PERFECT SIMULATION OF CONDITIONALLY SPECIFIED MODELS

May 26, 1998

Jesper Møller[1], *MaPhySto*[2]

Abstract: We discuss how the ideas of producing perfect simulations based on coupling from the past for finite state space models naturally extend to multivariate distributions with infinite or uncountable state spaces such as auto-gamma, auto-Poisson and auto-negative-binomial models, using Gibbs sampling in combination with sandwiching methods originally introduced for perfect simulation of point processes.

Keywords: Coupling from the past; exact simulation; Gibbs sampling; locally specified exponential family distributions; Markov chain Monte Carlo; Metropolis-Hastings algorithm; spatial statistics.

AMS 1991 MATHEMATICS SUBJECT CLASSIFICATION: 62H11, 62M30, 60K35.

1 Introduction

Since Propp and Wilson's (1996) seminal work on perfect simulation there has been an extensive interest in developing and applying their ideas in different contexts (see the survey in Propp and Wilson (1997)). Briefly, the main idea is to use coupling from the past (CFTP) and repeatedly use the same sampler for generating upper and lower Markov chains started further and further back in time until a pair of upper and lower chains coalesce at time 0, and then return the result as a perfect (or exact) simulation from a given target distribution (in Kendall (1996) and Kendall and Møller (1997) it is argued why the terminology 'perfect' is preferable). To do this Propp and Wilson (1996) assume that the state space is finite and equipped with a partial ordering so that the

[1]*Address for correspondence:* Department of Mathematics, Aalborg University, Fredrik Bajers Vej 7E, 9220 Aalborg, Denmark, Email:jm@math.auc.dk

[2]MaPhySto – Centre for Mathematical Physics and Stochastics, funded by a grant from The Danish National Research Foundation.

sampler is monotone and there exist a unique minimal element, in which the lower processes are started, and a unique maximal element, in which the upper processes are started. Then a chain produced by the sampler, started at any time $n \leq 0$ in an arbitrary initial state, sandwiches between that pair of lower and upper chains which was started at the same time n. Thereby it can be established under weak (ergodicity) conditions for the sampler that coalescence will happens for all sufficiently large n, and by considering a 'virtual simulation from time minus infinity' it follows that the output is a simulation from the target distribution.

These ideas have now been extended in various ways. Kendall (1996) and Häggström, van Lieshout and Møller (1996) outlined how to do perfect simulation for point processes, where the state space is uncountable. Especially, as Propp and Wilson (1996) in their examples required the target distribution to be attractive in a certain sense so that the Gibbs sampler becomes monotone, Kendall's work showed how to handle the opposite repulsive case. This has been further generalised in Kendall and Møller (1997), where the role of the minimal element (the empty point configuration) is emphasized (there exists no maximal element in a point process setting); see also Kendall (1997) and Häggström and Nelander (1997). An even more general approach, but for simulating multivariate continuous distributions has very recently been studied in Murdoch and Green (1997). This and the other mentioned papers will be commented further on in this paper.

The purpose of this paper is to show how these ideas can be further extended to produce perfect simulation of multivariate discrete (Section 2) and continuous (Section 3) target distributions, where the target distribution is naturally specified through its conditional distributions of one component given the others so that Gibbs sampling is the obvious way of producing samples. The examples of such distributions to be discussed will mainly be locally specified exponential family distributions (Besag (1974); Cressie (1993)) with applications in spatial statistics such as auto-binomial, auto-Poisson, auto-negative-binomial, auto-gamma models and certain pairwise-difference interaction models (Sections 2.3 and 3.3). Indeed many other examples of models could be included, e.g. combinations of the local characteristics from different types of models may specify a joint distribution from which we can make perfect simulations. The auto-gamma model has also been used in other papers on Markov chain Monte Carlo methods, in particular in connection to a Bayesian analysis of a dataset on pump reliability (Gelfand and Smith (1990); Murdoch and Green (1997)). In relation to this some empirical results will be reported at the end of Section 3.

The techniques used in Section 2 are much inspired by Kendall and Møller (1997) and some of the terminology and notation will be borrowed from that work. Compared to Propp and Wilson (1996, 1997) and Häggström and Nelander (1997) the extension is mainly that infinite discrete state spaces are covered as well provided the model is repulsive. For definiteness an auto-gamma model is considered in Section 3 and it is demonstrated how we can make perfect simulations with an arbitrarily good accuracy by Gibbs sampling. Also the possibility of using the Metropolis-Hastings algorithm will be investigated in Section 3.

Though the practicality of doing perfect simulation in general is yet not so clear, the present paper and the contributions mentioned above at least demonstrate that it is feasible to tackle many complex distributions by exploring the monotonicity properties of the model and the sampler. One important model, which I couldn't handle, is a conditional auto-regression as it seems impossible to bound the lower and upper processes by dominating chains (the construction of upper and lower processes will be similar to the coupling construction for the auto-binomial model in (ii), Section 2.3 - the problem is to realize (if possible) how to start these processes in the right manner - obviously, one needs both a 'positive' dominating chain and a 'negative' dominating chain).

2 Discrete multivariate distributions

Suppose we want to make simulations from a target distribution $\pi = \mathcal{D}(X)$ with $X = (X_1, \ldots, X_k)$ a k-dimensional discrete random vector, which is specified by its conditional distributions $\mathcal{D}(X_i|X_{-i})$, $i = 1, \ldots, k$, where $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$. Assume also that the support Ω of X is a subset of $\{0, 1, \ldots\}^k$ and that it contains the *minimal element* $0 = (0, \ldots, 0) \in \Omega$. Let $F_i(\cdot|x_{-i})$ denote the cumulative distribution function of $\mathcal{D}(X_i|X_{-i} = x_{-i})$ when $P(X_{-i} = x_{-i}) > 0$. We can then generate a Markov chain $X(t) = (X(t,1), \ldots, X(t,k))$, $t = 0, 1, \ldots$, started in $X(0) = 0$ and using cyclic Gibbs sampling by setting

$$X(t,i) = F_i^-(R(t,i)|X(t,i)_{\leftarrow}) \text{ for } i = 1, \ldots, k, \ t = 1, 2, \ldots$$

where

$$X(t,i)_{\leftarrow} = (X(t,1), \ldots, X(t,i-1), X(t-1,i+1), \ldots, X(t-1,k))$$

is the $k-1$ states of the components just before the ith update at time t and the $R(t,i)$ are *iid* uniform numbers between 0 and 1. Here the inverse F^- of a cumulative distribution function F is defined by $F^-(r) = \min\{s : F(s) \geq r\}$. In fact, if the Markov chain $X(t)$ is irreducible it is also aperiodic and it converges weakly towards X (see, for example, Roberts and Smith (1994)). In the sequel irreducibility of $X(t)$ is assumed.

