Statistica Sinica **12**(2002), 449-474

SPATIAL MIXTURE MODELS BASED ON EXPONENTIAL FAMILY CONDITIONAL DISTRIBUTIONS

Mark S. Kaiser, Noel Cressie and Jaehyung Lee

Iowa State University, The Ohio State University and Iowa State University

Abstract: Spatial statistical models are applied in many problems for which dependence in observed random variables is not easily explained by a direct scientific mechanism. In such situations there may be a latent spatial process that acts to produce the observed spatial pattern. Scientific interest often centers on the latent process and the degree of spatial dependence that characterizes it. Such latent processes may be thought of as spatial mixing distributions. We present methods for the specification of flexible joint distributions to model spatial processes through multi-parameter exponential family conditional distributions. One approach to the analysis of these models is Monte Carlo maximum likelihood, and an approach based on independence pseudo-models is presented for formulating importance sampling distributions that allow such an analysis. The methods developed are applied to a problem of forest-health monitoring, where the numbers of affected trees in spatial field plots are modeled using a spatial beta-binomial mixture.

Key words and phrases: Beta-binomial, hierarchical models, Markov random fields, Monte Carlo maximum likelihood.

1. Introduction

Observations of random variables taken in a geographical context often exhibit spatial dependence even though there is no obvious spatial mechanism to explain it. For example, data on infant mortalities categorized as due to Sudden Infant Death Syndrome (SIDS) in North Carolina from 1974 to 1984 exhibit spatial structure (e.g., Cressie and Chan (1989)) even though SIDS is not an infectious disease. Similarly, the rate of lip-cancer incidence in Scotland has been analyzed using spatial statistical models (e.g., Clayton and Kaldor (1987); Stern and Cressie (1999)). The presence of spatial dependence in such applications can often be due to an unspecified or missing covariate.

Two common ways to model spatial dependence are through the covariances between spatial random variables (geostatistical models), or through the set of conditional distributions of each spatial variable given all others (Markov random field models). We are concerned in this article with spatial mixture (i.e., hierarchical) models for non-Gaussian data. There have been recent papers that address this problem using geostatistical models; specifically DeOliveira, Kedem, and Short (1997) and Diggle, Tawn, and Moyeed (1998) use a fully Bayesian analysis, and Heagerty and Lele (1998) and Cressie (2000) use an empirical Bayes analysis. In contrast, this article addresses the problem using Markov random field (MRF) models. Up to now, models for spatial dependence have generally assumed a Gaussian MRF after a nonlinear transformation (e.g., Clayton and Kaldor (1987); Breslow and Clayton (1993); Knorr-Held and Besag (1998)). Our basic objective in this article is to present results that allow the construction of spatial mixture models that do not restrict the spatial mixing distribution to be Gaussian. The data model is one of conditionally independent random variables, conditional on parameters that are distributed according to a spatial process on a lattice. It is this latent spatial process that is the focus of our investigation, and in what is to follow we model this process as a finite MRF having exponential family conditional distributions. We do not restrict the conditional distributions to be Gaussian or to follow a one-parameter auto-model (e.g., Besag (1974)).

The remainder of the article is organized as follows. After laying out the general structure of spatial mixture models in Section 2, we address the formulation of flexible mixing distributions in Section 3. Results are given that allow multi-parameter exponential families, such as the beta, to be used in forming multivariate mixing distributions on the basis of conditional specifications. Our estimation approach is empirical Bayesian; Section 4 contains a Monte Carlo method for maximum likelihood estimation of unknown parameters for a general class of spatial mixture models. In Section 5, a spatial beta-binomial model is fitted to the number of trees in spatial field plots that exhibit foliar damage. Discussion and concluding remarks are in Section 6.

2. Spatial Mixtures

We assume a finite collection of random variables $\boldsymbol{Y} \equiv \{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ with geographical locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$. For example, in Section 5, $\boldsymbol{s}_i \equiv (u_i, v_i)$, where u_i denotes the longitude and v_i the latitude of the centroid for a forest monitoring plot in the northeastern United States. Given a set of parameters $\boldsymbol{\theta} \equiv \{\theta(\boldsymbol{s}_i) : i = 1, \ldots, n\}$, the density or mass function of $Y(\boldsymbol{s}_i)$ is assumed to depend only on $\theta(\boldsymbol{s}_i)$ and is denoted $f_i(y(\boldsymbol{s}_i)|\theta(\boldsymbol{s}_i))$, for $y(\boldsymbol{s}_i) \in \Omega_i$. Often, we will have $\Omega_1 = \ldots = \Omega_n$, but this is not necessary and will not be true for the example presented in Section 5. Let $\Omega \equiv \Omega_1 \times \ldots \times \Omega_n$ denote the joint sample space.

In all that is to follow, the response vector $\boldsymbol{Y}$ is taken to have conditionally independent components given $\boldsymbol{\theta}$, so that the data model is given by,

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_i f_i(y(\boldsymbol{s}_i)|\theta(\boldsymbol{s}_i)); \quad \boldsymbol{y} \in \Omega. \tag{1}$$

The collection of factors that influence the distribution of $\boldsymbol{Y}$ are assumed to include one or more spatial processes, to be modeled by assigning a joint distribution to $\boldsymbol{\theta}$ that incorporates dependence. In this article, we formulate such distributions through conditional specification of the density or mass functions $\{g_i(\theta(\boldsymbol{s}_i)|\{\theta(\boldsymbol{s}_j) : j \neq i\}); \theta(\boldsymbol{s}_i) \in \Theta_i : i = 1, \ldots, n\}$. We assume a neighborhood structure on the spatial locations is available and define, for $i = 1, \ldots, n$, $\boldsymbol{\theta}(N_i) \equiv \{\theta(\boldsymbol{s}_j) : \boldsymbol{s}_j$ is a neighbor of $\boldsymbol{s}_i\}$. Then the Markov random field (MRF) assumption on the spatial process $\{\theta(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ is expressed through $g_i(\theta(\boldsymbol{s}_i)|\{\theta(\boldsymbol{s}_j) : j \neq i\}) = g_i(\theta(\boldsymbol{s}_i) \mid \boldsymbol{\theta}(N_i); \boldsymbol{\lambda})$; $\theta(\boldsymbol{s}_i) \in \Theta_i$, where $\boldsymbol{\lambda}$ is a p-dimensional parameter associated with the joint probability measure of $\boldsymbol{\theta}$.

In Section 3, we show how the $\{g_i(\cdot|\cdot) : i = 1, \ldots, n\}$ that come from exponential families may be used, under appropriate conditions, to specify the joint distribution for $\boldsymbol{\theta}$. For the purpose of this article, we write this joint distribution through a negpotential function of $\boldsymbol{\theta}$ as,

$$g(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \frac{\exp\{Q(\boldsymbol{\theta}|\boldsymbol{\lambda})\}}{\int_{\Theta} \exp\{Q(\boldsymbol{t}|\boldsymbol{\lambda})\}\, d\boldsymbol{t}}; \qquad \boldsymbol{\theta} \in \Theta. \tag{2}$$

Thus, from (1) and (2), the marginal distribution of $\boldsymbol{Y}$ is the spatial mixture model,

$$m(\boldsymbol{y}|\boldsymbol{\lambda}) = \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta}; \qquad \boldsymbol{y} \in \Omega, \tag{3}$$

in which the degree of spatial dependence is captured by the parameter $\boldsymbol{\lambda}$. Notice that the data $\boldsymbol{y}$ depend on $\boldsymbol{\lambda}$ through the intermediary latent spatial process $\boldsymbol{\theta}$. Prediction of $\boldsymbol{\theta}$ depends on knowledge of $\boldsymbol{\lambda}$; estimation of $\boldsymbol{\lambda}$ is one of the problems we address.

3. Joint Mixing Distributions

The development of MRFs through specification of conditional distributions can be found in the seminal paper by Besag (1974), who took as an example the ‘auto-models’ obtained from one-parameter exponential family conditional distributions. Kaiser and Cressie (2000) provide some generalizations of the basic theory, and give a theorem that allows construction of a joint distribution, up to an unknown normalizing constant, from any set of specified conditional distributions that satisfy minimal regularity conditions. In particular, Theorem 3 of Kaiser and Cressie (2000) allows construction of multi-parameter MRF auto-models for $\boldsymbol{\theta}$.

Given that a neighborhood structure has been specified through the sets $\{N_i : i = 1, \ldots, n\}$, let $\boldsymbol{\theta}^* \equiv \{\theta^*(\boldsymbol{s}_1), \theta^*(\boldsymbol{s}_2), \ldots, \theta^*(\boldsymbol{s}_n)\}$ be any particular value in the support Θ. Although not necessary in general, it will be convenient here to assume that the components of $\boldsymbol{\theta}^*$ are all equal. We further assume that the 'positivity condition' of Besag (1974) holds, namely that $\boldsymbol{\theta} \in \Theta = \Theta_1 \times \ldots \times \Theta_n$, although this is stronger than what is needed (see Kaiser and Cressie (2000)).

Besag (1974) showed that an existing joint density or mass function for $\boldsymbol{\theta}$ may be written as in equation (2), where $Q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \sum_{\mathcal{C}} H_{\mathcal{C}}(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and $\mathcal{C}$ denotes the set of all cliques, defined as singletons or sets of locations such that all members of a set are neighbors of all other members of the same set. Theorem 3 of Kaiser and Cressie (2000) shows that a MRF joint density or mass function can be constructed from a set of conditional specifications, if and only if the $H_{\mathcal{C}}$ functions are invariant to permutation of their associated indices and the denominator of (2) is finite. In the case of 'pairwise-only dependence', where only cliques containing two or fewer locations contribute to $Q(\boldsymbol{\theta}|\boldsymbol{\lambda})$, $\mathcal{C} = \{i\}$ or $\mathcal{C} = \{i,\, j\}$ and one may define

$$H_i(\theta(\boldsymbol{s}_i)|\boldsymbol{\lambda}) \equiv \log\left[\frac{g_i(\theta(\boldsymbol{s}_i)|\boldsymbol{\theta}^*(N_i);\boldsymbol{\lambda})}{g_i(\theta^*(\boldsymbol{s}_i)|\boldsymbol{\theta}^*(N_i);\boldsymbol{\lambda})}\right];\ \ i = 1,\ldots,n, \tag{4}$$

$$H_{i,j}(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j)|\boldsymbol{\lambda}) \equiv$$

$$\log\left[\frac{g_i(\theta(\boldsymbol{s}_i)|\theta(\boldsymbol{s}_j),\boldsymbol{\theta}^*(N_{i-j});\boldsymbol{\lambda})}{g_i(\theta^*(\boldsymbol{s}_i)|\theta(\boldsymbol{s}_j),\boldsymbol{\theta}^*(N_{i-j});\boldsymbol{\lambda})}\right] - H_i(\theta(\boldsymbol{s}_i)|\boldsymbol{\lambda});\ \ i \neq j \in N_i. \tag{5}$$

In (5), N_{i-j} denotes the set of neighbors of site $\boldsymbol{s}_i$, excluding $\boldsymbol{s}_j$, and $H_{i,j}$ is defined to be zero if $j \notin N_i$.

In the pairwise-only dependence case, third- and higher-order $H_{\mathcal{C}}$ functions do not contribute to $Q(\cdot)$, and permutation invariance of the $H_{\mathcal{C}}$ functions is equivalent to each function $H_{i,j}(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j)|\boldsymbol{\lambda})$ being symmetric in its indices. Under such symmetry, the spatial mixing distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ may be written as in (2), with

$$Q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \sum_{1\leq i\leq n} H_i(\theta(\boldsymbol{s}_i)|\boldsymbol{\lambda}) + \sum_{1\leq i<j\leq n} H_{i,j}(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j)|\boldsymbol{\lambda}), \tag{6}$$

provided $Q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is integrable over Θ.

