

Stochastic models in reservoir characterization and Markov random fields for compact objects

Håkon Tjelmeland

**Stochastic models in reservoir characterization and
Markov random fields for compact objects**

Dr. Ing. Thesis

**Department of Mathematical Sciences
Norwegian University of Science and Technology
1996**

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree “Doktor Ingeniør” (Dr.Ing.) at the Norwegian University of Science and Technology in Trondheim, Norway. It is based on my work during the last four years. This period, I have been so fortunate to be a member of the statistics group at the Department of Mathematical Sciences, which has provided me with a stimulating working environment, and for this I am truly grateful. I will also thank my supervisor Henning Omre for inspiring discussions and support during my studies and express special thanks to Julian Besag at the University of Washington for statistical guidance and pleasant collaboration, both during my stay in Seattle the autumn of 1994 and later.

Thanks also to the staff of the Department of Mathematical Sciences for always being very helpful. The Norwegian Research Council is gratefully acknowledged for financial support.

Håkon Tjelmeland,
Trondheim, April 1996

Thesis outline

This thesis consists of four parts; an overview over Markov chain Monte Carlo methods followed by three papers, in which stochastic simulation is an important element. The papers are:

- I. Markov random fields with higher order interactions.
(with Julian Besag)
- II. Empirical comparison of MPLE and MLE for binary Markov random fields.
- III. A Bayesian framework for integrated reservoir characterization.
(with Henning Omre)

The two first papers are directed to a statistical audience, whereas the last is primarily written for the geostatistical community. The papers can be read independently of each other but the connection between the two first, makes it natural to read I before II.

Stochastic simulation by Markov chain Monte Carlo (MCMC) is extensively used in all three papers of this thesis. The introductory section contains an overview over MCMC algorithms. The use of MCMC started in statistical physics, but is today an important tool in many areas of statistics; see, for example, Besag and Green (1993), Smith and Roberts (1993) and references therein.

The work presented in the three papers is motivated by problems encountered in modeling of geology for evaluation of petroleum reservoirs. The spatial distribution of sedimentary facies, or different rock types, has a major influence on the fluid flow and thereby on the production of oil and gas. Two classes of phenomena can be defined, event and mosaic phenomena; see Haldorsen and Damsleth (1990), Omre (1991) and Ripley (1992). In the first class, one facies dominates and constitutes a background matrix, in which objects of other facies are embedded, and a typical example is small shale units in a sand matrix (Haldorsen and Lake, 1984). Mosaic phenomena represent packing of different facies objects without any rock type constituting a background.

Markov random field (MRF) is a natural choice of model for the spatial distribution for sedimentary facies, especially for mosaic phenomena. MRFs are frequently used as prior distributions in Bayesian image analysis and it often works well in image restoration problems; see Geman and Geman (1984), Besag (1986) and Chellappa and Jain (1993). The attention has mainly been restricted to formulations with pairwise interaction only, although it is well known that these are unable to reproduce the large-scale structures of typical spatial scenes. However, in image analysis the observations usually cover the entire area of interest and contain enough information to eliminate, from the posterior distribution, the long-range correlations present in the prior. In reservoir characterization, the situation is different. The observations are sparse and leave large volumes with little data. It then becomes essential to use a prior, in which the properties of the phenomenon under study is reflected. In paper I, we consider MRFs with higher-order interactions and demonstrate how to formulate models that can represent various spatial structures. Interpretations of the parameters are given and

realizations from the models are generated by Markov chain Monte Carlo. For example, models are defined, realizations from which have compact objects of one color and are without the unrealistic long-range correlations encountered in the Ising and Potts models. Potential applications are illustrated by two examples. The first concerns Bayesian image analysis and confirms that pairwise-interaction priors perform poorly for many image functionals, even if it works well for restoration. In the second example, a MRF is defined for a geological dataset and maximum likelihood parameter estimates are found by Markov chain Monte Carlo.

Paper II concerns parameter estimation in MRFs. Parameter estimation by maximum likelihood is complicated in many spatial models, including MRFs, because the normalizing constants are unavailable in closed form. The maximum pseudo-likelihood (Besag, 1975) has therefore been suggested as a computational simpler alternative. However, with the increased computer resources available, it has also become possible to compute good approximations to the maximum likelihood estimator by use of Markov chain Monte Carlo (Geyer and Thompson, 1992). The paper presents simulation experiments for the Ising model and the foreground/background model introduced in paper I, realizations from which have compact objects of black on a background of white. In each experiment, a realization is generated from the specified model and the maximum pseudo-likelihood (MPLE) and maximum likelihood (MLE) estimators are computed. For the Ising model, the results indicate that both MPLE and MLE give equally good results. Thus, the extra computational cost associated with MLE is not worthwhile. However, for the foreground/background model the results varies with the parameter values. In models with only weak interactions, both MPLE and MLE gave satisfactory results but with MLE slightly better than MPLE. With strong interactions, only MLE provided reliable estimates.

The last paper, which is primarily directed to the geostatistical community, discusses integration of information from different sources for reservoir characterization. Two classes of information are available, (i) general geologic knowledge about the type of reservoir to which the reservoir of study belongs and (ii) observations from the specific reservoir of interest. To combine the different sources of information, a Bayesian approach is adopted. The setting has many similarities with Bayesian image analysis but important differences also exist. First, relevant observations in reservoir characterization are of several types with different support and of varying precision. The observations fall naturally in three groups, measurements in wells, data from seismic surveys and the production history. Observations in wells and the production history are often quite precisely observed, whereas seismic data are typically associated with a high degree of measurement error. Furthermore, well observations usually have little blurring, whereas in seismic data, and especially for the production history, the blurring is severe. Second, the observations in reservoir characterization often carry little information for parts of the reservoir volume and this makes the choice of prior distribution more important than in image analysis. Third, the focus in image analysis often is to make a best guess on the true image, whereas the primary objective of reservoir characterization is to estimate a complex functional of the reservoir characteristics, namely the future production of oil and gas. For given reservoir properties, the production is given as the solution of a set of complex partial differential equations and must be found by numerical simulation. Thus, the possibility to generate realizations from the posterior distribution of the reservoir characteristics is essential. For the posterior to give a realistic representation of the uncertainty of future production, it is important that the prior gives a realistic description of the prior uncertainty in the reservoir characteristics. In the paper, it is argued that this is most easily obtained by choosing a hierarchical model for the prior. The requirement that sampling from

the posterior must be possible, then put constraints on the choice of the prior and these are discussed in the paper. In particular, two examples are discussed, one in the Gaussian family, for which substantial analytical treatment of the posterior is possible, and one with marked point processes where all inference must be based on stochastic simulation.

