Von Mises type statistics for single site updated local interaction random fields

Priscilla E. Greenwood
University of British Columbia

Ian W. McKeague*
Florida State University

Wolfgang Wefelmeyer
University of Siegen

Abstract

Random field models in image analysis and spatial statistics usually have local interactions. They can be simulated by Markov chains which update a single site at a time. The updating rules typically condition on only a few neighboring sites. If we want to approximate the expectation of a bounded function, can we make better use of the simulations than through the empirical estimator? We describe symmetrizations of the empirical estimator which are computationally feasible and can lead to considerable variance reduction. The method is reminiscent of the idea behind generalized von Mises statistics. To simplify the exposition, we consider mainly nearest neighbor random fields and the Gibbs sampler.

Key words and Phrases. Asymptotic relative efficiency, Gibbs sampler, Ising model, Markov chain Monte Carlo, Metropolis algorithm, parallel updating, variance reduction.

---

*Research partially supported by NSF Grant ATM-9417528.

[1] All three authors were partially supported by NSERC, Canada.

[2] *AMS 1991 subject classifications.* Primary: 62M40, 65U05; secondary: 60J05, 62G20, 62M05

[3] *Running title:* Von Mises statistics for random fields.

[4] P. E. Greenwood, Department of Mathematics, University of British Columbia, 1984 Mathematics Road, Vancouver BC, V6T 1Y4, Canada. E-mail: pgreenw@math.ubc.ca.

[5] I. W. McKeague, Department of Statistics, Florida State University, Tallahassee, FL 32306-3033, USA. E-mail: mckeague@stat.fsu.edu; Phone: (850) 644-6690; Fax: (850) 644-5471.

[6] W. Wefelmeyer, FB 6 Math, University of Siegen, Hoelderlin Street 3, 57068 Siegen, Germany. E-mail: wefelmeyer@mathematik.uni-siegen.d400.de.

1 Introduction

Suppose we want to calculate the expectation πf of a bounded function f under a distribution π on a K-dimensional space, based on i.i.d. copies $X^1, \ldots, X^n$ from π. In the nonparametric setting, with nothing known about π, the *empirical estimator* $E_n f = \sum_{i=1}^n f(X^i)/n$ has minimal asymptotic variance; see Bickel, Klaassen, Ritov and Wellner (1993). If the components of π are known to be independent, then $E_n f$ is no longer optimal, and a better estimator of πf is the *generalized von Mises statistic*

$$M_n f = \frac{1}{n^K} \sum_{i_1,\ldots,i_K=1}^{n} f(X_1^{i_1}, \ldots, X_K^{i_K}).$$

Since it is the expectation of f under the product of the marginal empiricals, $M_n f$ has again minimal asymptotic variance if nothing is known about the marginal distributions. Note that the terms $(X_1^{i_1}, \ldots, X_K^{i_K})$ have law π: They are obtained by mixing the components from the different i.i.d. copies $X^i = (X_1^i, \ldots, X_K^i)$. In other words, the von Mises statistic is obtained by replacing values of the components by values with different time indices. This works because there are no interactions either among the K components or among values with different time indices.

The purpose of this paper is to show that there is a way of extending the idea behind the von Mises statistic to evolving random fields that are known to have only local interactions in space and time. Specifically, we consider Markov chain Monte Carlo (MCMC) samplers for random fields with nearest neighbor interactions on a lattice. We mainly study the Gibbs sampler, but our approach naturally extends to other samplers (e.g., Metropolis) that are important in image analysis and spatial statistics for computing features of the underlying random field models (see, e.g., Geman and Geman, 1984, Besag and Green, 1993, Johnson, 1994, Künsch, Geman and Kehagias, 1995, Mollié, 1996).

A well-studied aspect of MCMC is the rate of convergence of the chain to stationarity (see, e.g., Schervish and Carlin, 1992, and Rosenthal, 1995). Here, however, we are interested in estimator variance, which can be more relevant than convergence rate as an optimality criterion (Grenander, 1993, p. 394), especially when the random field has relatively weak interactions so the sampler is rapidly mixing and convergence rate is no longer an issue. Indeed, simulation results of Johnson (1996) that apply in typical image analysis and spatial statistics applications of the Gibbs sampler, show that convergence to stationarity is achieved fairly quickly provided the interactions in the underlying random field are moderate. Many authors have utilized estimator variance as an optimality criterion in the MCMC context, see Peskun (1973), Frigessi, Hwang and Younes (1992), Green and Han (1992), Liu, Wong and Kong (1994), Geyer (1995), and Casella and Robert (1996).