Here we specify the conditional distributions associated with $\boldsymbol{\theta}$ to be of multiparameter exponential family form with densities or mass functions, for $i = 1,\ldots,n$,

$$g(\theta(\boldsymbol{s}_i)|\boldsymbol{\theta}(N_i);\,\boldsymbol{\lambda}) =$$

$$\exp\Big[\sum_{k=1}^{q}\{A_{i,k}(\boldsymbol{\theta}(N_i);\boldsymbol{\lambda})T_{i,k}(\theta(\boldsymbol{s}_i))\} - B_i(\boldsymbol{\theta}(N_i);\boldsymbol{\lambda}) + C_i(\theta(\boldsymbol{s}_i))\Big];\quad \boldsymbol{\theta} \in \Theta. \tag{7}$$

To be useful in statistical model building, we must have explicit forms for the functions $A_{i,k}(\boldsymbol{\theta}(N_i);\boldsymbol{\lambda})$ that play the role of natural parameters in (7). Besag (1974) gave a necessary parameterization for one-parameter exponential family conditional distributions (i.e., $q = 1$), but an analogous result is not possible with multi-parameter families, as noted by Cressie and Lele (1992). We give three ways that $A_{i,k}(\boldsymbol{\theta}(N_i);\boldsymbol{\lambda})$ may be parameterized so that the functions $H_{i,j}$ of equation (5) are symmetric in the indices i and j. Upon substituting (7) into (4) and (5), we obtain for all i,

$$\begin{aligned} &H_i(\theta(\boldsymbol{s}_i)|\boldsymbol{\lambda}) \\ &= \sum_{k=1}^{q} [A_{i,k}(\boldsymbol{\theta}^*(N_i);\boldsymbol{\lambda})\{T_k(\theta(\boldsymbol{s}_i)) - T_k(\theta^*(\boldsymbol{s}_i))\}] + C_i(\theta(\boldsymbol{s}_i)) - C_i(\theta^*(\boldsymbol{s}_i)), \end{aligned} \tag{8}$$

and for all i and j,

$$\begin{aligned} H_{i,j}(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j)|\boldsymbol{\lambda}) = \sum_{k=1}^{q} \Big[&\{A_{i,k}(\theta(\boldsymbol{s}_j),\boldsymbol{\theta}^*(N_{i-j});\boldsymbol{\lambda}) - A_{i,k}(\theta^*(\boldsymbol{s}_j),\boldsymbol{\theta}^*(N_{i-j});\boldsymbol{\lambda})\} \\ &\times\{T_k(\theta(\boldsymbol{s}_i)) - T_k(\theta^*(\boldsymbol{s}_i))\}\Big]. \end{aligned} \tag{9}$$

We present three propositions that yield functions $H_{i,j}(\cdot)$ symmetric in the indices i and j.

Proposition 1. *Let* $\boldsymbol{\lambda} = \{\alpha_{i,k}, \eta_{i,j} : i, j = 1,\ldots,n; i \neq j; k = 1,\ldots,q\}$, *and specify*

$$A_{i,k}(\boldsymbol{\theta}(N_i);\boldsymbol{\lambda}) = \alpha_{i,k} + \sum_{j\in N_i} \Big\{\eta_{i,j} \sum_{h=1}^{q} T_h(\theta(\boldsymbol{s}_j))\Big\}. \tag{10}$$

If $\eta_{i,j} = \eta_{j,i}$ *for all* $i \neq j$, *then* $H_{i,j}(\cdot) = H_{j,i}(\cdot)$.

Proof. Substitution of (10) into (9) gives

$$\begin{aligned} &H_{i,j}(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j)|\lambda) \\ &= \eta_{i,j}\Big[\sum_{k=1}^{q}\{T_k(\theta(\boldsymbol{s}_j)) - T_k(\theta^*(\boldsymbol{s}_j))\}\Big]\Big[\sum_{k=1}^{q}\{T_k(\theta(\boldsymbol{s}_i)) - T_k(\theta^*(\boldsymbol{s}_i))\}\Big]. \end{aligned}$$

The result follows because $\eta_{i,j}$ is symmetric in i and j.

Proposition 2. *Let* $\boldsymbol{\lambda} = \{\alpha_{i,k},\ \eta_{i,j,k} : i,\ j = 1,\ldots,n; i \neq j; k = 1,\ldots,q\}$, *and specify*

$$A_{i,k}(\boldsymbol{\theta}(N_i);\boldsymbol{\lambda}) = \alpha_{i,k} + \sum_{j\in N_i} \eta_{i,j,k}\, T_k(\theta(\boldsymbol{s}_j)). \tag{11}$$

If $\eta_{i,j,k} = \eta_{j,i,k}$ *for all* $i \neq j$, *and* $k = 1,\ldots,q$, *then* $H_{i,j}(\cdot) = H_{j,i}(\cdot)$.

Proof. Substitution of (11) into (9) gives

$$H_{i,j}(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j)|\boldsymbol{\lambda})=\sum_{k=1}^{q}\eta_{i,j,k}\left[\{T_k(\theta(\boldsymbol{s}_j))-T_k(\theta^*(\boldsymbol{s}_j))\}\{T_k(\theta(\boldsymbol{s}_i))-T_k(\theta^*(\boldsymbol{s}_i))\}\right].$$

The result follows because $\eta_{i,j,k}$ is symmetric in i and j.

Proposition 3. *Let* $\boldsymbol{\lambda} = \{\alpha_{i,k}, \eta_{i,j,k} : i, j = 1,\ldots,n; i \neq j; k = 1, 2\}$. *Specify*

$$\begin{aligned} A_{i,1}(\boldsymbol{\theta}(N_i),\boldsymbol{\lambda}) &= \alpha_{i,1} + \sum_{j\in N_i}\eta_{i,j}T_2(\theta(\boldsymbol{s}_j)),\\ A_{i,2}(\boldsymbol{\theta}(N_i),\lambda) &= \alpha_{i,2} + \sum_{j\in N_i}\eta_{i,j}T_1(\theta(\boldsymbol{s}_j)).\end{aligned} \tag{12}$$

If $\eta_{i,j} = \eta_{j,i}$ *for all* $i \neq j$, *then* $H_{i,j}(\cdot) = H_{j,i}(\cdot)$.

Proof. Substitution of (12) into (9) gives,

$$H_{i,j}(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j)|\boldsymbol{\lambda}) = \eta_{i,j}\,\{T_1(\theta(\boldsymbol{s}_j)) - T_1(\theta^*(\boldsymbol{s}_j))\}\{T_2(\theta(\boldsymbol{s}_i)) - T_2(\theta^*(\boldsymbol{s}_i))\}.$$

The result follows because $\eta_{i,j}$ is symmetric in i and j.

We have presented Proposition 3 in the case of a two-parameter exponential family. More general versions are possible, with Proposition 3 used to model pairs of natural parameters and any remaining parameters modeled using a combination of Propositions 1 and 2. Also note that one might set $\eta_{i,j,k} = 0$ for certain values of k (and all i, j) in Proposition 2. In the case of two-parameter exponential families (e.g., the Gaussian), setting $\eta_{i,j,2} \equiv 0$ would give the parameterization for one-parameter exponential families that Besag (1974) showed was necessary.

Typically, we reduce the number of free parameters in (10), (11), or (12) by placing additional restrictions on $\{\alpha_{i,k} : i = 1,\ldots,n; k = 1,\ldots,q\}$ and $\{\eta_{i,j,k} : i, j = 1,\ldots,n; i \neq j; k = 1,\ldots,q\}$. For example, in the application of Section 5 there are two sufficient statistics ($q = 2$); there, we take $\alpha_{i,k} \equiv \alpha_k$ and use Proposition 3 with $\eta_{i,j} \equiv \eta$.

From Kaiser and Cressie (2000), Theorem 3, if the denominator of (2) is finite for the $Q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ implied by (10), (11), or (12), then a valid MRF model has been constructed. Therefore, the parameter space of α's and η's consists of all such in (10), (11), or (12) that yield a finite denominator in (2). General results are not apparent, and conditional specifications in particular exponential families (e.g., beta or gamma) need to be considered separately. Nevertheless, Propositions 1, 2, and 3 provide considerable flexibility in the development of statistical models. We investigate one type of exponential family, the beta family, in more detail in Section 5.

Each of (10), (11), or (12) imply a distinct form for the negpotential function $Q(\boldsymbol{\theta}|\boldsymbol{\lambda})$. For a model constructed using (10) in Proposition 1, we have (up to an additive constant),

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = &\sum_{i=1}^{n} \Big\{ C_i(\theta(\boldsymbol{s}_i)) + \sum_{k=1}^{q} \alpha_{i,k} T_k(\theta(\boldsymbol{s}_i)) \Big\} \\ &+ \sum\sum_{1\le i<j\le n} \Big[\eta_{i,j} \Big\{ \sum_{k=1}^{q} T_k(\theta(\boldsymbol{s}_i)) \Big\} \Big\{ \sum_{h=1}^{q} T_h(\theta(\boldsymbol{s}_j)) \Big\} \Big], \end{aligned} \tag{13}$$

where, in the double summation, $\eta_{i,j} = \eta_{j,i}$ for all i, j, and $\eta_{i,j} = 0$ if $\boldsymbol{s}_j \notin N_i$. For a model constructed using (11) in Proposition 2, we have (up to an additive constant),

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = &\sum_{i=1}^{n} \Big\{ C_i(\theta(\boldsymbol{s}_i)) + \sum_{k=1}^{q} \alpha_{i,k} T_k(\theta(\boldsymbol{s}_i)) \Big\} \\ &+ \sum\sum_{1\le i<j\le n} \sum_{k=1}^{q} \left\{ \eta_{i,j,k} T_k(\theta(\boldsymbol{s}_i)) T_k(\theta(\boldsymbol{s}_j)) \right\}, \end{aligned} \tag{14}$$

where, in the double summation, $\eta_{i,j,k} = \eta_{j,i,k}$ for all i, j, k, and $\eta_{i,j,k} = 0$ if $\boldsymbol{s}_j \notin N_i$. For a model constructed using (12) in Proposition 3, and $q = 2$, we have (up to an additive constant),

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = &\sum_{i=1}^{n} \Big\{ C_i(\theta(\boldsymbol{s}_i)) + \sum_{k=1}^{2} \alpha_{i,k} T_k(\theta(\boldsymbol{s}_i)) \Big\} \\ &+ \sum\sum_{1\le i<j\le n} \eta_{i,j} \left\{ T_1(\theta(\boldsymbol{s}_i)) T_2(\theta(\boldsymbol{s}_j)) + T_2(\theta(\boldsymbol{s}_i)) T_1(\theta(\boldsymbol{s}_j)) \right\}, \end{aligned} \tag{15}$$

where, in the double summation, $\eta_{i,j} = \eta_{j,i}$ for all i, j and $\eta_{i,j} = 0$ if $\boldsymbol{s}_j \notin N_i$.

4. Estimation

Our approach to estimation of $\boldsymbol{\lambda}$ in this article is maximum likelihood. Recall from Section 1 that the underlying spatial process $\boldsymbol{\theta}$ is the focus of interest. Estimation of its parameters $\boldsymbol{\lambda}$ is informative for the spatial properties of the process. While it would be entirely reasonable to develop a hierarchical Bayesian model including a prior for $\boldsymbol{\lambda}$, we believe it can be helpful in the early stages of investigation to avoid questions of prior influence on estimation. Therefore, we take an empirical Bayesian approach and estimate $\boldsymbol{\lambda}$ by (Monte Carlo) maximum likelihood. In this section, we give a new method that is applicable to general MRF spatial mixture models, with mixing distributions that include those given by (13), (14), and (15).

