Statistica Sinica **21** (2011), 331-347

EXTENDING PSEUDO-LIKELIHOOD FOR POTTS MODELS

Saisuke Okabayashi, Leif Johnson and Charles J. Geyer

University of Minnesota

Abstract: We propose a conditional composite likelihood based on conditional probabilities of parts of the data given the rest for spatial lattice processes, in particular for Potts models, which generalizes the pseudo-likelihood of Besag. Instead of using conditional probabilities of single pixels given the rest (like Besag), we use conditional probabilities of multiple pixels given the rest. We find that our maximum composite likelihood estimates (MCLE) are more efficient than maximum pseudo-likelihood estimates (MPLE) when the true parameter value of the Potts model is the phase transition parameter value. Our MCLE are not as efficient as maximum likelihood estimates (MLE), but MCLE and MPLE can be calculated exactly, whereas MLE cannot, only approximated by Markov chain Monte Carlo.

Key words and phrases: Conditional composite likelihood, exponential family, ising model, maximum likelihood, peeling.

1. Introduction

Composite likelihood (Lindsay (1988)) is a generalization of the pseudo-likelihood of Besag (1975, 1977), that was proposed as a method of parameter estimation for models with complicated dependence structure for which the likelihood function could not be calculated exactly, or even approximated well in a reasonable amount of time. Examples of such models are spatial lattice processes (Besag (1975)), social network (exponential random graph) models (Strauss and Ikeda (1990); van Duijn, Gile, and Handcock (2009)), spatial point processes (Baddeley and Turner (2000), and references cited therein), and the DNA fingerprint model of Geyer and Thompson (1992). These models differ from the kind of applications that have recently made composite likelihood popular (Varin, Reid, and Firth (2011)) in that no marginal distributions, even univariate marginal distributions, can be calculated exactly or even approximated well in a reasonable amount of time. Hence no marginal composite likelihood scheme is practicable, but conditional composite likelihood schemes are practicable, at least for some of these models.

Here we investigate conditional composite likelihood for spatial lattice processes, in particular for Potts models (Potts (1952)), which generalize Ising models (Ising (1925)). These are models for random images like Figure 1, the random

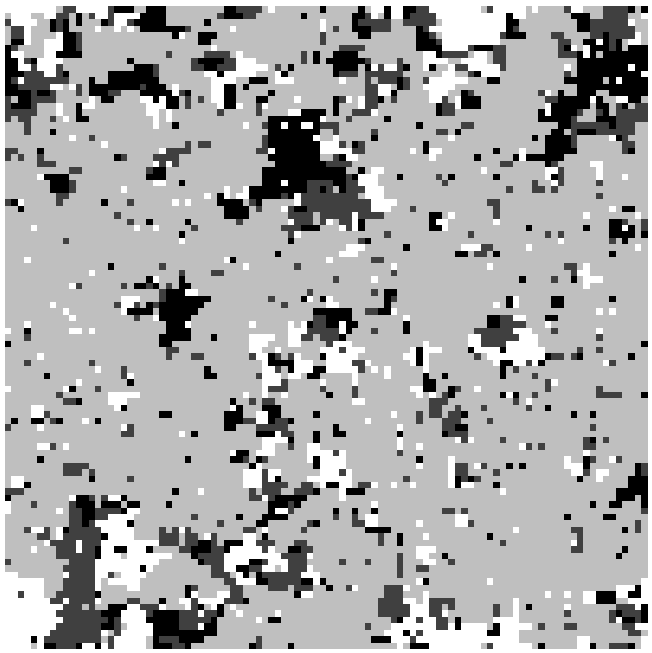

Figure 1. A realized sample from a four-color (white, light gray, dark gray, and black) Potts model with $\theta = (0, 0, 0, 0, \log(1 + \sqrt{4}))^T$.

variables being the colors of the pixels of the image (Ising models are two-color Potts models). These models have the spatial Markov property that two subsets x_A and x_B of the variables (two subregions of the image) are conditionally independent given the rest of the variables so long as no pixel in A is adjacent to any pixel in B. This means that the only subsets of interest are connected ones, which we call *windows*, and we only consider rectangular windows. Then the number of windows does not grow with the size of the windows, and this makes conditional composite likelihood for moderately large window sizes computationally feasible. (Some of the other models with complicated dependence structure mentioned above have spatial Markov properties that may make conditional composite likelihood feasible, but we do not investigate them here).

So how well does composite likelihood do in the context in which it originally arose (as pseudo-likelihood of Besag)? We show here that composite likelihood does improve Besag pseudo-likelihood, although it is still, in certain situations, inferior to maximum likelihood approximated using Markov chain Monte Carlo (MCMC). Inferiority of Besag pseudo-likelihood to maximum likelihood, in certain situations, has been demonstrated by others (Geyer and Thompson (1992); van Duijn, Gile, and Handcock (2009); Geyer (1990)). In particular, Geyer (1990, Chap.6) demonstrated that maximum pseudo-likelihood estimators (MPLE) oc-

casionally overestimate the dependence parameters in Ising models so much that data simulated at these MPLE bear no resemblance to the actual data.

