HAL

archives-ouvertes.fr

Shadow simulated annealing algorithm : a new tool for global optimisation and statistical inference

Radu Stoica, Madalina Deaconu, Anne Philippe, Lluis Hurtado-Gil

Shadow Simulated Annealing algorithm: a new tool for global optimisation and statistical inference

R. S. Stoica[1], M. Deaconu[2], A. Philippe[3], L. Hurtado[4]

[1] Université de Lorraine, CNRS, IECL, F-54000, Nancy, France
[2] Université de Lorraine, CNRS, Inria, IECL, F-54000, Nancy, France
[3] Université de Nantes, Laboratoire de Mathématiques Jean Leray, Nantes, France
[4] Departamento de Matemática Aplicada y Estadística, Universidad CEU San Pablo, 28003 Madrid, Spain

Abstract: This paper develops a new global optimisation method that applies to a family of criteria that are not entirely known. This family includes the criteria obtained from the class of posteriors that have normalising constants that are analytically not tractable. The procedure applies to posterior probability densities that are continuously differentiable with respect to their parameters. The proposed approach avoids the re-sampling needed for the classical Monte Carlo maximum likelihood inference, while providing the missing convergence properties of the ABC based methods. Results on simulated data and real data are presented. The real data application fits an inhomogeneous area interaction point process to cosmological data. The obtained results validate two important aspects of the galaxies distribution in our Universe : proximity of the galaxies from the cosmic filament network together with territorial clustering at given range of interactions. Finally, conclusions and perspectives are depicted.

1 Introduction

A large class of the mathematical questions issued from data sciences can be formulated as an optimisation problem. The complexity of the data structures requires more and more elaborate models with an important number of parameters controlling different aspects outlined by the data observation

process. Within this context, a typical a question is, what is the most probable model able to reproduce the behaviour exhibited by the analysed data. Clearly, the possible answer to this question assumes the existence of a class of models together with priors associated to its parameters.

A natural way to answer this question is the computation of the global maximum of the induced posterior distribution. The classical simulated annealing framework proposes the solution to this problem, provided sampling from the posterior distribution is possible.

Sampling posterior distributions is still a challenging mathematical problem. This is due to the fact that sometimes, the proposed simulation algorithms may be required to make computations of analytical intractable quantities. An example of such quantities is the evaluation of normalisation constants of the probability densities describing the considered models.

If only parameters estimation is considered, the common solution adopted is to provide maximum likelihood computations based on Monte Carlo simulations. This framework allows the user to benefit of the whole theoretical power of the likelihood inference, if one afford the price to pay in terms of computational costs. The computational cost may be excessively high whenever the initial condition is to far away from the desired solution. Furthermore this phenomenon introduces numerical instability of the proposed solution. The only reliable strategy within this context is to re-sample the model, as often as possible, hence increasing the computational cost.

Since less than a decade, a new methodological framework for statistical inference, the Approximate Bayesian Computation (ABC) allows to extend the Monte Carlo likelihood based inference, by providing solutions for sampling approximately from the posterior distribution. The authors in [17] proposed a new algorithm, ABC Shadow, that overcomes the main drawback of the previously mentioned ABC methods: the ABC Shadow allows the output distribution of the algorithm, to be as closed as desired to the aimed posterior distribution.

The work presented in this paper leads directly to a new method of parameter estimation based on a simulated annealing algorithm. To better outline its interest, let us consider the following example, inspired by applications in spatial data analysis.

Let **y** be an object pattern that is observed in a compact window $W \subset \mathbb{R}^d$. The observed pattern is supposed to be the realisation of a spatial process. Such a process is given by the probability density

$$p(\mathbf{y}|\theta) = \frac{\exp[-U(\mathbf{y}|\theta)]}{\zeta(\theta)} \tag{1}$$

with $U(\mathbf{y}|\theta)$ the energy function and $\zeta(\theta)$ the normalising constant. The model given by (1) may be considered as a Gibbs process, and it may represent a random graph, a Markov random field or a marked point process. Let $p(\theta|\mathbf{y})$ be the conditional distribution of the model parameters or the posterior law

$$p(\theta|\mathbf{y}) = \frac{\exp[-U(\mathbf{y}|\theta)]p(\theta)}{Z(\mathbf{y})\zeta(\theta)}, \tag{2}$$

where $p(\theta)$ is the prior density for the model parameters and $Z(\mathbf{y})$ the normalising constant. The posterior law is defined on the parameter space Θ. For simplicity, the parameter space is considered to be a compact region in $\mathbb{R}^r$ with r the size of the parameter vector. Let ν be the corresponding Lebesgue measure. The parameter space is endowed with its Borel algebra $\mathcal{T}$.

In the following, it is assumed that the probability density $p(\mathbf{y}|\theta)$ is strictly positive and continuously differentiable with respect to θ. This hypothesis is strong but keeps realistic, since it is often required by practical applications.

This paper constructs and develops a simulated annealing method to compute :

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} p(\theta|\mathbf{y}).$$

The difficulty of the problem is due to the fact that the normalising constant $\zeta(\theta)$ is not available in analytic closed form. Hence, special strategies are required to sample from the posterior distribution (2) in order to implement an optimisation procedure.

The plan of the paper is as follows. First, an Ideal Markov chain (IC) is constructed. This chain has as equilibrium distribution the posterior distribution of interest. So, in theory, this chain can be used to sample from the distribution of interest. Next, an Ideal Simulated Annealing (ISA) process built using a non-homogeneous IC chain is constructed. The convergence properties of the process are also given. Despite the good theoretical properties, the ISA process cannot be used to build algorithms for practical use,

since it requires the computation of normalising constants of the form $\zeta(\theta)$. The fourth section presents a solution to this problem. First an approximate sampling mechanism called the Shadow chain (SC) is presented. The SC is able to follow closely within some fixed limits the IC. Based on the SC chain, a Shadow Simulated Annealing (SSA) process is built. This process is controlled by two parameters, evolving slowly to zero. This double control allows to derive convergence properties of the process towards the global optimum we are interested in. These theoretical results allow the construction of a SSA algorithm. The algorithm is applied to simulated and real data, during the fifth section. The real data application fits an inhomogeneous point process with interactions to a cosmological data set, in order to obtain essential characteristics of the galaxies distribution in our Universe. At the end of the paper, conclusions and perspectives are formulated.

2   Ideal Chain for posterior sampling

In theory, Markov Chain Monte Carlo algorithms may be used for sampling $p(\theta|\mathbf{y})$. For instance, let us consider the general Metropolis Hasting algorithm. Assuming the system is in the state θ, this algorithm first chooses a new value ψ according to a proposal density $q(\theta \to \psi)$. The value ψ is then accepted with probability $\alpha_i(\theta \to \psi)$ given by

$$\alpha_i(\theta \to \psi) = \min_{\psi} \left\{ 1, \frac{p(\psi|\mathbf{y})}{p(\theta|\mathbf{y})} \frac{q(\psi \to \theta)}{q(\theta \to \psi)} \right\}. \tag{3}$$

The transition kernel of the Markov chain simulated by this algorithm is given by, for every $A \in \mathcal{T}$

$$\begin{aligned} P_i(\theta, A) &= \int_A \alpha_i(\theta \to \psi) q(\theta \to \psi) \mathbf{1}_{\{\psi \in A\}} d\psi \\ &+ \mathbf{1}_{\{\theta \in A\}} \left[1 - \int_A \alpha_i(\theta \to \psi) q(\theta \to \psi) d\psi \right]. \end{aligned} \tag{4}$$

Let us recall that the transition kernel of a Markov chain may act on both functions and measures, as it follows

$$P_i f(x) = \int_\Theta P_i(x, dy) f(y), \quad \mu P_i(A) = \int_\Theta \mu(d\theta) P_i(\theta, A). \tag{5}$$

The conditions that the proposal density $q(\theta \to \psi)$ has to meet, so that the simulated Markov chain has a unique equilibrium distribution

$$\pi(A) = \int_A p(\theta|\mathbf{y}) d\nu(\theta),$$

are rather mild [24]. Furthermore, if q and π are bounded and bounded away from zero on the compact Θ, then the simulated chain is uniformly ergodic ([24] Prop. 2, [13] Thm. 2.2). Hence, there exist a positive constant M and a positive constant $\rho < 1$ such that

$$\sup_{\theta \in \Theta} \| P_i^n(\theta, \cdot) - \pi(\cdot) \| \leq M\rho^n, \quad n \in \mathbb{N}.$$

For a fixed $\delta > 0$, a parameter value $\nu \in \Theta$ and a realisation $\mathbf{x}$ of the model $p(\cdot|\nu)$ given by (1), let us consider the proposal density

$$q(\theta \to \psi) = q_\delta(\theta \to \psi|\mathbf{x}) = \frac{f(\mathbf{x}|\psi)/\zeta(\psi)}{I(\theta, \delta, \mathbf{x})} \mathbf{1}_{b(\theta, \delta/2)}\{\psi\} \tag{6}$$

with $f(\mathbf{x}|\psi) = \exp[-U(\mathbf{x}|\psi)]$. Here $\mathbf{1}_{b(\theta, \delta/2)}\{\cdot\}$ is the indicator function over $b(\theta, \delta/2)$, which is the ball of center θ and radius $\delta/2$. Finally, $I(\theta, \delta, \mathbf{x})$ is the quantity given by the integral

$$I(\theta, \delta, \mathbf{x}) = \int_{b(\theta, \delta/2)} \frac{f(\mathbf{x}|\phi)}{\zeta(\phi)} \, d\phi.$$

This choice for $q(\theta \to \psi)$ guarantees the convergence of the chain towards π and avoids the evaluation of the normalising constant ratio $\zeta(\theta)/\zeta(\psi)$ in (3). We call the chain induced by these proposals the *ideal chain*. Nevertheless, the proposal (6) requires the computation of integrals such as $I(\theta, \delta, \mathbf{x})$, and this is as difficult as the computation of the normalising constant ratio. Later in the paper it will be shown how this construction allows a natural approximation of the ideal chain: the shadow chain.