4.1. Monte Carlo maximum likelihood

Consider a mixture model with $f(\boldsymbol{y}|\boldsymbol{\theta})$ as given in (1) and with the mixing distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ as given in (2). Without loss of generality, we can write,

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}) &= \frac{\exp\{Q_1(\boldsymbol{y}|\boldsymbol{\theta})\}}{\displaystyle\int_\Omega \exp\{Q_1(\boldsymbol{t}|\boldsymbol{\theta})\}\,d\boldsymbol{t}} \equiv \frac{\exp\{Q_1(\boldsymbol{y}|\boldsymbol{\theta})\}}{k_1(\boldsymbol{\theta})}, \\ g(\boldsymbol{\theta}|\boldsymbol{\lambda}) &= \frac{\exp\{Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})\}}{\displaystyle\int_\Theta \exp\{Q_0(\boldsymbol{t}|\boldsymbol{\lambda})\}\,d\boldsymbol{t}} \equiv \frac{\exp\{Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})\}}{k_0(\boldsymbol{\lambda})}, \end{aligned} \tag{16}$$

where $\boldsymbol{y} \in \Omega$ and $\boldsymbol{\theta} \in \Theta$. The log likelihood formed from the marginal density or mass function of $\boldsymbol{Y}$, given in (3), may then be written as

$$L(\boldsymbol{\lambda}) = \log\Big\{\int_\Theta \exp[Q_1(\boldsymbol{y}|\boldsymbol{\theta}) + Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda}) - \log\{k_1(\boldsymbol{\theta})\}]\,d\boldsymbol{\theta}\Big\} - \log\{k_0(\boldsymbol{\lambda})\}, \tag{17}$$

where $\boldsymbol{\lambda}$ is the p-dimensional parameter over which (17) is to be maximized.

One way to accomplish estimation for such a model is through the Monte Carlo maximum likelihood (MCMLE) approach of Geyer and Thompson (1992), which is based on maximizing a sequence of Monte Carlo approximations to $L(\boldsymbol{\lambda})$. The log likelihood (17) is of the general form considered in the context of constrained and missing data problems by Gelfand and Carlin (1993); in our case $\boldsymbol{\theta}$ constitutes the missing 'data', $\boldsymbol{y}$ the observed data, and $\boldsymbol{\lambda}$ the parameter of interest. The log likelihood (17) may be viewed as the log of a ratio of normalizing constants, since the integral appearing in the first term of (17) is the normalizing constant for the density $p(\boldsymbol{\theta}|\boldsymbol{y};\boldsymbol{\lambda})$. Monte Carlo estimation of a ratio of normalizing constants has been addressed by a number of authors (e.g., Newton and Raftery (1994); Chib (1995); Meng and Wong (1995); Ogata (1996)). Gelfand and Carlin (1993) consider estimating such a ratio of integrals for the entire likelihood surface. Our interest is rather in finding the value of $\boldsymbol{\lambda}$ that maximizes the likelihood using Monte Carlo methodology. We develop a new method for the selection of importance sampling distributions for MRF spatial mixture models based on independence pseudo-models.

A key to successful application of MCMLE is the choice of appropriate sampling distributions (from which to generate Monte Carlo samples) for evaluation of the two n-dimensional integrals ($k_0(\boldsymbol{\lambda})$ is an n-dimensional integral) appearing in (17). The strategy proposed by Geyer and Thompson (1992), and elaborated on in subsequent articles for non-mixture models (e.g., Geyer (1994); Geyer (1996)) and missing-information models (Gelfand and Carlin (1993)), is to choose a sampling distribution so that a given Monte Carlo approximation to (17) does not depend on $\boldsymbol{\lambda}$ (in practice, the current estimate of $\boldsymbol{\lambda}$). One sample may then be used both to form, and to maximize over $\boldsymbol{\lambda}$, a given approximation to $L(\boldsymbol{\lambda})$. In the approaches proposed in the literature, a sampling distribution is formed using the statistical model under analysis, evaluated at parameter values other than the current estimate, or as a finite mixture using several such parameter values (e.g., Torrie and Valleau (1977); Marinari and Parisi (1992); Gelfand and Carlin (1993); Geyer (1994), (1996)). Our experience with spatial mixture models, which involve complex patterns of dependence, is that using importance sampling distributions formed from the model itself can produce Monte Carlo estimates of the integrals in (17) that can take 'jumps' in value even after large Monte Carlo samples (e.g., of size $200,000$) have been used. Such jumps seem to be caused by the failure of the sampling mechanism to mix rapidly over the relevant sample space. We believe this failure is related to the presence of complex dependence in the sampling distribution, due to the spatial dependence in the mixing distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$.

Green (1992) suggested that sampling distributions having forms other than that of the statistical model might be appropriate in some problems. We pursue this idea here for our situation, where we have MRF spatial mixture models. The two terms in the log likelihood (17) will be approximated by selecting samples of size M from each of two known sampling distributions having densities $m_0(\boldsymbol{\theta}|\psi_0)$, $\boldsymbol{\theta} \in \Theta$, and $m_1(\boldsymbol{\theta}|\psi_1)$, $\boldsymbol{\theta} \in \Theta$, which we describe below. Let $\{\boldsymbol{\theta}_r^{(0)} : r = 1, \ldots, M\}$ denote the samples from $m_0(\boldsymbol{\theta}|\psi_0)$ and $\{\boldsymbol{\theta}_r^{(1)} : r = 1, \ldots, M\}$ denote the samples from $m_1(\boldsymbol{\theta}|\psi_1)$. Note that each element of these sets is n-dimensional. A Monte Carlo approximation to (17) is then

$$L_M(\boldsymbol{\lambda}) = \log\Big\{\frac{1}{M}\sum_{r=1}^{M}\Big(\frac{1}{m_1(\boldsymbol{\theta}_r^{(1)}|\psi_1)}\exp\Big[Q_1(\boldsymbol{y}|\boldsymbol{\theta}_r^{(1)}) + Q_0(\boldsymbol{\theta}_r^{(1)}|\boldsymbol{\lambda}) - \log\{k_1(\boldsymbol{\theta}_r^{(1)})\}\Big]\Big)\Big\} - \log\Big\{\frac{1}{M}\sum_{r=1}^{M}\Big(\frac{1}{m_0(\boldsymbol{\theta}_r^{(0)}|\psi_0)}\exp\Big[Q_0(\boldsymbol{\theta}_r^{(0)}|\boldsymbol{\lambda})\Big]\Big)\Big\}. \quad (18)$$

For the moment consider the second term in (18), which requires samples from m_0. Our strategy for defining m_0 is to construct an "independence pseudo-model", which relies on first having a sample from the MRF $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ for some starting value of $\boldsymbol{\lambda}$. Such a sample is easily produced using a Gibbs algorithm and the conditional specifications $\{g_i(\theta(\boldsymbol{s}_i)|\boldsymbol{\theta}(N_i); \boldsymbol{\lambda}) : i = 1, \ldots, n\}$. Given a sample of size S_0 from $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$, let $\{\hat{\mu}_0(\boldsymbol{s}_i), \hat{\sigma}_0^2(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ denote the sample means and sample variances at each location resulting from these S_0 simulated values. Choose, for $i = 1, \ldots, n$, densities $m_{0,i}(\theta(\boldsymbol{s}_i)|\psi_{0,i})$, $\theta(\boldsymbol{s}_i) \in \Theta_i$, by selecting values of $\psi_{0,i}$ to produce expectations and variances equal to $\hat{\mu}_0(\boldsymbol{s}_i)$

and $\hat{\sigma}_0^2(\boldsymbol{s}_i)$, respectively. Recommendations for the choice of $m_{0,i}$ are given below. Then, for $\psi_0 \equiv (\psi_{0,1},\ldots,\psi_{0,n})$, construct the sampling density

$$m_0(\boldsymbol{\theta}|\psi_0) = \prod_{i=1}^{n} m_{0,i}(\theta(\boldsymbol{s}_i)|\psi_{0,i}); \qquad \boldsymbol{\theta} \in \Theta. \tag{19}$$

We call such a sampling distribution an independence pseudo-model because it is constructed as a product of densities, which corresponds to a 'model' for $\boldsymbol{\theta}$ that assumes independence. This structure renders sampling relatively easy. Nevertheless, $m_0(\boldsymbol{\theta}|\psi_0)$ does reflect dependence in the actual model up to the effect of that dependence on the first two moments of the marginal distribution for each location. The principal value of forming the sampling density $m_0(\boldsymbol{\theta}|\psi_0)$ as in (19) is that its product form allows sampled values to cover the range of possible realizations in Θ more rapidly than they would by using the actual model.

A similar strategy may be used for selection of $m_1(\boldsymbol{\theta}|\psi_1)$, although completely general prescriptions for doing so are not available. For many models, such as the spatial beta-binomial model of Section 5, it will be reasonable to choose data model densities $f_i(y(\boldsymbol{s}_i)|\theta(\boldsymbol{s}_i))$ and conditional densities of the spatial mixing distribution $g_i(\theta(\boldsymbol{s}_i)|\{\theta(\boldsymbol{s}_j) : j \neq i\})$ to be conjugate forms. Suppose that we can write the exponent of the joint data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ in (16) as

$$Q_1(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big[\sum_{k=1}^{q} \{h_k(y(\boldsymbol{s}_i))T_k(\theta(\boldsymbol{s}_i))\} + R_i(y(\boldsymbol{s}_i))\Big],$$

and that $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ in (16) has conditional distributions that are of multiparameter exponential family form. Then the integral in (17), to be approximated using samples from $m_1(\boldsymbol{\theta}|\psi_1)$, has integrand:

$$\begin{aligned}
&\exp\left[Q_1(\boldsymbol{y}|\boldsymbol{\theta}) + Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda}) - \log\{k_1(\boldsymbol{\theta})\}\right] \\
&\propto \exp\Big[\sum_{i=1}^{n}\Big\{\sum_{k=1}^{q}\left[\{h_k(y(\boldsymbol{s}_i)) + \alpha_{i,k}\}\, T_k(\theta(\boldsymbol{s}_i))\right] + C_i(\theta(\boldsymbol{s}_i))\Big\} \\
&\qquad + \sum\sum_{1\le i<j\le n} \{W(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j))\} - \log\{k_1(\boldsymbol{\theta})\}\Big],
\end{aligned} \tag{20}$$

where $\sum\sum W(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j))$ corresponds to the double summation in the negpotential functions (13), (14), or (15), depending on how one has specified the spatial mixing distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$. For example, for a model formulated using Proposition 3, the terms of this double summation would be $W(\theta(\boldsymbol{s}_i),\theta(\boldsymbol{s}_j)) = \eta_{i,j}\,\{T_1(\theta(\boldsymbol{s}_i))T_2(\theta(\boldsymbol{s}_j)) +\ T_2(\theta(\boldsymbol{s}_i))T_1(\theta(\boldsymbol{s}_j))\}$. The importance of (20) lies in its similarity to the negpotential functions (13), (14), and (15), suggesting that a preliminary sample for selection of $m_1(\boldsymbol{\theta}|\psi_1)$ may be produced from a Gibbs algorithm using the conditionals $g_i(\theta(\boldsymbol{s}_i)|\{\theta(\boldsymbol{s}_j) : i \neq j\})$ with $\alpha_{i,k}$ of (10), (11), or (12) replaced with $\alpha_{i,k} + h_k(y(\boldsymbol{s}_i))$. Given a sample of size S_1 produced in this manner, the first two sample moments are computed for each location, yielding the values $\{\hat{\mu}_1(\boldsymbol{s}_i), \hat{\sigma}_1^2(\boldsymbol{s}_i) : i = 1, \ldots, n\}$. A sampling distribution $m_1(\boldsymbol{\theta}|\psi_1)$ is then formed in like manner to (19), namely

$$m_1(\boldsymbol{\theta}|\psi_1) = \prod_{i=1}^{n} m_{1,i}(\theta(\boldsymbol{s}_i)|\psi_{1,i}); \quad \boldsymbol{\theta} \in \Theta, \tag{21}$$

where the set of parameters $\{\psi_{1,i} : i = 1, \ldots, n\}$ are chosen to match expectations and variances with the sample moments.

