Statistica Sinica **25** (2015), 403-421
doi:http://dx.doi.org/10.5705/ss.2013.281w

ASYMPTOTICALLY EFFICIENT PARAMETER ESTIMATION IN HIDDEN MARKOV SPATIO-TEMPORAL RANDOM FIELDS

Tze Leung Lai and Johan Lim

Stanford University and Seoul National University

Abstract: Estimation of the parameters of Markov random field models for spatial and temporal data arises in many applications. There are computational and statistical challenges in developing efficient estimators because of the complexity of the joint distribution of the spatio-temporal models, especially when they involve hidden states that also need to be estimated from the observations. We develop composite likelihood estimators that are analytically and computationally tractable, and show that they are asymptotically efficient under some mild correlation decay assumptions.

Key words and phrases: Block maximum likelihood estimator, composite likelihood, correlation decay, hidden Markov random fields, image analysis, locally asymptotically normal family, ρ-mixing.

1. Introduction

Spatial data that are observed over time arise in many statistical applications to disciplines that range from economics, political science and social networks to neuroscience, earth sciences, engineering, epidemiology and imaging. Spatio-temporal models for these data are inherently complex as they need to reflect the dependence not only among different sites in space but also among different time points. The likelihood functions are often specified up to normalizing constants, which depend on the unknown parameters and are difficult to compute.

Motivated by applications to image analysis, we consider a class of spatio-temporal models that belong to the general framework of hidden Markov random fields. A random field of spatial and temporal data is characterized by a time index $t = 1, \ldots, T$, and a multidimensional spatial index $\omega \in \Omega \subset \mathbb{R}^d$. Let $Y_{t\omega}$ denote the observation at time t and site ω. This represents a noisy distortion of the underlying signal $Z_{t\omega}$. Let $\mathbf{Y}_t = \big(Y_{t\omega},\ \omega \in \Omega\big)$ and $\mathbf{Z}_t = (Z_{t\omega},\ \omega \in \Omega)$. Given $\mathbf{Z}_t$, the $Y_{t\omega}$ are assumed to be conditionally independent. A parametric spatio-temporal random field model is often assumed on the underlying signal $\mathbf{Z}_t$, showing its joint distribution over different sites and how it evolves over time. Because the observations $Y_{t\omega}$ are sampled at a finite collection of sites in

practice, we assume throughout the sequel that Ω is a finite set. Moreover, since we focus on applications to image analysis, we assume that the sites ω belong to a lattice, as in the case of pixels of an image, which we can assume without loss of generality (by rescaling if necessary) to be the integer lattice $\mathbb{Z}^d$. Thus, Ω is the intersection of a bounded region with $\mathbb{Z}^d$. Commonly used models for spatial interaction are Markov random fields, reviewed below in Section 1.1, and those for time evolution assume Markovian dynamics. Because $\mathbf{Y}_t$ instead of the actual signal $\mathbf{Z}_t$ is observed, we have a hidden Markov random field (HMRF), in which parameter estimation poses computational and statistical challenges, as will be explained in Section 1.2 in connection with an overview of methods in the literature to address these challenges.

In Section 2, we propose a block maximum likelihood estimator (MLE) that we show to be asymptotically efficient under certain correlation decay conditions. Section 3 describes implementation details and illustrates how the block MLE works in an example from brain imaging that involves both spatial and temporal data. Some concluding remarks and discussion are given in Section 4.

1.1. Markov random fields for spatial data

In a spatial Markov random field, the sites in Ω are related to one another by a neighborhood system $\mathcal{N} = \big\{N(\omega), \omega \in \Omega\big\}$, where $N(\omega)$ is the set of sites $v \neq \omega$ that are neighbors of ω. For any two sites $\omega, v \in \Omega$, $\omega \sim v$ means that $\omega \in N(v)$ and $v \in N(\omega)$. The set of neighbors of ω is often defined via a metric d, e.g., $N(\omega) = \{v : d(\omega, v) \leq r\big\}$. The pair $\big(\Omega, \mathcal{N}\big)$ defines an undirected graph in the sense that Ω contains the nodes and $\mathcal{N}$ determines the edges between the nodes. In the graph $\big(\Omega, \mathcal{N}\big)$, a clique is defined as a set of sites that consists of either a single site or sites which are neighbors of each other.

For fixed t, the set of random variables $Z_{t\omega}, \omega \in \Omega$, is called a Markov random field, with respect to the neighborhood system $\mathcal{N}$, if the conditional distribution of $Z_{t\omega}$ given $\mathbf{Z}_t - \{Z_{t\omega}\}$ depends only on $\{Z_{tv}, v \in N(\omega)\}$. One can specify the distribution of the Markov random field via a specification of these conditional distributions. Alternatively, one can specify the joint distribution of $\mathbf{Z}_t$ by a density function $f(\mathbf{z})$ that is proportional to $\prod_{c\in\mathcal{C}} \eta_c(\mathbf{z})$, where $\mathcal{C}$ is the collection of all possible cliques. Letting $\eta_c = \exp(-\psi_c)$, $f(\mathbf{z})$ can therefore be written in the form

$$h(\mathbf{z}) = \kappa^{-1} \exp\Big\{ -\sum_{c\in\mathcal{C}} \psi_c(\mathbf{z})\Big\}, \tag{1.1}$$

in which κ is the normalizing constant, and ψ_c is called a potential function. When $\mathbf{Z}_t$ is discrete, the normalizing constant is a sum and the distribution is called a Gibbs distribution. The equivalence between Markov random field and

Gibbs distribution under a positivity condition is known as the Hammersley-Clifford theorem; see Besag (1974). In statistical mechanics, the Gibbs distribution has a density function of the form

$$h(\mathbf{z}) = \frac{1}{\kappa(\beta)} \exp\Big\{ -\beta\Psi(\mathbf{z})\Big\}, \tag{1.2}$$

where $\Psi(\mathbf{z})$ represents the energy, $1/\beta$ represents the temperature, and $\kappa(\beta)$ is called the partition function. In particular, $\Psi(\mathbf{z}) = \sum_{\omega\in\Omega} z_\omega \sum_{v\in N(\omega)} \alpha_{\omega v} z_v$ with $z_\omega \in \{-1, 1\}$ corresponds to the Ising model in the statistical physics of magnetism, spin glass, lattice gas and in applications to neuroscience and image analysis. For the Potts model that is discussed in Section 1.2, $\Psi(\mathbf{z}) = \sum_{\omega\in\Omega} \psi(\omega) + \sum_{\omega\in\Omega} \sum_{v\in N(\omega)} \alpha_{\omega v} \mathrm{I}_{\{z_\omega = z_v\}}$. From the Gibbs specification (1.1) of the density of $\mathbf{Z}_t$, one can retrieve the conditional density of $Z_{t\omega}$ in the conditional distribution specification:

$$\phi\big(z\big|\mathbf{Z}_t - \{Z_{t\omega}\}\big) = \frac{\exp\Big\{ -\sum_{c\in\mathcal{C}} \psi_c(\mathbf{z})\Big\}}{\sum_{\mathbf{z}'} \exp\Big\{ -\sum_{c\in\mathcal{C}} \psi_c(\mathbf{z}')\Big\}}, \tag{1.3}$$

where $\mathbf{z}$ is the vector that has value z at site ω and the value of $\mathbf{Z}_t$ at the other sites, and $\mathbf{z}'$ is a similar vector that does not impose restrictions on the value at site ω. The Ψ and ψ_c in (1.2) and (1.3) are typically specified by a parametric model. For example, the β in the Gibbs distribution (1.2) is a parameter.

Besag (1974, Sec. 4) gives an overview of the *auto-models* for which the cliques $c \in \mathcal{C}$ consist of at most two sites and the energy function has the decomposition

$$\sum_{c\in\mathcal{C}} \psi_c(\mathbf{z}) = \sum_{\{\omega\}\in\mathcal{C}} z_\omega \Psi_1(z_\omega) + \sum_{\{\omega,v\}\in\mathcal{C}} \Psi_2(z_\omega, z_v), \tag{1.4}$$

where Ψ_1 and Ψ_2 are functions that may depend on some unknown parameters. In particular, he considers $\Psi_2(z_\omega, z_v) = \beta_{\omega v} z_\omega z_v$, in which $\beta_{\omega v}$ is an unknown parameter. Because of the computational difficulty in computing the MLE of the parameter vector $\boldsymbol{\theta}$, he proposes to work with the pseudo-likelihood

$$\prod_{\omega\in\Omega} \phi_{\boldsymbol{\theta}}(Z_{t\omega}|\mathbf{Z}_t - \{Z_{t\omega}\}), \tag{1.5}$$

formed by taking the product of the conditional densities (1.3) because "the lattice models under the conditional probability approach yield naturally to a very simple parameter estimation procedure (the coding techniques)" that is described in his Section 6.1; see Besag (1974, p.223).