3 Ideal Simulated Annealing process

3.1 Principle

The construction and the properties of a *Simulated Annealing* (SA) process in a general state space were investigated by [6, 7].

The SA process is built with the following ingredients: a function to optimise h, a Markov transition kernel P with equilibrium distribution $\pi \propto \exp(-h)$ and a cooling schedule for the temperature parameter T.

The function h is assumed to be continuous differentiable in θ and its global maximum is θ_{opt}. In the present situation $h : \Theta \to \mathbb{R}$ is obtained by taking

the logarithm of (2):

$$h(\theta) = U(\mathbf{y}|\theta) + \log \zeta(\theta) - \log p(\theta) + \log Z(\mathbf{y}). \tag{7}$$

It represents the loglikelihood function to which the prior term $\log p(\theta)$ is added. Clearly, if $p(\theta)$ is the uniform distribution over Θ, then maximising h leads to the maximum likelihood estimation. The transition kernel P is given by (4). The cooling schedule for the temperature is a logarithmic one, and it results from the proofs of the SA convergence.

The SA process simulates iteratively a sequence of distributions π_n

$$\pi_n(A) = \int_A p_n(\theta) d\nu(\theta) = \int_A \frac{\exp(-h(\theta)/T_n)}{c_n} d\nu(\theta) \tag{8}$$

with $p_n(\theta) = \frac{\exp(-h(\theta)/T_n)}{c_n}$ and $c_n = \int_\Theta \exp(-h(\theta)/T_n) d\nu(\theta)$, while T_n goes slowly to zero.

Each distribution π_n is simulated using a transition kernel P_n having it as equilibrium distribution. The transition kernel P_n is obtained, by modifying (4) in order to sample from π_n. The kernels sequence $(P_n)_{n\geq 0}$ induces an inhomogeneous Markov chain.

At low temperatures, the process converges weakly towards the global optimum of h, that is

$$\pi_n \xrightarrow[n\to\infty]{} \delta_{\theta_{opt}} \tag{9}$$

with $\delta_{\theta_{opt}}$ the Dirac measure in θ_{opt}, while considering the Hausdorff topology [6, 7].

3.2 Definition, properties and convergence

In order to define the SA process and to present its main convergence result, the Dobrushin coefficient, its properties and some extra-notations are introduced.

Let $\Theta = (\Theta, \mathcal{T}, \upsilon)$ be a state space with υ a probability measure on $\mathcal{T}$ and let $\Upsilon = \Upsilon(\Theta)$ be the set of all probability measures on the space $(\Theta, \mathcal{T})$. Throughout the entire, the norm $\| \cdot \|$ is the total variation norm, that is

for any $\mu \in \Upsilon$:

$$
\begin{aligned}
\| \mu \| &= \sup_{A \in \mathcal{T}} |\mu(A)| \\
&= \sup_{|g|<1} |\mu(g)| = \sup_{|g|<1} | \int_{\Theta} g(\theta)\mu(d\theta)|.
\end{aligned}
\tag{10}
$$

Let P be a transition kernel on Θ. The Dobrushin coefficient is defined as :

$$
c(P) = \sup_{\theta,\psi \in \Theta} \| P(\theta,\cdot) - P(\psi,\cdot) \| = \sup_{\theta,\psi \in \Theta} \sup_{A \in \mathcal{T}} |P(\theta, A) - P(\psi, A)|.
\tag{11}
$$

Lemma 1. *The Dobrushin coefficient, defined by (11), verifies the following properties:*

(i) $0 \leq c(P) \leq 1$,

(ii) $\| \mu P - \lambda P \| \leq c(P) \| \mu - \lambda \|$,

(iii) $c(P_1 P_2) \leq c(P_1)c(P_2)$.

A proof of the previous result is given below. The reader may also refer to [8, 25, 10, 29] for more details.

Proof: (i). Clearly $c(P) \geq 0$ by definition (11). The $c(P) \leq 1$ is also immediate by using the property that P is a transition kernel. For all $A \in \mathcal{T}$ and all $\theta, \psi \in \Theta$, $P(\theta, A)$, $P(\psi, A) \in [0,1]$ so $|P(\theta, A) - P(\psi, A)| \leq 1$. Thus, by taking the supremum in A and θ, ψ it results that $c(P) \leq 1$.

(ii). Let us recall first a classical result. If μ and ν are from Υ and g is a smooth function we have, by using (5):

$$
|\mu(g) - \nu(g)| \leq \frac{1}{2} \sup_{\theta,\psi \in \Theta} |g(\theta) - g(\psi)| \| \mu - \nu \|.
\tag{12}
$$

Let us prove now the inequality. By the definition of the total variation norm (10), the formula (5) and the definition (11) we get:

$$
\begin{aligned}
\| \mu P - \nu P \| &= \sup_{g, |g| \leq 1} |(\mu P)g - (\nu P)g| \\
&\leq \sup_{g, |g| \leq 1} \left\{ \frac{1}{2} \sup_{\theta,\psi \in \Theta} |Pg(\theta) - Pg(\psi)| \| \mu - \nu \| \right\} \\
&\leq c(P) \| \mu - \nu \|,
\end{aligned}
\tag{13}
$$

by applying (12).

(iii). It can be directly obtained, by using (11) and (ii) and lastly (12)

$$
\begin{aligned}
c(P_1 P_2) &= \sup_{\theta,\psi \in \Theta} \|P_1 P_2(\theta,.) - P_1 P_2(\psi,.)\| \\
&= \sup_{\theta,\psi \in \Theta} \|P_1(\theta,.)P_2 - P_1(\psi,.)P_2\| \\
&\leq c(P_1)c(P_2).
\end{aligned}
$$

$\square$

An immediate consequence of the previous result is that

$$
c(P_1 P_2 \cdots P_n) \leq \prod_{i=1}^{n} c(P_i). \tag{14}
$$

Assume that $(P_j)_{j=1}^{\infty}$ is a sequence of transition kernels on $(\Theta, \mathcal{T})$ and that μ_0 is a given initial distribution. The sequence $(P_j)_{j=1}^{\infty}$ and the distribution μ_0 define a discrete time *non-homogeneous Markov process* with the state space Θ. Here, we are interested in the successive distributions $\mu_j = \mu_{j-1}P_j$ of the random variables θ_j. Throughout the paper for $m+1 \leq k$, the following notation is used

$$
P^{(m,k)} = P_{m+1} P_{m+2} \cdots P_k,
$$

so that one has $\mu_k = \mu_m P^{(m,k)}$.

In the following it is assumed that the function to optimise $h : \Theta \to \mathbb{R}$ is $\mathcal{T}$−measurable, positive and bounded. For the sake of simplicity it is assumed that the global minimum of h is scaled to zero. The global maximum is obtained by considering the function $-h$. It is denoted by

$$
M_h = \{\theta \in \Theta | h(\theta) = 0\} \tag{15}
$$

the *minimum set* of h and by $\triangle h = \sup_{\theta \in \Theta} h(\theta)$ the maximum variation of h.

Definition 1. *Let h be the function to be optimised in the state space $(\Theta, \mathcal{T}, \upsilon)$, $(P_j)_{j \in \mathbb{N}}$ a sequence of transition kernels, $T_1 \geq T_2 \geq \ldots \geq T_{i-1} \geq T_i \to_{i \to \infty} 0$. A simulated annealing process is the non-homogeneous Markov process $(P_i)_{i \in \mathbb{N}}$ on $(\Theta, \mathcal{T}, \upsilon)$ defined by*

$$
\begin{aligned}
P_j(\theta, A) &= \int_A \alpha_j(\theta \to \psi) q(\theta \to \psi) \mathbf{1}_{\{\psi \in A\}} d\psi \\
&\quad + \mathbf{1}_{\{\theta \in A\}} \left[1 - \int_A \alpha_j(\theta \to \psi) q(\theta \to \psi) d\psi \right],
\end{aligned} \tag{16}
$$

with

$$\alpha_j(\theta \to \psi) = \min\left\{1, \left[\frac{\exp(-h(\psi))}{\exp(-h(\theta))}\right]^{1/T_j} \frac{q(\psi \to \theta)}{q(\theta \to \psi)}\right\}.$$

Again there is a lot of freedom in choosing the proposal distribution. Similarly to (6), we consider the distribution

$$q(\theta \to \psi) = q_\delta(\theta \to \psi|\mathbf{x}) = \frac{[f(\mathbf{x}|\psi)/\zeta(\psi)]^{1/T_j}}{I_j(\theta, \delta, \mathbf{x})}\mathbf{1}_{b(\theta,\delta/2)}\{\psi\} \qquad (17)$$

for a fixed $\delta > 0$, a parameter value $\nu \in \Theta$ and a realisation $\mathbf{x}$ of the model $p(\cdot|\nu)$. The quantity $I_j(\theta, \delta, \mathbf{x})$ is given by the integral

$$I_j(\theta, \delta, \mathbf{x}) = \int_{b(\theta,\delta/2)} [f(\mathbf{x}|\phi)/c(\phi)]^{1/T_j} \, d\phi.$$

The SA process build with proposals (17) is called *Ideal Simulated Annealing* (ISA).