References

- Besag, J. (1975). "Statistical analysis of non-lattice data", *The Statistician*, **24**, 179-195.
- Besag, J. (1986). "On the statistical analysis of dirty pictures (with discussion)", *J. Royal Statist. Soc. B*, **48**, 259-302.
- Besag, J. and Green, P.J. (1993). "Spatial statistics and Bayesian computation", *J. Royal Statist. Soc. B*, **55**, 3-23.
- Chellappa, R. and Jain, A. (1993). *Markov random fields – Theory and applications*, Academic Press, Boston.
- Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Trans. PAMI*, **6**, 721-741.
- Geyer, C.J. and Thompson, E.A. (1992). "Constrained Monte Carlo maximum likelihood for dependent data (with discussion)", *J. Royal Statist. Soc. B*, **54**, 657-699.
- Haldorsen, H.H. and Damsleth, E. (1990). "Stochastic modeling". *J. of Petroleum Technology*, 404-412.
- Haldorsen, H.H. and Lake, L.W. (1984). "A new approach to shale management in field-scale models", SPEJ, August 1984, 447-457.
- Omre, H. (1991). "Stochastic models for reservoir characterization", in Kleppe, J. and Skjæveland, S.M. (eds.) *Recent advances in improved oil recovery methods for North Sea sandstone reservoirs*, Norwegian Petroleum Directorate, Stavanger, Norway.
- Ripley, B.D. (1992). "Stochastic models for the distribution of rock types in petroleum reservoirs", in Walden, A.T. and Guttorp, P. (eds.) *Statistics in the environmental & earth sciences*, Edward Arnold, London.
- Smith, A.F.M and Roberts, G.O. (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods", *J. Royal Statist. Soc. B*, **55**, 3-23.

Markov chain Monte Carlo

1 Introduction

With increasing computer power available the last decades, stochastic simulation has become an important tool in many areas of statistics. For example, many statistical problems can be formulated as computation of an expectation $\mu_f = E_\pi\{f(X)\}$ for some function $f(\cdot)$, where X has a specified distribution $\pi(x)$, $x \in \Omega$. If independent samples x^1, \dots, x^S from $\pi(x)$ can be generated, an unbiased estimator of μ_f is given by

$$\hat{\mu}_f = \frac{1}{S} \sum_{s=1}^S f(x^s) \quad (1)$$

and an unbiased estimator of the estimation variance is

$$\hat{\sigma}_{\hat{\mu}_f}^2 = \frac{1}{S-1} \sum_{s=1}^S (f(x^s) - \hat{\mu}_f)^2. \quad (2)$$

One should note that this formulation includes computation of probabilities for specific events, if $f(x)$ is an indicator function. Moreover, the same samples can be used to estimate expectations of several functions $f(x)$. In Bayesian analysis of complex systems, another important application of stochastic simulation is to validate the prior distribution. The prior is often specified in terms of local conditional distributions and properties of the corresponding joint distribution are unknown. To inspect realizations from different alternative priors can therefore be of great help in the model specification process.

Several techniques exist to produce independent samples from a given probability distribution; see the overview in Ripley (1987). However, for many distribution, especially high-dimensional ones, no viable methods exist. When independent sampling is infeasible, an almost just as good alternative is to generate dependent samples and this can be done by Markov chain Monte Carlo (MCMC). In MCMC, one specifies a Markov chain with state space Ω and limiting distribution $\pi(x)$ and simulates the chain for a fixed number of steps, known as the “burn-in” period, after which the samples are essentially distributed according to the target distribution, $\pi(x)$. One can then collect a sample every k -th iteration, say, to obtain x^1, \dots, x^S . The samples are no longer independent but from the correct distribution. An unbiased estimator for μ_f is still given by equation (1) but the estimator for the associated variance must be modified to take into account dependence between the samples.

The use of MCMC started in statistical physics, where the aim is to estimate macroscopic properties of a system of interacting particles. Spatial statistics, especially Bayesian image analysis, was the first area of statistics where MCMC was extensively used; see Grenander (1983) and Geman and Geman (1984). Several other areas of statistics have followed, such as Bayesian analysis, hierarchical modeling, likelihood inference and statistical tests; see, for example, Besag and Green (1993) and references therein.

The following sections describe and discuss different MCMC-techniques. For simplicity of explanation and to avoid technicalities, we assume the target distribution, $\pi(x)$, to be discrete over a finite sample space Ω . MCMC can, however, be used for distributions defined on more general state spaces. The associated theory is essentially the same but sums must be replaced by integrals with respect to the appropriate measures. But note that special care

may be necessary when the state space is a union of spaces of different dimensionality; see Green (1995). To settle thoughts, much of the discussion relate to a specific situation, which is also of interest in paper I and II. Consider a lattice with N sites, or pixels, labeled from one to N . To each pixel there is associated a stochastic variable, X_i for pixel i , which takes one out of K values, $X_i \in \Lambda = \{0, 1, \dots, K - 1\}$ say. Finally, the target distribution $\pi(x)$ is the probability distribution for $X = (X_1, \dots, X_N)$. The sample space is thus $\Omega = \Lambda^N$.

2 Markov chains

For a general introduction to Markov chains, see, for example, Cox and Miller (1970). Here, we summarize the basic notions necessary as a background for the discussion of MCMC-techniques. A sequence of stochastic variables $\{X^s\}_{s=0}^{\infty}$ with common state space Ω is said to be a (discrete time) Markov chain if

$$\text{Prob}\{X^{s+1} = x' | X^0 = x^0, X^1 = x^1, \dots, X^s = x\} = \text{Prob}\{X^{s+1} = x' | X^s = x\} \quad (3)$$

for all $s > 0$ and states $x^0, x^1, \dots, x^{s-1}, x, x' \in \Omega$. A Markov chain is said to be stationary if the right side of equation (3) is independent of the time variable s and we then write

$$P(x \rightarrow x') = \text{Prob}\{X^{s+1} = x' | X^s = x\}. \quad (4)$$

The sample space, Ω , is finite, let $|\Omega|$ denote the number of elements and number the elements from 1 to $|\Omega|$. The transition probabilities, $P(x \rightarrow x')$ for $x, x' \in \Omega$, can then be arranged in a transition matrix P , where element (i, j) is the probability for transition from state number i to state number j . However, we continue to use the notation $P(x \rightarrow x')$ for elements in the matrix P . A stationary Markov chain is fully specified by the transition matrix P and a distribution for the initial state, x^0 . Denote the latter by $v_0(x^0)$; $x^0 \in \Omega$.

A distribution $\pi(x)$ is said to be a stationary distribution for the Markov chain if

$$\pi(x') = \sum_{x \in \Omega} \pi(x) \cdot P(x \rightarrow x') \quad (5)$$

for all $x' \in \Omega$. It follows easily that a sufficient condition for this is the so-called detailed balance condition, given by

$$\pi(x') \cdot P(x' \rightarrow x) = \pi(x) \cdot P(x \rightarrow x') \quad (6)$$

for all $x, x' \in \Omega$. It can be shown that a Markov chain which fulfill this condition has unchanged statistical properties if the time axis is reversed and the chain is therefore said to be time reversible. Stronger theoretical results are available for this class of Markov chains and therefore sometimes preferred in MCMC.

A stationary distribution, $\pi(x)$, is unique if the Markov chain is aperiodic and irreducible. A sufficient condition for aperiodicity is existence of at least one state x with $P(x \rightarrow x) > 0$ and this is often simple to verify. For a Markov chain to be irreducible, there must be a positive probability for coming from any state x to any other state x' in a finite number of steps. For an aperiodic and irreducible Markov chain, the distribution of X^s converges to the unique stationary distribution $\pi(x)$, regardless of $v_0(x^0)$. In this case, the $\pi(x)$ therefore is also called the limiting distribution.