1.2. Composite likelihood for parameter estimation in HMRFs

We have focused on spatial data at a fixed time t and have assumed the signal $\mathbf{Z}_t$ to be observable on fitting Markov random fields to these data. We now consider the more general setting of hidden Markov random fields (HMRFs) of spatial-temporal data $\mathbf{Y}_t, 1 \le t \le T$, that are related to the signal $\mathbf{Z}_t$ via the conditional densities $g_{\boldsymbol{\theta}}(\cdot|\cdot)$ such that given $\{Z_{t\omega} : 1 \le t \le T, \omega \in \Omega\}$, the $Y_{t\omega}$ are conditionally independent with density function $g_{\boldsymbol{\theta}}(y|Z_{t\omega})$. Markovian transition over time is assumed for $\mathbf{Z}_t$ so that the conditional density $h_{\boldsymbol{\theta}}(\cdot|\mathbf{Z}_{t-1})$ of $\mathbf{Z}_t$ given $\mathbf{Z}_{t-1}$ is of the form (1.1), in which $\psi_c(\mathbf{z})$ is replaced by $\psi_c(\mathbf{z}; \mathbf{Z}_{t-1}, \boldsymbol{\theta})$ and the normalizing constant κ depends on $\mathbf{Z}_{t-1}$ and $\boldsymbol{\theta}$. Thus, the likelihood function has the form

$$f_{\boldsymbol{\theta}}(\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_T) = \sum_{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T} \prod_{t=1}^{T} \Big\{ \Big[\prod_{\omega \in \Omega} g_{\boldsymbol{\theta}}(Y_{t\omega}|z_{t\omega}) \Big] h_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{z}_{t-1}) \Big\}, \qquad (1.6)$$

in which $h_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{z}_{t-1}) = h_{\boldsymbol{\theta}}(\mathbf{z}_1)$ for the case $t = 1$. In many applications such as that in Section 3.2, $g_{\boldsymbol{\theta}}$ depends only on a sub-vector $\boldsymbol{\theta}^{(1)}$ of $\boldsymbol{\theta}$ and $h_{\boldsymbol{\theta}}$ depends on another subvector $\boldsymbol{\theta}^{(2)}$.

The likelihood function (1.6) involves summation over all possible values of $\mathbf{z}_1, \ldots, \mathbf{z}_T$ and computation of the normalizing constant $\kappa(\boldsymbol{\theta}, \mathbf{z}_{t-1})$ in $h_{\boldsymbol{\theta}}(\cdot|\mathbf{z}_{t-1})$ given by (1.1) with $\psi_c(\mathbf{z})$ replaced by $\psi_c(\mathbf{z}; \mathbf{z}_{t-1}, \boldsymbol{\theta})$. The computational task is formidable if $T|\Omega|$ is large. We use $|A|$ to denote the size (number of elements) of a finite set A. Computing the likelihood function (1.6) is only part of computational task for evaluating the MLE that involves also a maximization algorithm. Besides its computational difficulty, the statistical properties of the MLE in HMRFs are relatively unexplored because of the analytical intractability of the likelihood function. To circumvent this difficulty, composite likelihoods have been widely used in place of (1.6).

Varin, Reid, and Firth (2011) describe composite likelihood as a product of component likelihoods, each component of which is a marginal or conditional density. They give an overview of composite likelihood methods, including Besag's pseudo likelihood (1.5) for Markov random fields and its variants such as pairwise likelihoods. Their Section 4.1 points out that "motivation for the use of any version of composite likelihood is computation: to avoid computing, or, in some cases, modelling the joint distribution of a possibly high-dimensional response vector." Lindsay, Yi, and Sun (2011, p.73) point out two basic "first order" properties of the maximum composite likelihood estimator (MCLE). One is that the standard Kullback-Leibler inequality applies to each sub-likelihood, and therefore the maximum of the expected value (under the true parameter $\boldsymbol{\theta}_0$) of the logarithm $\ell(\boldsymbol{\theta})$ of the composite likelihood is attained at $\boldsymbol{\theta}_0$, which they call "Fisher consistent." Another is that under usual regularity conditions, the

estimating equation formed by setting the gradient vector $\dot{\ell}(\boldsymbol{\theta})$ to $\mathbf{0}$ is unbiased in the sense that $E_{\boldsymbol{\theta}_0}(\dot{\ell}(\boldsymbol{\theta}_0)) = \mathbf{0}$. They point out, however, that "the second order properties of likelihood are not possessed by composite likelihood except under special circumstances" and explain why "one cannot count on asymptotic efficiency for the maximum composite likelihood estimators." Okabayashi, Johnson, and Geyer (2011) consider the Potts model and extend Besag's pseudo-likelihood to composite likelihoods that involve more general forms of $\mathcal{C}$ than those in (1.4). Their simulation studies show some improvements over the case where $\mathcal{C}$ only consists of singletons but the MCLEs are not as efficient as the MLEs that they implement by the MCMC algorithm of Swendsen and Wang (1987). However, their MCLE and Besag's maximum pseudo-likelihood estimator "can be calculated exactly" without relying on MCMC approximations.

2. Block Likelihood and Asymptotic Efficiency

In this section, we introduce a new kind of composite likelihood, called the "block likelihood", and show how the block MLE addresses the difficulties with previous MCLEs mentioned in Section 1.2. We then develop an asymptotic theory for the block MLE, establishing its asymptotic normality and efficiency when the HMRF generating the spatial and temporal data satisfies certain correlation decay conditions.

2.1. Parameter estimation via block likelihood

As indicated in the second paragraph of Section 1, Ω is assumed to be the intersection of a bounded region in $\mathbb{R}^d$ with the interger lattice $\mathbb{Z}^d$. Let $n = T \times |\Omega|$. We partition the time set $\{1, \ldots, T\}$ into H_n disjoint subsets of consecutive integers so that the size of each set, except the last one, is $[T/H_n]$. Similarly, we partition Ω into K_n subsets so that each subset has $\big(1 + o(1)\big)|\Omega|/K_n$ points, except for the subset containing boundary points of Ω. In this way, the domain $\{1, \ldots, T\} \times \Omega$ is partitioned into $H_n K_n$ disjoint blocks $\Gamma_{hk}, 1 \le h \le H_n, 1 \le k \le K_n$. For $A = \Gamma_{hk}$, let $\mathbf{Y}_A = \big(Y_{t\omega}, (t, \omega) \in A\big)$. The marginal density $f_{\boldsymbol{\theta}}(\mathbf{Y}_A)$ of $\mathbf{Y}_A$ for each block has the same form as (1.6) except that the block size is $n/(H_n K_n)$ or smaller, and the block likelihood function has logarithm

$$\ell(\boldsymbol{\theta}) = \sum_{h=1}^{H_n} \sum_{k=1}^{K_n} \log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}}). \tag{2.1}$$

The considerable reduction of the sample size, from n for $\log f_{\boldsymbol{\theta}}\big(\mathbf{Y}_1, \ldots, \mathbf{Y}_T\big)$ to no more than $n/(H_n K_n)$ for each summand $\log f_{\boldsymbol{\theta}}\big(\mathbf{Y}_{\Gamma_{hk}}\big)$ of (2.1), makes computation of the block MLE $\widehat{\boldsymbol{\theta}}_n$ feasible by using an EM algorithm that will be described in Section 3.1.

