Summary

Computational efficiency is a bottleneck for the use of stochastic geometry models in high-level Bayesian object recognition. In this paper we examine various approaches to reversible jump MCMC that lead to higher acceptance rates than the traditional approach of using a uniform proposal density. This is achieved by exploiting information about the posterior gained through a deterministic image processing tool, the Hough transform. We discuss how to choose an appropriate scaling of our proposal distributions following suggestions made in (Brooks et al., 2003). The scaling is based on an examination of the acceptance probability at a chosen point. We then extend the approach from uni-modal proposal distributions to multi-modal distributions and choose the scaling based on an examination of the acceptance probability at a number of points.

KEYWORDS: Bayesian object recognition, Markov object process, reversible jump MCMC, scaling.

1 Introduction

In (Baddeley and van Lieshout, 1993) Markov object processes were introduced as prior models for Bayesian image analysis. In contrast to Markov random fields, their discrete analogues, these models provide feature-based rather than pixel-based descriptions and so are suitable for high-level vision tasks. For example, Markov object processes have been successfully used for the extraction of road networks from satellite images (Lacoste et al., 2003). While providing high-level models, Markov object processes are complex, in particular when conditioning on observed data. Thus inference has to be simulation based, typically relying on methods such as spatial birth-and-death processes or reversible jump Markov chain Monte Carlo (MCMC). However, such MCMC algorithms can be rather slowly mixing and it often takes ingenuity to develop algorithms that have practical run-times. For example, the authors in (Baddeley and van Lieshout, 1993) consider approximations to the target spatial birth-death process in order to produce a more viable sampling procedure.

In this paper we build on the ideas of (Baddeley and van Lieshout, 1993) but utilize a reversible jump MCMC algorithm. This allows us to use some
of the approximation methods proposed in (Baddeley and van Lieshout, 1993) but still to define a Markov chain whose equilibrium distribution is the exact target distribution. The approximations are produced by standard, deterministic image processing methods and our work is a step towards examining how such deterministic methods may be exploited systematically to produce better mixing reversible jump MCMC algorithms.

Recently Brooks et al. (2003) developed a framework for choosing more efficient jump mechanisms in reversible jump MCMC. The methods are based on Taylor expansions of acceptance probabilities and have close connections to Langevin algorithms. We show how our approach fits into this framework and explore some of the recommendations for choosing proposal distributions made by (Brooks et al., 2003). We then extend the approach by considering the acceptance probability at various local modes in the likelihood. This produces a multi-modal proposal distribution in contrast to the uni-modal ones used in (Brooks et al., 2003). To illustrate the various approaches we choose a simple disk detection problem and present a comparison of the methods.

2 Bayesian object recognition

In the following we illustrate our methods on a simple problem in high level vision: the recognition of an unknown number of (possibly overlapping) objects of simple geometric shape. Of course, more complex image analysis tasks can be addressed using reversible jump MCMC but the aim here is to provide a first evaluation of the methods proposed rather than problem-specific optimizations. While in this paper we examine an artificial image only, the specific example of disk detection is motivated by the real-world problem of extracting sweat pores from fingerprint images (Parsons et al., 2005). Figure 1 shows a collection of non-overlapping disks in a 128×128 pixel grey level image. The image has been corrupted by iid pixelwise additive Gaussian white noise with standard deviation $\sigma = 20$.

2.1 The likelihood model

We describe a configuration of $n$ objects by a list $x = \{(x_1, s_1), \ldots, (x_n, s_n)\}$ where $x_i, i \in \{1, \ldots, n\}$, denotes the location of the centre of the object in the bounded sampling window $W \subset \mathbb{R}^2$. For simplicity we assume here that the sampling window is the unit square. Each object $(x_i, s_i)$ is further specified by a set of parameters $s_i \in \mathbb{R}^d$ which determine its shape. For example, $s_i = (\theta_i, l_i)$ may parameterize the orientation and length of a line segment. In our test example, the objects are disks of fixed radius and so here a configuration of $n$ objects is fully specified by the list of disk centre locations $x = \{x_1, \ldots, x_n\}$.

Each object covers a spatial region within the sampling window $W$ which is denoted by $R(x_i, s_i)$. The subset of the sampling window which is covered by at least one of the $n$ objects in $x$ is then denoted as $R(x) = \bigcup_{i=1}^{n} R(x_i, s_i)$.

Next, the object configuration is displayed in an image and so we define a function $I_x$ that describes the pixel image of the object configuration $x$. We assume each object has the same grey level or colour $g_1$ while the background has grey level or colour $g_0$. The sampling window $W$ is divided into a grid of
Figure 1: An image showing a configuration of non-overlapping disks corrupted by iid additive Gaussian noise with standard deviation $\sigma = 20$.

pixels $\Lambda$ and the object configuration $x$ is displayed by the pixel image

$$I_x(t) = \begin{cases} g_1 & \text{if } t \in R(x) \\ g_0 & \text{if } t \notin R(x) \end{cases} \quad t \in \Lambda.$$  

The image function $I_x$ describes the pixel image of the object configuration $x$ without the presence of any noise. However, in practice we observe an image that is corrupted by noise. For example, if we assume iid additive Gaussian white noise, then the value of each pixel is distorted independently from any other pixel by a Normally distributed error of zero mean. The conditional density for observing the noisy image $J_x = y$ given the underlying, true object configuration is $x$ is now given by

$$f(y|x) = \prod_{t \in \Lambda} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} \left( y(t) - I_x(t) \right)^2 \right] \tag{1}$$

where $\sigma^2$ is the noise variance. In the following we will assume (1) as our noise model. This may seem oversimplistic, but the model can be extended easily to more complex settings, for example by incorporating blurring. The noisy image displayed in Figure 1 has parameters $g_0 = 0$ and $g_1 = 40$ on a 256 grey level scale and was corrupted by iid additive Gaussian white noise with $\sigma = 20$.