Conditional composite likelihoods and Besag pseudo-likelihoods are exactly computable for these models, unlike likelihoods, so maximum composite likelihood estimators (MCLE) are much easier to compute than maximum likelihood estimators (MLE). We compute MCLE on Potts models for different window sizes, and measure their efficiency against the MPLE and MLE.

Composite likelihood approaches have been applied to problems with dependent data: Lele (2006) applied marginal composite likelihoods to stochastic population dynamics models with sampling error (these are multivariate normal models); Hjort and Varin (2008) applied conditional pseudo-likelihood and marginal composite likelihood to finite-state space Markov chain models and found that marginal composite likelihood performed almost as well as maximum likelihood while conditional pseudo-likelihood performed poorly; and Mardia et al. (2009) showed that for what they call closed exponential families, MCLE are identical to MLE. However, the dependence structure for Potts models (as well as the other models described in the first paragraph of this section) is so complicated that a marginal composite likelihood approach is not possible as in the first two papers. In addition, the model we consider is not a closed exponential family in the sense of Mardia et al. (2009) and thus their results cannot be applied here.

2. Potts Model

An exponential family of distributions (Barndorff-Nielsen (1978); Geyer (2009)) on a discrete sample space $\mathcal{X}$ has log likelihood

$$\ell(\theta) = \langle t(x), \theta \rangle - m(\theta),$$

where $t(x)$ is a vector of canonical statistics, θ a vector of canonical parameters, and $\langle \cdot, \cdot \rangle$ denotes the bilinear form

$$\langle t(x), \theta \rangle = \sum_{i=1}^{d} t_i(x)\theta_i.$$

So that the probability mass function sums to 1, the cumulant function m must have the form

$$m(\theta) = \log\left(\sum_{x \in \mathcal{X}} h(x) e^{\langle t(x), \theta \rangle} \right), \tag{2.1}$$

where h is a nonnegative function (for Potts models h is identically equal to one and is omitted hereafter). In models with complicated dependence, the sum in (2.1) may have no simple expression and can only be evaluated by explicitly

doing the sum, which can be prohibitively expensive. It can be approximated by MCMC (Geyer and Thompson (1992)), permitting MCMC approximations of MLE.

Let L be a finite set on which a symmetric irreflexive relation $\sim$ is defined, and let C be another finite set. Usually L is very large and a regular lattice, and $\sim$ is the neighbor relation for the lattice. Usually C has only a few elements called colors. A discrete spatial lattice process has, for each i in L, a random element of C denoted x_i. For the toroidal square lattice that we consider here, variables x_i and x_j are neighbors with $i \sim j$ if they are adjacent to one another horizontally or vertically, but not diagonally.

Let x be the random vector having components x_i, $i \in L$. A Potts model is an exponential family with canonical statistic vector, $t(x)$, indexed by $D = C \cup \{*\}$, where $*$ is an object not in C. For $c \in C$, $t_c(x)$ is the number of pixels having color c, and $t_*(x)$ is the number of neighboring pairs of variables x_i and x_j with $i \sim j$ that have the same color. Then the components of $t(x)$ can be expressed as

$$\begin{aligned} t_c(x) &= \sum_{i \in L} I(x_i = c), \qquad c \in C, \\ t_*(x) &= \frac{1}{2} \sum_{\substack{(i,j) \in L^2 \\ i \sim j}} I(x_i = x_j), \end{aligned}$$

where $I(\,\cdot\,)$ denotes the function taking logical expressions to the numbers zero and one, false expressions to zero and true expressions to one.

As in any exponential family, increasing the value of one canonical parameter leaving the others fixed increases the expected value of the corresponding canonical statistic. Thus for $c \in C$, increasing the value of θ_c increases the expected number of variables x_i that have color c, and increasing the value of θ_* increases the expected number of pairs of variables x_i and x_j with $i \sim j$ that have the same color.

The parametrization we are using here treats the colors symmetrically but is not identifiable because of the constraint

$$\sum_{c \in C} t_c(x) = |L|,$$

where $|L|$ denotes the cardinality of L (the number of random variables x_i). This implies that the direction $\delta = (1, 1, \ldots, 1, 0)$ is a direction of constancy of the log likelihood, which means θ and $\theta + s\delta$ refer to the same distribution for all real s

(Geyer (2009, Thm. 1)). This lack of identifiability causes no problems so long as we are aware of it.

3. Composite Likelihood for Potts Models

It simplifies notation if we consider vectors to be the same as functions so x_i means the same thing as $x(i)$: the ith component of the vector x is the same as the function x evaluated at the "point" i. Thus the notation C^L denotes the set of all functions $L \to C$, this being no different from the set of all vectors having index set L and values in C.

For any $A \subset L$, let x_A denote the vector-function x restricted to the set A. We can think of it as the subvector $\{ x_i : i \in A \}$, but technically, being a function, it knows its domain A as well as its values x_i, $i \in A$, so it is not "just" a subvector. If A and B are disjoint subsets of L, let $x_A \cup x_B$ denote the function that is the union of the functions x_A and x_B, where now we are thinking set-theoretically of a function as a set of argument-value pairs, so the value of $x_A \cup x_B$ at the index point $i \in A \cup B$ is $x_A(i)$ if $i \in A$ and is $x_B(i)$ if $i \in B$.