We formulate our results in terms of the Gibbs sampler for a random field π with nearest neighbor interactions on a lattice with K sites. A common estimator for the expectation πf is again the empirical estimator $E_n f$, now based on a realization $X^0, \dots, X^n$ of the Gibbs sampler Markov chain, which has stationary law π. Our goal is to find an estimator with smaller asymptotic variance than $E_n f$. We propose a class of von Mises type estimators that average over terms of the form $f(X_1^{i_1}, \dots, X_K^{i_K})$, similar to $M_n f$. Here the sites on the lattice are labeled $1, \dots, K$. Once more we can use collections $i_1, \dots, i_K$ of time indices for which $(X_1^{i_1}, \dots, X_K^{i_K})$ has stationary law π. Now, however, because of the interactions, the time indices cannot be too far apart.

The collections of time indices that can be used will be characterized in Section 2 in terms of 'admissible update functions.' In Section 3 we use these functions to define the class of von Mises type estimators that are expected to improve upon the usual empirical estimator. In Section 4 we carry out a simulation study of the Ising model to assess the variance reduction of the competing estimators over that of the empirical estimator. Section 5 discusses extensions to other samplers and gives some recommendations on the choice of admissible update functions.

2 Admissible update functions

We begin by describing general Gibbs samplers with deterministic sweeps on a square lattice. Let $S = \{0, \ldots, k-1\}^d$ be a square lattice of dimension d. For simplicity, take k to be even. The lattice has $K = k^d$ sites. Let E denote an arbitrary measurable state space. The configuration space is the product space E^S. Let $\pi(dx)$ be the law of a random field on E^S. For a given site s, it can be factored into marginal and conditional distributions as $\pi(dx) = m_{-s}(dx_{-s})p_s(x_{-s}, dx_s)$, where x_{-s} is obtained from x by omitting x_s. The conditional distributions $p_s(x_{-s}, dx_s)$ are called the *local characteristics* of π. From an initial configuration X^0 a *Gibbs sampler* generates a Markov chain $X^0, X^1, \ldots$ with invariant law π (Geman and Geman, 1984). A Gibbs sampler with deterministic sweep is based on some specified ordering of the sites. It updates a configuration X according to the transition distribution $Q(X, dx) = \prod_s p_s(x_{<s}, X_{>s}, dx_s)$, where $x_{<s}$ is the subconfiguration of all sites that come before site s.

Let f be a bounded measurable function on E^S. The usual estimator for the expectation $\pi f = \int \pi(dx) f(x)$ of f under π is the empirical estimator $E_n f$. For each s, the partially updated configuration $(X_{\leq s}^{i+1}, X_{>s}^i)$ also has stationary law π and further 'empirical' estimators are

$$E_n^s f = \frac{1}{n} \sum_{i=1}^{n} f(X_{\leq s}^{i+1}, X_{>s}^i). \tag{2.1}$$

A possible improvement over any of these estimators is the average

$$\overline{E}_n f = \frac{1}{K} \sum_s E_n^s f; \tag{2.2}$$

see Geweke (1992) or Greenwood, McKeague and Wefelmeyer (1996) for discussion. Can we do better if the field has only local interactions? We show that this is the case. For simplicity, we restrict attention to nearest neighbor interactions.

The set of nearest neighbors of a site s is $\partial s = \{t : |t-s| = 1\}$ with $|t-s| = \sum_j |t_j - s_j|$. We use a free boundary, in which case the boundary sites have fewer than $2d$ neighbors. We assume *nearest neighbor interactions*, $p_s(x_{-s}, dx_s) = p_s(x_{\partial s}, dx_s)$, i.e., the local characteristics at site s depend only on the nearest neighbors of s.

A widely used updating scheme for nearest neighbor models *respects the checkerboard pattern* of the lattice in the sense that it updates first the sites with, say, even parity and then those with odd parity (e.g., Heermann and Burkitt, 1992). A single site s is updated using the local characteristic $p_s(x_{\partial s}, dx_s)$. Note that the sites in the neighborhood ∂s have opposite parity to s. Therefore, all even, or all odd, sites can be updated simultaneously. We write a configuration $x = (y_e, y_o)$, where y_e and y_o are the subconfigurations of x on the even and odd sites respectively. Let $X^0 = (Y^0, Y^1)$ be an initial configuration. The Gibbs sampler based on this updating scheme first creates a subconfiguration Y^2 on the even sites, then a subconfiguration Y^3 on the odd sites, and so on. Here, rather than counting the update of a complete configuration as a time step, we define a *full time step* to be the update of an even *or* an odd subconfiguration (by the clock of a parallel computer say). This means that the output of the Gibbs sampler is $Y^0, Y^1, Y^2, \ldots$, and the sequence of complete configurations is given by $X^0 = (Y^0, Y^1),\ X^1 = (Y^2, Y^3), \ldots$.