2.2. Asymptotic normality of block MLE

Block likelihood is a composite likelihood that treats the $\mathbf{Y}_A$ as independent random vectors for disjoint blocks A and approximates (1.6) by a product of H_nK_n factors, with the (h,k)th factor being the marginal density of $\mathbf{Y}_{\Gamma_{hk}}$. For $A=\Gamma_{hk}$, let $\mathbf{D}_A=\big(\partial/\partial\boldsymbol{\theta}\big)\log f_{\boldsymbol{\theta}}(\mathbf{Y}_A)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, where $\boldsymbol{\theta}_0$ denote the true parameter vector. For $\mathbf{a},\mathbf{b}\in\{1,\ldots,T\}\times\Omega$, let $\Delta(\mathbf{a},\mathbf{b})=\max_{1\le i\le d+1}|a_i-b_i|$, and define $\Delta_{A,B}=\inf\big\{\Delta(\mathbf{a},\mathbf{b}),\mathbf{a}\in A,\mathbf{b}\in B\big\}$ for subsets A and B of $\{1,\ldots,T\}\times\Omega$. To establish the asymptotic normality of the block MLE, we first prove that $\sum_{1\le h\le H_n}\sum_{1\le k\le K_n}\mathbf{D}_{\Gamma_{hk}}$ is asymptotically normal. This requires the Lindeberg condition, which would be necessary if the $\mathbf{D}_{\Gamma_{hk}}$ were independent, in which case the covariance matrix of the sum would be $-\sum_{1\le h\le H_n,1\le k\le K_n}E_{\boldsymbol{\theta}_0}\big(\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)\big)$, where $\mathbf{D}^2_A(\boldsymbol{\theta})=\big(\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}\big)\log f_{\boldsymbol{\theta}}(\mathbf{Y}_A)$. Accordingly, we make the following assumptions: as $n\to\infty$,

(A1) $-n^{-1}\sum_{h=1}^{H_n}\sum_{k=1}^{K_n}E_{\boldsymbol{\theta}_0}\big[\mathbf{D}^2_{\Gamma_{hk}}\big(\boldsymbol{\theta}_0\big)\big]\to\mathbf{V}$,

(A2) $n^{-1}\sum_{h=1}^{H_n}\sum_{k=1}^{K_n}E_{\boldsymbol{\theta}_0}\big[\|\mathbf{D}_{\Gamma_{hk}}\|^2\mathrm{I}\big(\|\mathbf{D}_{\Gamma_{hk}}\|>\epsilon_n\sqrt{n}\big)\big]\to 0$,

where $\mathbf{V}$ is a positive definite matrix and ϵ_n is sequence of positive numbers such that $\epsilon_n\to 0$. Although the usual Lindeberg condition assumes (A2) for every fixed $\epsilon>0$, this actually implies that we can choose $\epsilon_n\to 0$ slowly enough such that (A2) holds; see Durrett (2005, p.441). For $A=\Gamma_{hk}$, let

$$\mathbf{D}^{(n)}_A=\mathbf{D}_A\mathrm{I}\big(\|\mathbf{D}_A\|\le\epsilon_n\sqrt{n}\big)-E_{\boldsymbol{\theta}_0}\big\{\mathbf{D}_A\mathrm{I}\big(\|\mathbf{D}_A\|\le\epsilon_n\sqrt{n}\big)\big\}. \tag{2.2}$$

Since the $\mathbf{D}_{\Gamma_{hk}}$ are actually dependent random vectors, (A1) and (A2) are not sufficient for the CLT to hold for their sum. We therefore need additional assumptions, which we state in terms of the correlation decay for the truncated vectors $\mathbf{D}^{(n)}_A$ and $\mathbf{D}^{(n)}_B$, in terms of the distance $\Delta_{A,B}$ between the sets $A=\Gamma_{hk}$ and $B=\Gamma_{h'k'}$:

(A3) $\big\|E_{\boldsymbol{\theta}_0}\big(\mathbf{D}^{(n)}_A\otimes\mathbf{D}^{(n)}_B\big)\big\|_\infty\le n\big(H_nK_n\big)^{-1}\rho\big(\Delta_{A,B}\big)$,

(A4) $\big\|E_{\boldsymbol{\theta}_0}\big(\mathbf{D}^{(n)}_A\otimes[\mathbf{D}^{(n)}_B\otimes\mathbf{D}^{(n)}_{\tilde A}\otimes\mathbf{D}^{(n)}_{\tilde B}]\big)\big\|_\infty\le n^2\big(H_nK_n\big)^{-2}\rho\big(\min\{\Delta_{A,B},\Delta_{A,\tilde A},\Delta_{A,\tilde B}\}\big)$,

where $\otimes$ denotes the Kronecker product, the norm $\|\mathbf{M}\|_\infty$ of a matrix $\mathbf{M}$ is the maximum absolute value of the entries of $\mathbf{M}$, and

(A5) $\big\|E_{\boldsymbol{\theta}_0}(\mathbf{D}^{(n)}_A|\mathbf{D}^{(n)}_B:\Delta_{A,B}\ge m)\big\|\le n^{1/2}(H_nK_n)^{-1/2}\rho(m)$, and $\sum_{m=1}^\infty m^{d-1}\rho(m)<\infty$.

Bolthausen (1982) has proved a central limit theorem for stationary random fields under α- and ρ-mixing conditions that motivate (A3)$-$(A5). Instead of

assuming the random fields to be mixing, our approach assumes correlation decay conditions, similar to those of ρ-mixing, directly on the derivatives of the summands of (2.1). Moreover, without requiring the random field to be stationary, we assume (A1) and (A2) to derive the CLT for sums of (2.2) over h and k, when (A3)$-$(A5) are satisfied. These assumptions and their connections to mixing conditions that underlie our use of block likelihood are discussed further in Section 4. In the Appendix, we modify the arguments of Bolthausen (1982) to prove the following result and also discuss (A3)$-$(A5) in the context of mixing random fields.

Theorem 1. *Under* (A1)$-$(A5), $n^{-1/2}\sum_{h=1}^{H_n}\sum_{k=1}^{K_n}\mathbf{D}_{\Gamma_{hk}}(\boldsymbol{\theta}_0)$ *converges in distribution to the normal distribution with mean* $\mathbf{0}$ *and covariance matrix* $\mathbf{V}$ *as* $n\to\infty$.

We use Theorem 1 to derive the asymptotic normality of the block MLE $\widehat{\boldsymbol{\theta}}_n$, assuming $\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})$ to be twice continuously differentiable in a neighborhood of $\boldsymbol{\theta}_0$. By Taylor's theorem,

$$\mathbf{0}=\dot{\ell}_n(\widehat{\boldsymbol{\theta}}_n)=\dot{\ell}_n(\boldsymbol{\theta}_0)+\ddot{\ell}_n(\boldsymbol{\theta}_n)(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}_n$ lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$ and we use $\dot{\ell}_n$ to denote $\partial\ell_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\ddot{\ell}_n$ to denote the Hessian matrix of second derivatives. This can be rewritten in the form

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)=-\Big(n^{-1}\sum_{h=1}^{H_n}\sum_{k=1}^{K_n}\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_n)\Big)^{-1}\sum_{h=1}^{H_n}\sum_{k=1}^{K_n}\frac{\mathbf{D}_{\Gamma_{hk}}(\boldsymbol{\theta}_0)}{\sqrt{n}}. \tag{2.3}$$

Hence, by Theorem 1, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$ has a limiting $\mathcal{N}(\mathbf{0},\mathbf{V}^{-1})$ distribution if it can be shown that

$$-n^{-1}\sum_{h=1}^{H_n}\sum_{k=1}^{K_n}\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_n)\xrightarrow{P}\mathbf{V}. \tag{2.4}$$

If the $\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)$ were independent, then it could follow from (A1) and the law of large numbers that (2.4) holds with $\boldsymbol{\theta}_n$ replaced by $\boldsymbol{\theta}_0$ and under an additional assumption justifying a truncation argument. Following the theory for the independent case, we assume that for every $\epsilon>0$, there exists $\delta>0$ such that

$$\text{(B1)}\ \lim_{n\to\infty}\sum_{h=1}^{H_n}\sum_{k=1}^{K_n}\Big[P_{\boldsymbol{\theta}_0}\Big\{\sup_{\lambda\in\mathrm{B}_\delta(\boldsymbol{\theta}_0)}\Big\|\mathbf{D}^2_{\Gamma_{hk}}(\lambda)-\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)\Big\|_\infty>\epsilon n\big(H_nK_n\big)^{-1}\Big\}$$
$$+n^{-1}E_{\boldsymbol{\theta}_0}\Big\{\big\|\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)\big\|\mathrm{I}\Big(\big\|\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)\big\|_\infty>n\Big)\Big\}\Big]=0,$$

where $\mathrm{B}_\delta(\boldsymbol{\theta})=\{\boldsymbol{\lambda}\in\Theta:\|\boldsymbol{\lambda}-\boldsymbol{\theta}\|<\delta\}$. The first summand in (B1) is for approximating $\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_n)$ by $\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)$ when $\boldsymbol{\theta}_n$ converges to $\boldsymbol{\theta}_0$. To ensure the

consistency of $\widehat{\boldsymbol{\theta}}_n$, we need an assumption which is similar to the first summand in (B1) but with $\log f_{\boldsymbol{\lambda}}(\mathbf{Y}_{\Gamma_{hk}}) - \log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})$ in place of $\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\lambda}) - \mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)$. Specifically, we assume that for every $\epsilon > 0$ there exists $\delta > 0$ such that as $n \to \infty$,

$$\text{(B2)}\ \sum_{h=1}^{H_n}\sum_{k=1}^{K_n}\Big[P_{\boldsymbol{\theta}_0}\Big\{\sup_{\boldsymbol{\lambda}\in \mathrm{B}_\delta(\boldsymbol{\theta})}\Big|\log f_{\boldsymbol{\lambda}}(\mathbf{Y}_{\Gamma_{hk}}) - \log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})\Big| > \epsilon n(H_nK_n)^{-1}\Big\} + n^{-1}E_{\boldsymbol{\theta}_0}\Big\{|\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})|\mathrm{I}\big(|\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})| > n\big)\Big\}\Big] \to 0$$