The following result states the convergence of the SA process for the optimisation of our problem, that is whenever $\upsilon(M_h) = 0$ (see [6, 7]), where M_h is the set defined in (15).

Theorem 1. *Let $(P_j)_{j\in\mathbb{N}}$ be a SA process. Suppose that there exist sequences $(n_j)_{j\in\mathbb{N}}$ and $(r_j)_{j\in\mathbb{N}}$ and a number $0 < d < 1$ such that $(n_j)_{j\in\mathbb{N}}$ is increasing and $\lim_{j\to\infty}(j - r_j) = \infty = \lim_{j\to\infty} r_j$. Assume also that the following conditions hold:*

(i) For each $j \geq 1$

$$c(P^{(n_j, n_{j+1})}) \leq 1 - (1 - d)\exp\left[-\sum_{l=n_j+1}^{n_{j+1}} \frac{\Delta h}{T_l}\right]; \qquad (18)$$

(ii) $\displaystyle \lim_{k\to\infty} \sum_{j=r_k}^{k} \exp\left(-\sum_{l=n_j+1}^{n_{j+1}} \frac{\Delta h}{T_l}\right) = \infty;$

(iii) $\displaystyle \lim_{k\to\infty} \frac{\mathcal{L}_h(1/T_{n_{r_k}})}{\mathcal{L}_h(1/T_{n_{k+2}})} = 1,$ *where* $\mathcal{L}_h(1/T) = \int_\Theta \exp(-h(\theta)/T)\upsilon(d\theta).$

Then

$$\lim_{j\to\infty} \|\mu_j - \pi_j\| = 0. \qquad (19)$$

The proof of this theorem is an adaptation of the proof given in [6].

Proof: Suppose that $\varepsilon > 0$ is given. The condition (ii) implies that one can find integers k and k' such that for all $k > k'$ we have:

$$\sum_{j=r_k}^{k} \exp\left[-\Delta h \sum_{l=n_j+1}^{n_{j+1}} \frac{1}{T_l}\right] > \frac{-\log(\varepsilon/4)}{1-d}. \tag{20}$$

By using now the standard inequality $1 - x < \exp(-x)$ for $x > 0$ and the relation (20) it gets:

$$\prod_{j=r_k}^{k}\left[1 - (1-d)\exp\left(-\Delta h \sum_{l=n_j+1}^{n_{j+1}} \frac{1}{T_l}\right)\right]$$
$$< \exp\left[-(1-d)\sum_{j=r_k}^{k} \exp\left(-\Delta h \sum_{l=n_j+1}^{n_{j+1}} \frac{1}{T_l}\right)\right] \tag{21}$$
$$< \frac{\varepsilon}{4}.$$

The condition (i) is used to obtain

$$c\big(P^{(n_j,n_{j+1})}\big) \leq 1 - (1-d)\exp\left(-\Delta h \sum_{l=n_j+1}^{n_{j+1}} \frac{1}{T_l}\right).$$

By doing now the product of these relations for $j \in \{r_k, \ldots, k\}$ it results

$$\prod_{j=r_k}^{k} c\big(P^{(n_j,n_{j+1})}\big) \leq \prod_{j=r_k}^{k}\left[1 - (1-d)\exp\left(-\Delta h \sum_{l=n_j+1}^{n_{j+1}} \frac{1}{T_l}\right)\right]. \tag{22}$$

And now, by embedding (21) within (22) we obtain, for $k \geq k'$:

$$c(P_{n_{r_k}+1} P_{n_{r_k}+1} \ldots P_{n_{k+1}}) < \frac{\varepsilon}{4}. \tag{23}$$

The previous integers k and k' can also be chosen such that

$$\log\left(\mathcal{L}_h\left(\frac{1}{T_{n_{r_k}}}\right)\Big/\mathcal{L}_h\left(\frac{1}{T_{n_{k+2}}}\right)\right) < \frac{\varepsilon}{2}, \tag{24}$$

by using hypothesis (iii).

Following ([6], Thm. 5.1), for $0 < k \leq i \leq n$ we have

$$\sum_{i=1}^{n} \|\pi_i - \pi_{i-1}\| \leq 2 \log \left(\frac{\mathcal{L}_h(1/T_k)}{\mathcal{L}_h(1/T_n)} \right). \tag{25}$$

Thus by writing (25) with the adequate indexes and putting together the result in (24), it comes out that:

$$\sum_{l=n_{r_k}}^{n_{k+1}-1} \|\pi_l - \pi_{l-1}\| < \varepsilon. \tag{26}$$

Consider now two positive integers M and N such that $M \leq N - 1$ and recall the notation, for $m \leq k - 1$

$$P^{(m,k)} = P_{m+1} P_{m+2} \dots P_k.$$

Let us prove by induction that

$$\pi_M P^{(M,N)} - \pi_N = \sum_{l=M}^{N-1} (\pi_l - \pi_{l+1}) P^{(l,N)}. \tag{27}$$

For the verification step of this relation consider M fixed and take $N = M + 1$. This gives the following equality:

$$\pi_M P^{(M,M+1)} - \pi_{M+1} = (\pi_M - \pi_{M+1}) P^{(M,M+1)},$$

which is verified since π_{M+1} is a stationary distribution and thus $\pi_{M+1} = \pi_{M+1} P_{M+1}$.

Assuming now, that the relation (27) is verified for M and N fixed with $M \leq N - 1$, let us prove that (27) is also true for $N + 1$. This leads to:

$$\pi_M P^{(M,N+1)} - \pi_{N+1} = \sum_{l=M}^{N} (\pi_l - \pi_{l+1}) P^{(l,N+1)} \tag{28}$$

which is equivalent to

$$(\pi_M P^{(M,N)} - \pi_{N+1}) P_{N+1} = \sum_{l=M}^{N-1} (\pi_l - \pi_{l+1}) P^{(l,N)} P_{N+1} + \pi_N P_{N+1} - \pi_{N+1} P_{N+1},$$

that becomes

$$(\pi_M P^{(M,N)} - \pi_N)P_{N+1} = \sum_{l=M}^{N-1}(\pi_l - \pi_{l+1})P^{(l,N)}P_{N+1}$$

and this is true since (28) is satisfied. We admit now that (27) is valid.
Let us now complete the proof of the Theorem 1.
Let $m > n_{k'+1}$ and take k'' such that $n_{k''+1} < m \le n_{k''+2}$. As $(\pi_m)_{m\ge1}$ is a stationary distribution, we have :

$$
\begin{aligned}
\|\mu_m - \pi_m\| &= \|\mu_{n_{r_{k''}}}P^{(n_{r_{k''}},m)} - \pi_m\| \\
&\le \|(\mu_{n_{r_{k''}}} - \pi_{n_{r_{k''}}})P^{(n_{r_{k''}},m)}\| + \|\pi_{n_{r_{k''}}}P^{(n_{r_{k''}},m)} - \pi_m\|.
\end{aligned}
\tag{29}
$$

In this last inequality the first term on the right hand side can be controlled by using (13) and (23)

$$
\begin{aligned}
\|(\mu_{n_{r_{k''}}} - \pi_{n_{r_{k''}}})P^{(n_{r_{k''}},m)}\| &\le 2c(P^{(n_{r_{k''}},m)}) \\
&\le 2c(P^{(n_{r_{k''}},n_{k''+1})}) \\
&< \frac{\varepsilon}{2}.
\end{aligned}
\tag{30}
$$

The second term can be expressed in the following form by using (27):

$$
\begin{aligned}
\|\pi_{n_{r_{k''}}}P^{(n_{r_{k''}},m)} - \pi_m\| &\le \sum_{l=n_{r_{k''}}}^{m-1}\|(\pi_{l+1} - \pi_l)P^{(l,m)}\| \\
&\le \sum_{l=n_{r_{k''}}}^{n_{k''+2}-1}\|\pi_l - \pi_{l+1}\| \\
&< \frac{\varepsilon}{2},
\end{aligned}
\tag{31}
$$

by making use of the inequality (26). Combining now (30) and (31) in (29) gives finally for $m > n_{k''+1}$

$$\|\mu_m - \pi_m\| < \varepsilon.$$

As $\varepsilon > 0$ is arbitrary the result of the theorem follows. This concludes the proof. $\qquad\square$

As a consequence of Theorem 1, it is obtained:

Corollary 1. *Suppose that the conditions of Theorem 1 are satisfied and suppose also that we have the weak convergence $\lim\limits_{j\to+\infty} \pi_j = \pi$ where $\pi \in \Upsilon(\Theta)$. Then we have also the weak convergence*

$$\lim_{j\to+\infty} \mu_j = \pi.$$

The following important result is also stated.