To motivate the construction of our estimators, we assume for now that the initial configuration X^0 is distributed according to the stationary law π. Then the Gibbs sampler Markov chain $X^0, X^1, \ldots$ is stationary. Now suppose that we replace a component X_s^i of the configuration X^i by a future value X_s^{i+j}. Which replacements leave the joint law of the configuration unchanged? We have already seen in $E_n^s f$ an example of such replacements for the general case with possibly non-local interactions—we replaced the values X_t^i by X_t^{i+1} for $t \leq s$. We will see that for nearest neighbor models more general replacements are possible.

It is convenient to describe such replacements by an *update function* $I\colon S \to \{0, 1, \ldots\}$, with $I(s)$ even for s even, and odd otherwise. An update function I describes a new configuration $Z^I = (Y_s^{I(s)})_{s \in S}$ in terms of the observed chain $Y^0, Y^1, Y^2, \ldots$ by specifying, for each site s, the time index $i = I(s)$ of the value going into this configuration. For example, the initial configuration is $X^0 = (Y^0, Y^1) = Z^{I^0}$, where

$$I^0(s) = \begin{cases} 0, & s \text{ even}, \\ 1, & s \text{ odd}, \end{cases} \tag{2.3}$$

and X^i is obtained from X^0 by shifting I^0 to yield $X^i = Z^{I^0+2i}$ for $i = 1, 2, \ldots$. We say that an update function is *admissible* if its values at any two neighboring sites differ by 1. A *move* picks a site s, then replaces $I(s)$ by $I(s)+2$, leaving I unchanged otherwise. A move is *admissible* if it preserves admissibility of the update function, see Figure 1. Note that an admissible move can be made at site s if and only if $I(t) = I(s)+1$ for all $t \in \partial s$. Note also that I^0 is an admissible update function, and that all admissible s are built-up by applying finitely many admissible moves to I^0.

A *move* picks a site s, then replaces $I(s)$ by $I(s)+2$, leaving I unchanged otherwise. A move is *admissible* if it preserves admissibility of the update function, see Figure 1. Note that an admissible move can be made at site s if and only if $I(t) = I(s)+1$ for all $t \in \partial s$. Note also that I^0 is an admissible update function, and that all admissible s are built-up by applying finitely many admissible moves to I^0.

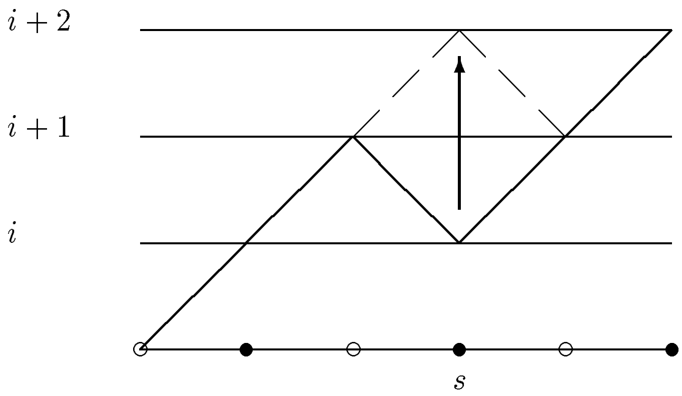

Figure 1: An admissible move at site s.

The following proposition shows that for an admissible update function I, the process Z^{I+2i}, $i = 0, 1, \ldots$ is distributed as output from another Gibbs sampler for π. The sweep of this new Gibbs sampler is *ordered* by I in the sense that it first updates the sites on the lowest level of I, then proceeds upwards layer by layer. In general, the new sweep does not respect the checkerboard pattern.