In this section we show how CFTP can be used for producing a perfect simulation from π within finite time. It is assumed that the conditional distribution functions satisfy a certain monotonicity condition with respect to the natural partial ordering $\leq$ on $\mathbf{R}^d$ $(d = 1, 2, \ldots)$ given by $(x_1, \ldots, x_d) \leq (y_1, \ldots, y_d)$ if $x_i \leq y_i$, $i = 1, \ldots, d$: We assume that for any i, $F_i(\cdot|X_{-i})$ is increasing in X_{-i}, that is,

$$F_i(\cdot|x_{-i}) \geq F_i(\cdot|\acute{x}_{-i}) \text{ if } x_{-i} \geq \acute{x}_{-i}, \ x, \acute{x} \in \Omega. \tag{1}$$

In analogy with Kendall (1996) and Kendall and Møller (1997) we refer to (1) as the *repulsive case* since X_i tends to be smaller as X_{-i} increases (another terminology is used in Häggström and Nelander (1997)). The opposite *attractive case* has earlier been studied (in the binary case where $X_i \in \{0, 1\}$) in Propp and Wilson (1996) and (for point processes) in Kendall (1996), Häggström, van Lieshout and Møller (1996) and Kendall and Møller (1997). This case will be commented on further in (ii), Section 2.3.

2.1 Perfect simulation algorithm (discrete repulsive models)

Let the situation be as described above and suppose that the *iid* uniform variates $R(t,i)$ are defined back in time $t = -1, -2, \ldots$, too. CFTP is then obtained by reusing these random numbers in the following construction of *lower* and *upper* processes $L_n(t) = (L_n(t,1), \ldots, L_n(t,k))$ and $U_n(t) = (U_n(t,1), \ldots, U_n(t,k))$, which are started at times n and generated forwards in time: For each integer $n \in \mathbf{Z}$, set

$$L_n(n) = 0, \quad L_n(t,i) = F_i^-(R(t,i)|U_n(t,i)_{\leftarrow}), \ i = 1, \ldots, k, \ t > n \tag{2}$$

and

$$U_n(n) = D(n), \quad U_n(t,i) = F_i^-(R(t,i)|L_n(t,i)_{\leftarrow}), \ i = 1, \ldots, k, \ t > n \tag{3}$$

with the *dominating chain* $D(n) = (D(n,1), \ldots, D(n,k))$ given by the mutually independent components

$$D(n,i) = F_i^-(R(n,i)|0_{-i}), \ i = 1, \ldots, k, \ n \in \mathbf{Z}.$$

Due to the conditioning in (2)-(3) we need to extend the definition of $F_i(\cdot|x_{-i})$ when $P(X_{-i} = x_{-i}) = 0$. For $i \in \{1, \ldots, k\}$ and $x_{-i} \in \{0, 1, \ldots\}^{k-1}$, define

$$F_i(\cdot|x_{-i}) = \max\{F_i(\cdot|y_{-i}) : y_{-i} \leq x_{-i}, \ P(X_{-i} = y_{-i}) > 0\}$$

which ensures that $F_i(x_i|x_{-i})$ is increasing in x_{-i}.

Now, in the *perfect simulation algorithm* we use a strictly decreasing sequence of non-positive starting times $n = n_j$ $(0 \geq n_1 > n_2 > \ldots)$ and repeat to generate lower and upper processes (L_n, U_n) until coalescence happens at time 0: For $j = 1, 2, \ldots$ set $n = n_j$ and generate $(L_n(t), U_n(t))$, $t = n, \ldots, 0$, until $L_n(0) = U_n(0)$; return then $Y = L_n(0)$ as a perfect simulation from the target distribution π (see Section 2.2).

In principle an arbitrary sequence of starting times $\{n_j\}$ as above may be used, but following the nearly optimal choice of doubling n (see Propp and Wilson (1996)) we propose to set $n_j = -2^j$ in applications. Notice that by the definition (2)-(3), when generating (L_n, U_n) for $n = -2^j$ and $j \geq 2$ we are reusing the random number stream $R(t,i), i = 1, \ldots, k, \ n/2 \leq t \leq 0$, used in the generation of $(L_{n/2}, U_{n/2}), \ldots, (L_0, U_0)$ together with the new random number stream $R(t,i), i = 1, \ldots, k, \ n \leq t < n/2$. Further comments are given in Section 2.3.

2.2 Theoretical results (discrete repulsive models)

The perfect simulation algorithm actually works (at least in theory) according to a general result presented in Kendall and Møller (1997, Theorem 1). In this section we restate and verify this for the present setup, partly for completeness and since we refer to the proof later on and partly because the results presented below extend those in Häggström and Nelander (1997). Actually, Häggström and Nelander assume that Ω is bounded by some maximum x^{max}; they start

each upper process in x^{max}, but compared to the perfect simulation algorithm in Section 2.1, this can only increase the coalescence time by at most one.