In traditional use of Markov chains, the transition probabilities, $P(x \rightarrow x')$, are typically given by the problem at hand and the goal is to find the corresponding limiting distribution. In MCMC the situation is reversed, the distribution $\pi(x)$ is dictated by the problem of interest and the goal is to construct a Markov chain with $\pi(x)$ as limiting distribution. One should note that there exist a large flexibility in the choice of transition matrix. There are $|\Omega|^2$ elements in P . Each row must sum to one and this gives $|\Omega|$ constraints. Equation (5) also impose $|\Omega|$ constraints and this leaves $|\Omega| \cdot (|\Omega| - 2)$ degrees of freedom in the specification of the transition matrix. Alternatively, if a reversible Markov chain is required, equation (6) impose $|\Omega| \cdot (|\Omega| - 1)/2$ constraints and $|\Omega| \cdot (|\Omega| - 1)/2$ degrees of freedom remain. There exist general prescriptions for the construction of suitable transition matrices and this is the topic of the next section. However, first we describe how different transition matrices with the same stationary distribution can be combined to form new matrices with unchanged stationary distribution.

Let P_1, \dots, P_M be M transition matrices for different Markov chains, all with $\pi(x)$ as a stationary distribution. These “base” transition matrices are, however, not required to be aperiodic or irreducible, which implies that corresponding limiting distributions need not exist. Based on P_1, \dots, P_M , one can construct new transition matrices, P , with the same stationary distribution. One possibility is, in each iteration, to draw at random with probabilities $\gamma_1, \dots, \gamma_M$ which transition matrix to use, i.e. to set

$$P = \sum_{m=1}^M \gamma_m \cdot P_m. \quad (7)$$

By insertion in equation (5) it follows easily that P has $\pi(x)$ as a stationary distribution. Moreover, the Markov chain is reversible if all base matrices are. An alternative construction, which also defines a Markov chain with $\pi(x)$ as a stationary distribution, is

$$P = P_1 \cdot \dots \cdot P_M. \quad (8)$$

However, in this case the Markov chain is in general not reversible, even if all the base matrices define reversible chains. If the chain is required to be reversible, one can instead use the product form

$$P = P_1 \cdot \dots \cdot P_M \cdot P_M \cdot \dots \cdot P_1, \quad (9)$$

which defines a reversible chain if the chains of the base matrices are reversible.

3 Metropolis-Hastings algorithm

Most MCMC-algorithms in use are based on a prescription given in Hastings (1970), in which it is specified how to construct reversible Markov chains with a specified stationary distribution, $\pi(x)$. The approach of Hastings is a generalization of the procedure in Metropolis et al (1953) and the transition probabilities are given as a product of two terms

$$P(x \rightarrow x') = Q(x \rightarrow x') \cdot \alpha(x \rightarrow x') \quad (10)$$

for $x \neq x'$, and

$$P(x \rightarrow x) = 1 - \sum_{x' \neq x} P(x \rightarrow x'). \quad (11)$$

Here, $Q(x \rightarrow x')$ are elements in a transition matrix Q on Ω , which must fulfill $Q(x \rightarrow x') > 0 \Leftrightarrow Q(x' \rightarrow x) > 0$ but is otherwise arbitrary. The $\alpha(x \rightarrow x')$ has interpretation of an acceptance probability and is given by

$$\alpha(x \rightarrow x') = \min \left\{ 1, \frac{\pi(x')}{\pi(x)} \cdot \frac{Q(x' \rightarrow x)}{Q(x \rightarrow x')} \right\}. \quad (12)$$

It is easily verified that the resulting transition matrix, P , fulfills the detailed balance conditions of equation (6) and thereby have the required stationary distribution. If the Markov chain is also aperiodic and irreducible, $\pi(x)$ is the limiting distribution and the corresponding Metropolis-Hastings algorithm is as follows.

1. Draw the initial state x^0 from $v_0(\cdot)$.
2. For $s = 1, 2, \dots$, do:
 - (a) draw a potential new state x' from the distribution $Q(x^{s-1} \rightarrow \cdot)$.
 - (b) compute the acceptance probability, $\alpha(x^{s-1} \rightarrow x')$, given by equation (12).
 - (c) With probability $\alpha(x^{s-1} \rightarrow x')$ accept x' as the new state, i.e. set $x^s = x'$, otherwise set $x^s = x^{s-1}$.

A very important property of the algorithm is that $\pi(x)$ appears only in the ratio $\pi(x')/\pi(x)$ in the expression for the acceptance probability and therefore needs to be known up to proportionality only. This is essential for the usability of the approach in areas as Bayesian analysis and spatial statistics, where the normalizing constants typically are unavailable. For example, consider the Ising model frequently used as prior distribution in Bayesian image analysis. The Ising model is within the family of stochastic fields described in the introduction, have $K = 2$ possible values for each of the N pixels and probability distribution given by

$$\pi(x) = c \cdot \exp \left\{ -\beta \cdot \sum_{i \sim j} I(x_i \neq x_j) \right\}, \quad (13)$$

where c is the normalizing constant, $I(\cdot)$ is the indicator function and the sum is over all pairs of nearest neighbor pixels i and j . The normalizing constant is given by

$$c = \left(\sum_{x \in \Omega} \exp \left\{ -\beta \sum_{i \sim j} I(x_i \neq x_j) \right\} \right)^{-1}, \quad (14)$$

where the number of terms in the outer sum is 2^N . A small image may consist of $N = 64 \times 64$ pixels and this gives 10^{1233} terms in the sum. Even if the normalizing constant in principle can be computed, it is impossible in practice.

The prescription of Hastings is often used to construct Markov chains which are not necessarily aperiodic or irreducible. It is rather used to define several transition matrices, all with $\pi(x)$ as a stationary distribution, which are combined to form aperiodic and irreducible chains. Below, the perhaps simplest such construction is discussed.

3.1 Single-site Metropolis-Hastings

Reconsider the random field distribution specified in the introduction. Thus, $\pi(x)$ is the distribution for a random vector $X = (X_1, \dots, X_N)$, where $X_i \in \Lambda = \{0, 1, \dots, K-1\}$ denotes the value of pixel i in a lattice of N pixels. For each pixel i , let Q_i be a transition matrix, which propose changes only for x_i , i.e.

$$Q_i(x \rightarrow x') = \begin{cases} \nu_i^x(x'_i) & \text{if } x = (x_1, \dots, x_N) \text{ and } x' = (x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_N) \\ 0 & \text{otherwise,} \end{cases} \quad (15)$$

where the univariate proposal distribution, $\nu_i^x(\cdot)$, may depend on the current state x . Furthermore, for each pixel i , use equations (10) and (11) to define a corresponding transition matrix P_i , with $\pi(x)$ as stationary distribution. The distribution $\nu_i^x(\cdot)$ may be chosen to have a very simple form, for example $\nu_i^x(x'_i) = 1/K$ for $x'_i \in \Lambda$ or

$$\nu_i^x(x'_i) = \begin{cases} \frac{1}{K-1} & \text{if } x'_i \neq x_i \\ 0 & \text{otherwise,} \end{cases} \quad (16)$$

where the latter has the intuitive appeal that the potential new state, x' , always differ from the current one; see also paper I and II where this version is used. Alternatively, the $\nu_i^x(\cdot)$ can be adapted to the target distribution and a natural choice then, is

$$\nu_i^x(x'_i) = \pi(x'_i | x_j; j \neq i). \quad (17)$$

This is known as the Gibbs sampler (Geman and Geman, 1984) or the heat-bath algorithm. It has the special feature that all acceptance probabilities are equal to one but is not the only single-site Metropolis-Hastings scheme with this property; see Barone and Frigessi (1989) and Green and Han (1992).