Proposition *Suppose π has nearest neighbor interactions and $X^i, i = 0, 1, \ldots$ is generated by a Gibbs sampler for π whose updating respects the checkerboard pattern of the lattice. If I is an admissible update function, then $Z^{I+2i}, i = 0, 1, \ldots$ is distributed as a Gibbs sampler for π having sweep ordered by I.*

Proof Apply moves to I site by site, in the order of the associated sweep. Then each move is admissible on the update function generated in the preceding step, and together they create $I + 2$. Fix a site s, and let I_s denote the update function obtained from I by applying the moves at the sites before s in the order of the new sweep. Suppose $I_s(s) = i$. Since I_s is admissible, $I_s(t) = i+1$ for $t \in \partial s$. The move at s replaces $I_s(s) = i$ by $i+2$. Recall that $Y_s^{i+2} = Z_s^{I+2}$ was generated using the conditional law $p_s(Y_{\partial s}^{i+1}, dx_s)$ which equals $p_s(Z_{-s}^{I_s}, dx_s)$. Hence Z^{I+2} is obtained from Z^I using the Gibbs sampler with the new sweep. $\square$

3 Von Mises type statistics

Let π be the law of a random field with nearest neighbor interactions, and consider a Gibbs sampler with an updating scheme that respects the checkerboard pattern of the lattice. Suppose the Gibbs sampler has generated configurations $X^0, \ldots, X^n$. Let f be a bounded measurable function on E^S.

Call an admissible update function I a *template* if it uses part of the initial configuration X^0, i.e., if $\min I$ equals 0 or 1. If the Gibbs samplers for π are ergodic, then each template I gives a strongly consistent estimator for πf,

$$E_n^I f = \frac{1}{n-h+2} \sum_{i=0}^{n-h+1} f(Z^{I+2i}).$$

Here h is the *height* of I, i.e., the number of full time steps it straddles. This means that $\max I$ equals $2h-2$ or $2h-1$.

If the Gibbs samplers for π are geometrically ergodic, then $E_n^I f$ is also asymptotically normal. Its variance, however, can be substantially different from that of the usual empirical estimator $E_n f$. Estimators with reduced variance might be obtained by averaging over some family $\mathcal{I}$ of templates:

$$E_n^{\mathcal{I}} f = \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} E_n^I f,$$

where $|\mathcal{I}|$ denotes the cardinality of $\mathcal{I}$. It seems difficult, however, to determine theoretically those classes of templates that always lead to a variance reduction.

Averaging over templates can be interpreted as symmetrizing $E_n f$, as in a generalized von Mises statistic. In general we expect such estimators to have smaller variance for larger families of templates. However, there is a trade-off in terms of computational cost: For high templates we would need to store more configurations, and for large families we would need to evaluate f more frequently, which could be critical in large lattices or when $f(x)$ is expensive to compute. What are good choices of $\mathcal{I}$?

Let us first consider the situation in which π is the law of an arbitrary random field, with not necessarily local interactions. Then we need a restricted definition of an admissible update function for the Proposition to hold. The corresponding templates will be just those used for the estimators $E_n^s f$ defined in (2.1), with sites $t \leq s$ updated. For s even the template I^s is

$$I^s(t) = \begin{cases} 0, & t > s,\, t \text{ even}, \\ 2, & t \leq s,\, t \text{ even}, \\ 1, & t \text{ odd}, \end{cases}$$

see Figure 2. The template for s odd is similar:

$$I^s(t) = \begin{cases} 1, & t > s,\, t \text{ odd}, \\ 3, & t \leq s,\, t \text{ odd}, \\ 2, & t \text{ even}. \end{cases}$$

As discussed in Section 5, for other templates I, the configurations Z^{I+2i} do not have stationary law π in general. As long as π is considered unknown, and we cannot exploit any special features of the function f, it makes sense to use the average of $E_n^s f = E_n^{I^s} f$ over all templates I^s. This is the estimator $\overline{E}_n f$ defined in (2.2). Note, however, that we do not have strict optimality, not even asymptotically, because the asymptotic covariance matrix of the $E_n^s f$ is not exactly circulant, in general, as pointed out in Greenwood et al. (1996).

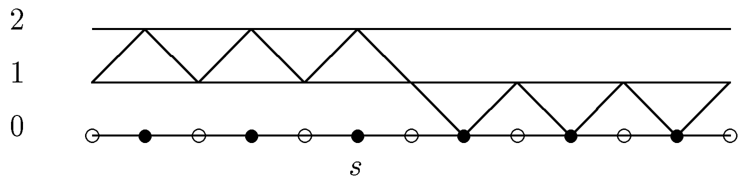

Figure 2: An admissible template when π is arbitrary; I^s for s even.

As mentioned earlier, we recover the usual empirical estimator $E_n f$ by taking $\mathcal{I}$ to be the one template I^0 defined in (2.3). An intermediate choice of $\mathcal{I}$ is I^0 together with the template

$$I^1(s) = \begin{cases} 2, & s \text{ even}, \\ 1, & s \text{ odd}, \end{cases}$$

corresponding to a half-updated configuration. Greenwood et al. (1996) show that one always obtains an asymptotic improvement over $E_n f$ by using both I^0 and I^1. This choice of $\mathcal{I}$ yields the estimator $G_n f = \frac{1}{2}(E_n^{I^0} f + E_n^{I^1} f)$.