For $n \in \mathbf{Z}$ let $X_n(t)$ denote the *target chain* defined by cyclic Gibbs sampling and started in the minimum 0 at time n:

$$X_n(n) = 0, \quad X_n(t,i) = F_i^-(R(t,i)|X_n(t,i)_{\leftarrow}), \ i = 1, \ldots, k, \ t > n.$$

Proposition 1 *For all times $m \leq n \leq u \leq t$ we have that the following sandwiching properties hold:*

$$L_n(t) \leq X_n(t) \leq U_n(t) \leq D(t), \tag{4}$$

$$L_n(t) \leq L_m(t) \leq U_m(t) \leq U_n(t), \tag{5}$$

$$L_m(t) = U_m(t) \quad \textit{if} \quad L_n(u) = U_n(u). \tag{6}$$

Proposition 1 is easily verified by induction. By (6), if coalescence happens at time u in the nth pair of lower and upper processes (when $n \leq u \leq 0$), we have also coalescence from time u to time 0 in the mth pair of lower and upper processes (whenever $m \leq n$), and so $Y = L_m(0) = U_m(0)$. Consequently, in order to verify that coalescence happens within finite time in the perfect simulation algorithm and the output Y follows the target distribution, it suffices to consider the case where $n_j = -j$ and define

$$N = \sup\{n \leq 0 : L_n(0) = U_n(0)\},$$

so $-N$ is the number of pairs of lower and upper processes needed for obtaining coalescence at time 0 when all the non-positive numbers are used as starting times (we set $\sup \emptyset = -\infty$).

Theorem 1 *With probability one, $N > -\infty$, and $Y = L_N(0)$ follows π.*

Proof: Clearly with probability one, $D(t) = 0$ for some $t \leq 0$, and so by (4), $N > -\infty$ almost surely. Hence we can define a 'virtual simulation' $Y = \lim_{n \to -\infty} L_n(0)$ from time minus infinity, and using (4)–(6), we get with probability one that

$$Y = L_N(0) = X_N(0) = U_N(0) = \lim_{n \to -\infty} X_n(0).$$

This together with the fact that $X_n(0)$ converges weakly to X as n tends to minus infinity (since, by stationarity of $R(t)$, $\mathcal{D}(X_n(0)) = \mathcal{D}(X(-n))$) imply that Y follows $\pi = \mathcal{D}(X)$.

Remarks: By stationarity of $R(t)$,

$$\mathcal{D}(L_n(t), U_n(t)) = \mathcal{D}(L_0(t-n), U_0(t-n)) \ \text{ if } \ t \geq n. \tag{7}$$

Hence, as observed in Propp and Wilson (1996), $\mathcal{D}(-N) = \mathcal{D}(M)$, where $M = \inf\{n \geq 0 : L_0(n) = U_0(n)\}$ is the first time of coalescence when the lower and

upper processes are started at time 0. But notice that $L_0(M)$ is in general a biased sample from π (Propp and Wilson (1996)).

The proof of Theorem 1 is only based on the sandwiching properties and the fact that 0 is an ergodic atom for the dominating chain. In the terminology of Foss and Tweedie (1997), $-N$ is a succesful backward coupling time, so the proof that $Y \sim \pi$ follows also from Foss and Tweedie's Theorem 3.1. Clearly, if $P(\cdot,\cdot)$ is the transition probability matrix for a target chain, then by (1), $P(x,0) \geq \prod_1^k F_i(0|0_{-i}) > 0$ for any $x \in \Omega$. Hence the state space Ω is a small set or equivalently $P(\cdot,\cdot)$ is uniformly ergodic (see, e.g., Meyn and Tweedie (1993, Theorem 16.0.2)). Although this gives uniform ergodicity, it is not for the same reason as in Foss and Tweedie (1997, Theorem 4.2), since $-N$ is not 'vertical' in the sense of Foss and Tweedie: in fact, if the state space Ω is infinite, we cannot achieve a backward coupling by considering sample paths generated by the Gibbs sampler and started at any state in Ω and any time (≤ 0). However, due to the stochastic domination given by D, we need only to consider target chains started in 0.

The following lemma can be applied for establishing that the models considered in Section 2.3 below are attractive or repulsive.

Lemma 1 *Let λ denote counting measure on $\{0,1,\ldots\}$ (or Lebesgue measure on $(0,\infty)$). Suppose that $F_i(\cdot|x_{-i})$ (whenever well-defined for x_{-i}) has density*

$$f_i(x_i|x_{-i}) \propto b_i(x_i)\exp(\theta_i(x_{-i})x_i) \tag{8}$$

with respect to λ, where $b_i(\cdot) \geq 0$ and $\theta_i(\cdot)$ are measurable real functions. Then $F_i(\cdot|x_{-i})$ is decreasing (increasing) in x_{-i} if $\theta_i(x_{-i})$ is increasing (decreasing) in x_{-i}.

Proof: Letting $a(\theta) = \int_0^\infty \exp(\theta y)\mathrm{d}\nu(y)$ with $\mathrm{d}\nu(y) = b_i(y)\mathrm{d}\lambda(y)$, then $\Theta = \{\theta : a(\theta) < \infty\}$ is an interval. For $\theta \in \Theta$, let $F_\theta(\cdot)$ denote the cumulative distribution function with density $\exp(\theta y)/a(\theta)$ with respect to ν. For any $r \geq 0$ with $\nu([0,r]) > 0$, it is easily seen by differentation that $g_r(\theta) = \int_r^\infty \exp(\theta y)\mathrm{d}\nu(y) / \int_0^r \exp(\theta y)\mathrm{d}\nu(y)$ is increasing in $\theta \in \Theta$ (if θ is an endpoint of Θ, we consider just left or right derivatives of $g_r(\theta)$). Hence $F_\theta(r) = 1/(1+g_r(\theta))$ is decreasing in θ if $\nu([0,r]) > 0$, and $F_\theta(r) = 0$ otherwise, whereby the lemma is verified.

2.3 Comments and examples (discrete attractive and repulsive models)

(i) Some empirical findings for simple examples of binary repulsive models (i.e. when $\Omega \subseteq \{0,1\}^k$) such as the hard-core model, Ising anti-ferromagnet, and random cluster model (with the parameter for the number of connected components chosen so that the random cluster model is repulsive) are reported in Häggström and Nelander (1997).