for all $\boldsymbol{\theta} \in \Theta$. As noted in the third paragraph of Section 1.2, the Kullback-Leibler inequality $n^{-1}\sum_{h,k} E_{\boldsymbol{\theta}_0}\big[\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})\big] \le n^{-1}\sum_{h,k} E_{\boldsymbol{\theta}_0}\big[\log f_{\boldsymbol{\theta}_0}(\mathbf{Y}_{\Gamma_{hk}})\big]$, $\boldsymbol{\theta} \ne \boldsymbol{\theta}_0$, holds for block (and other composite) likelihoods. It will be assumed in the sequel that the inequality is strict in the sense that the left-hand side is bounded away from the right-hand side for all large n. It will also be assumed that Θ is compact. The second summand in (B1) and that in (B2) enable us to truncate $\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)$ and $\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})$ similarly to (2.2) so that we can work with

$$\begin{aligned}
\overline{\mathbf{D}}^2_{\Gamma_{hk},n} &= \mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)\mathrm{I}\left(\left\|\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)\right\|_\infty \le n\right) - E_{\boldsymbol{\theta}_0}\left\{\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)\mathrm{I}\left(\left\|\mathbf{D}^2_{\Gamma_{hk}}(\boldsymbol{\theta}_0)\right\|_\infty \le n\right)\right\},\\
\overline{\ell}_{\Gamma_{hk}}(\boldsymbol{\theta}) &= \log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})\mathrm{I}\big(|\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})| \le n\big)\\
&\quad - E_{\boldsymbol{\theta}_0}\big\{\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})\mathrm{I}\big(|\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})| \le n\big)\big\}.
\end{aligned} \tag{2.5}$$

Since the $\overline{\mathbf{D}}^2_{\Gamma_{hk},n}$ and $\overline{\ell}_{\Gamma_{hk},n}$ are actually dependent, we impose correlation decay conditions analogous to (A3):

$$\text{(B3)}\ \max\Big\{\Big\|E_{\boldsymbol{\theta}_0}\big(\overline{\mathbf{D}}^2_{A,n}\otimes\overline{\mathbf{D}}^2_{B,n}\big)\Big\|_\infty, E_{\boldsymbol{\theta}_0}\big(\overline{\ell}_A(\boldsymbol{\theta})\overline{\ell}_B(\boldsymbol{\theta})\big)\Big\} \le n\Big(H_nK_n\Big)^{-1}\tilde{\rho}\big(\Delta_{A,B}\big)$$

for $\boldsymbol{\theta} \in \Theta$, $A = \Gamma_{hk}$ and $B = \Gamma_{h'k'}$, where $\tilde{\rho}$ satisfies $\sum_{m=1}^\infty m^{d-1}\tilde{\rho}(m) < \infty$. In the Appendix, we prove the following results on the consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$.

Theorem 2. *Assume* (A1)–(A5) *and* (B3). *Assume also that for every* $\epsilon > 0$ *there exists* $\delta > 0$ *such that* (B1) *and* (B2) *hold. Then* $\widehat{\boldsymbol{\theta}}_n$ *is consistent and* (2.4) *holds. Moreover,* $\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)$ *has a limiting* $\mathcal{N}\big(\mathbf{0}, \mathbf{V}^{-1}\big)$ *distribution as* $n \to \infty$.

2.3. Asymptotic efficiency of block MLE

We now prove the asymptotic efficiency of the block MLE under the assumptions of Theorem 2 and an additional assumption (C) below. A seemingly insurmountable difficulty in proving asymptotic efficiency is the analytical intractability of the full likelihood (1.6), which is often used to prove that the inverse of the Fisher information matrix is asymptotically minimal among the

covariance matrices of regular estimators in parametric models; see Andersen et al. (1993, p.600). We circumvent the difficulty by assuming that additional sparse observations $Z_{t\omega}$ of the signal are available at (t,ω) belonging to the boundary $\partial\Gamma_{hk}$ of each block. Specifically, letting $B = \bigcup_{1\le h\le H_n, 1\le k\le K_n} \partial\Gamma_{hk}$, we assume that $\mathbf{Z}_B = \{Z_{t\omega}, (t,\omega)\in B\}$ is also observed besides the $Y_{t\omega}, 1\le t\le T, \omega\in\Omega$. Since the $\mathbf{Y}_{\Gamma_{hk}}$ are conditionally independent given $\mathbf{Z}_B$, the "enriched" likelihood function has logarithm $\tilde{\ell}_n(\boldsymbol{\theta})$ that can be expressed as the sum

$$\tilde{\ell}_n(\boldsymbol{\theta}) = \sum_{h=1}^{H_n}\sum_{k=1}^{K_n} \log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}}|\mathbf{Z}_{\partial\Gamma_{hk}}) + \log f_{\boldsymbol{\theta}}(\mathbf{Z}_B), \tag{2.6}$$

noting that the conditional distribution of $\mathbf{Y}_{\Gamma_{hk}}$ given $\mathbf{Z}_B$ depends only on $\mathbf{Z}_{\partial\Gamma_{hk}}$.

The block MLE $\widehat{\boldsymbol{\theta}}_n$ has already been shown in Theorem 2 to be consistent and asymptotically normal. We can establish its asymptotic efficiency by showing that the enriched MLE that maximizes (2.6) is asymptotically efficient and also asymptotically equivalent to the block MLE $\widehat{\boldsymbol{\theta}}_n$ that maximizes (2.1). Choosing H_n and K_n appropriately to ensure sparsity of $\mathbf{Z}_B$, one expects in view of (1.1) and (1.6) that there exists $\epsilon_n\to 0$ such that $\sqrt{n}\epsilon_n\to\infty$ and

$$\text{(C)}\ n^{-1}\Big\{ \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\epsilon_n} \big\|\mathbf{D}\log f_{\boldsymbol{\theta}}(\mathbf{Z}_B)\big\|^2 + \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\epsilon_n} \sum_{h=1}^{H_n}\sum_{k=1}^{K_n} \big\|\mathbf{D}\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}}|\mathbf{Z}_{\partial\Gamma_{hk}}) - \mathbf{D}\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})\big\|^2 \Big\} \xrightarrow{P_{\boldsymbol{\theta}_0}} 0.$$

This condition will be discussed in Section 4 and is used to establish (a) local asymptotic normality (LAN) of the enriched log-likelihood (2.6), and (b) the asymptotically equivalent local behavior of (2.1) and (2.6) in an ϵ_n-neighborhood of $\boldsymbol{\theta}_0$. We first prove (a) and (b) and then use them to establish the asymptotic efficiency of $\widehat{\boldsymbol{\theta}}_n$ via the Hájek-LeCam theory. Note that although the theorem assumes additional observations $\mathbf{Z}_B$ besides $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_T$, $\widehat{\boldsymbol{\theta}}_n$ is only based on $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_T$ and the Hájek-LeCam theory is used to show that it is asymptotically efficient even when the additional observations $\mathbf{Z}_B$ are available and it discards them.

Theorem 3. *Under* (C) *and the assumptions of Theorem* 2, $\widehat{\boldsymbol{\theta}}_n$ *is asymptotically efficient.*