We return to nearest neighbor interactions. To make use of the nearest neighbor assumption, we must go beyond the templates just described. For large lattices it may not be computationally feasible to use all templates. If one uses only a few templates, they should be well spaced to reduce correlation between different $E_n^I f$. The higher the templates we allow, the better we can space them. However, high templates require more storage: To calculate $E_n^I f$ for a template I of height h, we must store h configurations at a time.

A good choice of templates also depends critically on the function f. We illustrate this with a simple case, a function $f(x) = f_{st}(x_s, x_t)$, with fixed sites s and t exactly r bonds apart. The r-th nearest neighbor correlation in the simulation study in Section 4 is an average of functions of this form. For such a function, two templates give asymptotically equivalent estimators $E_n^I f_{st}$ if on s and t the differences between the templates are the same. This leaves $r+1$ essentially different templates,

$$\begin{array}{lll} I(s) = I^0(s) & \text{and} & I(t) = I^0(t); \\ I(s) = I^0(s) & \text{and} & I(t) = I^0(t) + 2i,\ I(t) \leq r+1; \\ I(t) = I^0(t) & \text{and} & I(s) = I^0(s) + 2i,\ I(s) \leq r+1. \end{array}$$

What are good choices of templates for functions f_{st}?

Let us again first consider the situation where π is the law of an arbitrary random field, with not necessarily local interactions. There are essentially two templates. If r is even, they are

$$\begin{array}{lllr} I(s) = I(t) = 0 & \text{and} & I(s) = 2,\ I(t) = 0,\ s \leq t, & \text{if } s \text{ and } t \text{ are even,} \\ I(s) = I(t) = 1 & \text{and} & I(s) = 3,\ I(t) = 1,\ s \leq t, & \text{if } s \text{ and } t \text{ are odd.} \end{array}$$

If r is odd, s is even and t is odd, the two templates are

$$\begin{array}{ll} I(s) = 0, & I(t) = 1; \\ I(s) = 2, & I(t) = 1. \end{array}$$

We write $G_n^{st} f_{st}$ for the average of $E_n^I f_{st}$ over the two templates corresponding to s and t. If r is odd, we have $G_n^{st} f_{st} = G_n f_{st}$, and the two templates I^0 and I^1 are the best choice. The estimator $G_n^{st} f_{st}$ is not quite equal to $\overline{E}_n f_{st}$, which does not exploit any special features of the function. Simulations in Greenwood et al. (1996) show that the latter can be slightly worse. A possible explanation is that it does not assign equal weights to the two essentially different templates.

We return again to nearest neighbor interactions. If $r = 1$, we cannot do better than use $G_n f_{st}$. The nearest neighbor assumption cannot be exploited. If $r = 2$, we can do better than $G_n^{st} f_{st}$, the best non-nearest neighbor choice. The full von Mises type estimator is based on three templates. For even s and t they are

$$\begin{array}{ll} I(s) = I(t) = 0; & \\ I(s) = 2, & I(t) = 0; \\ I(s) = 0, & I(t) = 2, \end{array}$$

see Figure 3. For odd s and t they are

$$\begin{array}{ll} I(s) = I(t) = 1; & \\ I(s) = 3, & I(t) = 1; \\ I(s) = 1, & I(t) = 3. \end{array}$$

If $r = 3$, the best non-nearest neighbor choice is again $G_n f_{st}$, based on the two templates I^0 and I^1. The full von Mises type estimator is based on four templates. We expect the two highest of these templates on their own to be better than the two templates I^0 and I^1, because they are better spaced. If r is too large to use all $r+1$ templates, we suggest choosing a well-spaced subset to reduce correlation between the $E_n^I f_{st}$. Good choices are the two highest templates, or the two highest together with the two lowest.

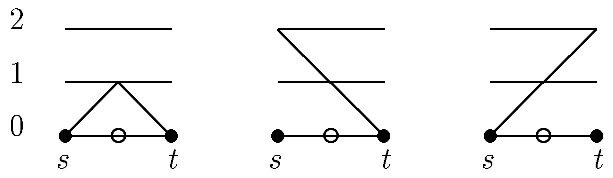

Figure 3: The three essentially different templates for $r = 2$ and s, t even.

4 Ising model simulations

In this section we evaluate the performance of some of the proposed von Mises type estimators when applied to Gibbs and Metropolis samplers for the Ising model.