(ii) Notice that L_n and U_n are not individually Markov chains in the repulsive

case. In the attractive case where for any i, $F_i(\cdot|X_{-i})$ is decreasing in X_{-i}, we need an upper bound x^{max} on Ω and we redefine the lower and upper chains by setting $L_n(n) = D(n), U_n(n) = x^{max}$ and interchanging the role of $L_n(t,i)_{\leftarrow}$ and $U_n(t,i)_{\leftarrow}$ when conditioning in (2)-(3). Then both L_n and U_n become Markovian, and the results in Proposition 1 and Theorem 1 are still valid except that (4) has to be modified so that D becomes a *lower dominating* chain (correspondingly the role of the minimum 0 and the maximum x^{max} are now interchanged). Propp and Wilson (1996,1997) report on simulation studies of the attractive Ising model (defined on very huge lattices) and its accompanying random cluster model.

Another model is the *auto-binomial model*. Here $\mathcal{D}(X_i|X_{-i} = x_{-i})$ is a binomial distribution with parameters n_i and $p_i(x_{-i})$, where $\log(p_i(x_{-i})/(1 - p_i(x_{-i}))) = \beta_i + \sum_{j:j\neq i}\beta_{ij}x_j$ and β_i, $\beta_{ij} = \beta_{ji}$ are real parameters. The model is attractive if all $\beta_{ij} \geq 0$ and repulsive if all $\beta_{ij} \leq 0$. For illustrative purposes, if the β_{ij} have different signs, then we can start the upper processes in $x^{max} = (n_1, \ldots, n_k)$ and the lower ones in 0, and set

$$\begin{aligned} L_n(t,i) = F_i^-(R(t,i)|&L_n(t,j)\mathbf{1}_{[\beta_{ij}\geq 0]} + U_n(t,j)\mathbf{1}_{[\beta_{ij}\leq 0]}, 1 \leq j < i; \\ &L_n(t-1,j)\mathbf{1}_{[\beta_{ij}\geq 0]} + U_n(t-1,j)\mathbf{1}_{[\beta_{ij}\leq 0]}, i < j \leq k), \\ U_n(t,i) = F_i^-(R(t,i)|&L_n(t,j)\mathbf{1}_{[\beta_{ij}\leq 0]} + U_n(t,j)\mathbf{1}_{[\beta_{ij}\geq 0]}, 1 \leq j < i; \\ &L_n(t-1,j)\mathbf{1}_{[\beta_{ij}\leq 0]} + U_n(t-1,j)\mathbf{1}_{[\beta_{ij}\geq 0]}, i < j \leq k), \end{aligned}$$

when $t > n$, where $\mathbf{1}_{[\cdot]}$ denotes indicator function.

Yet another interesting model is the 'pairwise-difference prior' model with (in the present setup) finite state space $\Omega = \{0, \ldots, m\}^k$ and

$$f_i(x_i|x_{-i}) \propto \exp(-\sum_{j:j\neq i}\beta_{ij}(x_i - x_j)^2)$$

where $\beta_{ij} = \beta_{ji}$ are real parameters (see, for example, Besag (1989) and Green (1996)). In applications of image analysis one takes all $\beta_{ij} \geq 0$ so that the model becomes attractive. When doing conventional MCMC forwards simulations from this model, unless m is small enough, it is often more convenient to use another Metropolis-Hastings algorithm than the Gibbs sampler.

Note that if we multiply the pairwise-difference prior density (or any other density with 'full conditionals' of the form (8)) with a 'likelihood term' $\prod_i g_i(x_i)$ in order to get a posterior density, the monotonicity properties of the prior and the posterior are the same.

(iii) Examples where $\Omega = \{0, 1, \ldots\}^k$ is infinite are provided by the *auto-Poisson model*, where $\mathcal{D}(X_i|X_{-i} = x_{-i})$ is a Poisson distribution with mean $\lambda_i(x_{-i}) = \exp(\beta_i + \sum_{j:j\neq i}\beta_{ij}x_j)$ and parameters $\beta_i \in \mathbf{R}$ and $\beta_{ij} = \beta_{ji} \leq 0$, and an *auto-negative-binomial model* given by that

$$f_i(x_i|x_{-i}) = \begin{pmatrix} \alpha_i + x_i - 1 \\ \alpha_i - 1 \end{pmatrix} \lambda_i(x_{-i})^{x_i}(1 - \lambda_i(x_{-i}))^{\alpha_i}$$

with the parameter $\alpha_i > 0$ and where now $\beta_i < 0$. Both models are repulsive; the joint distribution is not well-defined if we allow some β_{ij} to be positive (Besag (1974)), so it is not possible to include the attractive case.

Of course higher order interaction terms may be included in the auto-binomial, auto-Poisson and auto-negative-binomial models. For example, we can extend the exponent in the definition of $p_i(x_{-i})$ and $\lambda_i(x_{-i})$ with the term $\sum_{j<k:i\notin\{j,k\}} \beta_{ijk} x_j x_k$, where β_{ijk} does not depend on the ordering of i,j,k; for the auto-Poisson and auto-negative-binomial models, $\beta_{ijk} \leq 0$.

(iv) Suppose that $X_1,\ldots,X_l$ are conditionally independent given $X_{l+1},\ldots,X_k$ and that $X_{l+1},\ldots,X_k$ are conditionally independent given $X_1,\ldots,X_l$, where $1 \leq l < k$. Then in the repulsive case (1), an alternative perfect simulation algorithm is provided by replacing (2)-(3) by

$$L_n(n) = (0,\ldots,0,D(n,l+1),\ldots,D(n,k)),\ L_n(t,i) = F_i^-(R(t,i)|L_n(t,i)_{\leftarrow}),$$
$$U_n(n) = (D(n,1),\ldots,D(n,l),0,\ldots,0),\qquad U_n(t,i) = F_i^-(R(t,i)|U_n(t,i)_{\leftarrow}),$$

for $t > n$, whereby L_n and U_n become individually Markov chains with state space Ω. This perfect simulation algorithm shares the properties in Proposition 1 and Theorem 1 if the partial order $\leq$ used in (4)-(5) is replaced by another partial order $\preceq$ defined by

$$x \preceq y \Longleftrightarrow x_i \leq y_i \text{ for } i = 1,\ldots,l,\ x_j \geq y_j \text{ for } j = l+1,\ldots,k.$$

A comparative study of this perfect simulation algorithm, the algorithm in Section 2.1, and certain perfect simulation algorithms based on extensions of Fill (1997) will appear in Møller and Schladitz (1998).

