}(\kappa, \Theta_0, \Sigma_{W,0}, \Sigma_{W,c}) = 0$. The latter equation can be solved iteratively. However, to terms of the first order, the minimizer of (A.4) solves

$$0 = \left\{ \frac{\partial \widetilde{\text{MASKL}}(0,0)}{\partial \kappa} \right\} + \left\{ \frac{\partial^2 \widetilde{\text{MASKL}}(0,0)}{\partial \kappa \partial \kappa} \right\} \kappa,$$

which leads to a linear equation $\kappa = \mathcal{W}^{-1} \widetilde{\mathcal{H}}^{-1} \times (\Theta_0, \Sigma_{W,0}, \Sigma_{W,c}) \times \widetilde{\mathcal{J}}(\Theta_0, \Sigma_{W,0}, \Sigma_{W,c})$. All terms mentioned above are straightforward to compute and estimate.

References

- Apanasovich TV, Sheather S, Lupton JR, Popovic N, Turner ND, Chapkin RS, Carroll RJ. Testing for spatial correlation in nonstationary binary data with application to aberrant crypt foci in colon carcinogenesis. *Biometrics* 2003;59:752–761. [PubMed: 14969452]
- Bird RP. Role of aberrant crypt foci in understanding the pathogenesis of colon cancer. *Cancer Letters* 1995;93:55–71. [PubMed: 7600544]
- Bird RP, Good CK. The significance of aberrant crypt foci in understanding the pathogenesis of colon cancer. *Toxicology Letters* 2000;112-113:395–402. [PubMed: 10720758]
- Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological* 1999;61:265–285.
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993;88:9–25.

- Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* 1995;82:81–91.
- Christensen OF, Waagepetersen RP. Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics* 2002;58:280–286. [PubMed: 12071400]
- Cox DR, Reid N. A note on pseudo-likelihood constructed from marginal densities. *Biometrika* 2004;91:729–737.
- Cressie, NAC. *Statistics for Spatial Data*. Wiley; New York: 1993.
- Curriero FC, Lele S. A composite likelihood approach to semivariogram estimation. *Journal of Computational and Graphical Statistics* 1999;4:9–28.
- Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics with discussion. *Applied Statistics* 1998;47:299–326.
- Fuentes M. Spectral methods for nonstationary spatial processes. *Biometrika* 2002;89:197–210.
- Gallant, RA. *Nonlinear Statistical Models*. Wiley; New York: 1987.
- Godambe VP. An optimum property of regular maximum likelihood equation. *The Annals of Statistics* 1960;31:1208–1211.
- Guan Y. A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association* 2006;101:1502–1512.
- Handcock MS, Stein ML. A Bayesian analysis of kriging. *Technometrics* 1993;35:403–410.
- Handcock MS, Wallis JR. An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association* 1993;89:368–378.
- Heagerty PJ, Lele SR. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* 1998;93:1099–1111.
- Kuk AYC, Nott DJ. A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters* 2000;47:329–335.
- Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* 1996;91:1007–1016.
- Lindsay BG. Composite likelihood methods. *Contemporary Mathematics* 1988;80:221–239.
- McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* 1997;92:162–170.
- Müller WG. Least-squares fitting from the variogram cloud. *Statistics and Probability Letters* 1999;43:93–98.
- Neuhaus, JM.; Segal, MR. An assessment of approximate maximum likelihood estimators in generalized linear mixed models. In: Gregoire, TG.; Brillinger, DR.; Diggle, PJ.; Russek-Cohen, E.; Warren, WG.; Wolfinger, RD., editors. *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions (Lecture Notes in Statistics, Volume 122)*. Springer; New York: 1997. p. 11–22.
- Nott DJ, Rydén T. Pairwise likelihood methods for inference in image models. *Biometrika* 1999;86:661–676.
- Opsomer J, Wang Y, Yang Y. Nonparametric regression with correlated errors. *Statistical Science* 2001;16:134–153.
- Renard D, Molenberghs G, Geys H. A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis* 2004;44:649–667.
- Ruppert D. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* 2002;11:735–757.
- Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric Regression*. Cambridge University Press; Cambridge, U.K.: 2003.
- Sherman M, Apanasovich TV, Carroll RJ. On estimation in binary autologistic spatial models. *Journal of Statistical Computation and Simulation* 2004;76:167–179.
- Stein, ML. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag Inc.; Berlin: 1999.
- Varin C, Vidoni P. Pairwise likelihood inference for general state-space models. *Econometric Reviews*. 2007in press

- Varin C, Høst G, Skare Ø. Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics and Data Analysis* 2005;49:1173–1191.
- Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* 2002;97:1042–1054.
- Zhang H. On estimation and prediction for spatial generalized linear mixed models. *Biometrics* 2002;58:129–136. [PubMed: 11890307]
- Zhao Y, Joe H. Composite likelihood estimation in multivariate data analysis. *The Canadian Journal of Statistics* 2005;33:335–356.

Figure 1.

Q-Q plots for $\{\tilde{\mu}_{0.5}^{(i)}\}_{i=1}^{1000}$ for $\psi = 0.5, \Omega(\rho) = 0.5$ (left panel), and $\psi = 0.8, \Omega(\rho) = 0.6$ (right panel).

Figure 2.
The image of the colon with aberrant crypts displayed as larger dark and distended regions with raised "bumps."

Figure 3.

A typical rat colon. This shows how the data are made available to us: the black regions are Peyer's Patches, where aberrant crypt foci are not readable. The number on the horizontal axis means the physical distance from the distal part of the colon in centimeters.

Figure 4. Estimated probabilities of ACF formation with 95% confidence limits. The solid line corresponds to the irradiated group and the dashed line corresponds to the nonirradiated group. X is the normalized distance from the distal part of the colon.

Table 1

Results of the simulation to compare crossvalidation with our smoothing parameter selector MASKL. The methods are as follows: “MASKL” refers to the method based on minimization of MASKL (criterion described in Section 3), “CV” refers to the method based on crossvalidation, “ISB” is integrated squared bias and “IMSE” is integrated mean squared error of the probability function. Our method, MASKL, clearly dominates crossvalidation in terms of mean squared error.

Method	ISB(10^{-2})	IMSE (10^{-2})
$\psi = 0.5, \Omega(\rho) = 0.5$		
MASKL	0.003	0.246
CV	0.005	0.553
$\psi = 0.8, \Omega(\rho) = 0.6$		
MASKL	0.002	0.469
CV	0.003	1.425

Table 2

Results of the simulation for 1000 data sets under the first scenario. The first column is the method: “MASKL” refers to the method based on minimization of MASKL (criterion described in Section 3). Columns 2-4 and 5-7 are mean, 2.5th, 97.5th percentiles for ψ and $\Omega(\rho)$, respectively.

Method	$\hat{\psi}$	$\hat{\psi}.025$	$\hat{\psi}.975$	$\alpha\hat{\rho}$	$\alpha\hat{\rho}.025$	$\alpha\hat{\rho}.975$
MASKL	0.440	0.362	0.533	0.570	0.452	0.656
MASKL	0.764	0.664	0.972	0.562	0.454	0.758