Proof. We first show that the enriched parametric family with log-likelihood (2.6) satisfies

$$\tilde{\ell}_n(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) - \tilde{\ell}_n(\boldsymbol{\theta}_0) - n^{-1/2}\mathbf{u}^{\mathrm{T}}\mathbf{D}\tilde{\ell}_n(\boldsymbol{\theta}_0) + \frac{\mathbf{u}^{\mathrm{T}}\mathbf{V}\mathbf{u}}{2} \xrightarrow{P_{\boldsymbol{\theta}_0}} 0 \tag{2.7}$$

uniformly in $\mathbf{u} \in \mathcal{K}$, for every compact subset $\mathcal{K}$ of $\mathbb{R}^d$, and that $n^{-1/2}\mathbf{D}\tilde{\ell}_n(\boldsymbol{\theta}_0)$ has a limiting $\mathcal{N}(\mathbf{0}, \mathbf{V})$ distribution as $n \to \infty$. By Taylor's theorem,

$$\tilde{\ell}_n(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) - \tilde{\ell}_n(\boldsymbol{\theta}_0) = n^{-1/2}\mathbf{u}^{\mathrm{T}}\mathbf{D}\tilde{\ell}_n(\boldsymbol{\theta}^*_{\mathbf{u}}), \tag{2.8}$$

where $\boldsymbol{\theta}^*_{\mathbf{u}}$ lies between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}$. Combining (2.6) and (C) with (2.1), it follows that $\mathbf{D}\tilde{\ell}_n(\boldsymbol{\theta}^*_{\mathbf{u}}) = \mathbf{D}\ell_n(\boldsymbol{\theta}^*_{\mathbf{u}}) + o_p(n^{-1/2})$, in which $o_p(n^{-1/2})$ is uniform in $\mathbf{u} \in \mathcal{K}$. Hence, applying Taylor's theorem to $\mathbf{D}\ell_n(\boldsymbol{\theta}^*_{\mathbf{u}})$ yields

$$\mathbf{D}\tilde{\ell}_n(\boldsymbol{\theta}^*_{\mathbf{u}}) = \mathbf{D}\ell_n(\boldsymbol{\theta}_0) + \mathbf{D}^2\ell_n(\boldsymbol{\theta}_0)\frac{\mathbf{u}}{\sqrt{n}} + o_p(n^{-1/2}), \tag{2.9}$$

in which $o_p(n^{-1/2})$ is uniform in $\mathbf{u} \in \mathcal{K}$. Under (A1)−(A5), the proof of Theorem 1 in the Appendix shows that

$$n^{-1}\mathbf{D}^2\ell_n(\boldsymbol{\theta}_0) \xrightarrow{P} \mathbf{V}, \tag{2.10}$$

where $\mathbf{V}$ is given in (A1). Combining (2.9) in the case $\mathbf{u} = \mathbf{0}$ with Theorem 1 shows that $n^{-1/2}\mathbf{D}\tilde{\ell}_n(\boldsymbol{\theta}_0)$ has a limiting $\mathcal{N}(\mathbf{0}, \mathbf{V})$ distribution. Moreover, (2.7) follows from (2.8)−(2.10), proving the LAN property for the enriched log-likelihood (2.6).

In the Appendix, we show that under the assumptions of Theorem 2,

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{u}, \mathbf{V}^{-1}) \text{ under } \ P_{\boldsymbol{\theta}_0 + \mathbf{u}/\sqrt{n}}, \tag{2.11}$$

for every $\mathbf{u} \in \mathbb{R}$, where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. Hence, $\widehat{\boldsymbol{\theta}}_n$ is a regular estimator in the Hájek-LeCam theory of asymptotic efficiency in LAN models, and is asymptotically efficient by the Hájek convolution theorem since $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ and $n^{-1/2}\mathbf{D}\tilde{\ell}_n(\boldsymbol{\theta}_0)$ have the same $\mathcal{N}(\mathbf{0}, \mathbf{V})$ limiting distribution; see Andersen et al. (1993, pp.598-600).

3. Implementation and Simulation Study

In this section we describe an EM algorithm to compute the block MLE that we use in conjunction with the block Gibbs sampler to estimate the unobserved states $Z_{t\omega}$ in the hidden Markov model. We then consider an application of the algorithm to estimate the hidden neuronal inputs in a task-based functional MRI experiment, and illustrate the performance of the estimates with simulated data.

3.1. EM algorithm and estimation of states by block Gibbs sampler

The term $f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}})$ in (2.1) has the form (1.6) but with (t, ω) restricted to the block Γ_{hk}. Using independence of different blocks as the working model, we can apply the EM algorithm to maximize the corresponding log-likelihood

(2.1). It consists of an expectation step (E-step) and a maximization step (M-step) for each iteration. The E-step uses the current estimate $\tilde{\boldsymbol{\theta}}_{old}$ to substitute for $\boldsymbol{\theta}$ in the conditional expectation of the complete-data log-likelihood: $\ell_c(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_{old}) = \sum_{h,k} E_{\tilde{\boldsymbol{\theta}}_{old}}\big[\log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}}, \mathbf{Z}_{\Gamma_{hk}})\big|\mathbf{Y}\big]$, where $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_T)$. The M-step can be carried out by taking partial derivatives of ℓ_c with respect to the components of $\boldsymbol{\theta}$ and solving the estimating equation $\dot{\ell}_c(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_{old}) = \mathbf{0}$ to obtain an updated estimate $\tilde{\boldsymbol{\theta}}_{new}$ of $\boldsymbol{\theta}$. Thus the E-step is used for the conditional expectation $\sum_{h,k} E_{\tilde{\boldsymbol{\theta}}_{old}}\big[(\partial/\partial\boldsymbol{\theta}) \log f_{\boldsymbol{\theta}}(\mathbf{Y}_{\Gamma_{hk}}, \mathbf{Z}_{\Gamma_{hk}})\big|\mathbf{Y}\big]$, which can be computed by the Gibbs sampler; see Chapter 6 of Liu (2001).

The Gibbs sampler provides a powerful tool to estimate the posterior distribution of the unobserved state $\mathbf{Z}_{\Gamma_{hk}}$ given the observed data $\mathbf{Y}_{\Gamma_{hk}}$ in an HMRF. As pointed out by Liu (2001, pp.130-131), the Gibbs sampler is a Markov chain that converges geometrically to its stationary distribution, which is the posterior distribution of interest. The convergence rate depends on how the $Z_{t\omega}$ correlate with each other, leading Liu, Wong, and Kong (1994) to improve the efficiency of the Gibbs sampler by grouping highly correlated components to sample the groups (blocks) iteratively from their joint conditional distribution by the block Gibbs sampler. In the present context, this corresponds to dividing Γ_{hk} into smaller sub-blocks. Thus, the blocking idea is useful not only for parameter estimation but also for state estimation. Liu (2001) also describes other Markov chain Monte Carlo (MCMC) methods such as the Swendsen-Wang algorithm and generalized Gibbs to evaluate the posterior distribution of $Z_{t\omega}$.

In view of Theorem 2 and (2.4), we can use the inverse of the observed Fisher information matrix $n\widehat{\mathbf{V}} = -\sum_{h=1}^{H_n}\sum_{k=1}^{K_n} \mathbf{D}^2_{\Gamma_{hk}}(\widehat{\boldsymbol{\theta}}_n)$ to approximate the asymptotic covariance matrix of the block MLE $\widehat{\boldsymbol{\theta}}_n$. In particular, the standard errors of the components of $\widehat{\boldsymbol{\theta}}_n$ can be evaluated by taking the square roots of the diagonal elements of $\big(n\widehat{\mathbf{V}}\big)^{-1}$. Since the EM algorithm is used to compute $\widehat{\boldsymbol{\theta}}_n$, we can evaluate $\widehat{\mathbf{V}}$ by the formula

$$\widehat{\mathbf{V}} = -n^{-1} E_{\widehat{\boldsymbol{\theta}}_n}\left[\mathbf{S}\mathbf{S}^{\mathrm{T}} + \sum_{h=1}^{H_n}\sum_{k=1}^{K_n} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} \log f_{\boldsymbol{\theta}}\big(\mathbf{Y}_{\Gamma_{hk}}, \mathbf{Z}_{\Gamma_{hk}}\big)\Big|\mathbf{Y}\right] \tag{3.1}$$

introduced by Louis (1982), in which $\mathbf{S} = \sum_{h=1}^{H_n}\sum_{k=1}^{K_n} \big(\partial/\partial\boldsymbol{\theta}\big) \log f_{\boldsymbol{\theta}}\big(\mathbf{Y}_{\Gamma_{hk}}, \mathbf{Z}_{\Gamma_{hk}}\big)$ and the partial derivatives are taken at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n$.