To infer the unobserved object configuration $x$ given the observed, noisy image $y$ we examine the likelihood function:

$$L(x|y) \propto f(y|x).$$

However, estimates based on the likelihood function like the maximum likelihood estimate (MLE), can have undesirable properties, see for example (Baddeley and van Lieshout, 1993). We avoid this by choosing a Bayesian paradigm which is described in the next section.
2.2 The prior and posterior model

Commonly, we have some knowledge a priori about the likely configuration $x$. For example, we may know that the objects rarely occlude each other or we may have a priori information about connectivity or certain smoothness properties.

A systematic way of introducing this information is to specify it through a probability density $\pi(\cdot)$ on the space $\Omega$ of finite object configurations on $\mathbb{R}^2$. Here, we consider Markov object processes as prior models which can be defined through a probability density $\pi(\cdot)$ with respect to a unit rate Poisson process on the sampling window $W$. The density satisfies the following properties:

1. $\pi(\cdot)$ is hereditary, that is if $\pi(x) > 0$ then $\pi(y) > 0$ for all $y \subset x$;

2. the density satisfies the following spatial Markov property: if $\pi(x) > 0$ then $\pi(x \cup \{\xi\})/\pi(x)$ depends only on $\xi$ and the set $\{x_i \in x : \xi \sim x_i\}$ where $\sim$ is a symmetric relation between objects.

(Note that we use $x_i$ or small, Greek letters to denote a single object and $x$ to denote an object configuration.) We further restrict our attention to locally stable Markov object processes. Define the Papangelou conditional density of an object process with density $\pi(\cdot)$ as

$$\lambda(x, \xi) = \begin{cases} \frac{\pi(x \cup \{\xi\})}{\pi(x)} & \text{if } \pi(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

A Markov object process is locally stable if its Papangelou conditional intensity is uniformly bounded, that is there exists $\lambda^* > 0$ such that

$$\lambda(x, \xi) \leq \lambda^* \quad \text{for all } x \subset W, \xi \in W.$$

Having chosen the prior density $\pi(\cdot)$ we then condition on the observed data $y$ and use Bayes theorem to derive the conditional density for object configurations $x$ given an observed, noisy image $y$ as

$$\pi(x|y) \propto \pi(x)L(x|y).$$

We call the density $\pi(x|y)$ the posterior density and all inference is now based on this posterior model.

In our example we choose a Strauss process as prior model for the location of disks. The density $\pi(\cdot)$ of a Strauss process is given by

$$\pi(x) = \alpha \beta^n(x) \gamma^t(x).$$

Here $\alpha$ is the normalising constant and $\beta > 0$. The interaction parameter $\gamma$ lies between zero and one, $n(x)$ is the number of points in $x$ and $t(x)$ is the number of neighbour pairs in $x$. A point pair in $x$ is a neighbour pair if its points are less than the distance $R$ apart from each other. If $\gamma = 0$ then the above process is a hardcore model in which points are almost surely than the interaction radius apart from each other. For $0 < \gamma < 1$ the Strauss process favours point patterns in which only a few point pairs are neighbours.

If we choose $R/2$ equal to the disk radius $r$ then our prior favours object configurations where only a few objects occlude each other. In our test examples we assume that objects never occlude each other and so we choose $R = 2r$ and...
Choosing a Strauss process as our prior model compensates for some of the undesirable properties of the likelihood function. The likelihood function has no penalty for the number of objects in the true configuration \( x \). Thus when maximizing the likelihood we obtain configurations with a large number of occluding objects, representing an “over-fitting” of the image. Using a Strauss prior induces a penalty on the number of occluding objects (and, to a lesser extent, a penalty on an excessively large number of objects). Thus maximizing the log posterior density can be interpreted as a penalized log likelihood maximization with the penalization factor being proportional to the number of neighbour pairs. This produces the so-called MAP (maximum a posteriori) estimate which leads to a more parsimonious image interpretation than the MLE.

As mentioned earlier we use a simple disk detection problem as our test example. The sampling window is the unit square and we choose a Strauss hardcore process as prior model. The parameters of the prior model are \( \beta = 30 \), \( \gamma = 0 \) and \( R = 0.013 \). As we assume a fixed radius \( r = 0.013 \) for the disks a configuration \( x \) of \( n \) objects is fully specified by the list of disk centres \( x = \{x_1, \ldots, x_n\} \). We have

\[ R(x_i) = \{\xi \in W : ||\xi - x_i|| < r\}. \]

In summary, the posterior density given we observe the noisy image \( y \) is equal to

\[ \pi(x|y) = \alpha \beta^{n(x)} 1_{[t(x)=0]} \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{d}{2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{t \in \Lambda} (y(t) - I_x(t))^2 \right]. \] (2)

### 2.3 Sampling algorithms

The idea of all MCMC algorithms is to simulate a Markov chain whose equilibrium distribution coincides with the target distribution. Then, after the chain has run for a sufficiently long time, samples from the Markov chain have approximately the target distribution.

#### 2.3.1 Spatial birth-and-death processes

The use of spatial birth-and-death processes to sample from models in stochastic geometry goes back to (Preston, 1977). These continuous-time Markov jump processes are spatio-temporal point or object processes. At distinct time instances the process undergoes a change. This change can be either a birth, that is an object is added to the current pattern, or a death, that is an object is deleted from the current point configuration. The stochastic dynamics of the spatial birth-death-process can be defined through a birth and a death rate. The birth rate is a measurable function \( b : \mathbb{R}^d \times \Omega \to [0, \infty) \) such that \( \int_B b(x, \xi) d\xi < \infty \), where \( B \) is a bounded Borel set on the state space \( \Omega \), the family of finite object patterns on \( \mathbb{R}^d \). Given the current configuration \( x \) at time \( t \), the probability of a birth in \( B \) during the short time interval \([t, t+s)\) is given by \( s \int_B b(x, \xi) d\xi + o(s) \).

The death rate is a measurable function \( d : \mathbb{R}^d \times \Omega \to [0, \infty) \). Given the current configuration \( x \cup \{\xi\} \) at time \( t \), the probability that \( \xi \) is deleted during a time interval \([t, t+s)\) is given by \( s d(x, \xi) + o(s) \). Conditions on \( b \) and \( d \) which ensure the existence and ergodicity of the spatial birth-and-death process...
can be found in (Preston, 1977). The rate of convergence of ergodic spatial birth-and-death processes is examined in (Lotwick and Silverman, 1981; Møller, 1989).