Let $f_{A,\theta}$ denote the conditional probability mass function of a Potts model for x_A given the rest of the variables $x_{L \setminus A}$,

$$f_{A,\theta}(x_A \mid x_{L \setminus A}) = \frac{e^{\langle t(x_A \cup x_{L \setminus A}), \theta \rangle}}{\sum_{y \in C^A} e^{\langle t(y \cup x_{L \setminus A}), \theta \rangle}}. \tag{3.1}$$

Considering x_A random and the rest fixed, (3.1) is the likelihood of an exponential family (not the original family but a conditional family derived from it), and has all the properties of such. If $\mathcal{A}$ is a family of subsets of L, then

$$\ell_{\mathcal{A}}(\theta) = \sum_{A \in \mathcal{A}} \log f_{A,\theta}(x_A \mid x_{L \setminus A}) \tag{3.2}$$

is a composite log likelihood (CLL) (Lindsay (1988)) that inherits some properties of a log likelihood. In particular, (3.2) is concave, the maximizer is unique if it exists modulo the non-identifiability described at the end of the preceding section, and partial derivatives of (3.2) set equal to zero are unbiased estimating equations. If $\mathcal{A}$ is the set of all singletons $\mathcal{A} = \{ \{i\} : i \in L \}$, then (3.2) is log Besag pseudo-likelihood.

In addition to calculating and comparing MCLE and MLE, we wish to compare their efficiencies. Calculating asymptotic efficiencies, however, presents some issues. The derivatives of (3.2) are

$$\nabla \ell_{\mathcal{A}}(\theta) = \sum_{A \in \mathcal{A}} s_A(\theta), \tag{3.3a}$$

$$\nabla^2 \ell_{\mathcal{A}}(\theta) = - \sum_{A \in \mathcal{A}} \operatorname{Var}_\theta \{ t(x) \mid x_{L \setminus A} \}, \tag{3.3b}$$

where $s_A(\theta) = t(x) - E_\theta\{t(x) \mid x_{L\setminus A}\}$. The asymptotic variance of the composite likelihood estimator is given by the Godambe-Huber-White sandwich estimator, $W(\theta)^{-1}V(\theta)W(\theta)^{-1}$, where

$$V(\theta) = \operatorname{Var}_\theta\{\nabla \ell_{\mathcal{A}}(\theta)\} \tag{3.4a}$$

$$= \sum_{A\in\mathcal{A}} \sum_{B\in\mathcal{A}} \operatorname{Cov}_\theta\big[s_A(\theta), s_B(\theta)\big], \tag{3.4b}$$

$$W(\theta) = -E_\theta\{\nabla^2 \ell_{\mathcal{A}}(\theta)\} \tag{3.4c}$$

$$= \sum_{A\in\mathcal{A}} E_\theta\big[\operatorname{Var}_\theta\{t(x) \mid x_{L\setminus A}\}\big]. \tag{3.4d}$$

The unconditional expectations and covariances here cannot be calculated exactly but can be estimated by MCMC. The conditional expectations and covariances here can be calculated exactly if the set A is small enough so that the explicit sum with $|C|^{|A|}$ terms that appears in the denominator of the conditional probability mass function (3.1) can be computed.

The asymptotic efficiency of the MLE, of course, is inverse Fisher information $J(\theta)^{-1}$, where

$$J(\theta) = \operatorname{Var}_\theta\{t(x)\}, \tag{3.5}$$

and can be estimated by MCMC.

In spatial applications, however, such as the one we consider, the size of the lattice is typically fixed and so it is unclear if asymptotic efficiency is truly relevant. In fact, it is doubtful that the distribution for the MLE for the lattice sizes we consider will look normally distributed. Combined with the computational limitations described above, we compare empirical standard errors for MCLE and MLE, and also include asymptotic standard errors for MLE for reference.

4. Simulation and Results

In order to deal with the non-identifiability of the symmetric parametrization used above we fix one of the color parameters, say the first, at zero, which is the same as dropping this parameter from the parameter vector. Set

$$g(A) = \sum_{y\in C^A} e^{\langle t(y\cup x_{L\setminus A}),\theta\rangle}. \tag{4.1a}$$

We can then express (3.2) as

$$\ell_{\mathcal{A}}(\theta) = \sum_{A\in\mathcal{A}} \langle t(x_A \cup x_{L\setminus A}), \theta\rangle - \sum_{A\in\mathcal{A}} \log(g(A)). \tag{4.1b}$$

However, (4.1a) overflows for all but the most modestly sized images, rendering direct computation of (4.1b) impossible. We correct this by fixing a base case $\xi_A \in C^A$ for each $A \in \mathcal{A}$ and rewrite (3.1) as

$$f_{A,\theta}(x_A \mid x_{L\setminus A}) = \frac{e^{\langle t(x_A \cup x_{L\setminus A}),\theta\rangle}}{g(A)} = \frac{e^{\langle t(x_A \cup x_{L\setminus A}) - t(\xi_A \cup x_{L\setminus A}),\theta\rangle}}{g'(A)}, \tag{4.2}$$

where

$$g'(A) = \sum_{y \in C^A} e^{\langle t(y \cup x_{L\setminus A}) - t(\xi_A \cup x_{L\setminus A}),\theta\rangle}, \tag{4.3a}$$

so that

$$\ell_{\mathcal{A}}(\theta) = \sum_{A\in\mathcal{A}} \langle t(x_A \cup x_{L\setminus A}) - t(\xi_A \cup x_{L\setminus A}), \theta\rangle - \sum_{A\in\mathcal{A}} \log(g'(A)). \tag{4.3b}$$

The latter is computationally tractable, all terms in (4.3a) and (4.3b) being calculated without overflow.