3.2. Illustrative example

In task-based functional MRI (fMRI) experiments, BOLD (blood oxygen level-dependent) signals, which are measured in response to input stimuli (or task), are temporally delayed and distorted due to various technical reasons

Table 1. Block MLEs, with estimated standard errors in parentheses, and error rates in estimating binary pixel values for one simulated dataset.

block size	μ_{-1} $(=0)$	μ_1 $(=2)$	σ^2 $(=1)$	β $(=0.10)$	α $(=0.10)$	FPR	FNR	ERR	CPU time (mins)
12×12 $\times 30$	0.0050 (0.0050)	2.0119 (0.0051)	0.9989 (0.0056)	0.0998 (0.0012)	0.1003 (0.0032)	0.1414	0.1395	0.1404	113.3
24×24 $\times 45$	0.0049 (0.0049)	2.0121 (0.0049)	0.9987 (0.0054)	0.0996 (0.0012)	0.1003 (0.0032)	0.1415	0.1397	0.1406	141.1
48×48 $\times 90$	0.0049 (0.0050)	2.0120 (0.0051)	0.9986 (0.0057)	0.0996 (0.0011)	0.1004 (0.0031)	0.1413	0.1395	0.1404	275.0

(Ward and Mazaheri (2006)). Estimation of the neuronal input stimulus function from an fMRI experiment provides insights into actual brain activity during task activation (Höjen-Sörensen, Hansen, and Rasmussen (2000)). The fMRI data of a slice of the brain consist of a time series, indexed by $t = 1, \ldots, T$, of two-dimensional noisy images $\mathbf{Y}_t = \{Y_{t\omega}, \omega \in \Omega\}$ that are distortions of binary spatio-temporal hidden variables $Z_{t\omega}$, where $Z_{t\omega} = 1$ if the voxel ω is on stimulus at time t and $Z_{t\omega} = -1$ otherwise. We model $\mathbf{Z} = \big\{Z_{tw}, \omega \in \Omega, t = 1, \ldots, T\big\}$ by a Markov random field undergoing Markovian dynamics over time so that with $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$, $\boldsymbol{\theta}^{(1)} = (\mu_1, \mu_{-1}, \sigma^2)$ and $\boldsymbol{\theta}^{(2)} = (\alpha, \beta)$, the $g_{\boldsymbol{\theta}}$ and $h_{\boldsymbol{\theta}}$ in (1.6) are given by

$$h_{\alpha,\beta}(\mathbf{Z}) = \frac{1}{\kappa(\alpha,\beta)} \exp\Big\{ \sum_{t=1}^{T} \sum_{\omega\in\Omega} Z_{t\omega}\Big(\beta \sum_{v\in N(\omega)} Z_{tv} + \alpha Z_{t-1,\omega}\Big)\Big\}, \tag{3.2}$$

where $Z_{0\omega} = 0$, and $g_{\boldsymbol{\theta}}(\cdot|Z_{t\omega})$ is the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with $\mu = \mu_1$ if $Z_{t\omega} = 1$ and $\mu = \mu_{-1}$ if $Z_{t\omega} = -1$. The neighborhood of an interior point ω of Ω consists of four sites nearest to ω in the two-dimensional lattice Ω. Here (3.2) basically augments the Ising model in the second paragraph of Section 1.1 by adding the autoregressive term $\alpha Z_{t-1,\omega}$ at each voxel ω and time t, with the logarithmic link function for the binary time series. We simulated data from the above model with $\alpha = 0.1$, $\beta = 0.1$, $\mu_1 = 2, \mu_{-1} = 0$, and $\sigma^2 = 1$. Figure 1 plots a simulated dataset consisting of $(\mathbf{Z}_t, \mathbf{Y}_t)$ at $t = 50$ and 51. It shows the spatial and temporal dependencies clearly in $\mathbf{Z}$ but not in $\mathbf{Y}$.

Table 1 gives the block MLEs of $\alpha, \beta, \mu_{-1}, \mu_1$, and σ^2 for a single simulated dataset for $1 \le t \le 90$ and $\Omega = \{1, \ldots, 48\} \times \{1, \ldots, 48\}$. The size of each

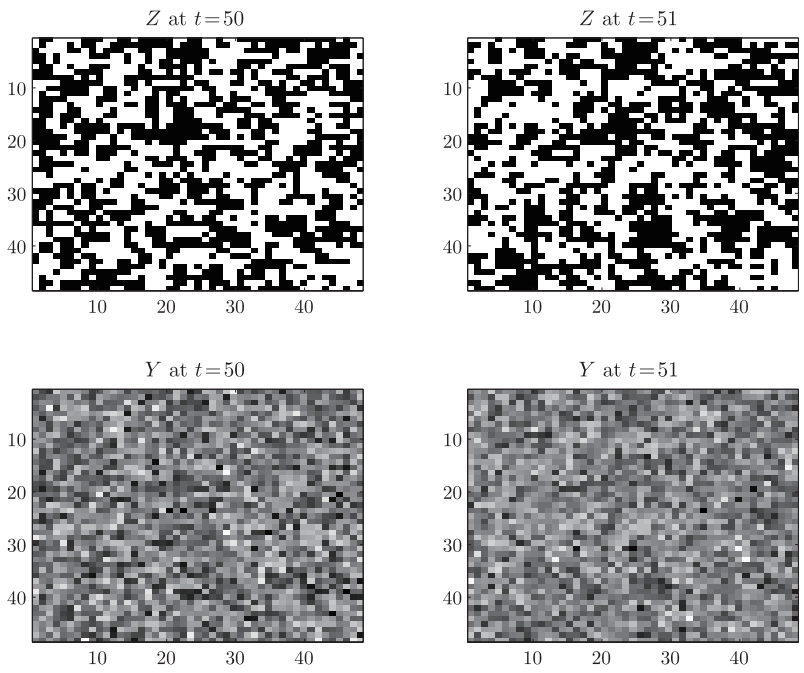

Figure 1. True ($\mathbf{Z}_t$) and observed ($\mathbf{Y}_t$) images at $t = 50$ and $t = 51$.

block varies from $(12 \times 12) \times 40$ to $(48 \times 48) \times 90$ (corresponding to the full MLE), with $(24 \times 24) \times 45$ as an intermediate choice. Thus, $n = (48)^2 90$ and $(H_n, K_n) = (3, 4^2), (2, 2^2), (1, 1)$ for the three cases in the table. The convergence criterion used by the EM algorithm in Section 3.1 is an upper bound of 10^{-5} for the size of each parameter increment. The table also reports the standard errors of the estimates computed by Louis' formula that is described in the last paragraph of Section 3.1. Note that the 95% confidence intervals based on them and asymptotic normality contain the true parameter values.

The table also gives the CPU times carried out on a desktop PC (Intel Core i7-4770 CPU (3.40 GHz)) and misclassification rates that include the false positive rate (FPR), the false negative rate (FNR), and the total error rate (ERR). Letting $\#_{\mathrm{S}}$ (or $\#_{\mathrm{NS}}$) denote the number of voxels on (or not on) stimulus and $\#_{\mathrm{T}} = \#_{\mathrm{S}} + \#_{\mathrm{NS}}$, we have ERR=(number of misclassified pixels)$/\#_{\mathrm{T}}$, FPR=($\#_{\mathrm{NS}}$ that are misclassified as on stimulus)$/\#_{\mathrm{NS}}$ and FNR=($\#_{\mathrm{S}}$ that are misclassified as not on stimulus)$/\#_{\mathrm{S}}$. One sees that the error rates of the estimates of the pixel values are almost the same over the three block sizes, and that using the smallest block size results in 60% savings in CPU time over the block size that corresponds to the MLE.

Table 2 reports the results for 100 simulated datasets from the model. Because each result is the mean of 100 simulations, the standard error can be com-

Table 2. Means of block MLEs, error rates and CPU times in estimating binary pixel values for 100 simulated datasets, with standard errors given in parentheses. Also given in square brackets is the corresponding range over the 100 simulations.

block size	μ_{-1} (= 0)	μ_1 (= 2)	σ^2 (= 1)	β (= 0.10)	α (= 0.10)	FPR	FNR	ERR	CPU time (mins)
12×12	0.0017	2.0027	0.9991	0.0993	0.1018	0.1445	0.1427	0.1436	142.0
$\times 30$	(0.00453)	(0.00497)	(0.00559)	(0.00094)	(0.00241)	(0.00192)	(0.00190)	(0.00191)	(33.6)
	[-0.0010,	[1.990,	[0.989,	[0.097,	[0.097,	[0.140,	[0.138,	[0.137,	[62.4,
	0.0015]	2.017]	1.014]	0.102]	0.109]	0.148]	0.146]	0.147]	213.4]
24×24	0.0016	2.0029	0.9989	0.0990	0.1020	0.1425	0.1427	0.1436	150.0
$\times 45$	(0.00450)	(0.00492)	(0.00552)	(0.00093)	(0.00243)	(0.00189)	(0.00187)	(0.00188)	(28.4)
	[-0.0010,	[1.990,	[0.989,	[0.097,	[0.097,	[0.140,	[0.138,	[0.139,	[71.7,
	0.0015]	2.017]	1.014]	0.101]	0.109]	0.148]	0.146]	0.147]	220.3]
48×48	0.0016	2.0029	0.9988	0.0990	0.1020	0.1425	0.1427	0.1436	311.7
$\times 90$	(0.00449)	(0.00490)	(0.00552)	(0.00092)	(0.00239)	(0.00191)	(0.00189)	(0.00190)	(60.3)
	[-0.0010,	[1.991,	[0.989,	[0.097,	[0.097,	[0.140,	[0.138,	[0.139,	[190.2,
	0.0016]	2.017]	1.014]	0.101]	0.109]	0.148]	0.146]	0.147]	484.1]

puted directly from these simulations, unlike the standard errors of the block MLEs in Table 1 that use Louis' formula. Moreover, to speed up the computation for 100 datasets, we relaxed the convergence criterion for the EM algorithm to an upper bound of 10^{-4} for the size of each parameter increment. The biases and the standard errors of the block MLEs are almost same for the three block sizes. In all cases, the true parameter values are within the 95% confidence limits using these standard errors and the normal approximation. On the other hand, Table 2 which gives the minimum and maximum (in square brackets) for each entry besides the mean over the 100 simulations shows substantial savings in CPU time by using smaller block sizes than the full size of the MLE.