Corollary 2. *Suppose that the hypothesis of the Theorem 1 are satisfied for the simulated annealing process $(P_i)_{i\geq 1}$. Suppose also that h achieve its maximum in $\theta_{opt} \in \Theta$. Then we have the weak convergence*

$$\lim_{j\to+\infty} \mu_j = \delta_{\theta_{opt}}.$$

Remark 1. *The first two conditions in Theorem 1 ensure the weak ergodicity of the SA process. The condition (i) is fulfilled for transition kernels built with rather general proposal densities $q(\theta, \psi)\upsilon(d\psi)$ with $q : \Theta \times \Theta \to [0, \infty[$ measurable ([6], Lemma 4.1, Thm. 4.2). The second condition (ii) if fulfilled if a logarithmic cooling schedule is used for the temperature*

$$T_j = \frac{K}{\log(j + 2)} \quad with \quad K > 0,$$

which is similar to the cooling schedules obtained for maximising Markov random fields or marked point process probability densities [4, 16]. The third condition (iii) is a bound for the sum of distances between two equilibrium distributions that correspond to two different temperatures ([6], Thm. 3.2). It is the key condition, that whenever $\upsilon(M_h) = 0$, it allows to reduce the weak convergence of the process to the weak convergence of the equilibrium distributions ([7], pp.44).

The previous results show that the ISA process given by (17) converges weakly towards the global optimum of the h function (7). Still, these results cannot be transformed directly into an optimisation algorithm to be used in practice, due to the need of computation of the normalising constants $\zeta(\theta)$. The next section shows how to overcome this drawback by building an alternative chain able to follow the path of the theoretical ISA process as close as desired.

4 Shadow Simulated Annealing process

The authors in [17] proposed an algorithm, *ABC Shadow* able to sample from posterior distributions (2). This section presents this sampling method, builds a SA process based on it and derives convergence results. This new process is called *Shadow Simulated Annealing* (SSA) process.

4.1 ABC Shadow sampling algorithm

The idea of the *ABC Shadow* algorithm [17] is to construct a *shadow Markov chain* able to follow the *ideal Markov chain* given by (16). The steps of the algorithm are given below.

Algorithm 1. ABC Shadow : *Fix δ and m. Assume the observed pattern is $\mathbf{y}$ and the current state is θ_0.*

1. *Generate $\mathbf{x}$ according to $p(\mathbf{x}|\theta_0)$.*

2. *For $k = 1$ to m do*

   - *Generate a new candidate ψ following the density $U_\delta(\theta_{k-1} \rightarrow \psi)$ defined by*
   $$U_\delta(\theta \rightarrow \psi) = \frac{1}{V_\delta} \mathbf{1}_{b(\theta,\delta/2)}\{\psi\},$$
   with V_δ the volume of the ball $b(\theta, \delta/2)$.

   - *The new state $\theta_k = \psi$ is accepted with probability $\alpha_s(\theta_{k-1} \rightarrow \psi)$ given by*

   $$
   \begin{aligned}
   \alpha_s(\theta_{k-1} \rightarrow \theta_k) &= \\
   &= \min\left\{1, \frac{p(\theta_k|\mathbf{y})}{p(\theta_{k-1}|\mathbf{y})} \times \frac{f(\mathbf{x}|\theta_{k-1})\zeta(\theta_k)\mathbf{1}_{b(\theta_k,\delta/2)}\{\theta_{k-1}\}}{f(\mathbf{x}|\theta_k)\zeta(\theta_{k-1})\mathbf{1}_{b(\theta_{k-1},\delta/2)}\{\theta_k\}}\right\} \\
   &= \min\left\{1, \frac{f(\mathbf{y}|\theta_k)p(\theta_k)}{f(\mathbf{y}|\theta_{k-1})p(\theta_{k-1})} \times \frac{f(\mathbf{x}|\theta_{k-1})}{f(\mathbf{x}|\theta_k)}\right\} \qquad (32)
   \end{aligned}
   $$

   otherwise $\theta_k = \theta_{k-1}$.

3. *Return θ_m.*

4. *If another sample is needed, go to step 1 with $\theta_0 = \theta_n$.*

The transition kernel of the Markov chain simulated by the previous algorithm is given by, for every $A \in \mathcal{T}$

$$
\begin{aligned}
P_s(\theta, A) \quad = \quad & \int_A \alpha_s(\theta \to \psi) U_\delta(\theta \to \psi) \mathbf{1}_{\{\psi \in A\}} d\psi \\
+ \quad & \mathbf{1}_{\{\theta \in A\}} \left[1 - \int_A \alpha_s(\theta \to \psi) U_\delta(\theta \to \psi) d\psi \right].
\end{aligned}
$$

The authors in [17] show also that since

$$
|P_i(\theta, A) - P_s(\theta, A)| \leq K_1 \delta
$$

with K_1 a constant depending on $\mathbf{x}, p$ and Θ, there exists $\delta_0 = \delta_0(\varepsilon, m) > 0$ such that for every $\delta \leq \delta_0$, we have

$$
|P_i^m(\theta, A) - P_s^m(\theta, A)| < \varepsilon
$$

uniformly in $\theta \in \Theta$ and $A \in \mathcal{T}$. If $p(x|\theta) \in \mathcal{C}^1(\Theta)$, then a description of $\delta_0(\varepsilon, n)$ can be provided.

The previous results state that for any ε and a given $\mathbf{x}$ we may find a δ such that the shadow and ideal chain may evolve as close as desired during a pre-fixed value of n steps. Under these assumptions, if $m \to \infty$ the Algorithm 1 does not follow closely the ideal chain started in θ_0, anymore. The ergodicity properties of the ideal chain and the triangle inequalities allow to give a bound for the distance the distance after n steps, between the shadow transition kernel and the equilibrium regime

$$
\|P_s^{(m)}(\theta, \cdot) - \pi(\cdot)\| \leq M(\mathbf{x}, \delta) \rho^m + \varepsilon.
$$

with $\rho \in (0, 1)$.

Iterating the algorithm more steps it is possible by re-freshing the auxiliary variable. This mechanism allows to re-start the algorithm for m steps more, and by this, to obtain new samples of the approximate distribution of the posterior.

4.2 An SA process based on the shadow chain

4.2.1 Process construction and properties

Theorem 2. *Let $(P_{i,j})_{j \geq 1}$ be an ISA process associated with the ideal transition kernel P_i (4) that samples from $\pi \propto \exp(-h)$ with h given by (7)*

using the proposals (6). The cooling schedule k_T is chosen with respect to the Theorem 1. According to the Algorithm 1, to each $(P_{i,j})$ a shadow chain $P_{s,j})$ (33) is attached. Then, there exists a sequence $\{\delta_j = \delta_j(\varepsilon, m), j \geq 1\}$ such that

$$|P_{i,j}^m(\theta, A) - P_{s,j}^m(\theta, A)| < \varepsilon$$

for all $j \geq 1$, uniformly in $\theta \in \Theta$ and $A \in \mathcal{T}$. If $p(x|\theta) \in \mathcal{C}^1(\Theta)$, then a description of $\delta_j(\varepsilon, n)$ can be provided.

Proof: The proof is obtained by considering for each step j of the algorithm, the Proposition 1 in [17]. □

The previous process given by the sequence the shadow chains P_s is named the Shadow Stochastic Annealing (SSA) process. Controlling the δ parameter in the same time with the temperature allows to obtain the following convergence properties.

Theorem 3. *Let us consider the assumptions of Theorem 2 fulfilled and let $\varepsilon > 0$ be a fixed value. Then, for the SSA process $(P_{s,j})_{j\geq 1}$, the following results hold :*

(i) *For each j, T_j, δ_j we have $\|P_{s,j}^m - \pi_j\| \leq \varepsilon$ and also for j big enough $\|P_{s,j}^m - \delta_{\theta_{opt}}\| \leq \varepsilon$.*

(ii) *Consider now for $j \geq 1$, $\varepsilon_j = \frac{\varepsilon}{j}$. For each j we can construct T_j, δ_j such that $\|P_{s,j}^m - \pi_j\| \leq \varepsilon_j$ and $\lim_{j \to +\infty} \|P_{s,j}^m - \delta_{\theta_{opt}}\| = 0$*

Proof: Theorem 2 proves that there exists a sequence $\{\delta_j = \delta_j(\varepsilon, m), j \geq 1\}$ such that

$$|P_{i,j}^m(\theta, A) - P_{s,j}^m(\theta, A)| < \varepsilon$$

for all $j \geq 1$, uniformly in $\theta \in \Theta$ and $A \in \mathcal{T}$.

Combining this with the result of the Corollary 2 allows to conclude.

We end the demonstration by the following observation : if $p(x|\theta) \in \mathcal{C}^1(\Theta)$, then a description of $\delta_j(\varepsilon, m)$ can be provided. □

Remark 2. *The first part of the preceding result can be interpreted as follows. Let $\{\theta_j, j \geq 1\}$ be a realisation of the SSA process. Then there exists*

a sequence $\{\delta_j = \delta_j(\varepsilon, m), j \geq 1\}$ *corresponding to each* θ_j *such that for* $j \geq j_{\max}(\varepsilon, m)$

$$\{\theta_j, j \geq j_{\max}\} \subset b(\theta_{opt}, \varepsilon)$$

where θ_{opt} *is the global optimum of the function* h.

Corollary 3. *We have*

$$\delta_j = \frac{K(x_j, T_j, n_j)}{j} \tag{33}$$

and $K(x_j, T_j, n_j)$

Remark 3. *The second part of the preceding result states that to a decreasing sequence of balls around the problem solution, it is possible to associate a sequence of* δ *parameters in order to get as close as desired to the global optimum of* h.

4.2.2 A new algorithm for global optimisation

The previous results justify the construction of the following algorithm.

Algorithm 2. Shadow Simulated Annealing (SSA) algorithm : *fix* $\delta = \delta_0$, $T = T_0$, n *and* $k_\delta, k_T : \mathbb{R}^+ \to \mathbb{R}^+$ *two positive functions. Assume the observed pattern is* $\mathbf{y}$ *and the current state is* θ_0.