3 Perfect simulation of auto-gamma models

In the sequel we use notation as in Section 2, but it is now assumed that each conditional distribution of the target $\pi = \mathcal{D}(X)$ is given by $\mathcal{D}(X_i|X_{-i} = x_{-i}) = \Gamma(\alpha_i, \gamma_i(x_{-i}))$, the gamma distribution with shape parameter $\alpha_i > 0$ and inverse scale parameter

$$\gamma_i(x_{-i}) = \beta_i + \sum_{j:j\neq i} \beta_{ij} x_j,$$

where the parameters $\beta_i > 0$ and $\beta_{ij} = \beta_{ji} \geq 0$. Thus $X = (X_1,\ldots,X_k)$ is stochastically dominated by k independent gamma variates with parameters $(\alpha_1,\beta_1),\ldots,(\alpha_k,\beta_k)$, respectively, and π has density

$$f(x) \propto \exp(-\sum_i \beta_i x_i - \sum_{i<j} \beta_{ij} x_i x_j) \prod_1^k x_i^{\alpha_i - 1} \tag{9}$$

with respect to Lebesgue measure on $\Omega = (0,\infty)^k$ (the density is only well-defined for nonnegative β_{ij}). Note that the dependence is expressed through the scale parameter γ_i (compare with Cressie (1993, p. 440) who instead considers a dependence through the shape parameter).

The auto-gamma model is also repulsive as $F_i(\cdot|x_{-i})$ increases as x_{-i} increases, so like in Section 2 we can use cyclic Gibbs sampling based on 'inversion' when simulating from π. However, it is more obvious to utilize the fact

that the conditional distributions are scale models: So let $G(t,i) \sim \Gamma(\alpha_i, 1),\ i = 1,\ldots,k,\ t \in \mathbf{Z}$, denote mutually independent gamma variates (for the simulation of the gamma distribution, see, e.g., Ripley, 1987). Then for the Markov chain $X(t) = (X(t,1),\ldots,X(t,k)),\ t = 0,1,\ldots$, we now set

$$X(0) = 0,\quad X(t,i) = G(t,i)/\gamma_i(X(t,i)_{\leftarrow}) \ \text{ for } \ i = 1,\ldots,k,\ t = 1,2,\ldots$$

whereby it is easily verified (see, for example, Roberts and Smith (1994)) that $X(t)$ converges weakly to X.

Apart from one major point, CFTP and the perfect simulation algorithm in Section 3.1 below follow the same lines as in Section 2.1: As we are using Gibbs sampling on continuous distributions we never get coalescence at time 0 of the lower and upper processes (Proposition 2), but we shall establish in Theorem 2 that with an arbitrarily good accuracy we can deliver a perfect simulation from π within finite time. These aspects are further discussed in Section 3.3.

3.1 Perfect simulation algorithm (auto-gamma)

Define the lower and upper processes by

$$L_n(n) = 0,\quad L_n(t,i) = G(t,i)/\gamma_i(U_n(t,i)_{\leftarrow}),\ i = 1,\ldots,k,\ t > n \tag{10}$$

and

$$U_n(n) = D(n),\quad U_n(t,i) = G(t,i)/\gamma_i(L_n(t,i)_{\leftarrow}),\ i = 1,\ldots,k,\ t > n \tag{11}$$

where the dominating chain $D(t)$ is defined by the mutually independent gamma variates

$$D(t,i) = G(t,i)/\beta_i,\ i = 1,\ldots,k,\ t \in \mathbf{Z}.$$

Let $n_1 > n_2 > \ldots$ be a given strictly decreasing sequence of non-positive integers. Then, in the *perfect simulation algorithm*, for $j = 1,2,\ldots$ set $n = n_j$ and generate $(L_n(t), U_n(t)),\ t = n,\ldots,0$, until $U_n(0,i) - L_n(0,i) \le \epsilon,\ i = 1,\ldots,k$; return $Z = (L_n(0) + U_n(0))/2$ as a *perfect simulation from π with accuracy provably within ϵ of π*, where $\epsilon > 0$ is a 'user-specified parameter'.

3.2 Theoretical results (auto-gamma)

In this section we verify that for any given $\epsilon > 0$ the stopping time

$$N(\epsilon) = N(\epsilon, \{n_j\}) = \sup\{n_j : U_{n_j}(0,i) - L_{n_j}(0,i) \le \epsilon,\ i = 1,\ldots,k\}$$

is almost surely finite, and that there is a coupling with a random vector W so that $\mathcal{D}(W) = \pi$ and the output $Z = (L_{N(\epsilon)}(0) + U_{N(\epsilon)}(0))/2$ satisfies that

$$|Z(i) - W(i)| \le \epsilon,\ i = 1,\ldots,k. \tag{12}$$

This justifies that the 'accuracy is ϵ'.

Let $X_n(t)$ denote the *target chain* defined by cyclic Gibbs sampling and started in the minimum 0 at time n:

$$X_n(n) = 0,\quad X_n(t,i) = G(t,i)/\gamma_i(X_n(t,i)_{\leftarrow}),\ i = 1,\ldots,k,\ t > n.$$

Proposition 2 *For all times $m \leq n \leq u \leq t$ we have that the sandwiching properties (4)-(6) remain valid in the present situation. However, with probability one, $N(0) = -\infty$, and if $n \leq s < t$, then*

$$[U_n(s,i) - L_n(s,i) \leq \epsilon] \not\Rightarrow [U_n(t,i) - L_n(t,i) \leq \epsilon]. \tag{13}$$

Proof: The sandwiching properties are straightforwardly verified by induction. Clearly, $P(U_n(t,i) > L_n(t,i)$ for all i and $n \leq t) = 1$, so $N(0) = -\infty$ almost surely. Since $U_n(t,i) - L_n(t,i)$ equals

$$D(t,i)\frac{\sum_{j=1}^{i-1} \beta_{ij}(U_n(t,j) - L_n(t,j)) + \sum_{j=i+1}^{k} \beta_{ij}(U_n(t-1,j) - L_n(t-1,j))}{\gamma_i(L_n(t,i)_{\leftarrow})\gamma_i(U_n(t,i)_{\leftarrow})}$$

where the gamma variate $D(t,i)$ is independent of the fraction, it follows that (13) holds.