The state space is $E = \{+1, -1\}$, representing two spin orientations. Under the Gibbs distribution π, a configuration $x \in E^S$ has mass proportional to $\exp(-H(x))$, where the energy function H is given by $H(x) = -\beta \sum_{\langle s,t\rangle_1} x_s x_t$. Here β is the inverse temperature and the sum is over unordered pairs $\langle s,t\rangle_1$ of nearest neighbors s, t. The updating scheme first updates the even sites in a raster scan, then the odd sites. For more about the Ising model see, e.g., Winkler (1995).

We consider the *r-th nearest neighbor correlation* $f(x) = \sum_{\langle s,t\rangle_r} x_s x_t / N$. The sum is over all unordered pairs $\langle s,t\rangle_r$ of r-th nearest neighbors s, t, and N is the number of these pairs. This function is an average over functions $f_{st}(x_s, x_t) = x_s x_t$. Functions of this form were discussed above. We denote $\pi f = \rho_r$, the *expected* r-th nearest neighbor correlation. The decay of ρ_r as r increases is an important descriptor of the spatial structure of the random field and is widely studied by physicists.

By $M_n f$ we denote the full von Mises type estimator, averaging on each pair $\langle s,t\rangle_r$ over all $r+1$ different templates. By $H_n f$ we denote the von Mises type estimator which uses only the two highest templates on each pair $\langle s,t\rangle_r$. We compare these estimators with $\overline{E}_n f$ and with $G_n^* f = \sum_{\langle s,t\rangle_r} G_n^{st} f_{st}/N$, the average over the best non-nearest neighbor choices G_n^{st} of Section 3. For r odd, the latter reduces to $G_n f$.

Variance reductions of these estimators over the usual empirical estimator are reported in Tables 1–4. The results are for a 100×100 lattice with free boundary, and are based on 1000 runs of the Gibbs and Metropolis samplers. For the Metropolis sampler, the proposal at each site is a spin flip. The variance σ^2 of the usual empirical estimator $E_n f$ is given in units of 10^{-6}. Each run of the sampler consisted of a burn-in of 20 sweeps followed by $n = 10$ sweeps used for estimation of ρ_r. For each entry in the tables, the estimator was calculated on the basis of these 10 sweeps for each of 1000 independent runs, and then the empirical variance over the runs was calculated. Simulation results of Johnson (1996) show that 20 sweeps provide an adequate burn-in for the values of β considered here.

Table 1. $\beta = 0.1$

	Gibbs sampler				Metropolis sampler			
	$r=2$	$r=3$	$r=4$	$r=5$	$r=2$	$r=3$	$r=4$	$r=5$
ρ_r	0.015	0.0024	0.0004	0.00007	0.015	0.0024	0.0004	0.00007
σ^2	3.39	2.26	1.60	1.24	9.57	2.49	5.00	1.55
$\overline{E}_n f$	12%	31%	18%	28%	28%	53%	49%	28%
$G_n^* f$	35%	38%	25%	35%	66%	72%	57%	72%
$H_n f$	50%	48%	48%	47%	31%	−33%	63%	−58%
$M_n f$	57%	63%	69%	74%	93%	87%	98%	94%

Table 2. $\beta = 0.2$

	Gibbs sampler				Metropolis sampler			
	$r=2$	$r=3$	$r=4$	$r=5$	$r=2$	$r=3$	$r=4$	$r=5$
ρ_r	0.07	0.02	0.007	0.003	0.07	0.02	0.007	0.003
σ^2	7.85	5.43	4.23	3.21	8.98	3.65	4.78	2.31
$\overline{E}_n f$	4%	8%	8%	8%	16%	20%	30%	10%
$G_n^* f$	16%	13%	13%	9%	44%	23%	33%	22%
$H_n f$	37%	37%	45%	38%	67%	2%	43%	22%
$M_n f$	33%	34%	46%	43%	81%	79%	90%	88%

Table 3. $\beta = 0.3$

	Gibbs sampler				Metropolis sampler			
	$r=2$	$r=3$	$r=4$	$r=5$	$r=2$	$r=3$	$r=4$	$r=5$
ρ_r	0.17	0.09	0.05	0.03	0.17	0.09	0.05	0.03
σ^2	24.5	24.8	23.4	20.5	15.8	14.7	15.0	12.0
$\overline{E}_n f$	2%	1%	3%	0%	8%	3%	10%	1%
$G_n^* f$	2%	2%	2%	1%	10%	4%	8%	3%
$H_n f$	11%	13%	19%	17%	35%	40%	49%	41%
$M_n f$	9%	10%	14%	12%	33%	35%	49%	49%