The Gibbs sampler is perhaps the most popular MCMC-algorithm. It is intuitive and seems to perform reasonably well for a large variety of target distributions. This makes it suitable for use in general simulation software; see BUGS (Thomas et al, 1992; Gilks et al, 1994). However, it should be noted that for any particular target distribution, more efficient algorithms than Gibbs sampler can easily be devised. Variations of Gibbs sampler where values in blocks of pixels are changed simultaneously are also defined; see the discussion in Besag et al (1995).

The use of the transition matrix P_i is often termed as an “update” of x_i or as a “visit” to pixel i . Different combinations of the base transition matrices therefore correspond to alternative visiting schedules. For example, the P in equation (7), with $M = N$ and $\gamma_m = 1/N$, corresponds to visit the pixels in a random order, whereas the product form in equation (8) uses a systematic scan through the pixels. The P in equation (9) goes systematic through the pixels, first in one order and then reversed. Yet another possibility, which produces a reversible chain, is proposed in Mezei (1981) and visits the pixels in an order which is randomly selected before each cycle.

If the target distribution has high correlations between elements in x or is strongly multi-modal, algorithms with only single-site updates often get very slow convergence. The chains are said to be “slow mixing”. To improve the convergence properties one must consider chains with simultaneous updates of several components. In the multi-modal case, the ideal proposal matrix, Q , gives jumps between modes. However, it is also essential that the proposal

distributions, $Q(x \rightarrow \cdot)$, are easy to sample from and the acceptance probabilities fast to compute. Finally, the acceptance probabilities must obviously not become too low. To specify Q -matrices with these properties is often very hard and requires a profound understanding of the target distribution in question. Some distributions have particular structures, which may suggest a solution; see, for example, paper III where the hierarchical structure of the prior makes it natural to use changes on two different scales to sample from the posterior. But more general strategies have also been suggested in the literature and this is the topic of the following sections.

3.2 Auxiliary variables method

For some target distributions, $\pi(x)$, the introduction of auxiliary variables makes it possible to construct simple Markov chains that propose simultaneous changes in many components of x . Let u denote a vector of auxiliary variables with state space Υ . The joint distribution of x and u is then specified by $\pi(x)$ and the conditional distribution for u given x , $\pi(u|x)$, through

$$\pi(x, u) = \pi(x) \cdot \pi(u|x) \quad (18)$$

for $(x, u) \in \Omega \times \Upsilon$. The auxiliary variables method proceeds by constructing a Markov chain with $\pi(x, u)$ as limiting distribution. It uses a transition matrix $P = P_1 \cdot P_2$, where P_1 and P_2 update x and u , respectively. By a reasonable choice of $\pi(u|x)$, it is easy to construct P_2 to make simultaneous changes in several components of u . For example, by drawing the new u from $\pi(u|x)$ itself. However, for the approach to be of any value, the corresponding $\pi(x, u)$ must also make it possible to construct a P_1 that make simultaneous changes in many components of x . No general prescription exists for how to do this. It requires a deep insight in the particular target distribution and this is the challenge and limitation in use of the method. The potential of the approach was first shown in the paper of Swendsen and Wang (1987). Several generalizations and variations have later been proposed; see Edwards and Sokal (1988), Wolff (1989), Besag and Green (1993), Higdon (1993), Møller (1993) and Geman (1993). Here, we describe the original algorithm of Swendsen and Wang, which illustrates very nicely how auxiliary variables can make it possible to specify simple transition matrices with simultaneous changes in many components of x .

The Swendsen-Wang algorithm simulates from the Ising model specified in equation (13) and introduces one auxiliary variable for each term in the sum in this expression. Let u_{ij} denote the variable corresponding to pixels i and j and let u be the vector of all auxiliary variables. The u_{ij} 's take one out of two values, zero or one, and are assumed to be conditionally independent for given x . If $x_i = x_j$, u_{ij} is zero or one with probabilities $e^{-\beta}$ and $1 - e^{-\beta}$, respectively. When $x_i \neq x_j$, u_{ij} equals zero with probability one, i.e.

$$\pi(u|x) = \prod_{i \sim j} \left[\left((1 - e^{-\beta})^{u_{ij}} \cdot (e^{-\beta})^{(1-u_{ij})} \right)^{I(x_i=x_j)} \cdot I(x_i \neq x_j \cup u_{ij} = 0) \right]. \quad (19)$$

Using $\pi(x|u) \propto \pi(x) \cdot \pi(u|x)$, it follows from straight forward calculations

$$\begin{aligned} \pi(x|u) &\propto \prod_{i \sim j} \left[(e^{\beta} - 1)^{u_{ij} I(x_i=x_j)} \cdot I(x_i \neq x_j \cup u_{ij} = 0) \right] \\ &\propto \prod_{i \sim j} I(x_i \neq x_j \cup u_{ij} = 0), \end{aligned} \quad (20)$$

where, to obtain the last expression, one must use

$$I(x_i = x_j \cup u_{ij} = 0) = 1 \left. \vphantom{I(x_i = x_j \cup u_{ij} = 0)} \right\} \Rightarrow I(x_i = x_j) = 1. \quad (21)$$

It follows from equation (20) that $\pi(x|u)$ is a uniform distribution over all states x where $x_i = x_j$ whenever $u_{ij} = 1$. The interpretation of $u_{ij} = 1$ therefore is existence of a bond between pixels i and j , which restricts them to have the same value. Sampling from $\pi(x|u)$ can be done by first forming clusters of pixels connected by bonds and thereafter draw values independently at random for each cluster.

The Swendsen-Wang algorithm is a block Gibbs sampler for $\pi(x, u)$, where P_1 draws the new x from $\pi(x|u)$ and P_2 the new u from $\pi(u|x)$. As for the single-site Gibbs sampler, all acceptance probabilities are equal to one. The resulting Markov chain is obviously aperiodic and it is also irreducible since it is possible to come from any state (x, u) to any other state (x', u') in two steps, by going via a state with all u_{ij} 's equal to zero.

3.3 Multi-grid method

The multi-grid method (Goodman and Sokal, 1989) gives, as the auxiliary variables method, a framework for construction of transition matrices that make large changes in the state vector, x , without obtaining too small acceptance probabilities. But unlike the use of auxiliary variables, no extended sample space is introduced. As suggested by its name, the method is originally proposed for distributions of variables related to pixels on a grid, but its applicability is not restricted to this situation. Here, we first describe the method in a general setting before concentrating on a specific example.

The multi-grid method is a specific type of Metropolis-Hastings algorithms, where some extra concepts and notation are introduced in order to enable large changes of the state vector. Several levels, $l = 0, 1, \dots, L$, are considered in the multi-grid method and to each level, l , there is associated a state vector v_l which takes values in a sample space Ω_l . In particular, level 0 is the level of interest and one has $v_0 = x$ and $\Omega_0 = \Omega$. Moreover, for each l there exists a partition of Ω_l into $\Gamma_l^r, r \in \mathcal{R}$, where \mathcal{R} is some index set, i.e. $\bigcup_{r \in \mathcal{R}} \Gamma_l^r = \Omega_l$ and $\Gamma_l^r \cap \Gamma_l^{r'} = \emptyset$ for $r \neq r'$. Finally, for each Γ_l^r there exists a corresponding function $v_{l+1} = g_l^r(v_l)$ which is one-to-one and on for $v_l \in \Gamma_l^r$ and $v_{l+1} \in \Omega_{l+1}$. Thus, $g_l^r(\cdot)$ is invertible and we let $v_l = h_l^r(v_{l+1})$ denote the corresponding inverse function.