If the spatial birth-and-death process satisfies the following the detailed balance condition then its stationary distribution has density $f$:

$$f(x) b(x, \xi) = f(x \cup \{\xi\}) d(\xi, x)$$

(3)

Here $f(x \cup \{\xi\})$ is assumed to be positive. A possible choice for the birth rate and death rate is

$$b(x, \xi) = 1 \quad \text{and} \quad d(x, \xi) = \lambda^{-1}(x; \xi)$$

where $\lambda(\cdot, \cdot)$ is the Papangelou conditional intensity of the target process. However, the Papangelou conditional intensity usually has very sharp peaks. For example, Figure 2 shows the Papangelou conditional intensity $\lambda(\theta, \cdot)$ for the posterior model defined in (2) with parameters $\beta = 30, \gamma = 0, \sigma = 20$ and the observed image displayed in Figure 1.

As also observed by Ripley (1977), the roughness of the Papangelou conditional intensity leads to a large number of points being deleted shortly after they are born. To improve speed of convergence, Ripley suggested the following birth and death rate:

$$b(x, \xi) = \lambda(x, \xi) \quad \text{and} \quad d(x, \xi) = 1.$$
A birth-death-process with the above birth and death rate can be sampled by censoring the births in a space-time Poisson process with intensity $\lambda^*$, the uniform bound on the Papangelou conditional intensity of the target process. However, due to the sharp peaks of the Papangelou conditional intensity the bound $\lambda^*$ will tend to be very large and so the majority of objects born in the dominating Poisson process will be censored. Baddeley and van Lieshout (1993) suggested the following method for more efficient approximate sampling of the birth rate. First the global maximum of the birth rate $b_{\text{max}} = \max_{\xi \in S} b(x, \xi)$ is determined. Now birth proposals are sampled from the set

$$S_x = \{ \xi \in S : b(x, \xi) \geq \epsilon b_{\text{max}} \}, \quad 0 < \epsilon < 1.$$  

To reduce the cost of determining $b_{\text{max}}$ and $S_x$ the authors employ a multi-resolution search strategy. Computational cost reduces as $\epsilon$ tends to one. However, the resulting birth-death-process is an approximation to the desired birth-death-process and the larger $\epsilon$ the lesser the accuracy of the approximation.

### 2.3.2 Reversible jump MCMC

Geyer and Møller (1994) developed an alternative to the spatial birth-and-death process approach: a Metropolis-Hastings algorithm which samples a point process with a random number of points. The algorithm is an example of a reversible jump MCMC algorithm formally introduced in (Green, 1995).

Let $\pi(\cdot)$ be the target density and consider the state space $E = \{ x : \pi(x) > 0 \}$. We follow the presentation in (Møller, 1999) which only considers births and deaths as transitions. This is a special case of the algorithm in (Geyer and Møller, 1994) which also allows the replacement of a point as a transition. Let $X_n = x$ be the current point configuration. With probability $p(x)$ the birth of a point $\xi$, which is sampled according to a density $b(x, \cdot)$, is proposed; the birth is accepted with a probability $\min\{1, \alpha(x, \xi)\}$. Alternatively the death of a point $\eta \in x$, which is chosen randomly with probability $d(x \setminus \eta, \eta)$, is proposed and its death is accepted with probability $\min\{1, \alpha(x \setminus \eta, \eta)^{-1}\}$. The acceptance probabilities are specified by the Metropolis-Hastings ratio

$$\alpha(x, \xi) = \lambda(x, \xi) \frac{1 - p(x)}{p(x)} \frac{d(x, \xi)}{b(x, \xi)}.$$  

(4)

To ensure that the constructed chain is time-reversible the following conditions. need to hold

$$p(x) > 0, \quad p(x \cup \{\xi\}) < 1, \quad d(x, \xi) > 0, \quad b(x, \xi) > 0,$$

where $x \cup \{\xi\} \in E$. If furthermore $p(\emptyset) < 1$ then the chain is aperiodic, $\pi$-irreducible, where $\pi$ is the distribution of the point process with density $\pi(\cdot)$, and positive Harris recurrent. If the point process with density $\pi(\cdot)$ is locally stable and $\pi(\cdot)$ is hereditary then the algorithm is geometrically ergodic. If there is an $m \in \mathbb{N}$ such that $\pi(x) = 0$ whenever $x$ contains more than $m$ points, then the algorithm is uniformly ergodic.

A common choice is to propose a death or a birth with equal probability $1/2$. Both birth proposal and death proposals are uniformly distributed, that is

$$b(x, \xi) = 1_{[\xi \in W]} \quad \text{and} \quad d(x, \xi) = \frac{1}{(n(x) + 1)} 1_{[\xi \in x]}.$$
Recall that the sampling window $W$ is the unit square. The Metropolis-Hastings ratio $\alpha(x, \xi)$ is then given by

$$\alpha(x, \xi) = \frac{\lambda(x, \xi)}{n(x) + 1} \mathbb{1}_{\xi \in W}.$$  

Similar to the birth-death-process described earlier, this algorithm suffers from slow mixing if $\lambda(x, \xi)$ has sharp peaks. A large proportion of points $\xi$ sampled uniformly on $W$ will have small $\lambda(x, \xi)$ and so their birth will have a very low acceptance probability. For example, Figure 2 shows $\log \lambda(\emptyset, \xi)$ for the noisy image in Figure 1, clearly displaying some sharp peaks. Using uniform birth and death proposal distributions to infer objects in Figure 1, the average acceptance probability is in stationarity is as low as $0.005$.

The relation between spatial birth-death processes and reversible jump MCMC is explored in (Cappé et al., 2003). The results therein assume certain continuity conditions that are not satisfied in our context, but it seems likely that similar results could be derived for our case. Perfect simulation algorithms exist both for the spatial birth-death-process (Kendall and Théronnes, 1999) and the above Metropolis-Hastings algorithm (Kendall and Møller, 2000).