4. Discussion

Comparison of the full likelihood (1.1) and the pseudo-likelihood (1.5) shows that they are not expected to be asymptotically equivalent, explaining why the MCLE has been found to be less efficient than the full MLE in the simulation studies of Okabayashi, Johnson, and Geyer (2011). The main idea underlying the block log-likelihood (2.1) is to keep it close to the full log-likelihood under the computational constraint that the block MLE is still computationally feasible. In fact, the asymptotic independence of suitably chosen blocks is expected

in stationary mixing random fields. For stationary mixing sequences of random variables, the asymptotic independence is used to establish the asymptotic normality of their sums via the CLT for sums of i.i.d. random variables; see Billingsley (1995, pp.366-367). Extensions of the CLT to stationary mixing random fields using this approach require more restrictive mixing conditions; see Remark 2 of Bolthausen (1982) who introduced another approach that is summarized in the proof of Theorem 1 in the Appendix. A recent review of the developments in the CLT for stationary mixing random fields is given by Wang and Woodroofe (2013), who also introduce a new condition under which the stationary random field has an m-dependent approximation and therefore satisfies the CLT.

Since our goal is to analyze the block MLE, we do not need the full force of stationarity and mixing conditions on σ-fields that are separated in space-time by the infimum of $\Delta_{A,B}$ over A belonging to one σ-field and B belonging to the other. We derive from the scratch the CLT for spatio-temporal random fields that may be non-stationary. In particular, we replace the ρ-mixing condition for stationary random fields in Bolthausen (1982) by the weaker and more direct correlation decay conditions (A3), (A4) and (B3) for the partial derivatives $\mathbf{D}^{(n)}_{\Gamma_{hk}}$ and $\overline{\mathbf{D}}^2_{\Gamma_{hk,n}}$. The CLT is usually proved by a truncation argument and analysis of the characteristic function of the truncated random variables. The assumptions (A1), (A2), (A5), (B1), and (B2) are related to this truncation argument.

Since we have an HMRF in which the $Y_{t\omega}$ are conditionally independent given $\mathbf{Z}$, the correlation decay conditions for $\mathbf{D}^{(n)}_{\Gamma_{hk}}$ and $\overline{\mathbf{D}}^2_{\Gamma_{hk,n}}$ are related to those for $\mathbf{Z}$. If $\mathbf{Z}$ has mixing properties, then one expects $Z_{t\omega}$, and therefore also $Y_{t\omega}$, to be asymptotically independent of $\mathbf{Z}_{\partial\Gamma_{hk}}$ if $(t,\omega) \in \Gamma_{hk}$ is sufficiently separated from B. This is the background for assumption (C), in which $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\epsilon_n} \|\mathbf{D} \log f_{\boldsymbol{\theta}}(\mathbf{Z}_B)\|^2 = o(n)$ is expected to hold when $\partial\Gamma_{hk}$ has much fewer sites than Γ_{hk} and $\mathbf{Z}$ is mixing. In particular, the MRF (3.2) undergoing Markovian dynamics is geometrically mixing if $\beta > \log 3$ and $\alpha > 0$; see Martin-Löf (1973).

Whereas the log-likelihood function (1.6) of $\boldsymbol{\theta}$ based on $\mathbf{Y}$ is intractable, the enriched log-likelihood function (2.6) can be expressed as a sum of tractable summands plus an asymptotically negligible term, in view of (C) and the assumptions (A1)$-$(A5) and (B1)$-$(B3). Under these assumptions, the block MLE based on $\mathbf{Y}$ is asymptotically equivalent to the MLE based on $(\mathbf{Y}, \mathbf{Z}_B)$, as shown in the proof of Theorem 3. The simulation study in Section 3.2 shows that the block MLE is indeed close to the full MLE and that both are close to the true parameter vector. We have also established in Section 2.3 the asymptotic efficiency of the block MLE, which is not shared by other maximum composite likelihood estimators in the literature.

The asymptotic theory in Section 2 and the simulation study in Section 3.2 show that there is much flexibility in the choice of the block sizes. Basically the block MLE chooses H_n to be a small fraction of T and K_n to be a small fraction of $|\Omega_n|$. We can also minimize the trace of the inverse of the observed Fisher information matrix, computed by Louis' formula (3.1), over a grid of such choices to determine the block size empirically. The results in Tables 1 and 2, with moderate values of diameter$(\Omega) = 48$ and $t = 90$ for which the full MLE is computationally tractable, show the robustness to the choice of the block sizes. They also show substantial savings in CPU time by using blocks for likelihood maximization via the EM algorithm in Section 3.1.

In Section 3, we have focused on standard algorithms to implement the block MLE. Here we describe other techniques to speed up the computation. An obvious technique for computing the block likelihoods is parallelization by using either graphical processing units (GPUs) or/and multiple CPUs. Whereas GPU greatly speeds up tasks that involve many relatively small computations, CPU can perform more complex tasks that require larger memory (Hockney and Jesshope (1981); Owens et al. (2008); Yu et al. (2014)). Another technique is to replace the Gibbs sampler by a faster algorithm for calculating MAP (maximum a posteriori) estimates of the states; see Greig, Porteous and Seheult (1989), Boykov and Kolmogorov (2004), Ravikumar and Lafferty (2006), Kumar and Zilberstein (2011), and Bhole et al. (2014). For example, Greig, Porteous and Seheult (1989), Boykov and Kolmogorov (2004) and Bhole et al. (2014) reformulate MAP estimation as the solution to a minimum-cut/maximum-flow algorithm on a graph, whose computational complexity in computing the MAP estimates of the states in the H_nK_n blocks in Section 3 is $O\big(n^4/(H_nK_n)^3\big)$.

The divide-and-conquer approach used by the block MLE has also appeared in other methods for the analysis of large spatio-temporal data sets. In particular, to detect spatio-temporal clusters in epidemiology and environmetrics, slices of spatio-temporal cylinders are used to scan an area of interest; see Kulldorff (1997), Kulldorff et al. (1998) and Tuia et al. (2008). The base of the cylinder is a circular zone in space, while the axis of the cylinder represents time. The objective of the data analysis is either retrospective identification of past clusters or prospective detection of emerging clusters. Unlike the highly complex spatio-temporal Markov random field models in image analysis, these spatio-temporal clusters are modeled by relatively simple Poisson or Bernoulli models. On the other hand, whereas the set Ω of sites considered herein is a lattice in $\mathbb{Z}^d$ that makes the choice of the blocks for the block MLE relatively straightforward, the geographical region of interest in the aforementioned applications of local spatio-temporal cylinders can be much more complicated, leading to statistical issues on how to define these cylinders that are not considered here.

Acknowledgement

Lai's research is supported by the National Science Foundation grant DMS-1106535 and Lim's research is supported by the National Research Foundation of Korea grant MSIP No. 2011-0030810.

Appendix: Proofs

Proof of Theorem 1. For notational simplicity, denote $P_{\boldsymbol{\theta}_0}$ and $E_{\boldsymbol{\theta}_0}$ by P and E. Let $\mathbf{a} \in \mathbb{R}^d - \{\mathbf{0}\}$ be nonrandom, and let $X_{hk} = \big(n/H_nK_n\big)^{-1/2}\mathbf{a}^{\mathrm{T}}\mathbf{D}^{(n)}_{\Gamma_{hk}}(\boldsymbol{\theta}_0)$, $S_n = \sum_{h=1}^{H_n}\sum_{k=1}^{K_n} X_{hk}$. In view of (A2), it suffices to prove that S_n, which has H_nK_n summands, has a limiting $\mathcal{N}(0, \mathbf{a}^{\mathrm{T}}\mathbf{V}\mathbf{a})$ distribution as $n \to \infty$. By (A1) and Lemma 2 of Bolthausen (1982), who cites Stein (1973) for the basic idea, we need only show that for any $\lambda \in \mathbb{R}$,

$$\lim_{n\to\infty} E\big\{(i\lambda - S_n)e^{i\lambda S_n}\big\} = 0. \tag{A.1}$$