1. *Generate* $\mathbf{x}$ *according to* $p(\mathbf{x}|\theta_0)$.

2. *For* $k = 1$ *to* m *do*

   • *Generate a new candidate* ψ *following the density* $U_\delta(\theta_{k-1} \to \psi)$ *defined by*

   $$U_\delta(\theta \to \psi) = \frac{1}{V_\delta} \mathbf{1}_{b(\theta, \delta/2)}\{\psi\},$$

   with V_δ *the volume of the ball* $b(\theta, \delta/2)$.

   • *The new state* $\theta_k = \psi$ *is accepted with probability* $\alpha_s(\theta_{k-1} \to \psi)$ *given by*

   $$\alpha_s(\theta_{k-1} \to \theta_k) =$$
   $$= \min\left\{1, \left[\frac{p(\theta_k|\mathbf{y})}{p(\theta_{k-1}|\mathbf{y})} \times \frac{f(\mathbf{x}|\theta_{k-1})}{f(\mathbf{x}|\theta_k)}\right]^{1/T} \times \frac{\mathbf{1}_{b(\theta_k, \delta/2)}\{\theta_{k-1}\}}{\mathbf{1}_{b(\theta_{k-1}, \delta/2)}\{\theta_k\}}\right\}$$
   $$= \min\left\{1, \left[\frac{f(\mathbf{y}|\theta_k)p(\theta_k)}{f(\mathbf{y}|\theta_{k-1})p(\theta_{k-1})} \times \frac{f(\mathbf{x}|\theta_{k-1})}{f(\mathbf{x}|\theta_k)}\right]^{1/T}\right\} \tag{34}$$

   otherwise $\theta_k = \theta_{k-1}$.

 3. *Return θ_m.*

 4. *Stop the algorithm or go to step 1 with $\theta_0 = \theta_n$, $\delta_0 = k_\delta(\delta)$ and $T_0 = k_T(T)$.*

It is easy to see that the SSA algorithm is identical to Algorithm 1 whenever δ and T remain unchanged. The SSA algorithm does not have access at the states issued from the ideal chain. When the algorithm is started, the distance between the ideal and the shadow chain depends on the initial conditions. Still, independently of these conditions, as the control parameters δ_j and T evolve, this distance evolves also, by approaching zero.

5 Applications

This section illustrates the application of the SSA algorithm. The next part of this section applies the present method for estimating the model parameters of three point processes: Strauss, area-interaction and Candy model. All these models are widely applied in domains such environmental sciences, image analysis and cosmology [26, 12, 27, 14, 15, 18, 23, 21]. Here the parameter estimation is done on simulated data and it is double aimed. The first purpose is to test the method on complicated models, that does not exhibit a closed analytic form for their normalising constants. The second one is to give to the potential user, some hints regarding the tuning of the algorithm. The last part of this section is dedicated to real data application: model fitting for the galaxy distribution in our Universe.

5.1 Simulated data: point patterns and segment networks

The SSA Shadow algorithm is applied here to the statistical analysis of patterns which are simulated from a Strauss model [9, 19]. This model describes random patterns made of points exhibiting repulsion. Its probability density with respect to the standard unit rate Poisson point process is

$$p(\mathbf{y}|\theta) \propto \beta^{n(\mathbf{y})}\gamma^{s_r(\mathbf{y})}$$
$$= \exp\left[n(\mathbf{y})\log\beta + s_r(\mathbf{y})\log\gamma\right]. \tag{35}$$

Here $\mathbf{y}$ is a point pattern in the finite window W, while $t(\mathbf{y}) = (n(\mathbf{y}), s_r(\mathbf{y}))$ and $\theta = (\log\beta, \log\gamma)$ are the sufficient statistic and the model parameter vectors, respectively. The sufficient statistics components $n(\mathbf{y})$ and $s_r(\mathbf{y})$ represent respectively, the number of points in W and the number of pairs

of points at a distance closer than r.

The Strauss model on the unit square $W = [0,1]^2$ and with density parameters $\beta = 100$, $\gamma = 0.5$ and $r = 0.1$, was considered. This gives for the parameter vector of the exponential model $\theta = (4.60, -0.69)$. The CFTP algorithm (see Chapter 11 in [12]) was used to get 1000 samples from the model and to compute the empirical means of the sufficient statistics $\bar{t}(\mathbf{y}) = (\bar{n}(\mathbf{y}), \bar{s}_r(\mathbf{y})) = (45.30, 17.99)$. The SA based on the ABC Shadow algorithm was run using $\bar{t}(\mathbf{y})$ as observed data, while considering the r parameter known.

The prior density $p(\theta)$ was the uniform distribution on the interval $[0,7] \times [-7,0]$. Each time, the auxiliary variable was sampled using 100 steps of a MH dynamics [26, 12]. The Δ and m parameters were set to $(0.01, 0.01)$ and 200, respectively. The algorithm was run for 10^6 iterations. The initial temperature was set to $T_0 = 10^4$. For the cooling schedule a slow polynomial scheme was chosen

$$T_n = k_T \cdot T_{n-1}$$

with $k_T = 0.9999$. A similar scheme was chosen for the Δ parameters, with $k_\Delta = 0.99999$. Samples were kept every 10^3 steps. This gave a total of 1000 samples.

The choice of the uniform prior was motivated by the fact that the posterior distribution using an uniform prior equals the likelihood distribution restricted to the domain of availability of the uniform distribution. Hence, following [1, 26, 12] and the argument in [17], since the ML estimate approaches almost surely the true model parameters whenever the number of samples increases, the expected results of the SSA algorithm should be close to the model parameters used for its simulation.

Figure 1 shows the time series of the SSA algorithm outputs applied to estimate the parameters of the previous Strauss process. The final values obtained for $\log \beta$ and $\log \gamma$ were 4.63 and -0.71, respectively. The Table 1 presents the empirical quartiles computed from the outputs of the algorithm. All these values are close to the true model parameters.

The area-interaction process introduced by [2] is able to describe point patterns exhibiting clustering or repulsion. Its probability density with respect
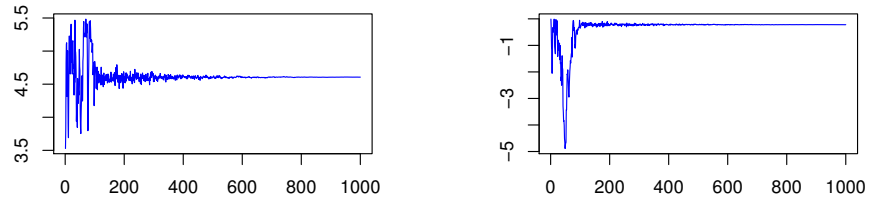
Figure 1: SSA outputs for the MAP estimates computation of the Strauss model parameters. The true parameters of the model were $\theta = (4.60, -0.69)$, while the estimates are $\widehat{\theta} = (4.63, -0.71)$.

Summary statistics SSA Strauss estimation			
Parameters	Q_{25}	Q_{50}	Q_{75}
SSA $\log \beta$	4.598	4.606	4.611
SSA $\log \gamma$	-0.728	-0.716	-0.708

Table 1: Empirical quartiles for the SSA Strauss model estimation.

to the standard unit rate Poisson point process is

$$p(\mathbf{y}|\theta) \propto \beta^{n(\mathbf{y})} \gamma^{a_r(\mathbf{y})} = \exp\left[n(\mathbf{y}) \log \beta + a_r(\mathbf{y}) \log \gamma\right].$$

with

$$a_r(\mathbf{y}) = -\frac{\nu\left[\cup_{i=1}^{n} b(y_i, r)\right]}{\pi r^2}.$$

The model vectors of the sufficient statistics and parameters are $t(\mathbf{y}) = (n(\mathbf{y}), a_r(\mathbf{y}))$ and $\theta = (\log \beta, \log \gamma)$, respectively. If $\log \gamma < 0$ the point pattern surface induced by the radii around the points tends to occupy the whole domain W. This leads to a regular or repulsive distribution of points. If $\log \gamma > 0$ the point pattern surface induced by the radii around the points tends to be reduced. This leads to an aggregate or clustered distribution of points. Hence, parameter estimation of this model is also a morphological indicator, while sampling its posterior allows to assess statistical significance of this tendency [17].

The following experiment was carried out, by considering the area-interaction model on the unit square $W = [0, 1]^2$ and with density parameters $\beta = 200$, $\gamma = e$ and $r = 0.1$. This gives for the parameters vector of the exponential model $\theta = (5.29, 1)$. A MH algorithm (see Chapter 7 in [12]) was used to get 1000 samples from the model and to compute the empirical means of the sufficient statistics $\bar{t}(\mathbf{y}) = (\bar{n}(\mathbf{y}), \bar{a}_r(\mathbf{y})) = (144.31, -78.88)$. As previously, the SSA algorithm was run using $\bar{t}(\mathbf{y})$ as observed data, while considering the r parameter known.

The prior density $p(\theta)$ was the uniform distribution on the interval $[0, 7] \times [-5, 5]$. Each time, the auxiliary variable was sampled using 250 steps of a MH dynamics [26, 12]. The Δ and m parameters were set to $(0.01, 0.01)$ and 100, respectively. The cooling schedule for the temperature, the descending scheme for δ, the number of iterations and the number of samples were chosen as in the previous experiment.