Remarks: As $N(0) = -\infty$ we have to take $\epsilon > 0$ in order to get a finite stopping time (as discussed in (iii), Section 3.3, there exist perfect Metropolis-Hastings algorithms with $N(0) > -\infty$). Further, (6) is really not useful anymore, and because of (13) we cannot stop the backward sampling the first time we get intermediate 'ϵ-coupling'. This is in contrast to the perfect simulation algorithm in the discrete case, but for the auto-gamma model it turns out that only the sandwiching properties (4)-(5) are needed in order to establish (12).

Note that the argument in Propp and Wilson (1996) for preferring the sequence $\{-2^j\}$ still applies in the present situation. Clearly, if we consider a subsequence $\{m_j\}$ of $\{n_j\}$, then for the corresponding stopping times we have that $N(\epsilon, \{m_j\}) \leq N(\epsilon, \{n_j\})$, so $-N(\epsilon, \{0, -1, \ldots\})$ is the smallest number of pairs of lower and upper processes needed for obtaining coalescence at time 0. Observe also that as (7) is satisfied, $\mathcal{D}(-N(\epsilon, \{0, -1, \ldots\})) = \mathcal{D}(M(\epsilon))$ with

$$M(\epsilon) = \inf\{n \geq 0 : L_0(n,i) = U_0(n,i) \leq \epsilon,\ i = 1, \ldots, k\}.$$

Theorem 2 *Let $\epsilon > 0$. With probability one, we have that $N(\epsilon) > -\infty$ and the limits*

$$L_{-\infty}(t) = \lim_{n \to -\infty} L_n(t), \quad U_{-\infty}(t) = \lim_{n \to -\infty} U_n(t) \tag{14}$$

exist and agree. Moreover,

$$L_{-\infty}(t) \ \textit{follows the auto-Gamma model} \ \pi \tag{15}$$

and (12) is satisfied with $W = L_{-\infty}(0)$.

Proof: By Proposition 2 and (5), the limits in (14) exist almost surely. Further, (5) also gives that $U_n(t) - L_n(t)$ decreases to $U_{-\infty}(t) - L_{-\infty}(t)$ as $n(\leq t)$ tends to minus infinity. To show that both $L_{-\infty}(t) = U_{-\infty}(t)$ and $N(\epsilon) > -\infty$ hold almost surely, it suffices to show that $E(U_0(t) - L_0(t))$ tends to 0 as $t \to \infty$, cf. (7). Hence, letting

$$l(tk+i) = L_0(t,i),\ u(tk+i) = U_0(t,i),\ d(tk+i) = D(t,i),\ t \geq 0$$

we just have to verify that

$$\lim_{t\to\infty} E(u(tk+i) - l(tk+i)) = 0. \tag{16}$$

Consider now any $t \geq 1$ and $i \in \{1,\ldots,k\}$. Set $\alpha = \min\beta_i > 0$, $\beta = \max\beta_{ij} > 0$, and

$$A(s) = \max\{D(s,1),\ldots,D(s,k)\}$$

$$B(s) = \frac{\beta(k-1)\max\{A(s-1),A(s)\}}{\alpha+\beta(k-1)\max\{A(s-1),A(s)\}}.$$

Further, for $i,j \in \{1,\ldots,k\}$ with $i \neq j$, let $\beta(i,j) = \beta_{ij}$ and set $\beta(i,i-j) = \beta(i,i-j+k)$ if $i < j$. We prove by induction that

$$(u(tk+i) - l(tk+i))/u(tk+i) \leq B(1)\cdots B(t) \tag{17}$$

for $t \geq 1$ and $i = 1,\ldots,k$: If $t = i = 1$, then by (10)-(11),

$$\frac{u(k+1) - l(k+1)}{u(k+1)} = \frac{\sum_{j=1}^{k-1}\beta(1,1-j)u(k+1-j)}{\beta_1 + \sum_{j=1}^{k-1}\beta(1,1-j)u(k+1-j)} \leq B(1)$$

where we have used that $l(k+1-j) = 0$, $u(\cdot) \leq d(\cdot)$, and the real function $a \to a/(a+b)$ is increasing for $b > 0$. So for any $t \geq 1$ and $i = 1,\ldots,k$ with $tk+i > k+1$ we get by similar arguments and by combining the induction hypothesis with the fact that $0 < B(t) < 1$,

$$\frac{u(tk+i) - l(tk+i)}{u(tk+i)} = \sum_{j=1}^{k-1}\frac{\beta(i,i-j)u(tk+i-j)}{\gamma_i(u(tk+i)_{\leftarrow})}\frac{u(tk+i-j) - l(tk+i-j)}{u(tk+i-j)}$$

$$\leq B(1)\cdots B(t-1)\frac{\sum_{j=1}^{k-1}\beta(i,i-j)u(tk+i-j)}{\beta_i + \sum_{j=1}^{k-1}\beta(i,i-j)u(tk+i-j)} \leq B(1)\cdots B(t)$$

where we set $B(1)\cdots B(t-1) = 1$ if $t = 1$. Thereby (17) is verified.