In view of (A5), we can choose $m = m_n$ such that $\rho(m)\sqrt{n} \to 0$ and $\sqrt{n}/m^d \to \infty$. Denoting Γ_{hk} by A, let $S_{hk,n} = \sum_{B=\Gamma_{h'k'}:\Delta_{A,B}\le m} X_{h'k'}$. As in Eq. 4 of Bolthausen (1982), we write

$$(i\lambda - S_n)e^{i\lambda S_n} = \mathrm{I} - \mathrm{II} - \mathrm{III}, \tag{A.2}$$

where $\mathrm{I} = i\lambda e^{i\lambda S_n}(1 - \sum_{h,k} X_{hk}S_{hk,n})$, $\mathrm{II} = e^{i\lambda S_n}\sum_{h,k} X_{hk}(1 - e^{-i\lambda S_{hk,n}} - i\lambda S_{hk,n})$, and $\mathrm{III} = \sum_{h,k} X_{hk}e^{i\lambda(S_n - S_{hk,n})}$. We can use (A3)–(A5) and arguments similar to those of Bolthausen (1982, p.1049) to bound $E(\mathrm{I}^2)$, $E(|\mathrm{II}|)$ and $|E(\mathrm{III})|$. In particular, denoting X_{hk} and $X_{h'k'}$ by X_A and X_B for $A = \Gamma_{hk}$ and $B = \Gamma_{h'k'}$, we have bounds similar to Bolthausen's for $|E(X_AX_B)|$ and $|E(X_AX_{A'}X_BX_{B'})|$, and can also follow Bolthausen to bound $|E(\mathrm{III})|$ by $|E(\mathrm{III})| \le \sum_{h,k}|E(X_{hk}|S_n - S_{hk,n})|$, to which we can apply (A5), noting that $\sum_{h,k}$ has H_nK_n terms and that $\sqrt{n}\rho(m) \to 0$ by the choice of $m = m_n$. Therefore, similar to Bolthausen (1982, p.1049), we can show that $E(\mathrm{I}^2)$, $E(|\mathrm{II}|)$ and $|E(\mathrm{III})|$ converge to 0 as $n \to \infty$, thereby proving (A.1).

Proof of Theorem 2. Let $\epsilon > 0$. It follows from (B2) and the compactness of Θ that there exist $\delta > 0$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J \in \Theta - \{\boldsymbol{\theta}_0\}$ such that $\bigcup_{j=1}^{J} \mathrm{B}_\delta(\boldsymbol{\theta}_j) = \Theta$ and

$$P\Big\{\max_{1\le h\le H_n, 1\le k\le K_n, 0\le j\le J}\ \sup_{\boldsymbol{\lambda}\in \mathrm{B}_\delta(\boldsymbol{\theta}_j)} \big|\log f_{\boldsymbol{\lambda}}(\mathbf{Y}_{\Gamma_{hk}}) - \log f_{\boldsymbol{\theta}_j}(\mathbf{Y}_{\Gamma_{hk}})\big| > \epsilon(H_nK_n)\Big\} \to 0 \tag{A.3}$$

as $n \to \infty$. For $0 \le j \le J$, it follows from (B3) that $n^{-1}\mathrm{Var}\big(\sum_{h=1}^{H_n}\sum_{k=1}^{K_n}\overline{\ell}_{\Gamma_{hk}}(\boldsymbol{\theta}_j)\big) = O(1)$, and therefore by Chebyshev's inequality,

$$n^{-1}\sum_{h=1}^{H_n}\sum_{k=1}^{K_n}\overline{\ell}_{\Gamma_{hk}}(\boldsymbol{\theta}_j) \xrightarrow{P} 0 \text{ as } n \to \infty, \text{ for } j = 0, \ldots, J. \tag{A.4}$$

Since $\liminf_{n\to\infty} n^{-1}\sum_{h=1}^{H_n}\sum_{k=1}^{K_n} E\big[\log f_{\boldsymbol{\theta}_0}(\mathbf{Y}_{\Gamma_{hk}}) - \log f_{\boldsymbol{\theta}_j}(\mathbf{Y}_{\Gamma_{hk}})\big] > 0$ for $1 \le j \le J$ (see the paragraph preceding Theorem 2), it follows from (A.4) together with (B2) and (A.3) that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$. A similar argument can be used to prove (2.4).

Proof of (2.11). By (2.7) and LeCam's third lemma (Andersen et al. (1993, p.596)), it suffices to show that under $P_{\boldsymbol{\theta}_0}$,

$$\begin{pmatrix} n^{-1/2}\boldsymbol{u}^{\mathrm{T}}\mathbf{D}\tilde{\ell}_n(\boldsymbol{\theta}_0) - \boldsymbol{u}^{\mathrm{T}}\mathbf{V}\boldsymbol{u}/2 \\ n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\begin{pmatrix} -\boldsymbol{u}^{\mathrm{T}}\mathbf{V}\boldsymbol{u}/2 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{u}^{\mathrm{T}}\mathbf{V}\boldsymbol{u} & \boldsymbol{u}^{\mathrm{T}} \\ \boldsymbol{u} & \mathbf{V}^{-1} \end{pmatrix}\right). \tag{A.5}$$

Making use of (2.3), (2.4), and Theorem 1, it can be shown that (A.5) indeed holds.

References

Andersen, P. K., Borgan, O, Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer, New York.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **23**, 192-236.

Bhole, C., Pal, C., Rim, D. and Wismüller, A. (2014). 3D segmentation of abdominal CT imagery with graphical models, conditional random fields and learning. *Machine Vision Appl.* **25**, 301-325.

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

Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields. *Ann. Probab.* **10**, 1047-1050.

Boykov, Y. and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal.* **26**, 1124-1137.

Durrett, R. (2005). *Probability: Theory and Examples*. 3rd edition. Duxburry Press, CA.

Greig, D., Porteous, B. and Seheult, A. (1989). Exact maximum A posteriori estimation for binary images. *J. Roy. Statist. Soc. Ser. B* **51**, 271-279.

Hockney, R. and Jesshope, C. R. (1981). *Parallel Computers*, Adam Hilger, Bristol.

Höjen-Sörensen, P., Hansen, L. K. and Rasmussen, C. E. (2000). Bayesian modeling of fMRI time series. *Adv. Neural Inform. Process. Syst.* **12**, 754-760.

Kulldorff, M. (1997). A spatial scan statistics. *Comm. Statist. Theor. Meth.* **26**, 1481-1496.

Kulldorff, M., Athas, W. F., Feurer, E. J., Miller, B. A. and Key, C. R. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *Amer. J. Public Health* **88**, 1377-1380.

Kumar, A. and Zilberstein, S. (2011). Message passing algorithms for quadratic programming formulations of MAP estimation. *Proc. 27th Conf. Uncertainty in Artificial Intelligence*, Barcelona, Spain.

Lindsay, B. G., Yi, G. and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statist. Sinica* **21**, 71-105.

Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Liu, J., Wong, W. H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27-40.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226-233.

Martin-Löf, A. (1973). Mixing properties, differentiability of the free energy and the central limit theorem for a pure phase in the Ising model at low temperature. *Comm. Math. Phy.* **32**, 75-92.

Okabayashi, S., Johnson, L. and Geyer, C. (2011). Extending pseudo-likelihood for Potts models. *Statist. Sinica* **31**, 331-347.

Owens, J.D., Houston, M., Luebke, D., Green, S., Stone, J. E. and Phillips, J. C. (2008). GPU computing. *Proc. IEEE* **96**, 879-899.

Ravikumar, P. and Lafferty, J. (2006). Quadratic programming relaxation for metric labeling and Markov random field MAP estimation. *Proc.* 23*rd Internat. Conf. Machine Learning*, Pittsburgh, PA.

Stein, C. (1973). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statist. & Probab.* **2**, 583-602.

Swendsen, R. and Wang, J. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58**, 86-88.

Tuia, D., Ratle, F., Lasaponara, R., Telesca, L. and Kanevski, M. (2008). Scan statistics analysis of forest fire clusters. *Commu. Nonlinear Sci. & Num. Simulations* **13**, 1689-1694.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21**, 5-42.

Wang, Y. and Woodroofe, M. (2013). A new condition for the invariance principle for stationary random fields. *Statist. Sinica* **23**, 1673-1696.

Ward, B. D. and Mazaheri, Y. (2006). State-space estimation of the input stimulus function using the Kalman filter: A communication system model for fMRI experiments. *J. Neuroscience Methods* **158**, 271-278.

Yu, D., Won, J-H., Lee, T., Lim, J. and Yoon, S. (2014). High-dimensional fused lasso regression using majorization-minimization and parallel processing. To appear in *J. Comput. Graph. Statist.*. DOI:10.1080/10618600.2013.878662.

Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA.

E-mail: lait@stanford.edu

Department of Statistics, Seoul National University, Seoul 151-747, Korea.

E-mail: johanlim@snu.ac.kr

(Received September 2013; accepted May 2014)