The forms of the component distributions $\{m_{0,i}(\theta(\boldsymbol{s}_i)|\psi_{0,i}),\ m_{1,i}(\theta(\boldsymbol{s}_i)|\psi_{1,i}) : i = 1, \ldots, n\}$ are very flexible since, in $m_0(\boldsymbol{\theta}|\psi_0)$ and $m_1(\boldsymbol{\theta}|\psi_1)$, the values ψ_0 and ψ_1 need not lie in the same parameter space as $\boldsymbol{\lambda}$ (e.g., Geyer (1994)). The distributions $\{m_{0,i}(\theta(\boldsymbol{s}_i)|\psi_{0,i}) : i = 1, \ldots, n\}$ may often be chosen to have the same form as $\{g_i(\theta(\boldsymbol{s}_i)|\boldsymbol{\theta}(N_i); \boldsymbol{\lambda}) : i = 1, \ldots, n\}$ but, particularly if the sample moments $\{\hat{\mu}_0(\boldsymbol{s}_i), \hat{\sigma}_0^2(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ do not exhibit a relation near that of the original conditional specifications, a different distribution with the same support may be chosen instead. For example, if the conditional densities $\{g_i(\theta(\boldsymbol{s}_i)|\boldsymbol{\theta}(N_i); \boldsymbol{\lambda}) : i = 1, \ldots, n\}$ were specified as gamma, but a substantial number of locations have sample moments such that $\hat{\sigma}_0^2(\boldsymbol{s}_i) > C\, \hat{\mu}_0^2(\boldsymbol{s}_i)$ for some constant C, inverse Gaussian distributions may be used for the $m_{i,0}(\cdot)$. Similar considerations apply to selection of the component distributions $\{m_{1,i}(\theta(\boldsymbol{s}_i)|\psi_{1,i}) : i = 1, \ldots, n\}$.

Given a pair of importance sampling distributions, samples $\{\boldsymbol{\theta}_r^{(0)} : r = 1, \ldots, M\}$ may be taken from $m_0(\boldsymbol{\theta}|\psi_0)$ and $\{\boldsymbol{\theta}_r^{(1)} : r = 1, \ldots, M\}$ taken from $m_1(\boldsymbol{\theta}|\psi_1)$, and a Monte Carlo approximation to the log likelihood can be computed as in (18). Any of a number of optimization techniques may be used to maximize this approximation. We consider Newton-Raphson, which involves taking first and second derivatives of the Monte Carlo log likelihood (18), see Appendix A. One advantage of using importance sampling distributions that do not depend on $\boldsymbol{\lambda}$ is that first and second derivatives of (18) are the corresponding Monte Carlo approximations of derivatives of the actual log likelihood (17). Such approximations may then be calculated from the *same* samples used in evaluation of (18). Further, an estimate of the observed information matrix is available along with the MCMLE of $\boldsymbol{\lambda}$, upon convergence of the Newton-Raphson MCMLE estimation procedure.

For notational purposes, define the first integrand in (17) as $I_1(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\lambda}) \equiv \exp[Q_1(\boldsymbol{y}|\boldsymbol{\theta}) + Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda}) - \log\{k_1(\boldsymbol{\theta})\}]$, and the second implicit integrand in (17) as, $I_0(\boldsymbol{\theta},\boldsymbol{\lambda}) \equiv \exp\{Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})\}$. Also define

$$J_{u,v}^{(k)}(\boldsymbol{y},\boldsymbol{\lambda}) \equiv \int_\Theta \frac{\partial Q_u(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\partial \lambda_k} I_v(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\lambda})\, d\boldsymbol{\theta}; \quad u,v = 0,1,\ k = 1,\ldots,p,$$

where it should be noted that I_0, and hence $J_{u,0}^{(k)}$, has no argument in $\boldsymbol{y}$. Then first derivatives of the log likelihood (17), for $k = 1,\ldots,p$, may be written as,

$$\frac{\partial L(\boldsymbol{\lambda}|\boldsymbol{y})}{\partial \lambda_k} = \left[\int_\Theta I_1(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\lambda})\, d\boldsymbol{\theta}\right]^{-1} J_{0,1}^{(k)}(\boldsymbol{y},\boldsymbol{\lambda}) - [k_0(\boldsymbol{\lambda})]^{-1}\, J_{0,0}^{(k)}(\boldsymbol{\lambda}), \tag{22}$$

and the second derivatives, for $k,h = 1,\ldots,p$, may be written as,

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\lambda}|\boldsymbol{y})}{\partial \lambda_k\, \partial \lambda_h} = &\int_\Theta \left[\frac{\partial^2 Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\partial \lambda_k\, \partial \lambda_h} + \frac{\partial Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\partial \lambda_k}\frac{\partial Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\partial \lambda_h}\right] I_1(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\lambda})\, d\boldsymbol{\theta}\\ &- \left[\int_\Theta I_1(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\lambda})\, d\boldsymbol{\theta}\right]^{-2} J_{0,1}^{(k)}(\boldsymbol{y},\boldsymbol{\lambda})\, J_{0,1}^{(h)}(\boldsymbol{y},\boldsymbol{\lambda})\\ &+ \int_\Theta \left[\frac{\partial^2 Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\partial \lambda_k\, \partial \lambda_h} + \frac{\partial Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\partial \lambda_k}\frac{\partial Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\partial \lambda_h}\right] I_0(\boldsymbol{\theta},\boldsymbol{\lambda})\, d\boldsymbol{\theta}\\ &- [k_0(\boldsymbol{\lambda})]^2\, J_{0,0}^{(k)}(\boldsymbol{\lambda})\, J_{0,0}^{(h)}(\boldsymbol{\lambda}).\end{aligned} \tag{23}$$

Monte Carlo approximation of all integrals appearing in (22) and (23) may be accomplished using the importance sampling distributions $m_0(\boldsymbol{\theta}|\psi_0)$ (for integrals involving $I_0(\boldsymbol{\theta},\boldsymbol{\lambda})$) and $m_1(\boldsymbol{\theta}|\psi_1)$ (for integrals involving $I_1(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\lambda})$). Following Geyer and Thompson (1992), we produce an MCMLE of $\boldsymbol{\lambda}$ by forming and maximizing a sequence of Monte Carlo log likelihoods as in (18). An outline of the entire algorithm is given in Appendix A.

4.2. Monte Carlo asymptotics

A number of large-sample properties relevant to assessing behavior of the MCMLE are readily available from the ‘independence pseudo-model’ formed by the two sampling distributions $m_0(\boldsymbol{\theta}|\psi_0)$ and $m_1(\boldsymbol{\theta}|\psi_1)$. In particular, evaluation of the Monte Carlo error may be accomplished on the basis of asymptotic normality of the difference between the Monte Carlo log likelihood in (18) and the actual log likelihood in (17).

Recall that $\{\boldsymbol{\theta}_r^{(\ell)} : r = 1, \ldots, M\}$ are a random sample from the importance sampling distribution $m_\ell(\boldsymbol{\theta}|\psi_\ell)$; $\ell = 0, 1$. Define

$$D_{\ell,r} \equiv \frac{1}{m_\ell(\boldsymbol{\theta}_r^{(\ell)}|\psi_\ell)} I_\ell(\boldsymbol{y}, \boldsymbol{\theta}_r^{(\ell)}, \boldsymbol{\lambda}); \quad r = 1, \ldots, , M, \ \ell = 0, 1,$$

where I_0 and I_1 are defined in Section 4.1. Given that the component distributions of m_1 and m_0 have finite means and variances, $D_{\ell,r}$, $r = 1, \ldots, M$, are independent and identically distributed with mean E_ℓ and variance V_ℓ, $\ell = 0, 1$. A Central Limit Theorem can be applied to (18) yielding,

$$M^{1/2} \left[L_M(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda})\right] \xrightarrow{d} N(0, V_2), \quad \text{as } M \to \infty, \tag{24}$$

where 'd' denotes convergence in distribution and $V_2 \equiv V_1/E_1^2 + V_0/E_0^2$. Note that V_2 will depend on $\boldsymbol{y}$, $\boldsymbol{\lambda}$, ψ_1, and ψ_0. The fundamental importance of (24) is that it allows an estimate of the Monte Carlo error in approximation of the log likelihood (17). Let the sample mean and sample variance of $D_{\ell,r}$, $r = 1, \ldots, M$, be denoted as $\bar{D}_\ell$ and $s^2_{D,\ell}$, respectively, $\ell = 0, 1$. Then the variance V_2 can be estimated as

$$\hat{V}_2 = \frac{s^2_{D,1}}{\bar{D}_1^2} + \frac{s^2_{D,0}}{\bar{D}_0^2}. \tag{25}$$

Geyer (1994, Theorem 7) gives sufficient conditions for asymptotic normality of the difference between the MCMLE estimate $\hat{\boldsymbol{\lambda}}_M$ and the maximum likelihood estimate $\hat{\boldsymbol{\lambda}}$. The most difficult of these conditions to verify in practice is asymptotic normality of the gradient of the Monte Carlo log likelihood. Define $\bigtriangledown L_M(\boldsymbol{x}) \equiv (\partial L_M(\boldsymbol{\lambda})/\partial\lambda_1 \mid_{\boldsymbol{\lambda}=\boldsymbol{x}}, \ldots, \partial L_M(\boldsymbol{\lambda})/\partial\lambda_p \mid_{\boldsymbol{\lambda}=\boldsymbol{x}})^T$. Then Geyer's condition becomes

$$M^{1/2} \bigtriangledown L_M(\hat{\boldsymbol{\lambda}}) \xrightarrow{d} N(0, \Sigma), \quad \text{as } M \to \infty, \tag{26}$$

for some covariance matrix Σ. Note that in (26) the gradient of the Monte Carlo log likelihood, $\bigtriangledown L_M(\cdot)$, is evaluated at the actual maximum likelihood estimate $\hat{\boldsymbol{\lambda}}$. Fortunately, the independence pseudo-model allows condition (26) to be easily checked, see Appendix B.

Given that the remaining conditions of Theorem 7 of Geyer (1994) are satisfied for our particular model, we may conclude that

$$M^{1/2}[\hat{\boldsymbol{\lambda}}_M - \hat{\boldsymbol{\lambda}}] \xrightarrow{d} N(0, H^{-1}\Sigma H^{-1}), \quad \text{as } M \to \infty, \tag{27}$$

where $H = - \bigtriangledown^2 L(\boldsymbol{\lambda})$ is the negative Hessian matrix of the true log likelihood and Σ is given by (B.5) in Appendix B. For practical use of this result we propose to estimate H with $- \bigtriangledown^2 L_M(\hat{\boldsymbol{\lambda}}_M)$, and likewise obtain empirical versions of Σ as described in Appendix B.

4.3. Maximum Likelihood asymptotics

The asymptotics of Section 4.2 are with respect to the Monte Carlo sample size M, for fixed $\boldsymbol{y}$, in contrast to traditional likelihood asymptotics, which are with respect to the statistical sample size n and random $\boldsymbol{y}$. In models with complex dependence structures, standard asymptotic results for maximum likelihood estimation may not apply. Nevertheless, as a first approximation, we could use H^{-1} in (27), the inverse of the negative Hessian matrix, as a covariance matrix for the maximum likelihood estimate $\hat{\boldsymbol{\lambda}}$. This matrix could be estimated through Monte Carlo approximation of the second derivatives of $L(\boldsymbol{\lambda})$ given in (23), yielding $\hat{H}^{-1} = (-\bigtriangledown^2 L_M(\hat{\boldsymbol{\lambda}}_M))^{-1}$. Alternatively, one could use a parametric bootstrap to estimate the sampling distribution of $\hat{\boldsymbol{\lambda}}$ (e.g., Geyer and Møller (1994)).