Since $d(tk+i)$ is independent of $B(1),\ldots,B(t-1)$ and $B(t) < 1$ we get from (17) that

$$E(u(tk+i) - l(tk+i)) \leq (\alpha_i/\beta_i)E(B(1)\cdots B(t-1)).$$

As the $A(s)$ are *iid* and $B(\cdot) < 1$, we have that

$$E(B(1)B(2)\cdots B(t-1)) \leq E(B(1)B(3)\cdots B(t-1)) = (EB(1))^{t/2}$$

for t even and where $EB(1) < 1$. So using again that $B(\cdot) < 1$, we conclude that $E(B(1)\cdots B(t-1)) \to 0$ as $t \to \infty$. This implies that (16) holds, and so we have shown that with probability one, $L_{-\infty}(t) = U_{-\infty}(t)$ and $N(\epsilon) > -\infty$.

To verify (15) is similar to the last part in the proof of Theorem 1 by using (4)-(5) and observing that the Markov chain $L_{-\infty}(t) = U_{-\infty}(t) = X_{-\infty}(t)$ is in equilibrium for all $t \in \mathbf{Z}$ as $\mathcal{D}(X_n(t)) = \mathcal{D}(X_0(t-n)),\ t \geq n$, where $X_0(t-n)$ converges weakly towards X as $n \to \infty$.

Finally, as $L_n(0) \leq W(0) \leq U_n(0)$ if $n \leq 0$, (12) follows from the triangle inequality.

3.3 Comments and examples (auto-gamma)

(i) In computing the state space is of course finite, and in order to avoid technical subtleties we may simply refer to the results for the discrete case as outlined in Section 2. If we take $\epsilon = 0$, the perfect simulation algorithm (Section 3.1) terminates when the lower and upper processes become equal because of rounding error on the machine; this may possibly cause substantial numerical errors so that we are not quite ensured that the accuracy is given by the precision "$\epsilon_{machine}$" of the numbers used in the computations. So it may be safer to keep the size of ϵ several orders of magnitude larger than $\epsilon_{machine}$.

(ii) The main example discussed in Murdoch and Green (1997) is how to do perfect simulations from a posterior distribution for a dataset on pump reliability (Gelfand and Smith (1990)), which is actually an auto-gamma model (9) with $k = 11$ and pairwise interactions between just one variate and each of the remaining variates. For this model I have produced two C-programs PerfectGammaBackwards.c (the algorithm in Section 3.1 for generating a perfect simulation) and PerfectGammaForwards.c (for forward runs and with output as shown below). These programs are available by anonymous ftp (ftp://ftp.math.auc.dk/pub/jm/); the accuracy ϵ is specified in the beginning of the programs. The average of $M(\epsilon)$ based on 10000 simulations for each value of ϵ was 9.3047 ($\epsilon = 10^{-3}$), 11.3170 ($\epsilon = 10^{-4}$), 13.3262 ($\epsilon = 10^{-5}$), 19.3508 ($\epsilon = 10^{-8}$), 31.3775 ($\epsilon = 10^{-14}$), 34.8263 ($\epsilon = 0$); the standard error of the average was 0.0050 ($\epsilon = 10^{-3}$), 0.0052 ($\epsilon = 10^{-4}$), 0.0054 ($\epsilon = 10^{-5}$), 0.0061 ($\epsilon = 10^{-8}$), 0.0072 ($\epsilon = 10^{-14}$), 0.0120 ($\epsilon = 0$).

The CFTP algorithms in Murdoch and Green (1997) use a larger number of paths than just the two paths for the lower and upper processes in the perfect simulation algorithm considered in Section 3.1, so it would not make much sense to use $M(\epsilon)$ in a comparative analysis. Duncan Murdoch has suggested to me that a hybrid algorithm could possibly do better than either theirs or mine, e.g. by first using my algorithm so that the lower and upper processes become close enough and then using their multigamma coupler.

(iii) It is at least theoretically feasible to make simulations with accuracy 0 by the following *perfect Metropolis-Hastings algorithm* by noticing that the density (9) of the auto-gamma model (as well as the auto-binomial, auto-Poisson, auto-negative-binomial and many other models) is of a particular form:

$$f(x) \propto h(x) \prod_1^k q_i(x_i)$$

where (in the case of the auto-gamma model) q_i is the density of $\Gamma(\alpha_i, \beta_i)$ and

$$h(x) = \exp(-\sum_{i<j} \beta_{ij} x_i x_j).$$

If we are using cyclic updates and the usual notation as introduced in Section 2, then it is now a Metropolis-Hastings algorithm which generates a Markov chain

$X(t)$, $t \geq 0$: At time t, when the ith coordinate is to be updated and the current state is given by $X(t-1,i)$ and $X(t,i)_{\leftarrow}$, first a proposal $D(t,i)$ is generated from q_i together with a uniform number $R(t,i)$ (between 0 and 1), and secondly we set $X(t,i) = D(t,i)$ if $R(t,i) \leq \min\{1, a_i(D(t,i), X(t-1,i), X(t,i)_{\leftarrow})\}$ and we retain $X(t,i) = X(t-1,i)$ otherwise (where all gamma variates $D(t,i)$ and random numbers $R(t,i)$ are assumed to be mutually independent). Here

$$a_i(d_i, x_i, x_{-i}) = \frac{f(x_1, \ldots, x_{i-1}, d_i, x_{i+1}, \ldots, x_k) q_i(x_i)}{f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k) q_i(d_i)} = \exp((x_i - d_i) \sum_{j: j \neq i} \beta_{ij} x_j)$$

is the Metropolis-Hastings ratio (Hastings, (1970)). This is larger than 1 if $d_i \leq x_i$ (i.e. moving down is always accepted), whilst it decreases from 1 to 0 as a function of x_{-i} when $d_i > x_i$. It is easily seen that $X(t)$ converges weakly to X.