The multi-grid method alternates between updates on the different levels by going up and down one level at a time. The sequence in which the different levels are visited is predetermined and a ‘‘W-cycle’’ is often used; see Figure 1. A fixed number of updates is done at each level. To describe how to go up or down one level, it suffices to specify how to change between levels 0 and 1. To go up or down between other levels, l and $l + 1$, is done similarly. Let $l = 0$ be the current level and let $v_0 = x$ be the current state vector with distribution, $\pi_0(v_0)$, equal to the target distribution. Let r be the index value, for which $v_0 \in \Gamma_0^r$, identify corresponding values v_0 and v_1 through the relation $v_1 = g_0^r(v_0)$ and let this induce a distribution $\pi_1(v_1)$ for $v_1 \in \Omega_1$ by

$$\pi_0(v_0 \in A | v_0 \in \Gamma_0^r) = \pi_1(v_1 \in \{u_1 : u_1 = g_0^r(u_0); u_0 \in A\}) \quad (22)$$

for all sets $A \subseteq \Gamma_0^r$. Note that, although suppressed in the notation, the distribution $\pi_1(v_1)$ is also dependent on r . Metropolis-Hastings steps, typically chosen so that each step changes

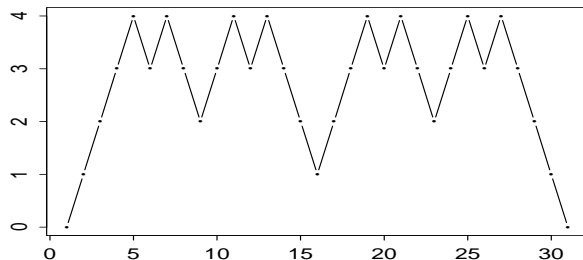


Figure 1: A typical W-cycle with $L = 4$ used in multi-grid algorithms.

only one element of v_1 , are then performed for v_1 with respect to the distribution $\pi_1(v_1)$. By the one-to-one relation between v_1 and v_0 , any change in v_1 has a corresponding change in v_0 . Moreover, from the relation in equation (22) it follows that for any update of v_1 which fulfill the detailed balance conditions with respect to $\pi_1(v_1)$, the corresponding update in v_0 also fulfill the detailed balance conditions with respect to $\pi_0(v_0)$. To return to level $l = 0$, one therefore should just compute $v_0 = g_0^r(v_1)$ with the updated v_1 .

As an example for how the multi-grid approach can be used in practice, we consider a target distribution from Bayesian image analysis and use a continuous distribution because multi-grid is most naturally used in this setting. Thus, let the elements in the state vector, x , correspond to N pixels in a two-dimensional rectangular lattice and let the target distribution be given by

$$\pi(x) \propto \exp \left\{ - \sum_{i=1}^n \frac{(y_i - x_i)^2}{2\sigma_i^2} - \beta \cdot \sum_{i \sim j} (x_i - x_j)^2 \right\} \quad (23)$$

for $x \in \Omega = R^N$, where the second sum is over all pairs of pixels i and j which are immediate neighbors. The x_i represents the (unknown) true value in pixel i , y_i is the observation in pixel i and is assumed to be Gaussian with expectation x_i and variance σ_i^2 . The true scene, x , is assigned a pairwise interaction prior; see also Green and Han (1992).

Now, consider one grid for each level l , where the grid on level $l + 1$ is coarsened by a factor of two in each coordinate direction relative to the grid on level l . Thus, $\Omega_l = R^{N_l}$ where $N_0 = N$ and $N_{l+1} = N_l/4$ for $l \geq 0$. Each pixel on level $l + 1$ has four associated pixels on level l as indicated by arrows in Figure 2. To connect the different levels, let element k of v_{l+1} be equal to the mean of the values in the four associated pixels on level l . To make this precise, let C_k^{l+1} denote the set of four pixels on level l associated to pixel k on level $l + 1$ and let $\kappa_l(i)$ be the pixel on level $l + 1$ associated to pixel i on level l . The Γ_l^r is then given by

$$\Gamma_l^r = \left\{ v_l = (v_{l1}, \dots, v_{ln_l}) : v_{li} - \frac{1}{4} \sum_{j \in C_{\kappa_l(i)}^{l+1}} v_{lj} = r_i ; i = 1, \dots, N_l \right\}, \quad (24)$$

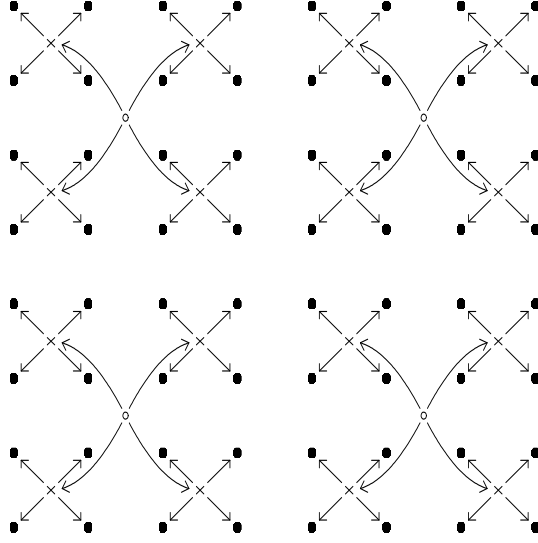


Figure 2: The multi-grid used in the example, with $L = 2$ and a 8×8 grid on level 0. The pixels on levels 0, 1 and 2 are marked with blobs, crosses and circles, respectively. The sets C_k^l , for pixels on levels 1 and 2, are indicated with arrows.

where

$$r \in \mathcal{R} = \left\{ r = (r_1, \dots, r_{N_l}) : \sum_{i \in C_k^{l+1}} r_i = 0; k = 1, \dots, N_{l+1} \right\}. \quad (25)$$

Element k of $g_l^r(\cdot)$ is given by

$$(g_l^r(v_l))_k = \frac{1}{4} \sum_{i \in C_k^{l+1}} v_{li}, \quad (26)$$

so that element i of the corresponding inverse function, $h_l^r(\cdot)$, becomes

$$(h_l^r(v_{l+1}))_i = r_i + v_{l+1, \kappa_l(i)}. \quad (27)$$

The linearity in equation (26) gives

$$\pi_{l+1}(v_{l+1}) \propto \pi_l(h_l^r(v_{l+1})), \quad (28)$$

from which it follows that $\pi_l(v_l)$ get the same form for all l , namely

$$\pi_l(v_l) \propto \exp \left\{ - \sum_{i=1}^{N_l} \frac{(y_{li} - v_{li})^2}{2\sigma_{li}^2} - \sum_{i \sim j} \beta_{lij} (z_{li} + v_{li} - v_{lj})^2 \right\}, \quad (29)$$

where the y_i 's, σ_i 's, β_{ij} 's and z_i 's are functions of corresponding quantities on level $l-1$ and the index value r . Of course, on level $l=0$ one has $y_i = y_i$, $\sigma_i = \sigma_i$, $\beta_{ij} = \beta$ and $z_i = 0$. An important property of equation (29) is that parameters in $\pi_l(v_l)$ can be computed when going up from level $l-1$, so that no further references to lower levels are needed for doing updates on level l . The benefit of the multi-grid approach is thereby twofold. First, one update of each pixel on a high level l is much faster than on level 0 because of lower dimensionality of the state vector. And note that higher levels are visited most frequently. Second, a change in one pixel on level l corresponds to changes in 4^l pixels on level 0.