To obtain asymptotic confidence intervals for $\boldsymbol{\lambda}$ based on the MCMLE, we need to consider asymptotics as both $M \to \infty$ and $n \to \infty$. For this purpose, we choose to feature the sample size n in the notation of this section. Let $\hat{\boldsymbol{\lambda}}^{(n)}$ be the MLE and $\hat{\boldsymbol{\lambda}}_M^{(n)}$ the MCMLE for a sample of size n.

Proposition 4. *Let $\{Y^{(n)} : n = 1, 2, \ldots\}$ be a sequence of random variables for which $Y^{(n)} \xrightarrow{d} Y \sim F$, as $n \to \infty$. For each n, let $\{X_M^{(n)} : M = 1, 2, \ldots\}$ be a sequence of random variables defined on the same probability space as $Y^{(n)}$. If, for any $\epsilon > 0$, $\lim_{n\to\infty} \limsup_{M\to\infty} Pr\left[\mid Y^{(n)} - X_M^{(n)} \mid \geq \epsilon\right] = 0$, then $X_M^{(n)} \xrightarrow{d} Y$.*

Proof. See Appendix C.

To apply the result to spatial mixture models, replace $Y^{(n)}$ with $n^{1/2}[\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}]$ and $X_M^{(n)}$ with $n^{1/2}[\hat{\boldsymbol{\lambda}}_M^{(n)} - \boldsymbol{\lambda}]$. While the proposition tells us that, for M sufficiently large, $\hat{\boldsymbol{\lambda}}_M^{(n)}$ and $\hat{\boldsymbol{\lambda}}^{(n)}$ have the same asymptotic distribution, in practice we must rely on observation of the relative magnitudes of estimates of $H^{-1}\Sigma H^{-1}$ and H^{-1}. That is, if $tr(H^{-1}\Sigma H^{-1})/tr(H^{-1})$ is estimated to be "small", then MCML can form the basis of our inference on $\boldsymbol{\lambda}$. In practice, we interpret "small" to mean that,

$$\frac{tr\{(-\bigtriangledown^2 L_M(\hat{\boldsymbol{\lambda}}_M^{(n)}))^{-1}\,\hat{\Sigma}\,(-\bigtriangledown^2 L_M(\hat{\boldsymbol{\lambda}}_M^{(n)}))^{-1}\}}{tr\{(-\bigtriangledown^2 L_M(\hat{\boldsymbol{\lambda}}_M^{(n)}))^{-1}\}} \leq 0.01. \tag{28}$$

If condition (28) does not hold, the Monte Carlo sample size M should be increased in the estimation procedure. If condition (28) is met, Wald-theory approximate confidence regions for subsets of the components of $\boldsymbol{\lambda}$ are formed using the MCMLE $\hat{\boldsymbol{\lambda}}_M^{(n)}$ and its estimated covariance matrix, $\hat{H}^{-1} = (-\bigtriangledown^2 L_M(\hat{\boldsymbol{\lambda}}_M^{(n)}))^{-1}$.

5. A Spatial Beta-Binomial Model

We apply the theory and methodology of the previous sections to a problem involving a spatially dependent latent process of probabilities of an event. It is natural to model probabilities with beta distributions, but the theory of Markov random fields has, until now, been unable to produce spatial versions. In this section, we define a spatial beta process and apply it to a problem concerned with forest health.

5.1. Motivating example – forest health

A cooperative, multi-agency program, jointly managed by the US Environmental Protection Agency and the US Forest Service, began in 1990 to monitor the condition of forests in the northeastern United States. The program consists of site visits to permanent sample plots selected on the basis of a systematic-grid survey design (Conkling and Byers (1992); Tallent-Halsell (1994)). One of the variables recorded during site visits is 'crown dieback', a measure of visible injury to the foliated portion of a tree. Crown dieback is related to the potential of a tree for carbon fixation and nutrient storage, and high values of dieback indicate physiological damage to photosynthetic cells (Pye (1988)). Crown dieback measures the response of a tree to environmental conditions and is not inherently a spatial phenomenon, that is, it is not infectious. But, environmental processes that are inherently spatial, such as atmospheric pollution and insect infestation, are often underlying causes of this type of damage to the photosynthetic portion of trees. Thus, it is appropriate to fit these data with a spatial mixture model.

We used one set of data from 1993, provided by the US Environmental Protection Agency, to illustrate the theory and methodology presented in this article. Because the spatial processes that influence crown dieback may vary for different species of trees (e.g., sensitivity of foliage to atmospheric pollution, specificity of insect infestation) we modeled one particular species, namely sweet birch, *Betula lenta*. Sweet birch is somewhat habitat specific, requiring moist woodland conditions to flourish. It occurs principally in a region from southern Maine and southwestern Quebec to Delaware and Kentucky (Gleason and Cronquist (1963)).

To define whether or not a tree was affected in terms of crown dieback, we computed the 75th percentile of crown dieback for records of all 722 trees of any species in the data set. This value provides a benchmark from which we can investigate foliar damage to sweet birch trees in the northeast United States. We defined a binary random variable for each sweet birch tree in a sample plot, equal to 1 if crown dieback for that tree was greater than the 75th percentile, and equal to 0 otherwise. Only sampling plots with at least five sweet birch trees were considered.

5.2. Model formulation

Let $m(\boldsymbol{s}_i)$ denote the number of birch trees in the sample plot with centroid $\boldsymbol{s}_i$. Let $Y(\boldsymbol{s}_i)$ denote a random variable representing the number of affected birch trees in that plot; $i = 1, \ldots, n$. The spatial locations of the $n = 36$ sample plots used in this analysis are given in Figure 1, and the data for those plots are presented in Table 1. Also listed in Table 1 is the neighborhood structure assumed for the analysis. Here site $\boldsymbol{s}_j$ was considered to be a neighbor of site $\boldsymbol{s}_i$ if $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq 48$ km, $i, j = 1, \ldots, n$. The value of 48km was selected based on a previous spatial analysis (using a number of neighborhood definitions) of locations at which sweet birch was found to occur. Notice from Table 1 that, using this neighborhood definition, some sites do not have any neighbors (e.g., site 1).

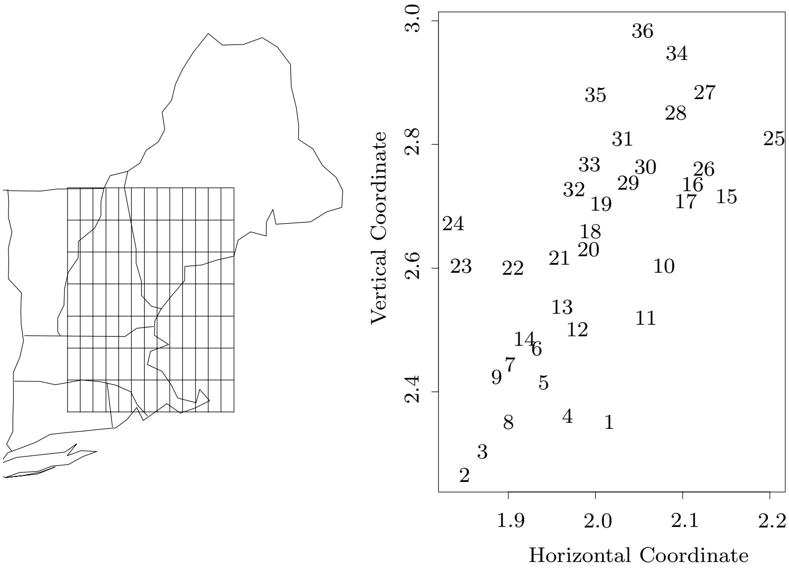

Figure 1. Locations of Forest Health Sampling Plots. Data were obtained in an area with coordinate system given in the map on the right (the scale is km $\times 10^3$). Numbers correspond to the site indices $\{\boldsymbol{s}_i\}$ in Table 1. The exact geographic location of the origin of this coordinate system was not given; the map on the left gives the region within which the area on the right lies.

Following the development in Sections 2 and 3, a spatial beta-binomial model for these data can be defined. Assume that, given values for $\boldsymbol{\theta} \equiv \{\theta(\boldsymbol{s}_i) : i =$

$1, \ldots, n\}$, the random variables $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ are independent binomials. Then the joint binomial data model is given by (1) with probability mass functions, for $y(\boldsymbol{s}_i) \in \Omega_i \equiv \{0, 1, \ldots, m(\boldsymbol{s}_i)\}$, and $i = 1, \ldots, n$,

$$f_i(y(\boldsymbol{s}_i)|\theta(\boldsymbol{s}_i)) = \frac{m(\boldsymbol{s}_i)!}{y(\boldsymbol{s}_i)!(m(\boldsymbol{s}_i) - y(\boldsymbol{s}_i))!}\{\theta(\boldsymbol{s}_i)\}^{y(\boldsymbol{s}_i)}\{1 - \theta(\boldsymbol{s}_i)\}^{m(\boldsymbol{s}_i)-y(\boldsymbol{s}_i)}. \quad (29)$$

Table 1. Data from 36 field plots in the northeastern United States: number of birch trees, $m(\boldsymbol{s}_i)$, and number of damaged trees, $y(\boldsymbol{s}_i)$, in sampling plot located at $\boldsymbol{s}_i; i = 1, \ldots, 36$. Neighbors are specified as locations within 48 km of the location of interest.

$\boldsymbol{s}_i$	$m(\boldsymbol{s}_i)$	$y(\boldsymbol{s}_i)$	Neighbors	$\boldsymbol{s}_i$	$m(\boldsymbol{s}_i)$	$y(\boldsymbol{s}_i)$	Neighbors
1	5	3		19	14	2	32
2	11	2	3	20	22	3	18,21
3	6	2	2	21	6	2	18,20,22
4	13	3	5	22	7	1	21
5	7	2	4,6,8,9	23	16	12	24
6	7	1	5,7,14	24	7	3	23
7	13	2	6,9	25	8	1	
8	5	2	5,9	26	8	2	15,16
9	5	1	5,7,8	27	5	1	28
10	8	3		28	7	2	27
11	9	1		29	5	3	30
12	6	3		30	11	2	29,31
13	5	1		31	8	5	30
14	15	3	6	32	11	4	19
15	18	8	16,17,26	33	13	5	
16	6	3	15,17,26	34	26	20	
17	13	1	15,16	35	6	3	
18	7	2	20,21	36	11	2	

The $\{\theta(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ are assumed to follow a joint distribution that we construct from beta conditionals, with neighborhood structure given by Table 1. That is, the Markov random field is constructed from conditional densities of the form given in (7) with $q = 2$, $T_{i,1} = \log\{\theta(\boldsymbol{s}_i)\}$, $T_{i,2} = \log\{1 - \theta(\boldsymbol{s}_i)\}$, $C_i(\theta(\boldsymbol{s}_i)) = 0$, and $B_i(\boldsymbol{\theta}(N_i); \boldsymbol{\lambda}) = \log[\Gamma\{A_{i,1}(\boldsymbol{\theta}(N_i); \boldsymbol{\lambda}) + 1\}\,\Gamma\{A_{i,2}(\boldsymbol{\theta}(N_i); \boldsymbol{\lambda}) + 1\}] - \log[\Gamma\{A_{i,1}(\boldsymbol{\theta}(N_i); \boldsymbol{\lambda}) + A_{i,2}(\boldsymbol{\theta}(N_i); \boldsymbol{\lambda}) + 2\}]$. The natural parameter functions must satisfy $-1 < A_{i,k}(\boldsymbol{\theta}(N_i); \boldsymbol{\lambda}) < \infty$; $i = 1, \ldots, n$; $k = 1, 2$. Subject to suitable restrictions ensuring values in these ranges, any of the three propositions of Section 3 may be used to form valid models.