CFTP should now be obvious: Set $L_n(n) = 0$, $U_n(n) = D(n)$, and for $t > n$ set

$$L_n(t,i) = D(t,i) \text{ if } R(t,i) \leq a_i(D(t,i), L_n(t-1,i), U_n(t,i)_{\leftarrow})$$

and $L_n(t,i) = L_n(t-1,i)$ else, and

$$U_n(t,i) = D(t,i) \text{ if } R(t,i) \leq a_i(D(t,i), U_n(t-1,i), L_n(t,i)_{\leftarrow})$$

and $U_n(t,i) = U_n(t-1,i)$ else. Clearly, the usual sandwiching properties remain valid with $D(t)$ as the dominating chain, and for a perfect Metropolis-Hastings algorithm similar to the algorithm in Section 2.1 it is not hard to verify that the coalescence time is almost surely finite and the output follows the target distribution. (Sketch of proof: Let $\epsilon > 0$. Conditioning on the event that $D(t-1,1) \leq \epsilon, \ldots, D(t,k) \leq \epsilon$, then for $t > n$ the conditional probability of the event $L_n(t) = D(t)$ is greater than $\delta(\epsilon)^k$, where $\delta(\epsilon) = \inf\{a_i(d_i, x_i, x_{-i}) : 0 \leq x_i \leq d_i \leq \epsilon, 0 \leq x_j \leq \epsilon, i \neq j\} > 0$. From this we get that the coalescence time is almost surely finite; that the output follows the target distribution is verified in the same way as in the proofs of Theorems 1 and 2.)

However, in practise (even for the simple example considered in (ii) above) this perfect Metropolis-Hastings algorithm may become extremely slow as the lower processes get stuck in 0 for a very long time unless the parameters $\beta_{ij}/(\beta_i \beta_j)$, $i < j$, are sufficiently small or equivalently if the interactions are sufficiently weak. One may object that this is clearly due to the way proposals are generated. But we need a coupling construction, where it is feasible to generate a dominating chain, which is in equilibrium. Actually we are using the smallest possible dominating chain as this has both to dominate the proposals and the Markov chain $X_{-\infty}(t)$ started 'in equilibrium at time minus infinity' and generated by the Metropolis-Hastings algorithm (see the proofs of Theorems 1 and 2).

(iv) The coupling construction in Section 3.1 and the proofs of Proposition 2 and Theorem 2 are essentially only based on a few properties: (a) The 'local characteristics' $\mathcal{D}(X_i | X_{-i} = x_{-i})$ are scale families, where the inverse scaling factor $\gamma_i(x_{-i})$ is an increasing function of x_{-i} and the support is the positive half-line. Thereby lower and upper processes are naturally constructed (Section

3.1). (b) There is a (natural) dominating and stationary Markov chain $D(t)$ (in the proof of Theorem 2 we referred to that the $D(t,i)$ are *iid*, but this may of course be weakened). (c) $ED(t)$ is finite. Apart from (a)-(c) the fact that $\mathcal{D}(X_i|X_{-i} = x_{-i})$ is a gamma distribution was really never used.

Certain models can be transformed into a auto-gamma model such as auto-generalized-gamma and hierarchical models obtained by conditional independent normally distributed variates with the precisions (inverse variances) given by an auto-gamma model. It would be interesting to see if the ideas in Sections 3.1-3.2 become useful for other multivariate continuous distributions.

Acknowledgement: This research was supported by MaPhySto – Centre for Mathematical Physics and Stochastics, funded by a grant from the Danish National Research Foundation – and by the Danish Informatics Network in the Agricultural Sciences, funded by a grant from the Danish Research Councils. Many thanks to my colleagues in Aalborg, Martin B. Hansen and Uffe Kjærulff for programming assistance ((ii) in Section 3.3) and Bjarne Højgaard for stimulating discussions. Also the encouragement and helpful comments of Julian Besag, Olle Häggström, Duncan Murdoch, and James Propp are gratefully acknowledged.

References

[1] Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistical Society* **B 36**, 192–225.

[2] Besag, J.E. (1989). Towards Bayesian image analysis. *Applied Statistics* **16**, 395–407.

[3] Cressie, N. (1993). *Statistics for Spatial Data*, revised edn. New York: Wiley.

[4] Fill, J.A. (1997). An interruptible algorithm algorithm for perfect sampling via Markov chains. Preprint, Department of Mathematical Sciences, The John Hopkins University. To appear in *Annals of Applied Probability.*

[5] Foss, S.G. and Tweedie, R.L. (1997). Perfect simulation and backward coupling. *Stochastic Models.* (To appear)

[6] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

[7] Green, P.J. (1996). MCMC in image analysis. *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter), 381–399. London: Chapman and Hall.

[8] Häggström, O., van Lieshout, M.N.M. and Møller, J. (1996). Characterisation results and Markov chain Monte Carlo algorithms including exact

simulation for some spatial point processes. Research Report R-96-2040, Department of Mathematics, Aalborg University. To appear in *Bernoulli.*

[9] Häggström, O. and Nelander, K. (1997). Exact sampling from anti-monotone systems. Research Report 1997–03, Department of Mathematics, Chalmers University of Technology and Göteborg University.

[10] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

[11] Kendall, W.S. (1996). Perfect simulation for the area-interaction point process. Research Report 292, Department of Statistics, University of Warwick. To appear in *Probability Perspective* (eds. C.C Heyde and L. Accardi). Singapore: World Scientific Press.

[12] Kendall, W.S. (1997). On some weighted Boolean models. To appear in *Advances in Theory and Applications of Random sets* (ed. D. Jeulin), 105–120. Singapore: World Scientific Publishing Company.

[13] Kendall, W.S. and Møller, J. (1997). Perfect Metropolis-Hastings simulation of locally stable point processes. (In preparation)

[14] Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability.* London: Springer-Verlag.

[15] Murdoch, D.J. and Green, P.J. (1997). Exact sampling from a continuous state space. Research Report S–97–01, Department of Mathematics, University of Bristol. To appear in *Scandinavian Journal of Statistics.*

[16] Møller, J. and Schladitz, K. (1998). Extensions of Fill's algorithm for perfect simulation. (In preparation)

[17] Propp, J.G. and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.

[18] Propp, J.G. and Wilson, D.B. (1997). How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *Journal of Algorithms.* (To appear)

[19] Ripley, B.D. (1987). *Stochastic Simulation* New York: Wiley.

[20] Roberts, G.O. and Smith, A.F.M. (1994). Simple conditions for the convergence of the Gibbs sampler and the Metropolis-Hastings algorithms. *Stochastic Processes and their Applications.* **49**, 207–216.