How large improvements one can expect from the multi-grid approach discussed above depend on whether the distribution is dominated by the observations, y_i , or the pairwise interaction prior. If the variances, σ_i^2 , are sufficiently small, the long range correlations present in the prior are lost in the posterior and simultaneous changes in a large number of pixels is then probably not a good idea. However, if the observations are quite sparse, i.e. many of the variances, σ_i^2 , can be considered to be infinity, which is a typical situation in modeling of geology, the posterior also exhibit long range correlations and one should expect a substantial gain in the convergence rate from the multi-grid approach; see also Hurn (1995) where a multi-grid algorithm is used on a slightly different model.

In the literature it has been proposed to combine multi-grid ideas with the introduction of auxiliary variables. For example, Kandel et al (1988 and 1989) describe a multi-grid version of the Swendsen-Wang algorithm, where, with the notation introduced above, the index r when going up from level l to $l+1$ is also a function of auxiliary variables, u_l . A variant of this was tried for the Markov random field discussed in paper I but without substantial improvements of the convergence properties. A more general description of the approach is given in Besag and Green (1993), although the multi-grid aspect of the algorithm is not emphasized in their description.

3.4 Metropolis-coupled Markov chains and simulated tempering

Metropolis-coupled Markov chains and simulated tempering are two related methods for speeding up convergence when simulating from strongly multi-modal distributions. Less problem-specific information is needed for these methods than for auxiliary variables and multi-grid. However, the improvement in the convergence properties that can be expected is also lower than for the two methods previously discussed.

Both for Metropolis-coupled Markov chains and simulated tempering, a set of distributions $\pi_k(x)$; $k = 0, \dots, K$ are chosen, where $\pi_0(x)$ is equal to the target distribution, $\pi(x)$. All $\pi_k(x)$ have the same sample space Ω . The $\pi_0(x)$ and $\pi_K(x)$ are often called the ‘‘cold’’ and ‘‘hot’’ distributions, respectively. The distributions chosen should have small differences between any two neighbor-distributions, $\pi_k(x)$ and $\pi_{k+1}(x)$, and the hot distribution, $\pi_K(x)$, should be easy to sample from, either by direct simulation or by MCMC with single-site updates. Write the target distribution as $\pi(x) = c \cdot h(x)$. Then, a typical choice for $\pi_k(x)$ is given by

$$\pi_k(x) \propto h_k(x) = [h(x)]^{\frac{1}{T_k}}, \quad (30)$$

where $1 = T_0 < T_1 < \dots < T_K$.

Metropolis-coupled Markov chains, first discussed in Geyer (1991), simulate K processes in parallel. The state vector considered is $\chi = (x_0, x_1, \dots, x_K)$ with each $x_k \in \Omega$, and the joint distribution given by

$$\pi(\chi) = \pi_0(x_0) \cdot \pi_1(x_1) \cdot \dots \cdot \pi_K(x_K). \quad (31)$$

A Metropolis-Hastings algorithm is used to sample from this distribution and the transitions are of two types; changes in only one x_k and changes where it is proposed to swap two vectors x_k and x_{k+1} . For a change in only one x_k , the acceptance probability is as for simulation from $\pi_k(x_k)$. If x_k and x_{k+1} are proposed to be swapped with some fixed probability, the Metropolis-Hastings acceptance probability becomes

$$\min \left\{ 1, \frac{h_k(x_{k+1}) \cdot h_{k+1}(x_k)}{h_k(x_k) \cdot h_{k+1}(x_{k+1})} \right\}. \quad (32)$$

Note that this does not involve the normalizing constants. If the distributions $\pi_k(x); k = 0, 1, \dots, K$ are well chosen, swaps of x_k and x_{k+1} should often be accepted. This enables visits of x_0 to the different modes in $\pi_0(x_0)$ because the old x_0 , which is in one mode, is swapped with x_1 , which may be in another mode. The x_1 is likewise able to visit different modes by swaps with x_2 , and so on up to swaps between x_{K-1} and x_K . By assumption, the distribution of x_K is easy to sample from.

If x is a high-dimensional vector, simultaneous storing of K such vectors, as necessary in the Metropolis-coupled Markov chain approach, can be a problem. Simulated tempering (Marinari and Parisi, 1992; Geyer and Thompson, 1995) avoids this problem by regarding only one of the distributions $\pi_k(x)$ at a time but is alternating stochastically between them. This is done by defining a joint distribution for k and x by

$$\pi(k, x) \propto \gamma_k \cdot h_k(x) \quad (33)$$

for $k \in \{0, 1, \dots, K\}$ and $x \in \Omega$, where $\gamma_0, \dots, \gamma_K$ are constants that must be pre-computed. A Metropolis-Hastings algorithm is run for this distribution. As with Metropolis-coupled Markov chains, two types of transitions are used; changes in only x and changes in k by increasing or decreasing it by one unit. The acceptance probability for a change in x is as for simulation from $\pi_k(x)$. If a change from k to $l = k \pm 1$ is proposed with probability $\beta(k \rightarrow l)$, the acceptance probability becomes

$$\alpha((k, x) \rightarrow (l, x)) = \min \left\{ 1, \frac{\gamma_l \cdot h_l(x)}{\gamma_k \cdot h_k(x)} \cdot \frac{\beta(l \rightarrow k)}{\beta(k \rightarrow l)} \right\}. \quad (34)$$

If the γ_k 's are estimated to give an approximately uniform marginal distribution for k , such changes should often be accepted. This allows transitions between different modes in $\pi_0(x)$ by going via the hot distribution, $\pi_K(x)$. The Metropolis-Hastings updates ensure that x , after convergence, is distributed according to $\pi_0(x) = \pi(x)$ whenever $k = 0$.

Simulated tempering can easily be specified to give regeneration, i.e. to produce independent segments in the sample path. This can be used to improve estimation; see Ripley (1987) and Mykland et al (1995).

For both Metropolis-coupled Markov chains and simulated tempering, it remains the question of how to choose the number K and the distributions $\pi_k(x); k = 1, \dots, K$; alternatively the numbers T_1, \dots, T_K if equation (30) is used. And for simulated tempering the numbers $\gamma_k; k = 0, 1, \dots, K$ must also be decided. To find the latter so that the marginal for k becomes uniform is not as easy as it may first appear because the normalizing constants for the distributions $\pi_0(x), \dots, \pi_K(x)$ are typically unknown. This topic is discussed in the references mentioned and involve iterative adjustments in order to find suitable choices. Metropolis-coupled Markov chains and simulated tempering therefore are no simple solutions but seems to be a reasonable approach if no better problem-specific alternative can be found.