3 Proposals from the conditional posterior

3.1 Introduction

Suppose the current state of the Metropolis-Hastings chain $(X_t)_{t \geq 0}$ is $X_t = x = \{x_1, \ldots, x_n\}$. As suggested earlier we propose births or deaths with equal probability $1/2$. From the discussion in the previous section, to increase the average acceptance ratio we need to propose more points that lie in the peaks of the Papangelou conditional intensity. The most promising approach is to use the conditional posterior as proposal density for the location of a new object:

$$b(x, \xi) = \pi(x \cup \{\xi\}|x, y) = \frac{\pi(x \cup \{\xi\}|y)}{\int_W \pi(x \cup \{\xi\}|y)d\xi} \propto \lambda(x, \xi),$$

where $\lambda(x, \xi)$ is the Papangelou conditional intensity of the posterior model $\pi(x|y)$. Note that this approach is equivalent to the suggestion by Ripley (1977) in the context of birth-death-processes.

Let

$$C(x) = \int_W \lambda(x, \xi)d\xi$$

denote the inverse of the normalizing constant of the birth density $b(x, \xi)$. If objects proposed for deletion are chosen uniformly from the set of objects in the current configuration then the Metropolis-Hastings acceptance ratio $\alpha(x, \xi)$ is given by

$$\alpha(x, \xi) = \frac{C(x)}{n(x) + 1}.$$  

Notice that for any given current configuration $x$ the Metropolis-Hastings acceptance ratio is constant. In the terminology of (Brooks et al., 2003) which
will be introduced in Section 4, this algorithm is an infinite order conditional maximization approach.

As in the setting of spatial birth-death processes, sampling from the conditional posterior using rejection sampling with uniformly distributed proposals is computationally expensive when $\lambda(x, \xi)$ has sharp peaks. We address this issue in the next section.

Analogous to (Møller, 1999) quantitative bounds on the convergence rate can be obtained using the techniques developed in (Rosenthal, 1995). Unfortunately, these tend to be too weak to be of any practical use.

### 3.2 Approximate proposals from the conditional posterior

For our disk detection example we obtain the following expression assuming a Strauss process as prior model:

$$
\lambda(x, \xi) = \beta \gamma^{t(x, \xi)} \frac{L(x \cup \{\xi\}|y)}{L(x|y)}.
$$

where $t(x, \xi)$ is the number of neighbours of $\xi$ in $x$. Now consider the log-likelihood difference $l(x, \xi) = \log(L(x \cup \{\xi\}|y)) - \log(L(x|y))$. We have

$$
l(x, \xi) = -\frac{1}{2\sigma^2} \sum_{t \in A} \left( \left( y(t) - I_{x \cup \{\xi\}}(t) \right)^2 - \left( y(t) - I_x(t) \right)^2 \right)
$$

$$
= \frac{1}{\sigma^2} (g_1 - g_0) \left( \sum_{t \in R(\xi) \setminus R(x)} y(t) - \frac{(g_1 + g_0)}{2} |R(\xi) \setminus R(x)| \right),
$$

where $R(\xi) \setminus R(x)$ is the region covered by $R(\xi)$ but not by $R(x)$. Furthermore, $|R(\xi) \setminus R(x)|_\Lambda$ is the number of pixels in the region $R(\xi) \setminus R(x)$. As pointed out in (Baddeley and van Lieshout, 1993) the expression for the log-likelihood ratio is related to the Hough transform, a standard deterministic image processing tool. Algorithms for the efficient computation of discretizations of the Hough transform have been developed and can be found for example in (Illingworth and Kittler, 1987). Here we use similar techniques to compute a discretization of the Papangelou conditional intensity to devise a more efficient sampling procedure of the conditional posterior. We divide the sampling window into a grid of $m \times m$ equally sized square cells. For each cell $C_i$, $i = 1, \ldots, m^2$, we compute $\tilde{\lambda}(x, c_i)$ where $c_i$ is the centre point of the cell and

$$
\tilde{\lambda}(x, c_i) = \beta \gamma^{\tilde{t}(x, c_i)} \frac{L(x \cup \{c_i\}|y)}{L(x|y)}.
$$

Here $\tilde{t}(x, c_i)$ is the number of points in $x$ that are at a distance less than $R - \frac{1}{m\sqrt{2}}$ from $c_i$. Thus, if $\xi \in C_i$ then $t(x, \xi) \geq \tilde{t}(x, c_i)$. Now, let

$$(A \oplus b(0, R)) = \{x \in W : \min_{y \in A} |y - x| \leq R\}
$$

be the dilation of $A$ by a disk of radius $R$. Furthermore, set

$$
K = \max_{\xi \in C_i} |R(\xi) \setminus R(c_i)|_\Lambda
$$

$$
y_i = \max_{t \in C_i \oplus b(0, R)} y(t)
$$

$$
M_i = \max \left\{ 1, \exp \left( \frac{(g_1 - g_0)}{\sigma^2} K y_i \right) \right\},
$$

9
then
\[ \lambda(x, \xi) \leq \tilde{\lambda}(x, c_i)M_i \quad \text{for all } x \subseteq W, \xi \in C_i. \]

Let
\[ \tilde{C}(x) = \sum_{j=1}^{m^2} \tilde{\lambda}(x, c_j). \]

We can use \( \tilde{\lambda}(x, \cdot) \) in a rejection sampling scheme as follows. We choose cell \( C_i \) with probability \( \tilde{\lambda}(x, c_i)/\tilde{C}(x) \) and then sample \( \xi \) uniformly within \( C_i \). We then accept the proposed point with probability \( \lambda(x, \xi)/(M_i \tilde{\lambda}(x, c_i)) \).