This alternate way of expressing the CLL may look more familiar for those accustomed to the logistic regression expression for Besag pseudo-likelihood. In the case where each A is a singleton and C has only two colors, so each C^A has just two elements, denoted ξ_A and η_A, (4.3a) becomes

$$g'(A) = 1 + e^{\langle t(\eta_A \cup x_{L\setminus A}) - t(\xi_A \cup x_{L\setminus A}),\theta\rangle}, \tag{4.4a}$$

and (4.3b) becomes

$$\begin{aligned} \ell_{\mathcal{A}}(\theta) &= \sum_{A\in\mathcal{A}} \log\left(\frac{e^{\langle t(x_A \cup x_{L\setminus A}) - t(\xi_A \cup x_{L\setminus A}),\theta\rangle}}{1 + e^{\langle t(\eta_A \cup x_{L\setminus A}) - t(\xi_A \cup x_{L\setminus A}),\theta\rangle}}\right) \\ &= \sum_{A\in\mathcal{A}} y_A \log p_A + (1 - y_A)\log(1 - p_A), \end{aligned} \tag{4.4b}$$

where

$$\begin{aligned} y_A &= \begin{cases} 1, & x_A = \eta_A, \\ 0, & x_A = \xi_A, \end{cases} \\ p_A &= \operatorname{logit}^{-1}(\omega_A) = \frac{e^{\omega_A}}{1 + e^{\omega_A}}, \\ \omega_A &= \langle t(\eta_A \cup x_{L\setminus A}) - t(\xi_A \cup x_{L\setminus A}), \theta\rangle, \end{aligned}$$

and this has the form of log likelihood for logistic regression. The general CLL (4.3b) does not have this form.

The code used to calculate the CLL (4.3b) is available in the R package `potts` (Geyer and Johnson (2010)) available from CRAN. All Potts model simulations

used the MCMC code in this package, which uses the Swendsen-Wang algorithm (Swendsen and Wang (1987); Wang and Swendsen (1990)).

There is no theoretical need to do so, but in order to simplify computation we restrict $\mathcal{A}$ such that all $A \in \mathcal{A}$ have the same shape and size. We simulated samples from Potts models on two different square lattice sizes, 32×32 and 100×100, and with two, three, or four colors. Each time we used for the true value of the parameter

$$\theta = (\theta_1, \ldots, \theta_k, \theta_*)^T = (0, \ldots, 0, \log(1 + \sqrt{|C|}))^T.$$

We chose this value for θ because it corresponds to the phase transition in the lattice (Potts (1952)); see Figure 1 for a sample image from this model at this parameter value. Since $\theta_c = 0$ for all $c \in C$, there is no preference for one color over the others, though it may appear so in Figure 1. All images that result from permuting colors are equally likely and the marginal distribution of any single pixel has all colors equally likely. Geyer (1990, Chap. 6) showed this parameter value to be particularly difficult to estimate by maximum pseudo-likelihood for Ising models, and we believe this is the most difficult for composite likelihood too. Maximum pseudo-likelihood and maximum composite likelihood should have greater efficiency at other parameter values. We are especially interested in the estimates for the $*$-component, which is hardest to estimate.

For each realization, we calculated MCLE using several of the following different sizes and shapes of elements of $\mathcal{A}$. Due to memory and time constraints, we did not use all methods on each image.

Let N denote total number of pixels in the image. The models were

- MPLE. $\mathcal{A}$ was the collection of all singletons in L, and $|\mathcal{A}| = N$.
- Two. Each $A \in \mathcal{A}$ was two horizontally adjacent pixels, with no elements of $\mathcal{A}$ overlapping, and $|\mathcal{A}| = N/2$.
- Two Overlapping. Each $A \in \mathcal{A}$ was two horizontally adjacent pixels, and $|\mathcal{A}| = N$.
- Four. Each $A \in \mathcal{A}$ was a two by two section of the image, with no overlap between different A's, and $|\mathcal{A}| = N/4$.
- Four Overlapping. Each $A \in \mathcal{A}$ was a two by two section of the image, and $|\mathcal{A}| = N$.
- Nine. Each $A \in \mathcal{A}$ was a three by three section of the image, with no overlap between different A's, except where necessary at the edge of the image. We did not use images with rows and columns as multiples of 3, so we used the minimum overlap possible while still having each pixel in the image in at least one element of $\mathcal{A}$, and $|\mathcal{A}| = \lceil n_{row}/3 \rceil \lceil n_{col}/3 \rceil$.