Table 4. $\beta = 0.4$

	Gibbs sampler				Metropolis sampler			
	$r=2$	$r=3$	$r=4$	$r=5$	$r=2$	$r=3$	$r=4$	$r=5$
ρ_r	0.39	0.30	0.23	0.19	0.39	0.30	0.23	0.19
σ^2	122	204	273	285	93.2	148	197	241
$\overline{E}_n f$	0%	0%	0%	0%	0%	0%	1%	0%
$G_n^* f$	0%	0%	0%	0%	0%	0%	1%	0%
$H_n f$	1%	0%	2%	3%	1%	4%	8%	7%
$M_n f$	1%	0%	1%	1%	1%	3%	5%	4%

The expected r-th nearest neighbor correlation ρ_r naturally decreases with r; the decrease is stronger for small β, i.e., high temperature. In terms of the variance of the empirical estimator $E_n f$, the Metropolis sampler is better at low temperature ($\beta \geq 0.3$), and the Gibbs sampler is better at high temperature ($\beta = 0.1$), which is consistent with a comparison of their rates of convergence to π, see, e.g., Winkler (1995, p.153). The improvement of Metropolis over Gibbs at low temperature is about the same for different r, i.e., for different functions f, and it increases with β. For $\beta = 0.4$ the variance of $E_n f$ is about 25% less under the Metropolis sampler than what it is under the Gibbs sampler.

As anticipated, the full von Mises estimator $M_n f$ has the best general performance of all the estimators considered. At low temperature, however, it can be slightly worse than $H_n f$, even though that estimator is based on only two templates. To see why, note from the proof of the Proposition that a template produces output from a sampler with a sweep that does not in general respect the checkerboard pattern of the lattice, so the asymptotic variance of the corresponding empirical estimator can be different from that of the usual empirical estimator. High templates represent the strongest departures from checkerboard updating, and are the most likely to produce changes in the variance. An extreme example of this occurs with the Metropolis sampler at high temperature and for odd-order nearest neighbor functions, where $H_n f$ has variance much larger than that of the usual empirical estimator. Although $M_n f$ also uses high templates, the other templates compensate for the poor performance of the high templates in this case.

The full von Mises estimator $M_n f$ and the best non-nearest neighbor choice $G_n^* f$ have smaller variance under the Metropolis sampler than under the Gibbs sampler at all temperatures. This is in contrast with the usual empirical estimator, which has smaller variance under Gibbs than under Metropolis at high temperatures.

The improvement of $M_n f$ over $G_n^* f$ and $\overline{E}_n f$ is greater under Metropolis than under Gibbs, probably due to weaker correlations in the Metropolis chain and hence between $E_n^I f$ with distinct I. The improvement is less pronounced at lower temperatures, for which the correlations in the field and hence in the chain are stronger. The best overall performance is obtained from the Metropolis sampler in conjunction with $M_n f$.

5 Discussion

We have introduced a class of von Mises type statistics for estimating features of a local interaction random field via MCMC. Our simulation study has shown that the proposed estimators can produce significant variance reduction over the usual empirical estimator. We now discuss features of the random field and the sampler that are essential for our approach, and those that can be generalized. We also give some recommendations on how to choose the templates for a von Mises type estimator.

Admissible update functions. In general, if an update function I is not admissible, then the stationary law of Z^I is different from π, and the estimator $E_n^I f$ is not consistent for πf. Inadmissible update functions are obtained from admissible update functions by inadmissible moves. A move at a site s, from I to I' say, can be inadmissible in two ways: we may not have $I(t) = I(s) + 1$ for all $t \in \partial s$, or the move may be too high, $I'(s) > I(s) + 2$. In both cases, the value $Z_s^{I'(s)}$ is not generated by conditioning on $Z_t^{I(t)}, t \in \partial s$, but on future values, so $Z^{I'}$ is not distributed as it would be under the Gibbs sampler, i.e., the conditional distribution of $Z_s^{I'}$ given $Z_{-s}^{I'}$ fails to agree with the local characteristic $p_s(Z_{-s}^{I'}, dx_s)$ of π.

Respecting the checkerboard. We have restricted attention to sweeps that update first the even and then the odd sites, or vice versa. This is not necessary, but any sweep that does not respect the checkerboard pattern is worse in terms of statistical efficiency (there are fewer admissible update functions). An extreme case is a sweep which always picks a site adjacent to the previous one. Then the templates I^s give the only admissible update functions, and our approach gets nothing out of the nearest neighbor assumption.