Rather than using rejection sampling to sample from the conditional posterior and then perform a Metropolis-Hastings acceptance step, it is easier to use \( \tilde{\lambda}(x, \cdot) \) to propose a new object as described above and then simply continue immediately with a Metropolis-Hastings acceptance step. The Metropolis-Hastings ratio in this case is given by
\[ \alpha(x, \xi) = \frac{\lambda(x, \xi)}{(n(x) + 1)\tilde{\lambda}(x, c_i)} \]

where \( i \) is such that \( \xi \in C_i \). In equilibrium, this leads to an average acceptance probability of about 0.025, which is five-fold increase compared to the average acceptance probability using uniformly distributed birth proposals.

After a birth or death of a point \( \xi \) we need to compute \( \tilde{\lambda}(x \cup \{\xi\}, c_i) \) or \( \tilde{\lambda}(x \setminus \{\xi\}, c_i) \) respectively for \( i = 1, \ldots, m^2 \). However, in our setting we have
\[ \tilde{\lambda}(x, c_i) = \tilde{\lambda}(x \cup \{\xi\}, c_i) = \tilde{\lambda}(x \setminus \{\xi\}, c_i) \]
for all \( i \) such that \( R(\xi) \cap C_i = \emptyset \). Thus our approximation to the conditional posterior needs to be updated only for a small number of points.

4 The Brooks-Guidici-Roberts framework

In (Brooks et al., 2003) the authors introduce a general framework that gives guidance on how to choose proposal distributions in reversible jump MCMC algorithms. The suggestions are inspired by the properties of the Langevin algorithm. This algorithm exploits information on the gradient of the log target density and can be shown to have superior convergence properties compared to the commonly used random walk Metropolis-Hastings algorithm. Given the current state of the Markov chain is \( X_t = x \) the Langevin algorithm proposes a move to \( z \) which is sampled according to a Normal distribution of mean
\[ x + \frac{s^2}{2} \nabla \log \pi(x) \]
and variance \( s^2 \). Then the proposed state \( z \) is accepted or rejected according to the usual Metropolis-Hastings acceptance probability \( \min\{1, \alpha(x, z)\} \). Examination of the acceptance ratio \( \alpha(x, \xi) \) of the Langevin algorithm shows that not only there are states \( z \) such that \( \alpha(x, z) = 1 \) but more importantly that \( \frac{\partial \alpha(x, z)}{\partial z}|_{z=x} = 0 \).

The framework in (Brooks et al., 2003) aims to produce proposal distributions in reversible jump MCMC algorithms with similar properties to Langevin
proposals. The authors achieve this by first choosing specific points in the state space, so called centring points, at which the Metropolis-Hastings-Green acceptance ratio is examined. Two suggestions are made on how to choose centring points. They can be weak non-identifiability points, that is points at which the proposed sample point in the higher dimensional space is equivalent (in terms of the model defined) to the current state in the lower dimensional space. Alternatively, centring points can be modes of the conditional posterior. This centring method is called the conditional maximization approach. Given the centring point, the scaling, that is the spread of the proposal distribution, is chosen to satisfy certain criteria. The zeroth order approach chooses the scaling such that the Metropolis-Hastings-Green ratio at the centring point is equal to one. Higher order approaches choose the scaling such that higher order partial derivatives of the Metropolis-Hastings-Green ratio at the centring point become zero.

In the following we examine how conditional maximization centring may be used in our context. (Weak non-identifiability centring is not suitable for our problem.)

4.1 The zeroth order conditional maximization approach

In the following we describe how to implement a zeroth order conditional maximization approach in our context. The idea is to choose the spread of the proposal distribution such that the Metropolis-Hastings-Green acceptance ratio at the conditional posterior mode is equal to one.

Let \( X_t = x \) be the current state of our chain. As before we propose births and deaths with equal probability and sample death candidates uniformly from the existing configuration. As described in the previous chapter we compute a discrete approximation \( \tilde{\lambda}(x, \cdot) \) to \( \lambda(x, \cdot) \) and choose as our centring point

\[
\theta = \arg\max_{i=1, \ldots, m^2} \tilde{\lambda}(x, c_i).
\]

While \( \theta \) may not be identical to the global posterior mode, it will lie in the vicinity of a significant local mode of the conditional posterior density. To determine the centring point \( \theta \) we use a multiresolution approach as described in (Illingworth and Kittler, 1987). We first determine potential modes using an approximation on a coarse grid and then refine the grid in the vicinity of the potential mode to determine the location corresponding to the largest peak in \( \tilde{\lambda}(x, \cdot) \).

The zeroth order approach requires us to satisfy one equation and thus determines one parameter of the proposal distribution. To exploit this we choose the location of a new object according to an isotropic Gaussian distribution with mean \( \theta \) and covariance matrix \( sI_2 \) where \( I_2 \) denotes the identity matrix. Let \( \phi(\cdot) \) denote the univariate standard Normal density. Then the Metropolis-Hastings acceptance ratio is given by

\[
\alpha(x, \xi) = \frac{\lambda(x, \xi)}{\lambda(x, \theta)} \frac{1}{n(x) + 1} \frac{s^2}{\phi\left(\frac{\xi_1 - \theta_1}{s^2}\right) \phi\left(\frac{\xi_2 - \theta_2}{s^2}\right)}
\]

where \( \theta = (\theta_1, \theta_2), \xi = (\xi_1, \xi_2) \). Hence we achieve \( \alpha(x, \theta) = 1 \) by setting

\[
s^2 = \frac{(n(x) + 1)}{2\pi \lambda(x, \theta)}
\]

(5)
In the following we consider a higher order approach which allows us to use non-isotropic Gaussian proposal distributions.

4.2 The second order conditional maximization approach

As described earlier, to imitate the behaviour of the Langevin algorithm, the authors in (Brooks et al., 2003) suggest to choose the parameters of the proposal distribution such that not only the Metropolis-Hastings-Green acceptance ratio is one at the centring point, but also such that its (higher order) derivatives are equal to zero. In fact, the authors point out evidence to the fact that the flatness of the acceptance ratio at certain central moves is likely to be more important than having the acceptance ratio equal to one for the central move. Note that it is often computationally more convenient to set the derivatives of the logarithm of the acceptance ratio equal to zero.