Figure 2 and the Table 2 present the obtained results of the SSA algorithm applied to estimate the parameters of the previous area-interaction process. The final values of the algorithm's time series outputs were for $\log \beta$ and $\log \gamma$, 5.30 and 1.03, respectively. The empirical quartiles computed from the outputs of the algorithm indicate that the final results are close to the true model parameters.
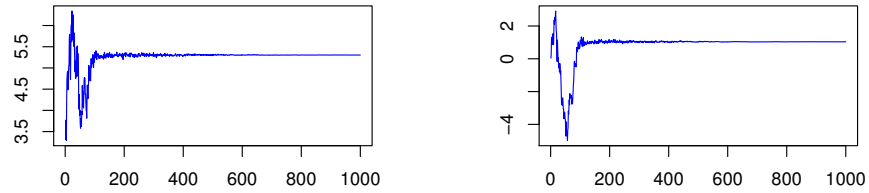
Figure 2: SSA outputs for the MAP estimates computation of the area-interaction model parameters. The true parameters of the model were $\theta = (5.29, -1.00)$, while the estimates are $\widehat{\theta} = (5.30, -1.03)$.

Summary statistics for SSA Area Interaction estimation			
Parameters	Q_{25}	Q_{50}	Q_{75}
SSA $\log \beta$	5.298	5.304	5.308
SSA $\log \gamma$	1.026	1.035	1.042

Table 2: Empirical quartiles for the SSA Area Interaction model estimation.

The Candy model is an object point process that simulates networks made of connected segments [27]. The model was successfully applied in image analysis and cosmology [14, 18].

A segment $y = (w, \xi, l)$ is given by its centre w, its orientation ξ and its length l. The orientation is a uniform random variable on $M = [0, \pi)$, while the length is a fixed value. The probability density of the considered Candy model, with respect to the standard unit rate Poisson point process, is

$$p(\mathbf{y}|\theta) \propto \exp\langle\theta_d n_d(\mathbf{y}) + \theta_s n_s(\mathbf{y}) + \theta_f n_f(\mathbf{y}) + \theta_r n_r(\mathbf{y})\rangle \qquad (36)$$

with $\mathbf{y}$ a segments configuration, $\theta = (\theta_d, \theta_s, \theta_f, \theta_r)$ and

$$t(\mathbf{y}) = (n_d(\mathbf{y}), n_s(\mathbf{y}), n_f(\mathbf{y}), n_r(\mathbf{y}))$$

the parameter and the sufficient statistic vectors, respectively. Each parameter controls its associate statistic. Here, n_d is the number of segments connected at both of its extremities or doubly connected, n_s is the number of segments connected at only one of its extremities or singly connected, n_f is the number of segments that are not connected or free and n_r the number of pairs of segments that are too close and not orthogonal.

The connectivity of two segments is defined by the relative position of their extremities and their relative orientation. Two segments with only one pair of extremities situated within the connection distance r_c and with absolute orientation difference lower than a curvature parameter τ_c are connected. Similar to the orientation difference, the orthogonality of two segments is controlled the parameter τ_r. For full details regarding the Candy model description and properties we recommend [27].

Figure 3 pictures a realisation of the Candy model on $W = [0, 3] \times [0, 1]$. The segment length is $l = 0.12$, the connection distance is $r_c = 0.01$, and the curvature parameters are $\tau_c = \tau_r = 0.5$ radians. The model parameters are $\theta_d = 10$, $\theta_s = 6$, $\theta_f = 2$ and $\theta_r = -1$. It can be noticed that with these parameters the model outcomes tend to form patterns of a rather connected segments. An adapted MH algorithm [27] was used to obtain 10000 samples of the previous model and to compute the vector of the empirical means of the sufficient statistics $\bar{t}(\mathbf{y}) = (\bar{n}_d(\mathbf{y}) = 55.67, \bar{n}_s(\mathbf{y}) = 50.26, \bar{n}_f(\mathbf{y}) = 10.90, \bar{n}_r(\mathbf{y}) = 40.24)$. These statistics were used as the data entry for a ABC SA algorithm.
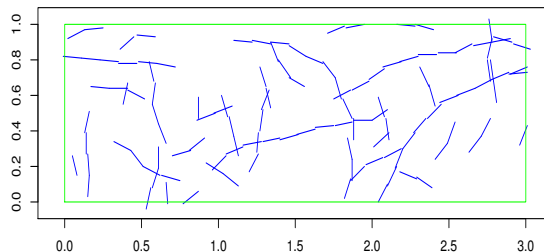
Figure 3: Realisation of the Candy model.

The prior density $p(\theta)$ was the uniform distribution on the interval $[0,12]^3 \times [-12,0]$. Each time, the auxiliary variable was sampled using 200 steps of the adapted MH dynamics [28]. The Δ and m parameters were set to $(0.01, 0.01, 0.01, 0.01)$ and 500, respectively. The other algorithm's parameters were chosen as in the previous examples.

The algorithm results of the SSA are shown in Figure 4 and Table 3. The final values of the algorithm's time series outputs were for θ_d, θ_s, θ_f and θ_r, 10.009,6.002,1.982 and -0.996 respectively. Together with the empirical quartiles given by the outputs of the algorithm, all these indicate that the algorithm outputs and the true model parameters are again rather close.

Summary statistics for the SSA Candy estimation			
Algorithm	Q_{25}	Q_{50}	Q_{75}
SSA θ_d	9.958	9.988	10.016
SSA θ_s	5.983	6.009	6.040
SSA θ_f	1.944	1.974	2.012
SSA θ_r	-1.017	-0.985	-0.962

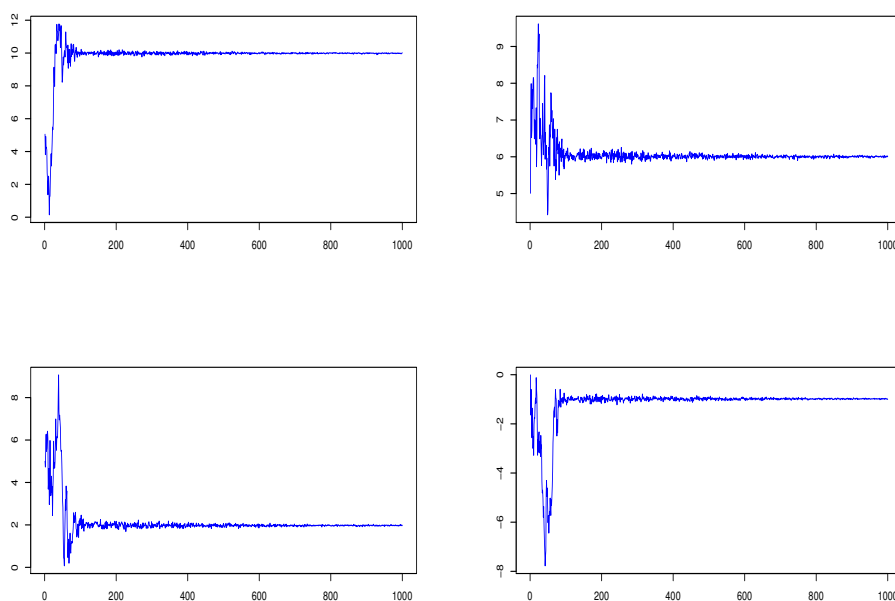Table 3: Empirical quartiles for the SSA Candy model estimation.

Figure 4: SSA outputs for the MAP estimation of the Candy model parameters. The true parameters of the model were $\theta = (10, 6, 2, -1)$, while the estimates are $\widehat{\theta} = (10.009, 6.002, 1.982, -0.996)$.

5.2 Cosmology real data application: a point process model for fitting the galaxies distribution

The galaxies are not spread uniformly in our Universe. Their position exhibits an intricate pattern made of filaments and clusters [11]. Knowing the positions of the galaxies, while assuming a Bayesian working framework, pattern detectors were built for filaments [18, 23, 22] and more recently for clusters [21].

The previous detectors do not assume any particular model for the spatial distribution of galaxies. Now, given the detected structure, fitting a statistic model to the observed field of galaxies becomes a natural question. As a continuation of the application presented in [17], we show that the ABC Shadow posterior sampling algorithm and the SSA algorithm for parameter estimation are tools to be considered in such a task.

Figure 5 pictures a sample of the considered data. It represents a cube of side 30 h^{-1} Mpc from SDSS (Sloan Digital Sky Survey) catalogue together with the induced filaments pattern [23]. Here the main axes of the filaments pattern are represented by a set of continuous curves that are called spines. More formally, the spines are ridge lines given by those regions where the filaments structures exhibited by the galaxies positions are the most aligned and the most connected. For full details regarding filaments detection and spines computation, the reader may refer to [23, 22].

Fitting a model to the galaxy distribution is extremely complex task [11]. The authors [23, 20, 22] suppose and infer that the galaxies tend to be close to the main axes of the filaments pattern, while forming clusters, similar to pearl on a necklace. Therefore, in the following, we consider a point process model that controls the number of galaxies, their proximity to the filaments network spine and their mutual interactions.