4 Simulated annealing

In contrast to the MCMC-methods discussed in the previous section, all of which use stationary Markov chains and Metropolis-Hastings updates, simulated annealing is based on the theory of non-stationary Markov chains. It was first introduced as a technique for optimization (Kirkpatrick et al, 1983; Geman and Geman, 1984) but, even if this is still the typical application, it can also be used for stochastic simulation.

When $\pi(x)$ is the target distribution of interest, simulated annealing defines a family of distributions $\pi_T(x)$ on a possibly extended sample space $\Omega_s \supseteq \Omega$, where $T > 0$ is a univariate parameter called temperature. The $\pi_T(x)$ must be chosen to fulfill

$$\lim_{T \rightarrow 0} \pi_T(x) = \begin{cases} \pi(x) & \text{for } x \in \Omega \\ 0 & \text{otherwise.} \end{cases} \quad (35)$$

Simulated annealing runs a Markov chain, where iteration s performs a Metropolis-Hastings update with respect to the distribution $\pi_{T(s)}(x)$, where $T(s)$ is a predefined annealing schedule with

$$\lim_{s \rightarrow \infty} T(s) = 0. \quad (36)$$

Based on slightly different assumptions for $\pi(x)$, $\pi_T(x)$ and $T(s)$ there exist several theorems stating that the distribution of x^s converges to the target distribution as $s \rightarrow \infty$, if $T(s)$ approaches zero slow enough; see Geman and Geman (1984), Gidas (1985), Hajek (1988) and Winkler (1990). However, to follow the annealing schedules permitted by the theorems require, in real applications, too much computer resources to be feasible. In practice, it therefore is necessary to let the temperature decrease faster than allowed by theory; see the discussion in Geman and Geman (1984). Another severe disadvantage of simulated annealing, when used for stochastic simulation, is that each run produce only a single sample from the target distribution. To obtain another sample, the entire procedure must be re-started.

In simulated annealing, it is not necessary that the Metropolis-Hastings updates define an irreducible chain for $\pi(x)$. It is sufficient that the chain is irreducible for $\pi_T(x)$ with $T > 0$. This may simplify the specification of a legal proposal matrix, Q , if, for example, $\pi(x)$ is a posterior distribution based on exact observations. This also makes the technique suitable for optimization problems. To see this, let the target distribution, for some specific function $h(x)$, be given by

$$\pi(x) = \begin{cases} c & \text{if } h(x) \geq h(u) \text{ for all } u \in \Omega_s \\ 0 & \text{otherwise,} \end{cases} \quad (37)$$

i.e. $\pi(x)$ takes some constant value c where $h(x)$ has its global maxima, and is zero otherwise. When the maxima for $h(x)$ are unknown, it is obviously impossible to define a irreducible Markov chain for $\pi(x)$ itself. But simulated annealing with

$$\pi_T(x) \propto [h(x)]^{\frac{1}{T}} \quad (38)$$

can still be used to produce a sample from the distribution.

5 Convergence analysis

Generation of samples from a target distribution, $\pi(x)$, by MCMC-methods relies on the fact that the distribution of X^s converges to $\pi(x)$ as $s \rightarrow \infty$. After some transient phase, the

samples, X^s , therefore are essentially distributed according to $\pi(x)$. In any application of the techniques, it is crucial to be able to decide the length of the transient phase. In this section, we discuss different approaches used to address this and related problems. To avoid technical difficulties, the attention is restricted to the case of a finite sample space and stationary, aperiodic, irreducible and reversible Markov chains. Rigorous discussion of the topic in a more general setting can be found in Tierney (1994).

Denote the eigenvalues of the transition matrix P by $1 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_{|\Omega|}$. The convergence of X^s to the target distribution, $\pi(x)$, is governed by

$$R = \max\{|\lambda_2|, |\lambda_{|\Omega|}|\} \quad (39)$$

and smaller R gives faster convergence. A related question is the rate of convergence of the empirical average in equation (1). It can be shown, see Green and Han (1992) and Besag and Green (1993), that also this is governed by the eigenvalues of P and one has, asymptotically in S ,

$$\mathbb{E} \left\{ |\hat{\mu}_f - \mu_f|^2 \right\} \sim \frac{\text{Var}_\pi(f(x))}{S} \cdot \tau(f), \quad (40)$$

where $\tau(f)$ is the integrated autocorrelation time, which can be written

$$\tau(f) = \sum_{k=2}^{|\Omega|} w_k \frac{1 + \lambda_k}{1 - \lambda_k}. \quad (41)$$

Here, $w_2, \dots, w_{|\Omega|}$ are weights that depend on the function f and the transition matrix P . Thus, for fast convergence of X^s , all eigenvalues, except $\lambda_1 = 1$, should be small in absolute value, whereas large negative eigenvalues help for the convergence of $\hat{\mu}_f$. Thereby, a reasonable strategy is to use two different transition matrices, one in the transient phase and another after convergence.

For target distributions where the use of MCMC-techniques is necessary, the calculation of eigenvalues is computational prohibitive. In practice, the results mentioned above therefore are only indicative for the convergence properties and it is necessary to have methods for estimation of the convergence of X^s and for $\tau(f)$. For the former, much research has recently focused on analytical bounds for the number of iterations necessary to achieve convergence, so that decisions on the length of the transient phase can be done “a priori”, i.e. before the simulation starts. More precisely, the question addressed is to find the minimum s , denoted by s^* , for which

$$\|v_s(\cdot) - \pi(\cdot)\| \leq \varepsilon, \quad (42)$$

where $v_s(\cdot)$ denotes the distribution of X^s , $\|\cdot\|$ is some norm and $\varepsilon > 0$ a tolerance level. To achieve this, one determines a function $K(s)$, for which

$$\|v_s(\cdot) - \pi(\cdot)\| \leq K(s) \quad (43)$$

for all s , and solves the equation $K(s) = \varepsilon$ with respect to s to find an upper bound for s^* . For results of this type, see, for example, references in Smith and Roberts (1993).

A slightly different strategy is used by Frigessi et al (1995). Their goal is not to find explicit bounds for s^* but rather to analyze the “computational complexity” of the simulation algorithms. The interest is in the behavior of s^* as a function of N when N goes to infinity. Although no computable bounds for s^* are supplied, it can be used to compare different algorithms.

Except in a few special cases, no a priori bounds of any practical value is available so far and for the practitioner this implies that the length of the transient phase must be estimated by some kind of output analysis. The attention is then limited to a few scalar functions, $f_1(x), \dots, f_p(x)$. The convergence of each function, $f_j(X^s)$, is bounded by the convergence of X^s and by using several functions one hopes that at least one of them converges almost as slow as X^s . Otherwise, the approach clearly give too optimistic results. A simple method, used in all three papers of this thesis, is to plot the scalars $f_j(x^s)$ as a function of iteration number s , and visually inspect them to see when the distributions seem to have stabilized. Recently, several more elaborate methods have also been proposed; see Cowles and Carlin (1994) and Brooks and Roberts (1995) for reviews and comparisons. However, no methods solely based on Markov chain output can give any guarantee for their results and should be used with care. Especially with multi-modal target distributions, convergence often is indicated prematurely if the chain is slow mixing. For example, no methods that only utilize simulation output, can detect a mode which is not yet visited. It therefore is often recommended to run several realizations of the Markov chain, starting from a distribution which is over-dispersed relative to the target distribution. Or, even better, if the location of the modes are known, to start runs from each of them. This approach is used to check convergence for the Markov random fields in paper I. Another possibility, if some properties of the target distribution is known analytically, is to compare the analytical values with corresponding estimates; see the example in paper III.