The first order conditional maximization entails choosing the proposal density such that

\[ \alpha(x, \xi) = 1 \text{ and } \frac{\partial \alpha(x, \xi)}{\partial \xi_j} \bigg|_{\xi = \theta, \xi_j = 0} = 0, \text{ for } j = 1, 2. \]

However, in our context this would produce equations that are analytically intractable. Instead we set the first and second order partial derivatives of the log acceptance ratio at \( \theta \) equal to zero.

We have

\[ \log(\alpha(x, \xi)) = \log(\lambda(x, \xi)) - \log(n + 1) - \log(f(\xi|\theta, \Sigma)) \]

where \( f(\cdot|\theta, \Sigma) \) is the bivariate Normal density with mean \( \theta \) and covariance matrix \( \Sigma \). Thus the first and second order conditions are given by the equations

\[ \frac{\partial \log(\lambda(x, \xi))}{\partial \xi_j} \bigg|_{\xi = \theta} = \frac{\partial \log(f(\xi|\theta, \Sigma))}{\partial \xi_j} \bigg|_{\xi = \theta} = 0 \quad (6) \]

\[ \frac{\partial^2 \log(\lambda(x, \xi))}{\partial \xi_i \partial \xi_k} \bigg|_{\xi = \theta} = \frac{\partial^2 \log(f(\xi|\theta, \Sigma))}{\partial \xi_i \partial \xi_k} \bigg|_{\xi = \theta} \text{ for } i, k \in \{1, 2\}. \quad (7) \]

Note that these equations are satisfied by fitting a bivariate Normal distribution with mean \( \theta \) to the conditional posterior distribution.

However, as our posterior density is a step-function its derivatives are not defined. Instead we replace derivatives by differences and thus implicitly approximate the posterior density by a continuous function. We took a similar approach in (Bhalerao et al., 2001) where we replaced the likelihood function by an expression that was determined by a Normal mixture distribution. While this meant we were sampling an approximate posterior distribution it had the advantage of having analytical expressions for the likelihood function instead of pixelwise products, significantly reducing the amount of computation needed to determine the Metropolis-Hastings-Green acceptance probabilities.

In the problem setting here, we fit a Normal distribution to the posterior at \( \theta \) by setting

\[ \Sigma = \left( -L''(x, \theta) \right)^{-1} \]

where \( L'' \) is the curvature matrix of the log posterior density at the mode \( \theta \) which has entries:

\[ L''_{ij} = \left( \hat{\lambda}(x, \theta^* + \delta e_i + \delta e_j) - \hat{\lambda}(x, \theta^* - \delta e_i + \delta e_j) \right. \]

\[ - \left. \hat{\lambda}(x, \theta^* + \delta e_i - \delta e_j) + \hat{\lambda}(x, \theta^* - \delta e_i - \delta e_j) \right) / 4\delta^2 \quad i, j = 1, 2. \]
Here for $i = 1, 2$, the vector $e_i$ is the unit vector in direction of the $i$th coordinate. For computational convenience we choose $\delta = 1/m$.

## 5 A multi-centring approach

The methods discussed so far improve convergence by having higher acceptance probabilities than uniformly distributed proposals. However, this comes at a computational expense. Firstly, the approximate Papangelou conditional intensity needs updating after each change in the current configuration. Fortunately, our posterior distribution satisfies a spatial Markov property which means any changes are only local. Also, multi-resolution strategies can be used to reduce the computational cost but these require careful implementation. Nevertheless, for more complex problems this approach is likely to become computationally too expensive. Furthermore, both the zeroth and the second order approach sample from a uni-modal proposal density. This is in contrast to sampling from the conditional posterior which is multi-modal. In the following we consider the use of a proposal density which is a mixture of uni-modal distributions whose modes are centring points.

Locations $\xi \in W$ at which the data provides large evidence for an object will usually correspond to local modes in $\lambda(x, \cdot)$ where $x \subseteq W$ and $\xi \notin x$. Thus it is reasonable to use a subset of the local modes in $\lambda(\emptyset, \cdot)$ as centring points when determining the scale for the proposal distribution of births. More formally, for a Strauss process we have that

$$\lambda(x, \xi) = \lambda(\emptyset, \xi)$$

for any $\xi$ such that $R(\xi) \cap R(x) = \emptyset$.

We compute the approximation $\lambda(\emptyset, \cdot)$ of the Papangelou conditional intensity $\lambda(\emptyset, \cdot)$. As this computation is done only once we can use a fine grid. We then employ a mode-finding algorithm like the mean-shift algorithm (Comaniciu and Meer, 2002) to determine $K$ major, but well separated modes $\theta_1, \ldots, \theta_K$.

Our proposal density for births is now given by the mixture density

$$b(x, \xi) = \frac{1}{\epsilon + \sum_{j=1}^{K} p_j(x)} \left[ \epsilon 1_{[\xi \in W]} + \sum_{i=1}^{K} p_i(x) f(\xi | \theta_i, s) \right].$$

Here $f(\xi | \theta, s)$ denotes a bivariate Normal density with mean $\theta$ and covariance $sI_2$. The weight $p_i(x)$ of the $i$th Normal component is chosen to be equal to $\lambda(x, \theta_i)$. One mixture component is a Uniform density whose weight is proportional to $\epsilon$. This is needed to keep the proposal density well defined as $\sum_{j=1}^{K} p_j(x) = 0$ if $\theta_i \notin x$ for all $i \in \{1, \ldots, K\}$. The parameter $\epsilon$ should be chosen small compared to $\lambda(\emptyset, \theta_i)$ for $i \in \{1, \ldots, K\}$.

Let $\Theta(x) = \{\theta_i : R(\theta_i) \cap R(x) = \emptyset, i = 1, \ldots, K\}$. Unfortunately, it is not possible to achieve an acceptance ratio

$$\alpha(x, \theta_i) = 1 \quad \text{for all } \theta_i \in \Theta(x).$$

Instead we choose a scaling, that is $s$ such that $\alpha(x, \theta_i)$ is close to one for all $\theta_i \in \Theta(x)$. We exclude any $\theta_i \notin \Theta(x)$ from the scaling considerations as these usually do not correspond to major local modes in $\lambda(x, \cdot)$. In fact, for a Strauss hardcore process we have $\lambda(x, \theta_i) = 0$ for $\theta_i \notin \Theta(x)$. 