Such a model can be represented by the point process given by the following probability density

$$p(\mathbf{y}|\theta, F) \propto \beta_1^{n(\mathbf{y})} \beta_2^{d_F(\mathbf{y})} \gamma^{-a_r(\mathbf{y})} \tag{37}$$

with the model parameters vector given by

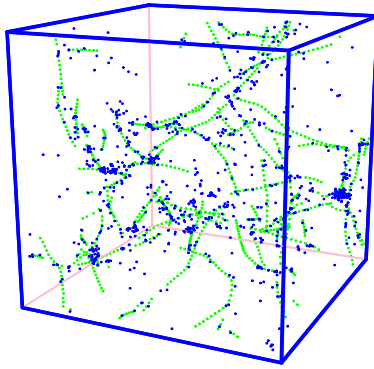$$\theta = (\log \beta_1, \log \beta_2, \log \gamma),$$

Figure 5: Galaxies positions (blue) and the induced filaments pattern or spines (green).

and the sufficient statistics vector

$$t(\mathbf{y}) = (n(\mathbf{y}), d_F(\mathbf{y}), a_r(\mathbf{y})).$$

The parameter $\beta_1 > 0$ controls the statistics $n(\mathbf{y})$ which represents the total number of galaxies in the configuration $\mathbf{y}$. The parameter $\beta_2 > 0$ controls the proximity of the galaxies centres from the observed filaments pattern F. Its associate sufficient statistic is :

$$d_F(\mathbf{y}) = - \sum_{i=1}^{n(\mathbf{x})} d(y_i, F)$$

with $d(y_i, F)$ the minimum distance from the galaxy position y_i to the spines pattern F. The parameter $\gamma > 0$ produces repulsive $(\gamma < 1)$ or clustering $(\gamma > 1)$ interactions among galaxies. The corresponding sufficient statistics is

$$a_r(\mathbf{y}) = \frac{3A(\mathbf{x})}{4\pi r^3}, A(\mathbf{y}) = \nu[\bigcup_{i=1}^{n(\mathbf{x})} b(y_i, r)].$$

Here $b(y, r)$ is the ball centred in y with radius r. So, $A(\mathbf{y})$ represents the volume of the object resulting from the set union of the spheres of radius r and that are centred in the points given by the configuration $\mathbf{y}$. The division of $A(\mathbf{y})$ by the volume of a sphere, allows to interpret the statistic $a_r(\mathbf{x})$ as the number of points or spheres needed to form a structure with volume $A(\mathbf{y})$.

The proposed model (37) is an inhomogeneous area-interaction process. The inhomogeneity governs the distribution of the galaxies with respect to the filaments pattern, while the area-interaction component controls the cluster formation among the galaxies. For further reading regarding the properties of the point process model we recommend [2].

The sufficient statistics of the model are computed from the considered data set for various interaction radii. They are represented in the Table 4. The total number of points in the observed volume and the distance to the filaments pattern do not depend on the interaction range.
First, for each radius value, the posterior distribution (2) associated to the model (37) was maximised using the SSA Shadow algorithm. The prior density $p(\theta)$ was the uniform distribution on the interval $[-50, 50]^3$. Each time, the auxiliary variable was sampled using 100 steps of a MH dynamics [26, 12]. The Δ and m parameters were set to $(0.01, 0.01, 0.01)$ and 100,

Data for the galaxy pattern						
r	0.5	1	1.5	2	2.5	3	3.5
$n(\mathbf{y}) = 1024,\ d_F(\mathbf{y}) = -\sum_{i=1}^{n(\mathbf{y})} d(y_i, F) = -1180.05$						
$a_r(\mathbf{y})$	724.29	484.01	357.16	263.10	195.08	142.86	105.30

Table 4: The observed sufficient statistics computed for the galaxy pattern, depending on the range parameter r.

respectively. The descending schedules for the temperature and the δ parameter, were fixed as in the previous examples. The algorithm was run for 10^6 iterations. Samples were kept every 10^3 steps. This gave a total of 1000 samples.

The obtained results are shown in the Figure 6. The right column of the figure presents for each radius, the boxplots of the outputs of the SSA algorithm, associated to each model parameter. The model parameter behaviour can be analysed while r increase. The β_1 parameter that is a baseline for the number of galaxies in the observed volume tend to stabilise. The evolution of the β_2 parameter indicates that the proximity to the filaments becomes more and more important while the considered interaction ranges increase. Almost the same thing can be stated for the behaviour of the γ parameter. Still, there is a difference between the β_2 and the γ parameters behaviour: the first one tends maybe to stabilise while the second one, exhibit maybe a decreasing tendency.

The left column in the Figure 6 shows the evolution of the SSA algorithm for $r = 2$. As for the previous cases, for high temperature values, the algorithm travels around the configuration space, while for low temperatures, it approaches the convergence regime. The obtained estimates for $\widehat{\theta} = (\widehat{\log \beta_1}, \widehat{\log \beta_2}, \widehat{\log \gamma})$ are $(-0.33, 0.98, 4.57)$.

Since the chosen prior distribution $p(\theta)$ was the uniform distribution over the compact parameter space Θ, the SSA output is the Maximum Likelihood Estimate (MLE) restricted to Θ. Hence, asymptotic errors may be derived following [5, 27, 17]. For each parameter, the asymptotic standard deviation and the Monte Carlo Standard error (MCSE) were computed respectively. The asymptotic standard deviation indicates the difference between the MLE estimate and the true model parameters. The SSA output can be
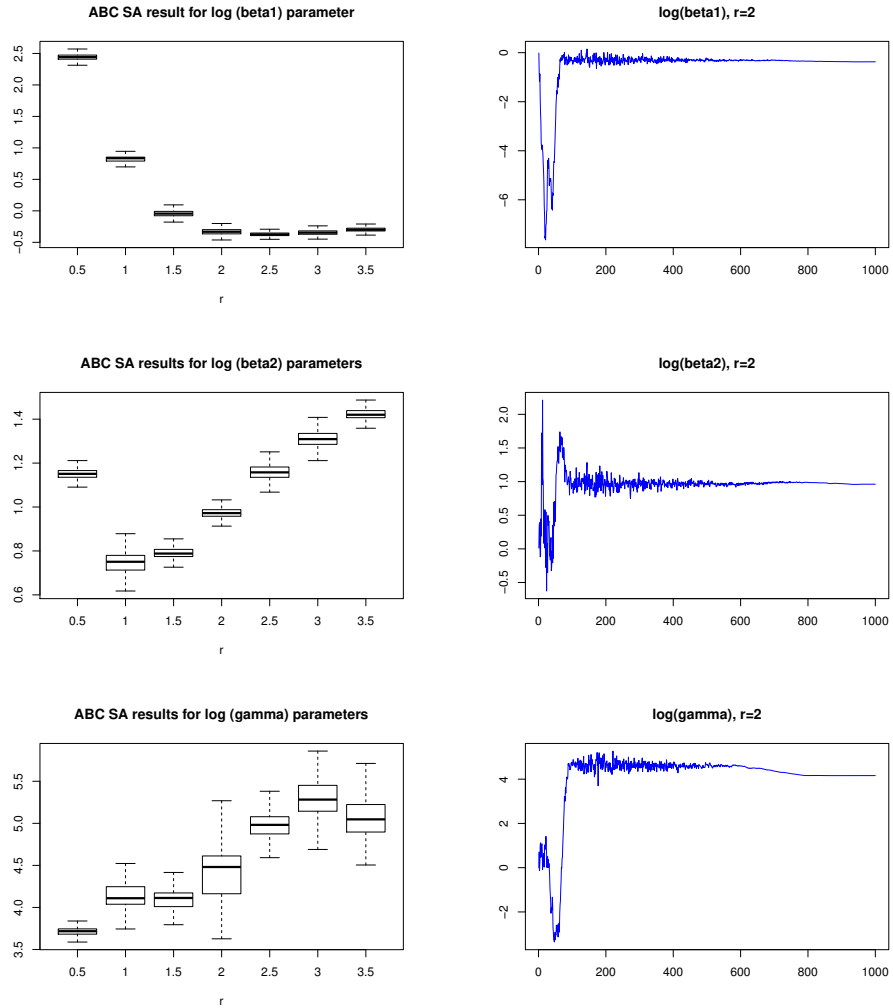
Figure 6: SSA outputs for the MAP estimates computation of the inhomogeneous area-interaction model parameters fitted to the considered SDSS sample. Right column: box plots of SSA algorithm outputs for each parameter depending on the interaction radius. Left column: the time series of the outputs of the SSA algorithm for $r = 2$.

also interpreted as a MCMCML estimate. Then the MCSE represents the difference between the MLE and its Monte Carlo counterpart. The obtained results are presented in Table 5, where for each parameter the first column indicates the asymptotic standard deviation and the second one, the MCSE. These values indicate a rather high quality of the estimation for models with interaction radius smaller than $3 \ h^{-1}$ Mpc. This may be explained by the fact that maybe a change of the clustering regime appears at these interaction ranges [3]. Such a change is also suggested in Figure 6 where the interaction parameter behaves like reaching a maximum around $3 \ h^{-1}$ Mpc.

Asymptotic errors for the SSA estimates					
r	$\sigma_{\log \beta_1}$	$\sigma_{\log \beta_1}^{MC}$	$\sigma_{\log \beta_2}$	$\sigma_{\log \beta_2}^{MC}$	$\sigma_{\log \gamma}$	$\sigma_{\log \gamma}^{MC}$
0.5	0.04	1e-4	0.03	7e-5	0.07	2e-4
1	0.03	1e-4	0.03	9e-5	0.09	5e-4
1.5	0.05	1e-4	0.04	1e-4	0.12	1e-3
2	0.05	1e-4	0.04	2e-4	0.17	2e-3
2.5	0.05	2e-4	0.04	2e-4	0.24	5e-3
3	0.05	2e-4	0.04	5e-4	0.43	0.016
3.5	0.05	2e-4	0.04	9e-4	0.66	0.039

Table 5: Asymptotics errors for the SSA MAP estimates of the model (37) fitted to the considered cosmological sample. For each corresponding radius a model was fitted, and the asymptotic errors were computed for each model. For the computation of the MCSE 15×10^3 samples from the fitted model were used.