- Nine Overlapping. Each $A \in \mathcal{A}$ was a three by three section of the image, and $|\mathcal{A}| = N$.
- Sixteen. Each $A \in \mathcal{A}$ was a four by four section of the image, with no overlap between different A's, and $|\mathcal{A}| = N/16$.

We refer to our MCLE for θ as $\tilde{\theta}$, and the MLE for θ as $\hat{\theta}$. Figure 2 shows the resulting distributions for the $*$-component of the estimators on the four color 32×32 and 100×100 lattices for MPLE, MCLE (Two and Four non-overlapping models only) and MLE. The distributions for MPLE and MCLE look symmetric and show little bias, with variability that decreases for larger window sizes. This decreasing trend is more evident on the larger 100×100 lattice, where the variability of all estimators is substantially reduced. The MLE exhibits noticeably smaller variability than the MCLE on both lattices, but a conspicuous amount of bias and skewness. The bias and skewness are dramatically reduced for the larger 100×100 lattice but nonetheless present, suggesting comparisons of MCLE should be done to the empirical MLE rather than the asymptotic MLE. The bias and skewness for the MLE are not present for the other components of θ (not shown). MLE were approximated here using 10 iterations of Newton-Raphson starting at the true value for θ, using 10,000 MCMC samples per iteration to approximate $\nabla \ell(\theta)$ and $\nabla^2 \ell(\theta)$.

We explore the efficiency of the MCLE by window size further. Figures 3, and 4 and Tables 1, 2, 3, 4, and 5 clearly show a downward trend in the empirical standard error of the $*$-component of the estimators as the window size increases, with overlapping windows slightly outperforming non-overlapping windows of the same size. The standard errors of the $*$-component decreased by about 10% in most cases when going from the MPLE to MCLE Two. Subsequent decreases in standard error varied from 6-10% with each increase in window size (of course, increases in window size are not regular). The MCLE Sixteen model on the two-color 100×100 lattice had a standard error that was about 7/10 of the MPLE, a notable improvement, but still 1.75 times that of the empirical MLE (Table 1). The MCLE performed better relative to the MLE on the smaller 32×32 two-color lattice, getting as low as 1.4 times the MLE standard error for the Nine Overlap model (Table 3).

Not surprisingly, the best standard error in the plots is for the asymptotic MLE. This is equal to inverse Fisher Information, the inverse of (3.5), estimated here using 1,000 MCMC samples of $t(x)$ that were spaced 500 iterations of the Swendsen-Wang algorithm apart. This spacing was empirically determined so that samples were nearly uncorrelated. The empirical MLE performed much closer to this on the larger 100×100 lattice compared to the 32×32 lattice.

The overall performance of the different estimators was measured here by the empirical root mean square error, $\text{RMSE}(\tilde{\theta}) = \sqrt{(1/n)\sum_{i=1}^{n}(\tilde{\theta}^i - \theta)^T(\tilde{\theta}^i - \theta)}$,

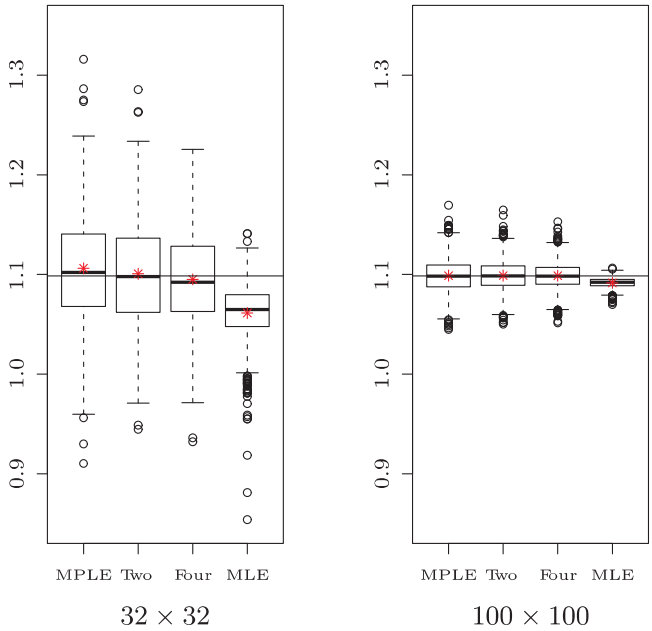

Figure 2. Boxplot of the $*$-component of estimators for a 4-color image, on 32×32 lattice (left) and 100×100 (right). Estimators are MPLE, MCLE (Two and Four), and MLE. The mean of the estimator is plotted with an asterisk and the true value of θ_*, $\log(1+\sqrt{4}) = 1.0986$, is shown by horizontal line.

Table 1. 2 color, 100×100 image. se: standard error for $*$-component of $\tilde{\theta}$ and $\hat{\theta}$, RMSE: root mean square error of $\tilde{\theta}$ and $\hat{\theta}$, n: number of simulations.

	MPLE	Two	Four	Nine	Sixteen	MLE
se	0.0192	0.0173	0.0160	0.0147	0.0137	0.0078
RMSE	0.0322	0.0279	0.0242	0.0222	0.0198	0.0106
n	946	946	946	946	944	931

where $\tilde{\theta}^i$ is the estimator from observation i out of n samples. Although the standard errors for MLE were much lower than for MCLE, this was not the case with RMSE. This is of course due to the bias in the distribution of the MLE on the 32×32 lattice, which is not present in the distribution of the MCLE (Figures 2). With the reduced bias on the 100×100 lattice, the MLE handily outperforms the best MCLE with respect to RMSE by about a factor of two (Tables 1 and 2).