13
For $\theta_i \in \Theta(x)$ we have

$$\alpha(x, \theta_i) = \frac{\lambda(x, \theta_i)}{\lambda(x, \theta_i)} \frac{1}{b(x, \theta_i)}$$

$$= \frac{\lambda(x, \theta_i)}{(n(x) + 1)} \frac{\epsilon + \sum_{j=1}^{K} p_j(x)}{(\epsilon + \sum_{j=1}^{K} p_j(x)) (n(x) + 1)}$$

$$\approx \frac{\lambda(x, \theta_i)}{(n(x) + 1)} \frac{\epsilon + \sum_{j=1}^{K} \lambda(x, \theta_j)}{\lambda(x, \theta_i) f(\theta_i | \theta_i, s)}$$

$$= \frac{2\pi s^2 (\epsilon + \sum_{j=1}^{K} \lambda(x, \theta_j))}{(n(x) + 1)}.$$

Thus, if we set

$$s^2 = \frac{n(x) + 1}{2\pi (\epsilon + \sum_{j=1}^{K} \lambda(x, \theta_j))}$$

we have $\alpha(x, \theta_i) \approx 1$ for all $\theta_i \in \Theta(x)$.

Note that the above approach is somewhat similar to the data-driven paradigm developed by Tu and Zhu (2002). Their method applied to our context would use a proposal distribution that is a Normal mixture distribution whose components are centred at the local modes of the classic Hough Transform

$$H(\xi) = \sum_{t \in R(\xi)} y(t).$$

No advice is given on how best to choose the scaling of the mixture components. The “data-driven” approach works well when the modes of the likelihood function coincide with the modes in the posterior, for example when using uniform priors as in (Tu and Zhu, 2002). However, for more complex prior models this can be inefficient. In our example, the data-driven algorithm tends to propose objects close to the local modes of $\lambda(\theta_i, \cdot)$. However, once an object is accepted in the vicinity of a mode, any object proposed close to that mode is likely to be rejected. Because the proposal density lacks adaptation it continues to propose these objects that have a high probability of rejection leading to inefficiencies and lower acceptance probabilities in equilibrium.

6 Comparison of algorithms

In this section we compare the following proposal distributions $b(x, \cdot)$ for the location of an object proposed to be added:

1. the Uniform distribution described at the end of Section 2.3.2;
2. the approximate conditional posterior density described in Section 3.2;
3. the zeroth order conditional maximization proposal distribution described in Section 4.1;
4. the second order conditional maximization proposal distribution described in Section 4.2 and
5. the multi-centring approach described in Section 5.

To compute the approximate conditional posterior density we use a $64 \times 64$ grid. For the two conditional maximization approaches we use an adaptive Hough Transform approach. We first compute an approximation on a $32 \times 32$ grid and then refine the grid by a factor of 2 around the mode of the coarse grid. In the multi-centring approach we use a $128 \times 128$ grid to approximate the Papangelou conditional intensity $\lambda(\emptyset, \cdot)$. We then determine a set of 25 major local modes $\Theta = \{\theta_1, \ldots, \theta_{25}\}$ using the mean-shift algorithm. Figure 3 shows $\log(\lambda(\Theta, \cdot))$ which is relatively flat in comparison to $\lambda(\emptyset, \cdot)$ displayed in Figure 2. All algorithms were started in the empty configuration and run for 100000 iterations. Figure 4 shows a typical sample from the posterior distribution overlayed onto the noisy image.

![Figure 3: The Papangelou conditional intensity $\lambda(\Theta, \cdot)$ for the posterior model defined in (2) with parameters $\beta = 30, \sigma = 20$ and the observed image displayed in Figure 1. The configuration $\Theta$ consists of 25 well-separated major local modes in $\lambda(\emptyset, \cdot)$.](image)

In Table 6 below we summarize some quantitative characteristics of the different MCMC algorithms. These are

1. the average acceptance rate $\alpha_{\text{initial}}$ in the initial 1000 iterations;
2. the average acceptance rate $\alpha_{\text{equi}}$ in the last 50000 iterations;
3. the autocorrelation at lag 1000;
4. the runtime of the algorithm in seconds on a laptop with a 2.00 GHz processor;
Table 1: Quantitative characteristics of the five algorithms.

<table>
<thead>
<tr>
<th>proposal distribution</th>
<th>$\alpha_{\text{initial}}$</th>
<th>$\alpha_{\text{equi}}$</th>
<th>ACF</th>
<th>runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform</td>
<td>0.02</td>
<td>0.005</td>
<td>0.46</td>
<td>600</td>
</tr>
<tr>
<td>approximate conditional posterior</td>
<td>0.08</td>
<td>0.025</td>
<td>0.12</td>
<td>1694</td>
</tr>
<tr>
<td>0th order conditional maximization</td>
<td>0.07</td>
<td>0.100</td>
<td>0.44</td>
<td>1612</td>
</tr>
<tr>
<td>2nd order conditional maximization</td>
<td>0.06</td>
<td>0.026</td>
<td>0.05</td>
<td>1652</td>
</tr>
<tr>
<td>multi-centring approach</td>
<td>0.07</td>
<td>0.040</td>
<td>0.36</td>
<td>726</td>
</tr>
</tbody>
</table>