Nearest neighbor interactions. We have considered a random field with nearest neighbor interactions, for which the lattice can be partitioned into two sets of sites such that each set contains no two neighboring sites. More general local interactions may also allow a similar partitioning, perhaps into more than two sets. The minimal number of such sets is the *chromatic number*; see Grenander (1993, p. 382). A good sweep would go through one such set at a time. Of course, the higher the chromatic number, the fewer admissible moves.

Gibbs samplers. When the Gibbs sampler updates a given site, it does not use the present value at that site. It is easy to check that the Gibbs sampler is the only single site updating scheme with this property which is also in detailed balance with π. The property is, however, not essential for our approach. We may also use single site updating rules that use the present value at the site as long as they are *local* in the sense that they condition only on values at nearby sites otherwise. Examples are the Metropolis algorithm and many other Glauber dynamics (reversible spin flip updates) for Ising models, see Neves and Schonmann (1992, p.334–336).

Nearest neighbor functions. For functions depending only on certain sites, one should pick templates differing on these sites. We have already discussed functions depending on just two sites. More generally, if $f(x) = f_A(x_A)$ depends only on the sites in A, two templates I and I' give essentially different estimators $E_n^I f$ and $E_n^{I'} f$ only if I and I' differ on A, and not just by a shift. If A contains many sites or sites lying far apart, there are many essentially different templates. Again, templates should be chosen well spaced to reduce correlation between the corresponding estimators.

Choice of templates. An optimal weighted average of the estimators $E_n^I f$ would require estimation of their asymptotic covariance matrix. This can be done using the method of batch means as discussed in Geyer (1992). Such a weighted average may not be worth the effort, however, and we suggest using simple averages over a well-chosen (possibly quite small) class of templates. Surprisingly, the asymptotic variance of the $E_n^I f$ can change considerably with the choice of template (see the Metropolis sampler results in Section 5), and can be higher or lower than that of the usual empirical estimator. The greatest variance reduction is achieved by using templates that are well-spaced (in terms of the area between the update functions) and thus less correlated. Estimation of the variances/covariances is feasible but again inconvenient. In general, we recommend the use of a combination of (well-spaced) high and low templates, to safeguard against placing excessive weight on templates with large variance.

Acknowledgements The authors thank an Associate Editor and two referees; their suggestions led to considerable improvements in the presentation.

References

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian inference. *Journal of the Royal Statistical Society Series B* **55**, 25–37.

Bickel, P. J. , Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Casella, G. and Robert, C. P. (1996). Rao–Blackwellization of sampling schemes.

Biometrika **83**, 81–94.

Frigessi, A., Hwang, C.-R. and Younes, L. (1992). Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. *Annals of Applied Probability* **2**, 610–628.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In: *Bayesian Statistics* **4** (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), 169–193, Oxford University Press.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science* **7**, 473–483.

Geyer, C. J. (1995). Conditioning in Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* **4**, 148–154.

Mollié, A. (1996). Bayesian mapping of disease. In: *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.), 359–379, Chapman and Hall, New York.

Green, P. J. and Han, X.-L. (1992). Metropolis methods, Gaussian proposals and antithetic variables. In: *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis* (P. Barone, A. Frigessi and M. Piccioni, eds.), 142–164, Lecture Notes in Statistics 74, Springer-Verlag, Berlin.

Greenwood, P. E., McKeague, I. W. and Wefelmeyer, W. (1996). Outperforming the Gibbs sampler empirical estimator for nearest neighbor random fields. *The Annals of Statistics* **24**, 1433–1456.

Grenander, U. (1993). *General Pattern Theory. A Mathematical Study of Regular Struc-*

tures. Clarendon Press, Oxford.

Heermann, D. W. and Burkitt, A. N. (1992). Parallel algorithms for statistical physics problems. In: *The Monte Carlo Method in Condensed Matter Physics* (K. Binder, ed.), 53–74, Springer-Verlag, Berlin.

Johnson, V. E. (1994). A model for segmentation and analysis of noisy images. *Journal of the American Statistical Association* **89**, 230–241.

Johnson, V. E. (1996). Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association* **91**, 154–166.

Künsch, H., Geman S. and Kehagias, A. (1995). Hidden Markov random fields. *The Annals of Applied Probability* **5**, 577–602.

Liu, J. S., 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.

Neves, E. J. and Schonmann, R. H. (1992). Behavior of droplets for a class of Glauber dynamics at very low temperature. *Probability Theory and Related Fields* **91**, 331–354.

Peskun, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika* **60**, 607–612.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association* **90**, 558–566.

Schervish, M. J. and Carlin, B. P. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics* **1**, 111–127.

Winkler, G. (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer-Verlag, Berlin.