For describing positive spatial dependence, a model formed from Proposition 3 is the most useful of three possibilities. Furthermore, by defining

$$\begin{aligned} A_{i,1}(\boldsymbol{\theta}(N_i);\, \boldsymbol{\lambda}) &= \alpha_1 - \eta \sum_{j \in N_i} \log\{1 - \theta(\boldsymbol{s}_j)\} \\ A_{i,2}(\boldsymbol{\theta}(N_i);\, \boldsymbol{\lambda}) &= \alpha_2 - \eta \sum_{j \in N_i} \log\{\theta(\boldsymbol{s}_j)\}, \end{aligned} \tag{30}$$

we reduce $\boldsymbol{\lambda}$ to $(\alpha_1, \alpha_2, \eta)$. To ensure integrability of the conditional density functions, we require $\eta \geq 0$ and $\alpha_k > -1$, $k = 1, 2$. To see that this model reflects positive spatial dependence, consider the conditional expectation for a particular random variable $\theta(\boldsymbol{s}_i)$ as a function of $\kappa_{i,1} \equiv \eta \sum_{j \in N_i} \log\{\theta(\boldsymbol{s}_j)\}$ and $\kappa_{i,2} \equiv \eta \sum_{j \in N_i} \log\{1 - \theta(\boldsymbol{s}_j)\} : E\{\theta(\boldsymbol{s}_i)|\boldsymbol{\theta}(N_i)\} = (\alpha_1 - \kappa_{i,2} + 1)/(\alpha_1 - \kappa_{i,2} + \alpha_2 - \kappa_{i,1} + 2)$, which is increasing in $\kappa_{i,1}$ and decreasing in $\kappa_{i,2}$. Large values of the elements in $\{\theta(\boldsymbol{s}_j) : j \in N_i\}$ increase $\kappa_{i,1}$ and decrease $\kappa_{i,2}$ and, thus, increase $E\{\theta(\boldsymbol{s}_i)|\boldsymbol{\theta}(N_i)\}$.

Using (8) and (9), this spatial model with beta conditionals has a joint density given by (2) with $\boldsymbol{\lambda} \equiv (\alpha_1, \alpha_2, \eta)$ and, modulo an additive constant not depending on $\boldsymbol{\lambda}$,

$$\begin{aligned} Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \sum_{1 \leq i \leq n} & [\alpha_1 \log\{\theta(\boldsymbol{s}_i)\} + \alpha_2 \log\{1 - \theta(\boldsymbol{s}_i)\}] \\ & - \eta \sum_{1 \leq i < j \leq n} [\log\{\theta(\boldsymbol{s}_i)\} \log\{1 - \theta(\boldsymbol{s}_j)\} + \log\{1 - \theta(\boldsymbol{s}_i)\} \log\{\theta(\boldsymbol{s}_j)\}]. \end{aligned} \tag{31}$$

5.3. MCML estimation

We write the spatial beta-binomial data model as in (16) with

$$\begin{aligned} Q_1(\boldsymbol{y}|\boldsymbol{\theta}) &= \exp\Big[\sum_{i=1}^{n} y(\boldsymbol{s}_i) \log\{\theta(\boldsymbol{s}_i)\} - (m(\boldsymbol{s}_i) - y(\boldsymbol{s}_i)) \log\{1 - \theta(\boldsymbol{s}_i)\}\Big], \\ k_1(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log\{m(\boldsymbol{s}_i)!\} - \log\{y(\boldsymbol{s}_i)!\} - \log\{(m(\boldsymbol{s}_i) - y(\boldsymbol{s}_i))!\}, \end{aligned} \tag{32}$$

and the function $Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is given in equation (31). Notice from (31) and (32) that, considered as a function of $\boldsymbol{\theta}$, $\exp\{Q_1(\boldsymbol{y}|\boldsymbol{\theta}) + Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda}) - k_1(\boldsymbol{\theta})\}$ has the same form as $Q_0(\boldsymbol{\theta}|\boldsymbol{\lambda})$, with the value of α_1 replaced by $\alpha_1 + y(\boldsymbol{s}_i)$ and the value of α_2 replaced by $\alpha_2 + m(\boldsymbol{s}_i) - y(\boldsymbol{s}_i)$, illustrating (20). Thus, to generate values from $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$, from which the sample moments $\{\hat{\mu}_0(\boldsymbol{s}_i), \hat{\mu}_1(\boldsymbol{s}_i), \hat{\sigma}_0^2(\boldsymbol{s}_i), \hat{\sigma}_1^2(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ are computed, Gibbs samples with the same form of conditional distributions may be used. In addition, the individual densities appearing in $m_0(\boldsymbol{\theta}|\psi_0)$ in (19) may be of the same form as those appearing in $m_1(\boldsymbol{\theta}|\psi_1)$ in (21). Here, beta densities were used throughout so that the sampling distributions used in evaluation of the Monte Carlo approximation (18) and its derivatives were, for $\ell = 0, 1$,

$$m_\ell(\boldsymbol{\theta}|\psi_\ell) = \frac{1}{k^*(\psi_\ell)} \exp\Big[\sum_{i=1}^{n} (\psi_{\ell,i,1} \log\{\theta(\boldsymbol{s}_i)\} + \psi_{\ell,i,2} \log\{1 - \theta(\boldsymbol{s}_i)\})\Big], \tag{33}$$

where $k^*(\psi_\ell) = \prod_{i=1}^{n} \Gamma(\psi_{\ell,i,1})\,\Gamma(\psi_{\ell,i,2})/\Gamma(\psi_{\ell,i,1} + \psi_{\ell,i,2})$. In (33), the parameters $\{\psi_{\ell,i} : i = 1, \ldots, n\}$ are chosen to match the first two sample moments at each location, $\{\hat{\mu}_\ell(\boldsymbol{s}_i), \hat{\sigma}^2_\ell(\boldsymbol{s}_i) : i = 1, \ldots, n\}$, that result from samples of size $200,000$ generated via the ℓth Gibbs sampler, for $\ell = 0, 1$.

Using $\boldsymbol{\lambda} \equiv (\alpha_1, \alpha_2, \eta)$ and importance sampling distributions as given in (33), the MCML estimation procedure of Section 4 was carried out on the forest-health data of Table 1. Results of this procedure are presented in Table 2. The starting value of $\boldsymbol{\lambda}^{(0)} = (3.582,\ 5.774,\ 3.733)$ was determined by maximizing a first approximation of the log likelihood formed from Laplace approximations to each of the two integrals of (17). Details of this procedure are contained in Lee (1997). The starting value $\boldsymbol{\lambda}^{(0)}$ was used to obtain sampling distributions, and values $\{\theta_r^{(1)} : r = 1, \ldots, M\}$ and $\{\theta_r^{(0)} : r = 1, \ldots, M\}$ were generated from these distributions using $M = 800,000$. The resulting Monte Carlo approximation (18) to the log likelihood, evaluated at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(0)}$, had the value -210.6008. Maximization of the Monte Carlo log likelihood in $\boldsymbol{\lambda}$, using 4 iterations of a numerical Newton-Raphson algorithm with convergence criterion $\delta_N = 10^{-6}$ (see Appendix A), yielded the new value $\boldsymbol{\lambda}^{(1)} = (4.118,\ 6.549,\ 4.484)$; the resulting Monte Carlo log likelihood value was $L_M^{(1)}(\boldsymbol{\lambda}^{(1)}) = -210.5746$. New sampling distributions were chosen using the value $\boldsymbol{\lambda}^{(1)}$, samples of size $800,000$ were again generated from these sampling distributions, and so forth. In this case, the MCML convergence criterion of $\xi_M = 0.005$ (see Appendix A) was met for $L_M^{(3)}(\boldsymbol{\lambda}^{(3)}) - L_M^{(3)}(\boldsymbol{\lambda}^{(2)}) = 0.0037$, where $\boldsymbol{\lambda}^{(3)} = (4.121, 6.524, 4.489)$. The MCMLE is thus taken as $\hat{\boldsymbol{\lambda}}_M = \boldsymbol{\lambda}^{(3)}$, and the Monte Carlo log likelihood value is $L_M(\hat{\boldsymbol{\lambda}}_M) = -210.5656$. For comparison, a traditional beta-binomial model, having the same conditional binomial model as used here and a mixing distribution constructed from independent and identical beta distributions, yielded a maximized log likelihood of -213.2654.

Table 2. Iteration history for estimation of parameters in the spatial beta-binomial model.

Cycle		Newton Raphson Iterations	Value of $\boldsymbol{\lambda}$	Value of $L_M(\boldsymbol{\lambda})$
1	Start		(3.582, 5.774, 3.733)	−210.6008
	End	4	(4.118, 6.549, 4.484)	−210.5746
2	Start		(4.118, 6.549, 4.484)	−210.5525
	End	2	(4.127, 6.433, 4.548)	−210.5458
3	Start		(4.127, 6.433, 4.548)	−210.5693
	End	2	(4.121, 6.524, 4.489)	−210.5656

5.4. Inference and interpretation of spatial dependence

Using $\hat{\boldsymbol{\lambda}}_M \equiv (\hat{\alpha}_{1,M},\, \hat{\alpha}_{2,M},\, \hat{\eta}_M)$, the estimated covariance matrix in (27) is

$$\begin{pmatrix} 0.021 & 0.033 & 0.039 \\ 0.033 & 0.055 & 0.061 \\ 0.039 & 0.061 & 0.078 \end{pmatrix}.$$

The estimated inverse negative Hessian matrix $\hat{H}^{-1} = (-\bigtriangledown^2 L_M(\hat{\boldsymbol{\lambda}}_M))^{-1}$ is

$$\begin{pmatrix} 5.165 & 7.250 & 6.793 \\ 7.250 & 12.023 & 7.372 \\ 6.793 & 7.372 & 13.657 \end{pmatrix}.$$

The estimated value of $tr(H^{-1}\Sigma H^{-1})/tr(H^{-1})$ as given in condition (28) is $0.005 < 0.01$, and hence we are willing to use $\hat{H}^{-1}$ as an approximation to the covariance matrix of maximum likelihood estimates. Estimated correlations among parameter estimators are quite high, with $r(\alpha_1, \alpha_2) = 0.92$, $r(\alpha_1, \eta) = 0.81$ and $r(\alpha_2, \eta) = 0.56$. These correlations are reflected in the joint confidence regions shown in Figure 2, where the three pairwise approximate 90% confidence ellipsoids formed from the Wald statistic (e.g., Serfling (1980, p.157)) are presented. In the lower right panel of Figure 2, univariate 90% marginal confidence intervals are also presented. There is evidence of spatial dependence in a comparison of the maximized log likelihoods for an independence model with 2 parameters (-213.2654) and the spatial beta-binomial model with 3 parameters (-210.5656); a standard likelihood ratio test results in a test statistic of 5.3996 and a nominal p-value of 0.0201. Figure 2 also shows that there is some evidence for positive spatial dependence, although the regions and interval for η include the value $\eta = 0$. Formal likelihood inference based on asymptotic distributions may not be entirely satisfactory for use with this spatial mixture model, and would tend to be overly conservative in favor of the hypothesis of no dependence. The true significance levels of such procedures is an area of investigation that we do not pursue here.