The algorithm that proposes new disks locations uniformly on the sampling window has the longest transient phase needing over 5000 iterations to get to a configuration that accounts for the disks that have strong support by the data. It is also the algorithm with the lowest acceptance rates, both initially and in equilibrium. Figure 5 shows the trace plot of the number of disks in the final 10000 iterations and illustrates the behaviour of this algorithm in stationarity. Once in a while it adds a disk, but this disk only survives a relatively short time before being deleted. Figure 5 also shows the autocorrelation function computed over the last 50000 iterations. It shows a reasonably fast decay due to the fact that a new disks is added from time to time and then removed very quickly. We also plot the deviance $-2 \log(\pi(x|y)) + 2 \log(\alpha)$ of the unnormalized posterior density. Compared to the algorithm using the approximate conditional posterior the deviance is much higher for Uniformly distributed proposals. Clearly, the algorithm has not yet reached the region around the posterior mode that is visited by the other four algorithms. Computationally this algorithm is the cheapest of the five methods considered.
When using the approximate conditional posterior to make proposals we achieve a very short transient phase of approximately 50 iterations. The acceptance rate in equilibrium is about five times higher than the one for the algorithm with uniform proposals, see also the trace of the number of points in the last 10000 iterations shown in Figure 6. In contrast to the algorithm with uniform proposals, disk are always proposed at locations with relatively large support by the data. While the algorithm moves quickly to high posterior density regions it explores much less the tails of the posterior distribution which is apparent from the deviance plot. The autocorrelation function decays much slower than for uniform proposals but is smaller at short lags, see Figure 6. Computationally this is the most expensive algorithm.

Like the previous algorithm the zeroth order conditional maximization algorithm has a very short transient phase of approximately 50 iterations. During this phase the scale is very small with $s \approx 3.26 \times 10^{-6}$. However, once in a configuration with high posterior density values, the standard deviation increases to about 0.16. The zeroth order conditional maximization approach has the highest acceptance rate, however the autocorrelation function, see Figure 7, decays slightly slower than for the approximate conditional posterior method. The reason for the large acceptance rate is that the scaling of birth proposals is chosen such as to achieve an acceptance ratio of one around the conditional posterior mode. While the algorithm makes proposals centred at the conditional mode as in the previous approach, these have a higher spread than the ones made by the approximate conditional posterior. This also overcomes problems that are due to the discretization of the conditional posterior density. The plot of the deviance in Figure 7 shows that the algorithm has the largest variation in the deviance, however, the deviance is generally larger than for the algorithm using the approximate conditional posterior or the one using the multi-centring approach. With respect to run-time, despite using an adaptive approach, this algorithm is similar to the approximate conditional posterior approach.

The transient phase is also very short for the second order conditional maximization approach at about 50 iterations. Here the scaling is determined by the second order properties of the conditional posterior. As can be seen from Figures 2 and 3 the conditional posterior has very sharp peaks which leads to the proposal distribution being highly concentrated, with standard deviations being on average as small as $7.36 \times 10^{-9}$. The algorithm alternately adds and deletes disks at about the same four locations. Even worse, due to discretization effects, it does not discover one disk with reasonably strong support by the data! The run-time is similar to the previous two approaches. Below, Figure 8 shows the trace plot and autocorrelation function of the number of points and the deviance.

The multi-centring approach also has a short transient phase at about 50 iterations. Its run-time is similar to the approach using uniform proposals, but of course, it requires on some computational effort prior to simulation in order to determine the local modes in $\lambda(\emptyset, \cdot)$. The algorithm has a relatively high acceptance rate. If a disk that has strong support by the data is removed, it is re-introduced reasonably quickly. On average, the deviance is lower than in the two conditional maximisation approaches but slightly higher than when using the approximate conditional posterior, see Figure 9. The figure also shows the trace and autocorrelation function of the number of disks.
7 Summary and Conclusions

One of the major obstacles in using Markov object processes for Bayesian object recognition has been the very low acceptance rates when sampling the corresponding posterior models using reversible jump MCMC. The traditional approach of using uniform proposals for object locations leads to a long transient phase and thus long run-times. A much more promising method is to use the conditional posterior as a proposal distribution. For our simple test problem, an approximation to the conditional posterior can be computed and used in a reversible jump algorithm. While this leads to a higher acceptance rate and a much shorter transient phase, it is computationally expensive.

We developed a “hybrid” method between the above two approaches by utilizing a mixture density whose components are centred around modes in the Papangelou conditional intensity $\lambda(\emptyset, \cdot)$. This produces higher acceptance rates and a shorter transient phase than uniformly distributed proposals. The runtime of the algorithm is low compared to the alternatives.

An open question is to how to choose the scale of the components of the mixture distribution. We used a similar approach to (Brooks et al., 2003) which ensures that the acceptance probability of objects with strong support by the data and that are not in the current configuration is close to one. Our simulation results suggest that this is a satisfactory approach.

We also considered two suggestions made in (Brooks et al., 2003): the zeroth order and the second order conditional maximization. These lead to a short transient phase but the computational cost is high. For the second order approach simulation results suggest that only a very limited number of object configurations are visited by the chain. This could be alleviated by introducing proposals that move objects. However, this would provide no answers as to how to choose proposals that change the number of objects. One reason for this undesirable behaviour is the fact that the scaling is chosen according to the properties of the acceptance probability at one single location. In our example, the acceptance probabilities have very sharp peaks, so that such a local approach is doomed to give misleading results. Our multi-centring approach circumvents this problem by considering the acceptance probability at various locations that have been chosen based on an a priori examination of the likelihood function.

References


O. Cappé, C. Robert, and T. Rydén. Reversible jump, birth-and-death and


Figure 5: The trace (above) and the autocorrelation function (middle) of the the number of disks using a uniform birth density. The bottom plot displays the deviance of the unnormalized posterior density.
Figure 6: The trace (above) and the autocorrelation function (middle) of the number of disks using the approximate conditional posterior as proposal distribution. The bottom plot displays the deviance of the unnormalized posterior density.
### Figure 7

The trace (above) and autocorrelation function (middle) of the number of disks using the zeroth order conditional maximization approach. The bottom plot displays the deviance of the unnormalized posterior density.
Figure 8: The trace (above) and autocorrelation function (middle) of the number of disks using the second order conditional maximization approach. The bottom plot displays the deviance of the unnormalized posterior density.
Figure 9: The trace (above) and autocorrelation function (middle) of the number of disks using the multi-centring approach. The bottom plot displays the deviance of the unnormalized posterior density.