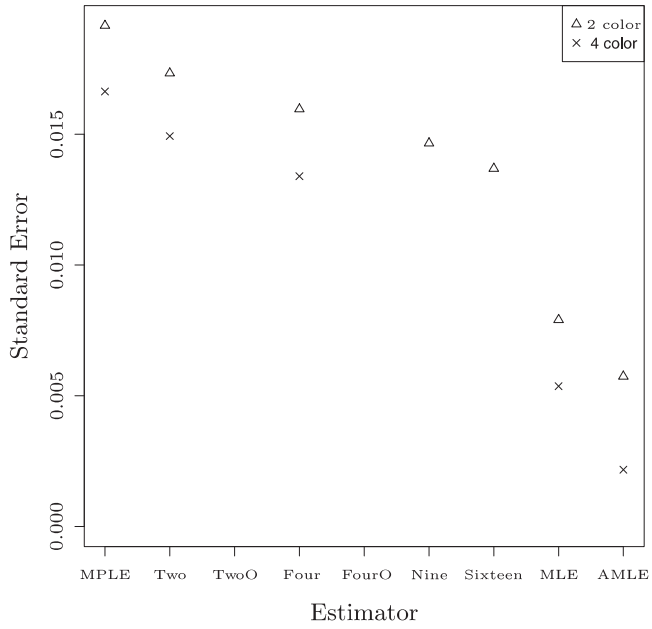

Figure 3. Standard errors of $*$-component for estimators with different window sizes on 100×100 Potts images. TwoO: Two overlapping, FourO: Four overlapping. Values for standard errors are the same as se values in Tables 1 and 2.

Table 2. 4 color, 100×100 image. se: standard error for $*$-component of estimator, RMSE: root mean square error of estimator, n: number of simulations.

	MPLE	Two	Four	MLE
se	0.0166	0.0149	0.0134	0.0052
RMSE	0.0571	0.0506	0.0441	0.0206
n	1893	1893	1893	750

Table 3. 2 color, 32×32 image. se: standard error for $*$-component of estimator, RMSE: root mean square error of estimator, n: number of simulations.

	MPLE	Two	Two overlap	Four	Four overlap	Nine	Nine overlap	MLE
se	0.0680	0.0638	0.0610	0.0591	0.0547	0.0550	0.0505	0.0359
RMSE	0.1367	0.1252	0.1166	0.1146	0.1007	0.1014	0.0906	0.0823
n	1135	1135	1135	1135	1135	1135	1135	1000

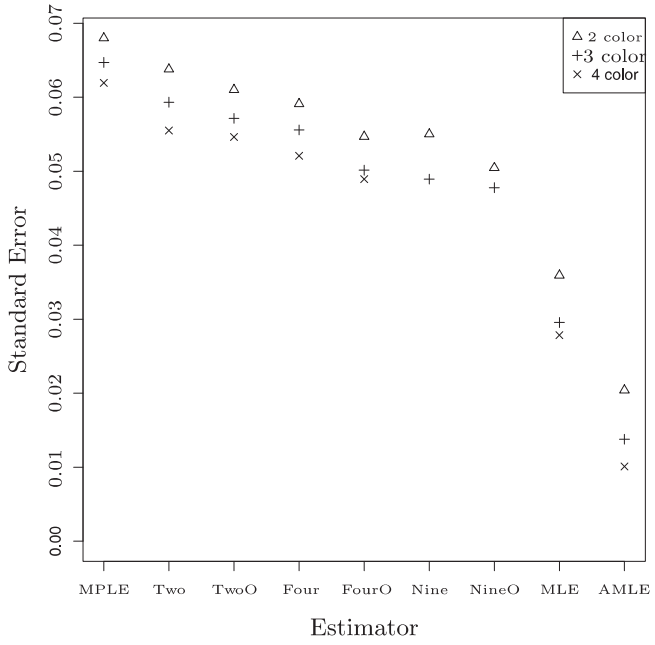

Figure 4. Standard errors of $*$-component for estimators with different window sizes on 32×32 Potts images. TwoO: Two overlapping, FourO: Four overlapping, NineO: Nine overlapping. Values for standard errors are the same as se values in Tables 3, 4, and 5.

Table 4. 3 color, 32×32 image. se: standard error for $*$-component of estimator, RMSE: root mean square error of estimator, n: number of simulations.

	MPLE	Two	Two overlap	Four	Four overlap	Nine	Nine overlap	MLE
se	0.0647	0.0593	0.0571	0.0556	0.0501	0.0489	0.0478	0.0296
RMSE	0.1770	0.1589	0.1535	0.1467	0.1339	0.1340	0.1223	0.1205
n	364	364	364	364	364	94	93	1000

5. Discussion

Since maximizing the CLL when the window A is chosen to be nearly the entire lattice is almost the same as finding the MLE, we anticipated more accurate estimates for larger size windows. Our choice of window size, however, was restricted by our computing resources; computing MCLE for a three by three window (our "Nine" model) for the four color 100×100 lattice requires more memory than our computers had available (but see the Appendix for another approach).