To investigate the level of spatial dependence represented by this fitted model, we examined sample correlations obtained from simulation and calculation. It is straightforward to show that

$$corr\{Y(\boldsymbol{s}_i),\, Y(\boldsymbol{s}_j)\} = \frac{m(\boldsymbol{s}_i)m(\boldsymbol{s}_j)\,\mathrm{Cov}\,\{\theta(\boldsymbol{s}_i), \theta(\boldsymbol{s}_j)\}}{v(\boldsymbol{s}_i)\,v(\boldsymbol{s}_j)},$$

where for $i = 1, \ldots, n$, $v(\boldsymbol{s}_i) \equiv \left[m(\boldsymbol{s}_i)E\{\theta(\boldsymbol{s}_i) - \theta(\boldsymbol{s}_i)^2\} + m(\boldsymbol{s}_i)^2\mathrm{Var}\,\{\theta(\boldsymbol{s}_i)\}\right]^{1/2}$. First and second moments of $\boldsymbol{\theta}$ were estimated from a Gibbs sample of size $2,000$

from $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$, using $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_M$. The results are summarized in Table 3, in which the smallest, largest, and median pairwise correlations are presented for both values of the mixing random variables $\{\theta(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ and the observable random variables $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$.

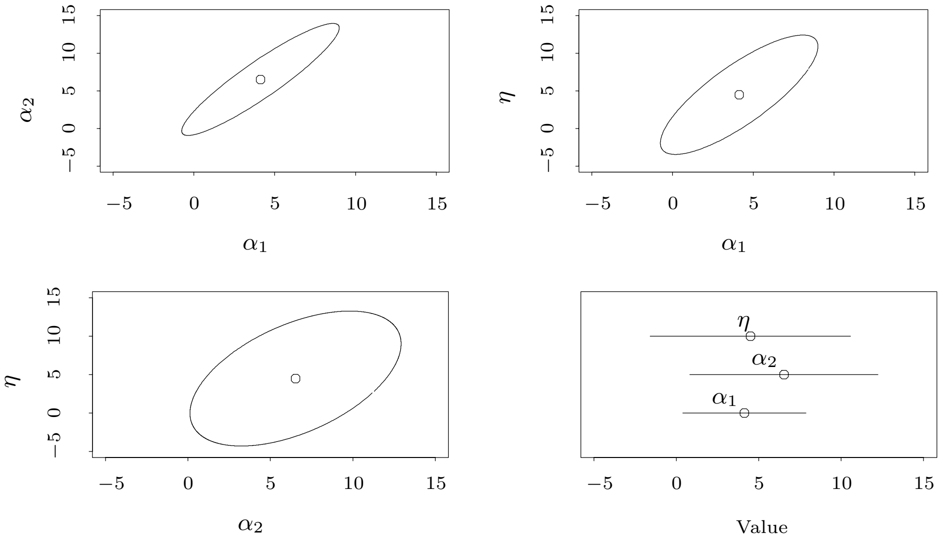

Figure 2. Approximate Confidence Regions and Intervals. Joint 90% confidence regions are shown for pairs of parameters in the spatial beta-binomial mixture model. Marginal 90% intervals are given in the lower right panel of the figure.

Table 3. Correlations (from simulations) for latent ($\boldsymbol{\theta}$) and observable ($\boldsymbol{Y}$) random variables.

			Correlation	
Site i	Site j	Distance (km)	$\{\theta(\boldsymbol{s}_i), \theta(\boldsymbol{s}_j)\}$	$\{y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)\}$
20	22	55.9	0.258 (L)	0.114
29	30	26.9	0.498 (M)	0.170
15	16	26.5	0.588 (H)	0.210
6	9	48.3	0.306	0.060 (L)
27	28	28.1	0.532	0.169 (M)
19	32	28.5	0.528	0.260 (H)

The correlations among the $\{(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)) : i < j;\, i = 1, \ldots, n\}$ are weaker than those among the $\{\, (\theta(\boldsymbol{s}_i), \theta(\boldsymbol{s}_j)) : i < j;\, i = 1, \ldots, n\}$, due to the presence of the binomial variability in the former. Further, the smallest, largest, and median values occur at different pairs of locations for these random variables because of the influence of the $\{m(\boldsymbol{s}_i) : i = 1, \ldots, n\}$, the number of birch trees at each location. Overall, the estimated spatial dependence using this spatial beta-binomial mixture model is seen to be substantial. Although correlations among pairs of observed random variables $\boldsymbol{Y}$ are indicative of the type of 'weak' spatial dependence typically expected in field studies, the estimated correlations among pairs of latent random variables $\boldsymbol{\theta}$ implies a much stronger spatial dependence in the latent spatial process.

6. Discussion

Although we have not addressed prediction, the issue of predicting the latent process $\boldsymbol{\theta}$ is an important one. By modeling $\boldsymbol{\theta}$ directly (rather than a transformation of it) through spatial mixture models, there is no need for 'correcting' back-transformed predictions to render them interpretable on a meaningful scale. With squared error loss, the optimal predictor for the latent process at a location $\boldsymbol{s}_0$ is $\hat{\theta}(\boldsymbol{s}_0) \equiv E[\theta(\boldsymbol{s}_0)|\boldsymbol{y}]$, where $\boldsymbol{y} \equiv \{y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$. Assuming that the location $\boldsymbol{s}_0$ belongs to the same model structure as the observed locations, the form of the predictor $\hat{\theta}(\boldsymbol{s}_0)$ is available from the conditional specifications used to formulate the joint mixing distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$. In addition, for the data models considered, $E\{Y(\boldsymbol{s}_0)|\theta(\boldsymbol{s}_0)\}$ will be a function of $\theta(\boldsymbol{s}_0)$, say $t(\theta(\boldsymbol{s}_0))$. This allows the development of an optimal predictor for the observable process $\boldsymbol{Y}$, which may be useful for cross-validation purposes. Such a predictor is $\hat{y}(\boldsymbol{s}_0) \equiv E[Y(\boldsymbol{s}_0)|\boldsymbol{y}] = E\left[E\{Y(\boldsymbol{s}_0)|\theta(\boldsymbol{s}_0)\} \mid \boldsymbol{y}\right] = E\left[t(\theta(\boldsymbol{s}_0)) \mid \boldsymbol{y}\right]$. Markov Chain Monte Carlo methods can be used to sample from the conditional density $p(\boldsymbol{\theta}|\boldsymbol{y})$ and thus allow evaluation of this predictor.

While the methodology presented in this article is computationally intensive, and despite a number of open questions, we have demonstrated how non-Gaussian latent spatial processes can be modeled, estimated, and used to investigate spatial dependence. This ability will increase the relevance of statistical models for many problems, including those in the environmental sciences and epidemiology.

Acknowledgements

This research was supported by the U.S. Environmental Protection Agency under Assistance Agreement CR 822919-01-0, the National Science Foundation under grant DMS 9816630 (Kaiser), and the Office of Naval Research under grant N00014-99-1-0001 (Cressie). The Centre de Recherche en Economie et Statistique at INSEE (France) provided support to Cressie during the completion of this paper.

Appendix A. Estimation Algorithm

Choose a starting value $\boldsymbol{\lambda}^{(0)}$. Set $L_M^{(0)}(\boldsymbol{\lambda}^{(0)}) = -\infty$ and begin with cycle $q = 1$.

1. Choose sampling distributions $m_0(\boldsymbol{\theta}|\psi_0^{(q)})$ and $m_1(\boldsymbol{\theta}|\psi_1^{(q)})$ as described in Section 4.1. Select Monte Carlo samples of size M from each distribution; in the example of Section 5 we use $M = 800,000$.
2. Maximize $L_M^{(q)}(\boldsymbol{\lambda})$ in $\boldsymbol{\lambda}$ by taking $\boldsymbol{\lambda}^{(w,q)} = \boldsymbol{\lambda}^{(q)}$ for $w = 0$.
   (a) Compute the Monte Carlo log likelihood (17), its first derivatives (22), and its second derivatives (23). Denote the Monte Carlo log likelihood as $L_M^{(q)}(\boldsymbol{\lambda}^{(w,q)})$, the vector of first derivatives as $\bigtriangledown L_M^{(q)}(\boldsymbol{\lambda}^{(w,q)})$ and the matrix of second derivatives as $H_M^{(q)}(\boldsymbol{\lambda}^{(w,q)})$.
   (b) Let
   $$\boldsymbol{\lambda}^{(w+1,q)} = \boldsymbol{\lambda}^{(w,q)} - [H_M^{(q)}(\boldsymbol{\lambda}^{(w,q)})]^{-1} \bigtriangledown L_M^{(q)}(\boldsymbol{\lambda}^{(w,q)}).$$

   Iterate steps 2(a) and 2(b) (over w) until $L_M^{(q)}(\boldsymbol{\lambda}^{(w+1,q)}) - L_M^{(q)}(\boldsymbol{\lambda}^{(w,q)}) \leq \delta_N$; in the example of Section 5, we use $\delta_N = 10^{-6}$. Each subsequent evaluation of $L_M^{(q)}(\cdot)$, $\bigtriangledown L_M^{(q)}(\cdot)$, and $H_M^{(q)}(\cdot)$ uses the same Monte Carlo samples selected in step 1. Let the value of $\boldsymbol{\lambda}$ that maximizes $L_M^{(q)}(\boldsymbol{\lambda})$ be denoted as $\boldsymbol{\lambda}^{(q)}$ and the value of the Monte Carlo log likelihood at this value be denoted as $L_M^{(q)}(\boldsymbol{\lambda}^{(q)})$.
3. If $L_M^{(q)}(\boldsymbol{\lambda}^{(q)}) - L_M^{(q)}(\boldsymbol{\lambda}^{(q-1)}) \leq \xi_M$, declare $\boldsymbol{\lambda}^{(q)}$ to be the MCMLE of $\boldsymbol{\lambda}$. Otherwise, update q to $(q+1)$ and return to step 1; in the example of Section 5 we used $\xi_M = 0.005$.

Appendix B. Asymptotic Normality of $\bigtriangledown L_M(\hat{\boldsymbol{\lambda}})$

Begin with the notation of Section 4.2 and let $D_{\ell,r}^{(k)} \equiv \frac{\partial Q_0(\boldsymbol{\theta}_r^{(\ell)}|\boldsymbol{\lambda})}{\partial \lambda_k} D_{\ell,r}$; $\ell = 0, 1$. Averaging over $r = 1, \ldots, M$ yields $\bar{D}_\ell^{(k)}$. Then using (18), components of $\bigtriangledown L_M(\boldsymbol{\lambda})$ may be written as Monte Carlo approximations to (22), for $k = 1, \ldots, p$,

$$\frac{\partial L_M(\boldsymbol{\lambda})}{\partial \lambda_k} = \frac{\bar{D}_1^{(k)}}{\bar{D}_1} - \frac{\bar{D}_0^{(k)}}{\bar{D}_0}. \tag{B.1}$$

If the following integrals exist for a particular model, we define for $\ell = 0, 1$ and $k = 1, \ldots, p$,

$$\begin{aligned} \mu_\ell^{(k)} &\equiv E\{D_{\ell,r}^{(k)}\} = \int_\Theta \frac{\partial Q_0(\boldsymbol{\theta}_r^{(\ell)}, \boldsymbol{\lambda})}{\partial \lambda_k} I_\ell(\boldsymbol{y}, \boldsymbol{\theta}_r^{(\ell)}, \boldsymbol{\lambda})\, d\boldsymbol{\theta}_r^{(\ell)}, \\ \mu_\ell^{(k,h)} &\equiv E\{D_{\ell,r}^{(k)}\, D_{\ell,r}^{(h)}\} = \int_\Theta \frac{\partial Q_0(\boldsymbol{\theta}_r^{(\ell)}, \boldsymbol{\lambda})}{\partial \lambda_k} \frac{\partial Q_0(\boldsymbol{\theta}_r^{(\ell)}, \boldsymbol{\lambda})}{\partial \lambda_h} \frac{I_\ell^2(\boldsymbol{y}, \boldsymbol{\theta}_r^{(\ell)}, \boldsymbol{\lambda})}{m_\ell(\boldsymbol{\theta}_r^{(\ell)} \mid \psi_\ell)}\, d\boldsymbol{\theta}_r^{(\ell)}, \\ \mu_\ell^{(0,k)} &\equiv E\{D_{\ell,r}\, D_{\ell,r}^{(k)}\} = \int_\Theta \frac{\partial Q_0(\boldsymbol{\theta}_r^{(\ell)}, \boldsymbol{\lambda})}{\partial \lambda_k} \frac{I_\ell^2(\boldsymbol{y}, \boldsymbol{\theta}_r^{(\ell)}, \boldsymbol{\lambda})}{m_\ell(\boldsymbol{\theta}_r^{(\ell)} \mid \psi_\ell)}\, d\boldsymbol{\theta}_r^{(\ell)}, \end{aligned} \tag{B.2}$$

where it should be noted that I_0 has no argument in $\boldsymbol{y}$.