The integrated autocorrelation time, $\tau(f)$, gives the Monte Carlo error for $\hat{\mu}_f$ and it can be expressed as in equation (41). However, in practice the computation of eigenvalues is prohibitive and one must resile to estimation of $\tau(f)$. This is a standard problem in time series analysis and several techniques exist. In paper II of this thesis, we use the method proposed in Hastings (1970). It is based on blocking of the time series of length $S = b \cdot k$ into b blocks of length k and the variance of $\hat{\mu}_f$ is estimated from the between-blocks mean square by

$$\hat{\sigma}_{\hat{\mu}_f}^2 = \frac{1}{b \cdot (b-1)} \sum_{s=1}^b \left[\left(\frac{1}{k} \sum_{t=1}^k f(x^{(s-1) \cdot k + t}) \right) - \hat{\mu}_f \right]^2. \quad (44)$$

Other methods can be found in Sokal (1989) and Geyer (1992).

References

- Barone P. and Frigessi, A. (1989). "Improving stochastic relaxation for Gaussian random fields", *Probab. Engng Inform. Sci.*, **4**, 369-389.
- Besag, J. and Green, P.J. (1993). "Spatial statistics and Bayesian computation", *J. Royal Statist. Soc. B*, **55**, 3-23.
- Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). "Bayesian computation and stochastic systems (with discussion)", *Statistical Science*, **10**, 3-66.
- Brooks S. and Roberts, G. (1995). "Diagnosing convergence of Markov chain Monte Carlo algorithms", Technical report 95-12, Statistical Laboratory, University of Cambridge, UK.
- Cowles, M.K. and Carlin, B.P. (1994). "Markov chain Monte Carlo convergence diagnostics: A comparative review", Technical report 94-008, University of Minnesota, USA.
- Cox, D.R. and Miller, H.D. (1970). *The Theory of Stochastic Processes*, Methuen & Co. Ltd.,

London.

- Edwards, R.G. and Sokal, D. (1988). “Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm”, *Phys. Rev. D*, **38**, 2009-2012.
- Frigessi, A., Martinelli, F. and Stander, J. (1995). “Computational complexity of Markov chain Monte Carlo methods for finite Markov random fields”, Technical report, Dipartimento di Matematica, Terza Università di Roma, Italy.
- Geman, D. (1993). “Contribution to discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods”, *J. Royal Statist. Soc. B*, **55**, 73-74.
- Geman, S. and Geman, D. (1984). “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”, *IEEE Trans. PAMI*, **6**, 721-741.
- Geyer, C.J. (1991). “Monte Carlo maximum likelihood for dependent data”, *Computer Science and Statistics: Proc. 23rd Symp. Interface*.
- Geyer, C.J. (1992). “Practical Markov chain Monte Carlo”, *Statistical Science*, **7**, 473-511.
- Geyer, C.J. and Thompson, E.A. (1995). “Annealing Markov chain Monte Carlo with applications to ancestral inference”, *J. Amer. Statist. Assoc.*, **90**, 909-920.
- Gidas, B. (1985). “Nonstationary Markov chains and convergence of the annealing algorithm”, *J. Statistical Physics*, **39**, 73-130.
- Gilks, W.R., Thomas, A. and Spiegelhalter, D.J. (1994). “A language and program for complex Bayesian modelling”, *The Statistician*, **43**, 169-178.
- Goodman, J. and Sokal, A.D. (1989). “Multigrid Monte Carlo method. Conceptual foundations”, *Phys. Rev. D*, **40**, 2035-2071.
- Green, P.J. (1995). “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination”, *Biometrika*, **82**, 711-732.
- Green, P.J. and Han, X. (1992). “Metropolis methods, Gaussian proposals and antithetic variables”, in Barone, P., Frigessi, A. and Piccioni, M. (eds.) *Stochastic models, statistical methods, and algorithms in image analysis*, Lecture Notes in Statistics, **74**, Springer Verlag.
- Grenander, U. (1983). *Tutorial in Pattern Theory*, Report, Div. Applied Mathematics, Brown. Univ.
- Hajek, B. (1988). “Cooling schedules for optimal annealing”, *Mathematics of Operations Research*, **13**, 311-329.
- Hastings, W.K. (1970). “Monte Carlo sampling methods using Markov chains and their applications”, *Biometrika*, **57**, 97-109.
- Higdon, D. (1993). “Contribution to discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods”, *J. Royal Statist. Soc. B*, **55**, 78.
- Hurn, M. (1995). “On the use of auxiliary variables in Markov chain Monte Carlo methods”, Technical report, 95:07, School of Mathematical Sciences, University of Bath, UK.
- Kandel, D., Domany, E., Ron, D., Brandt, A. and Loh, E. (1988). “Simulations without critical slowing down”, *Phys. Rev. Lett.*, **60**, 1591-1594.
- Kandel, D., Domany, E. and Brandt, A. (1989). “Simulations without critical slowing down: Ising and three-state Potts model”, *Phys. Rev. B*, **40**, 330-343.

- Kirkpatrick, S., Gellatt, C.D. and Vecchi, M.P. (1983). "Optimization by simulated annealing", *Science*, **220**, 671-680.
- Marinari, E. and Parisi, G. (1992). "Simulated tempering: a new Monte Carlo scheme", *Europhys. Lett.*, **19**, 451-458.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). "Equations of state calculations by fast computing machines", *J. Chem. Phys.* **21**, 1087-1092.
- Mezei, M. (1981). "On the selection of the particle to be perturbed in the Monte Carlo method", *Jour. Comp. Phys.*, **39**, 128-136.
- Mykland, P., Tierney, L. and Yu, B. (1995). "Regeneration in Markov chain samplers", *J. Amer. Statist. Assoc.*, **90**, 233-246.
- Møller, J. (1993). "Contribution to discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods", *J. Royal Statist. Soc. B*, **55**, 84-85.
- Ripley, B.D. (1987). *Stochastic Simulation*. Wiley, New York.
- Smith, A.F.M and Roberts, G.O. (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods", *J. Royal Statist. Soc. B*, **55**, 3-23.
- Sokal, A.D. (1989). *Monte Carlo methods in statistical mechanics: foundations and new algorithms*, Troisième cycle de la Physique en Suisse Romande lecture notes.
- Swendsen, R.H. and Wang, J.S. (1987). "Nonuniversal critical dynamics in Monte Carlo simulations", *Phys. Rev. Lett.* **58**, 86-88.
- Thomas, A., Spiegelhalter, D.J. and Gilks, W.R. (1992). "BUGS: A program to perform Bayesian inference using Gibbs sampling", in Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. (eds.) *Bayesian statistics 4*, Clarendon Press, Oxford.
- Tierney, L. (1994). "Markov chains for exploring posterior distributions (with discussion)", *Ann. Statist.*, **22**, 1701-1762.
- Winkler, G. (1990). "An ergodic L^2 -theorem for simulated annealing in Bayesian image reconstruction", *J. Appl. Prob.*, **28**, 779-791.
- Wolff, U. (1989). "Collective Monte Carlo updating for spin systems", *Phys. Rev. Lett.*, **62**, 361-364.