Table 5. 4 color, 32×32 image. se: standard error for $*$-component of estimator, RMSE: root mean square error of estimator, n: number of simulations.

	MPLE	Two	Two overlap	Four	Four overlap	MLE
se	0.0619	0.0555	0.0546	0.0521	0.0489	0.0278
RMSE	0.2265	0.2103	0.2064	0.1999	0.1910	0.1769
n	234	233	233	232	232	991

Our primary interest was in the efficiency of MCLE for different size windows. Since using non-overlapping instead of overlapping windows consumes far less computing resources, we focused on obtaining results for non-overlapping windows for all the lattices we considered. We were also able to produce results for overlapping windows in cases where the demand on computing resources were not overly burdensome.

Figures 3 and 4 show a clear downward trend in the standard error of the $*$-component of the estimators. This trend holds for the estimators of the other components of θ (results not shown).

The exact nature of the trend is not clear, though we believe that that the standard errors must approach that of the MLE when the window is nearly the entire lattice (if this could be done, which it cannot for large lattices). For each component $d \in C \cup \{*\}$ of the MCLE, we have the conjecture

$$\frac{\operatorname{Var}(\tilde{\theta}_d)}{\operatorname{Var}(\hat{\theta}_d)} \propto \frac{N}{|A|},$$

where N is the total number of pixels in the lattice. We do not have any strong reason for this conjecture other than intuition and studying graphs such as Figures 3 and 4. Further investigation is needed to determine the nature of the trend.

Like many problems, MCLE computation suffers from a memory and speed trade off. We chose to cache as much of the calculation as possible to increase speed. Our method creates a list of arrays, storing all the values of $t(y \cup x_{L \setminus A})$ for all $y \in C^A$ for all $A \in \mathcal{A}$. This requires a list of length $|\mathcal{A}|$, each element is an array of size $|C|^{|A|} \times |C|$. Our implementation prefers transparent correctness over computational efficiency. Runtimes for typical runs with 2 or 4 colors on 100×100 and 32×32 images are reported in Table 6. Surprisingly, computation of the estimators for slightly larger non-overlapping window sizes (1×2 and 2×2 windows for the two color Potts models) is quicker than computation of the MPLE. This is due to the reduced size of $\mathcal{A}$; for example, the non-overlapping

Table 6. Runtimes (in seconds) including creating the cache and optimizing the CLL. Times are from a 2GHz computer, using a non-parallelized version.

Colors	Size	MPLE	Two	Two ovrlp	Four	Four ovrlp	Nine	Nine ovrlp	Sixteen
2	100×100	121	67	142	59	264	531	5,133	40,629
4	100×100	219	187	515	977	4,034			
2	32×32	10	6	12	5	24	50	461	3,851
4	32×32	17	12	47	34	365			

1×2-window composite likelihood model has exactly half the number of components as the pseudo-likelihood model. Since each component is very simple, computations are still quite fast, and with fewer components, the overall computation time is less. The speed advantage disappears as $|C|$ increases, as the number of calculations required per component grows exponentially with $|C|$.

We recommend using as large a window size as one's patience and computing resources allow. In terms of efficiency, window size selection seems to clearly follow the guideline "bigger is better". The tradeoff, of course, is longer (sometimes substantially longer) runtimes, which may not be worth the diminishing gains in efficiency. For example, in the two color 100×100 lattice, the Nine non-overlap estimator has a standard error of 0.0147 and a non-parallelized runtime of just under 9 minutes. To get to the lower standard error of 0.0137 associated with the Sixteen non-overlap model, one would have to wait for nearly 11.5 hours. Is 11.5 hours too long? Many of us may say yes, but this becomes a rather subjective discussion involving the value one places on computing time and the importance of the accuracy of the results.

Two possible ways to improve computing speed are parallel computing and implementing a peeling algorithm.

When calculating the CLL, computation for each window $A \in \mathcal{A}$ can be done with no influence on computations for other windows. This makes the CLL easily parallizable. We were able to use the `mclapply` function from the `multicore` package (Urbanek (2009)) for R. This required almost no alteration of the existing code, and using 8 cores for processing achieved a five-fold speedup. However, this approach does not increase the computational feasibility of larger window sizes. To enable MCLE calculations on large window sizes we would need to spread $\mathcal{A}$ across multiple computers.

A peeling algorithm (Cannings, Thompson, and Skolnick (1978); Lauritzen and Spiegelhalter (1988)) could be implemented to efficiently calculate the composite likelihood in slices. For more detail, see the Appendix. In addition to making feasible computations of MCLE for larger size windows than we have

done here, this method could also be used to calculate conditional expectations and covariances. These in turn could be used to calculate asymptotic variances which could be compared more directly to that of the MLE than the empirical variances that we have computed here. Peeling code is, however, tricky and difficult to write and validate.

The composite likelihood approach adds an interesting new perspective to the parameter estimation discussion for exponential families with complex dependencies. While MCLE show a far larger standard error than the MLE, for problems where MCMC methods are difficult to implement, the composite likelihood approach with even moderate window size is a marked improvement over the pseudo-likelihood approach.