Let Γ_ℓ be the $(p+1)\times(p+1)$ matrix,

$$\Gamma_\ell = \begin{pmatrix} V_\ell & \mu_\ell^{(0,1)} & \mu_\ell^{(0,2)} \ldots, \mu_\ell^{(0,p)} \\ \mu_\ell^{(0,1)} & \mu_\ell^{(1,1)} & \mu_\ell^{(1,2)} \ldots, \mu_\ell^{(1,p)} \\ & & \vdots \\ \mu_\ell^{(0,p)} & \mu_\ell^{(1,p)} & \mu_\ell^{(2,p)} \ldots, \mu_\ell^{(p,p)} \end{pmatrix}, \quad \ell = 0,1, \tag{B.3}$$

where recall that $V_\ell \equiv \mathrm{Var}\,(D_{\ell,r})$.

Then the $2(p+1)$ vector $\tilde{D} \equiv (\bar{D}_0, \bar{D}_0^{(1)}, \ldots, \bar{D}_0^{(p)}, \bar{D}_1, \bar{D}_1^{(1)}, \ldots, \bar{D}_1^{(p)})^T$ is asymptotically normal with mean $\tilde{E} \equiv (E_0, \mu_0^{(1)}, \ldots, \mu_0^{(p)}, E_1, \mu_1^{(1)}, \ldots, \mu_1^{(p)})^T$ and variance

$$\frac{1}{M}\Gamma = \frac{1}{M}\begin{pmatrix} \Gamma_0 & \mathbf{0} \\ \mathbf{0} & \Gamma_1 \end{pmatrix}.$$

Define the transformation $g(\tilde{D}) = (g_1(\tilde{D}), \ldots, g_p(\tilde{D}))$ where, for $k = 1, \ldots, p$, $g_k(\tilde{D})$ gives the kth element of $\bigtriangledown L_M(\boldsymbol{\lambda})$ as in expression (B.1). Hence,

$$g(\tilde{E}) = \Big(\frac{\mu_1^{(1)}}{E_1} - \frac{\mu_0^{(1)}}{E_0}, \ldots, \frac{\mu_p^{(1)}}{E_1} - \frac{\mu_0^{(p)}}{E_0}\Big)^T = \Big(\frac{\partial L(\boldsymbol{\lambda} \mid \boldsymbol{y})}{\partial \lambda_1}, \ldots, \frac{\partial L(\boldsymbol{\lambda} \mid \boldsymbol{y})}{\partial \lambda_p}\Big)^T.$$

Also, we have that, for $\ell = 0,1$, and $k = 1, \ldots, p$,

$$G_\ell^{(0,k)} = \frac{\partial g_k(\tilde{D})}{\partial \bar{D}_\ell}\Big|_{\tilde{D}=\tilde{E}} = \frac{(-1)^\ell \mu_\ell^{(k)}}{E_\ell^2} \quad G_\ell^{(k,k)} = \frac{\partial g_k(\tilde{D})}{\partial \bar{D}_\ell^{(k)}}\Big|_{\tilde{D}=\tilde{E}} = \frac{(-1)^{\ell+1}}{E_\ell}. \tag{B.4}$$

Using the two $(p\times(p+1))$ matrices $G_\ell = [(G_\ell^{(0,1)}, \ldots, G_\ell^{(0,p)})^T : \mathrm{diag}(G_\ell^{(1,1)}, \ldots, G_\ell^{(p,p)})]$ to construct the $(p \times (2p+2))$ matrix $G = [G_1 : G_0]$; $\ell = 0,1$,

$$M^{1/2}[\bigtriangledown L_M(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda})] \xrightarrow{d} N(0, \Sigma) \quad \text{as } M \to \infty, \tag{B.5}$$

where $\Sigma = G\Gamma G^T$. The elements of Γ and G in expression (B.5) may be estimated by replacing the expectation operators in (B.2) by averages over sampled values $\boldsymbol{\theta}_r^{(\ell)}, r = 1, \ldots, M$, $\ell = 0, 1$.

Thus, because of our approach using independent pseudo-models, Theorem 7 of Geyer (1994) reduces from establishing Markov chain central limit theorems to simply insuring that the integrals in (B.2) exist. For many choices of m_0 and m_1 the existence of these integrals is easy to establish.

Appendix C. Proof of Proposition 4

The proof is a modification of the proof of Theorem 25.5 in Billingsley (1995).

Let $\{X \leq b\} \equiv \{\omega : X(\omega) \leq b\}$ for singletons ω of the probability space common to $X_M^{(n)}$ and $Y^{(n)}$, take $\epsilon > 0$, and suppose that $(x-\epsilon,\, x+\epsilon)$ is in the continuity set of F. Then

$$\begin{aligned} Pr\left[X_M^{(n)} \leq x\right] &= Pr\left[\left(X_M^{(n)} \leq x\right) \cap (|Y^{(n)} - X_M^{(n)}| \geq \epsilon)\right] \\ &\quad + Pr\left[\left(X_M^{(n)} \leq x\right) \cap (|Y^{(n)} - X_M^{(n)}| < \epsilon)\right]. \qquad \text{(C.1)} \end{aligned}$$

The first right hand side (rhs) term of (C.1) is bounded above by $Pr[|Y^{(n)} - X_M^{(n)}| \geq \epsilon]$. For the second rhs term, $\{(X_M^{(n)} \leq x) \cap (|Y^{(n)} - X_M^{(n)}| < \epsilon)\} \subset \{(X_M^{(n)} \leq x) \cap (Y^{(n)} - X_M^{(n)} < \epsilon)\} \subset \{Y^{(n)} < x + \epsilon\}$, hence the term is bounded above by $Pr[Y^{(n)} < x + \epsilon]$. Lower bounds may be developed in a similar manner: $Pr[X_M^{(n)} \leq x] \geq Pr[Y^{(n)} \leq x - \epsilon] - Pr[|Y^{(n)} - X_M^{(n)}| \geq \epsilon]$; $Pr[X_M^{(n)} \leq x] \leq Pr[Y^{(n)} \leq x + \epsilon] + Pr[|Y^{(n)} - X_M^{(n)}| \geq \epsilon]$. Let $M \to \infty$ followed by $n \to \infty$, and recall that $Y^{(n)} \overset{d}{\to} Y \sim F$. Then $F(x-\epsilon) - \lim_{n\to\infty} \liminf_{M\to\infty} Pr[|Y^{(n)} - X_M^{(n)}| \geq \epsilon] \leq \liminf_{n\to\infty} \liminf_{M\to\infty} Pr[X_M^{(n)} \leq x] \leq \limsup_{n\to\infty} \limsup_{M\to\infty} Pr[X_M^{(n)} \leq x] \leq \lim_{n\to\infty} \limsup_{M\to\infty} Pr[|Y^{(n)} - X_M^{(n)}| \geq \epsilon] + F(x+\epsilon)$. From the assumed condition, $F(x-\epsilon) \leq \liminf_{n\to\infty} \liminf_{M\to\infty} Pr[X_M^{(n)} \leq x] \leq \limsup_{n\to\infty} \limsup_{M\to\infty} Pr[X_M^{(n)} \leq x] \leq F(x+\epsilon)$. Now, let $\epsilon \to 0$; since x is a continuity point of F, we obtain the desired result.

References

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36**, 192-225.

Billingsley, P. (1995). *Probability and Measure.* 3rd edition. Wiley, New York.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 1313-1321.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671-681.

Conkling, B. L. and Byers, G. E. (Ed.) (1992). *Forest Health Monitoring Field Methods Guide*, EPA/600/X-92/073. U.S. Environmental Protection Agency, Las Vegas.

Cressie, N. (2000). Geostatistical methods for mapping environmental exposures. In *Spatial Epidemiology: Methods and Applications* (Edited by P. Elliot, J. C. Wakefield, N. Best and D. J. Briggs), 185-204. Oxford University Press, Oxford.

Cressie, N. and Chan, N. H. (1989). Spatial modeling of regional variables. *J. Amer. Statist. Assoc.* **84**, 393-401.

Cressie, N. and Lele, S. (1992). New models for Markov random fields. *J. Appl. Probab.* **29**, 877-884.

DeOliveira, V., Kedem, B. and Short, D. A. (1997). Bayesian prediction of transformed Gaussian random fields. *J. Amer. Statist. Assoc.* **92**, 1422-1433.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Appl. Statist.* **47**, 299-350.

Gelfand, A. E. and Carlin, B. P. (1993). Maximum-likelihood estimation for constrained- or missing-data models. *Can. J. Statist.* **21**, 303-311.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc. Ser. B* **56**, 261-274.

Geyer, C. J. (1996). Estimation and optimization of functions. In *Markov Chain Monte Carlo in Practice* (Edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), 241-258. Chapman and Hall, London.

Geyer, C. J. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.* **21**, 359-373.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54**, 657-699.

Gleason, H. A. and Cronquist, A. (1963). *Manual of Vascular Plants of Northeastern United States and Adjacent Canada.* D. Van Nostrand Co, New York.

Green, P. J. (1992). Discussion of "Constrained Monte Carlo Maximum Likelihood for Dependent Data". *J. Roy. Statist. Soc. Ser. B* **54**, 683-684.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *J. Amer. Statist. Assoc.* **93**, 1099-1111.

Kaiser, M. S. and Cressie, N. (2000). The construction of multivariate distributions from Markov random fields. *J. Multivariate Anal.* **73**, 199-220.

Knoor-Held, L. and Besag, J. E. (1998). Modelling risk from a disease in time and space. *Statist. Medicine* **17**, 2045-2060.

Lee, J. (1997). Specification of dependence structures and simulation-based estimation for conditionally specified statistical models. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Lett.* **19**, 451-458.

Meng, X. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6**, 831-860.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *J. Roy. Statist. Soc. Ser. B* **56**, 3-48.

Ogata, Y. (1996). Evaluation of spatial Bayesian models – two computational methods. *J. Statist. Plann. Inference* **51**, 1-18.

Pye, J. M. (1988). Impact of ozone on the growth and yield of trees: A review. *J. Environ. Qual.* **17**, 347-360.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Stern, H. and Cressie, N. (1999). Inference for extremes in disease mapping. In *Disease Mapping and Risk Assessment for Public Health* (Edited by A. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J.-F. Viel, and R. Bertollini), 63-84. Wiley, Chichester.

Tallent-Halsell, N.G. (Ed.) (1994). *Forest Health Monitoring. 1994 Field Methods Guide.* U.S. Environmental Protection Agency, Washington, D.C.

Torrie, G. M. and Valleau, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.* **23**, 187-199.

Department of Statistics, Snedecor Hall, Iowa State University, Ames, Iowa 50011-1210, U.S.A.

E-mail: mskaiser@iastate.edu

Department of Statistics, 1958 Neil Avenue, 404 Cockins Hall, The Ohio State University, Columbus, Ohio 43210-1247, U.S.A.

E-mail: ncressie@stat.ohio-state.edu

Department of Statistics, Snedecor Hall, Iowa State University, Ames, Iowa 50011-1210, U.S.A.

(Received March 2000; accepted May 2001)