In order to test the values and the significance of each of the model parameters, the ABC Shadow algorithm was used as in [17]. The algorithm was run for 10^5 iterations. Samples were kept every 100 steps. This gave a total of 1000 samples.

The results obtained for the interaction radius $r = 2$ are shown in the Figure 7. From the obtained results, the marginals of each parameter posterior are approximated using an Epanechnikov kernel, and on this basis, the computed MAP is $\widehat{\theta} = (-0.29, 0.95, 4.57)$. The results are coherent since they are originated from an approximate distribution while fitting into the confidence intervals induced by the SSA estimation and its associate asymptotic standard error.

In the following, based on the ABC Shadow approximation of the posterior
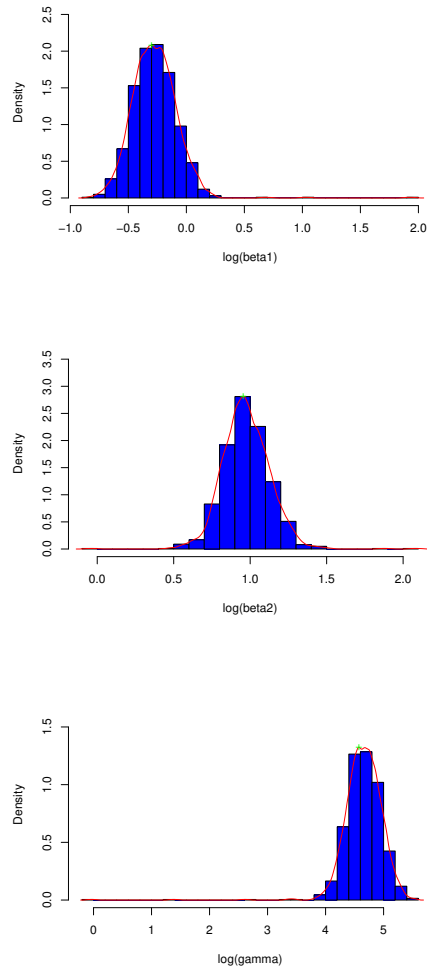
Figure 7: ABC Shadow outputs for the approximate posterior sampling of the inhomogeneous area interaction model fitted to the SDSS sample (the range parameter is $r = 2$). The obtained MAP estimate is $\widehat{\theta} = (-0.29, 0.95, 4.57)$.

of (37) described previously, two statistical tests were conducted.

First, a Student test was carried on to check whether the mean of the posterior distribution is different from the the SSA algorithm output. This test was conducted for each parameter, respectively. It used the marginal posterior samples given by the ABC Shadow. The obtained $p-values$ were all greater than 0.77, so there is no evidence that the approximated posterior mean is different from the SSA output. This is a rather encouraging result, since the ABC Shadow is an approximate method, while the SSA exhibits convergence.

Next, a second Student test was conducted in order to verify whether the obtained parameter values are significantly different from 0. The obtained $p-$value for $\log \beta_1$ was 0.16. For the parameters $\log \beta_2$ and $\log \gamma$ the corresponding $p - values$ were less than 10^{-5}. The result indicates that the β_1 parameter is not significantly different from 1. It also gives statistical significance of the following cosmological facts: the galaxies tend to be distributed close to the filaments while forming clusters, and only rarely being placed independently in our Universe.

6   Conclusion and perspectives

This paper presents a global optimisation method, the Shadow Simulated Annealing (SSA) algorithm, that can be applied to a family of criteria that are not fully known. It can be applied to maximise posterior probability densities exhibiting normalising constants that are not available in analytic closed form. The SSA algorithm can be used with those posteriors provided by probability densities that are continuously differentiable with respect to their parameters. This is rather a strong hypothesis but often encountered in practice.

The method provides convergence results towards the global optimum, that are equivalent with the results obtained for the simulated annealing whenever the optimisation criterium is entirely known [4, 6, 7].

This work opens perspectives from a mathematical and application point of view. The mathematical challenge is to extend the family of criteria to which these methods apply. From an applied point of view, a thorough sta-

tistical study of the galaxies distribution in our Universe, based on the tools presented in this work, it is currently carried on by part of the authors of the paper.

Aknowledgements

References

[1] A. J. Baddeley, E. Rubak, and R. Turner. *Spatial Point Patterns: Methodology and Applications with R.* Chapman and Hall/CRC Press, London, 2016.

[2] A. J. Baddeley and M. N. M. van Lieshout. Area-interaction point processes. *Annals of the Institute of Statistical Mathematics*, 47:601–619, 1995.

[3] M. Einasto, L.J. Liivamägi, E. Tempel, E. Saar, J. Vennik, P. Nurmi, M. Gramann, J. Einasto, E. Tago, P. Heinämäki, A. Ahvensalmi, and V.J. Martinez7. Multimodality of rich clusters from the sdss dr8 within the supercluster-void network. *Astronomy and Astrophysics*, 542, 2012.

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

[5] C. J. Geyer. Likelihood inference for spatial point processes. In O. Barndorff-Nielsen, W.S. Kendall, and M.N.M. van Lieshout, editors, *Stochastic Geometry, Likelihood and Computation*. CRC Press/Chapman and Hall, Boca Raton, 1999.

[6] H. Haario and E. Saksman. Simulated annealing process in general state space. *Advances in Applied Probability*, 23(4):866–893, 1991.

[7] H. Haario and E. Saksman. Weak convergence of the simulaed annealing process in general state space. *Annales Academiae Scientiarum Fennicae*, 17:39–50, 1992.

[8] M. Iosifescu and R. Theodorescu. *Random Processes and Learning*. Springer, 1969.

[9] F. P. Kelly and B. D. Ripley. A note on Strauss's model for clustering. *Biometrika*, 63(2):357–360, 1976.

[10] A. Makur. Introduction to Markov mixing. `http://www.mit.edu/~ a_makur/docs/Markov Chain Mixing Coefficients Ergodicity.pdf`, 2016.

[11] V. J. Martinez and E. Saar. *Statistics of the galaxy distribution*. Chapman and Hall, 2002.

[12] J. Møller and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC, Boca Raton, 2004.

[13] G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

[14] R. S. Stoica, X. Descombes, and J. Zerubia. A Gibbs point process for road extraction in remotely sensed images. *International Journal of Computer Vision*, 57:121–136, 2004.

[15] R. S. Stoica, E. Gay, and A. Kretzschmar. Cluster detection in spatial data based on Monte Carlo inference. *Biometrical Journal*, 49(2):1–15, 2007.

[16] R. S. Stoica, P. Gregori, and J. Mateu. Simulated annealing and object point processes : tools for analysis of spatial patterns. *Stochastic Processes and their Applications*, 115:1860–1882, 2005.

[17] R. S. Stoica, A. Philippe, P. Gregori, and J. Mateu. Abc shadow algorithm: a tool for statistical analysis of spatial patterns. *Statistics and Computing*, 27:1225–1238, 2017.

[18] R. S. Stoica, E. Tempel, L. J. Liivamägi, G. Castellan, and E. Saar. Spatial patterns analysis in cosmology based on marked point processes. In Statistics for astrophysics. Methods and applications of the regression, editors, *Statistics for astrophysics. Methods and applications of the regression*. European Astronomical Society Publication Series, EDP Sciences, 2015.

[19] D. J. Strauss. A model for clustering. *Biometrika*, 62:467–475, 1975.

[20] E. Tempel, R. Kipper, E. Saar, M. Bussov, A. Hektor, and J. Pelt. Galaxy filaments as pearl necklaces. *Astronomy and Astrophysics*, 572 (A8), 2014.

[21] E. Tempel, M. Kruuse, R. Kipper, T. Tuvikene, J. G. Sorce, and R. S. Stoica. Bayesian group finder based on marked point processes. method and application to the 2mrs data set. *Astronomy and Astrophysics*, 618 (A61):1–18, 2018.

[22] E. Tempel, R. S. Stoica, R. Kipper, and E. Saar. Bisous model - detecting filamentary pattern in point processes. *Astronomy and Computing*, 16:17–25, 2016.

[23] E. Tempel, R. S. Stoica, E. Saar, V. J. Martinez, L. J. Liivamägi, and G. Castellan. Detecting filamentary pattern in the cosmic web: a catalogue of filaments for the SDSS. *Monthly Notices of the Royal Astronomical Society*, 438(4):3465–3482, 2014.

[24] L. Tierney. Markov chains for exploring posterior distribution (with discussion). *The Annals of Statistics*, 22(4):1701–1762, 1994.

[25] M. N. M. van Lieshout. Stochastic annealing for nearest-neighbour point processes with application to object recognition. *Adv. in Appl. Probab.*, 26(2):281–300, 1994.

[26] M. N. M. van Lieshout. *Markov Point Processes and their Applications*. Imperial College Press, London, 2000.

[27] M. N. M. van Lieshout and R. S. Stoica. The Candy model revisited: properties and inference. *Statistica Neerlandica*, 57:1–30, 2003.

[28] M. N. M. van Lieshout and R. S. Stoica. Perfect simulation for marked point processes. *Computational Statistics and Data Analysis*, 51:679–698, 2006.

[29] Gerhard Winkler. *Image analysis, random fields and Markov chain Monte Carlo methods*, volume 27 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, second edition, 2003. A mathematical introduction, With 1 CD-ROM (Windows), Stochastic Modelling and Applied Probability.