Acknowledgement

The idea for this article arose while listening to Nancy Reid's talk about her article in this issue at the 2009 Joint Statistical Meetings.

Appendix

The peeling algorithm (Cannings, Thompson, and Skolnick (1978); Lauritzen and Spiegelhalter (1988)) can be used to efficiently evaluate all of the conditional probabilities, expectations, variances, and covariances in (3.2), (3.3a), (3.3b), (3.4b), and (3.4d).

Suppose we are given a family $\mathcal{B}$ of subsets of L and a family of functions φ_B, $B \in \mathcal{B}$, where φ_B is a function of the variables x_B only, and suppose we wish to evaluate

$$Q = \sum_{x \in C^L} \prod_{B \in \mathcal{B}} \varphi_B(x_B)$$

(where, as above, L is the whole lattice and x_B denotes the restriction of the function x to the set B). The peeling algorithm does this efficiently by summing out one variable at a time. Fix $i \in L$, and define

$$B^* = \bigcup \{ B \in \mathcal{B} : i \in B \} \Big\backslash \{i\},$$
$$\varphi_{B^*}(x_{B^*}) = \sum_{y \in C^{\{i\}}} \prod_{\substack{B \in \mathcal{B} \\ i \in B}} \varphi_B(y \cup x_{B \setminus \{i\}}).$$

Now define

$$\mathcal{B}_{\text{new}} = \{ B \in \mathcal{B} : i \notin B \} \cup \{B^*\},$$
$$L_{\text{new}} = L \setminus \{i\}.$$

Then

$$Q = \sum_{x \in C^{L_{\text{new}}}} \prod_{B \in \mathcal{B}_{\text{new}}} \varphi_B(x_B).$$

Thus we have a problem of the same abstract form. To give a name to the functions φ_B we call them *potentials*. Then the peeling algorithm can be phrased in words as follows. To evaluate the sum of a product of potentials, sum out one variable at a time. Each sum yields a problem of the same form (sum of a product of potentials). After each step we have one fewer variable and one new potential (φ_{B^*} above) involving all the variables that were neighbors of the variable summed out.

In order to carry out the peeling algorithm we must be able to represent each φ_{B*}, which is a function with finite domain and range, so it can be tabulated, taking space $|C|^{|B^*|}$. Depending on the "peeling sequence," which is the order in which the variables are summed out, the size of the maximal such table may grow too large for feasible computation or may stay reasonably small. The peeling algorithm works in practice only for "good" peeling sequences. For Potts models, the optimal peeling sequence for a rectangular region $A = \bigcup \mathcal{B}$ starts in one corner and proceeds along rows, starting at the beginning of the next row when one row is finished. Then the size of B^* is never larger than twice the number of pixels in a row.

All of the summations that seem to have $|C|^{|A|}$ terms in in (3.2), (3.3a), (3.3b), (3.4b), and (3.4d) can thus be done in $O(|C|^r)$ time and space, when A is rectangular and r is the number of rows or columns, whichever is smaller.

References

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Austral. N. Z. J. Statist.* **42**, 283-322. Addendum **44**, 503.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families*. John Wiley, Chichester.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179-195.

Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64**, 616-618.

Cannings, C., Thompson, E. A. and Skolnick, M. H. (1978). Probability functions on complex pedigrees. *Adv. Appl. Probab.* **10**, 26-61.

Geyer, C. J. (1990). Likelihood and exponential families. Unpublished Ph.D. Thesis, University of Washington. `http://purl.umn.edu/56330`.

Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession *Electron. J. Stat.* **3**, 259-289.

Geyer, C. J. and Johnson L. (2010). potts: Markov chain Monte Carlo for Potts models. R package version 0.5.

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.

Hjort, N. L. and Varin, C. (2008). ML, PL, QL in Markov Chain Models. *Scand. J. Statist.* **35**, 64-82.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Z. Phys.* **31**, 253-258.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **50**, 157-224.

Lele, S. R. (2006). Sampling variability and estimates of density dependence: a composite-likelihood approach. *Ecology* **87**, 189-202.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 221-239.

Mardia, K. V., Kent, J. T., Hughes, G. and Taylor, C. C. (2009). Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika* **96**, 975-982.

Potts, R. B. (1952). Some generalized order-disorder transformations. *Proc. Camb. Phil. Soc.* **48**, 106-109.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *J. Amer. Statist. Assoc.* **85**, 204-212.

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

Urbanek, S. (2009). Multicore: Parallel processing of R code on machines with multiple cores or CPUs. R package version 0.1-3.

van Duijn, M. A. J., Gile, K. J. and Handcock, M. S. (2009). A framework for the comparison of maximum pseudo likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks* **31**, 52-62.

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

Wang, J. S. and Swendsen, R. H. (1990). Cluster Monte Carlo algorithms. *Physica A* **167**, 565-579.

Department of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

E-mail: sai@stat.umn.edu

Department of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

E-mail: leif@stat.umn.edu

Department of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

E-mail: charlie@stat.umn.edu

(Received September 2009; accepted July 2010)