Exact simulation and unbiased estimation for diffusions

Gareth Roberts

University of Warwick
Chicheley Hall, July 2014

Including work with Alex Beskos, Paul Fearnhead, Krys Latuszynski, Omiros Papaspiliopoulos, and Giorgos Sermaidis

Diffusions

A d-dimensional diffusion is a continuous-time strong Markov process with continuous sample paths. We can define a diffusion as the solution of the Stochastic Differential Equation (SDE):

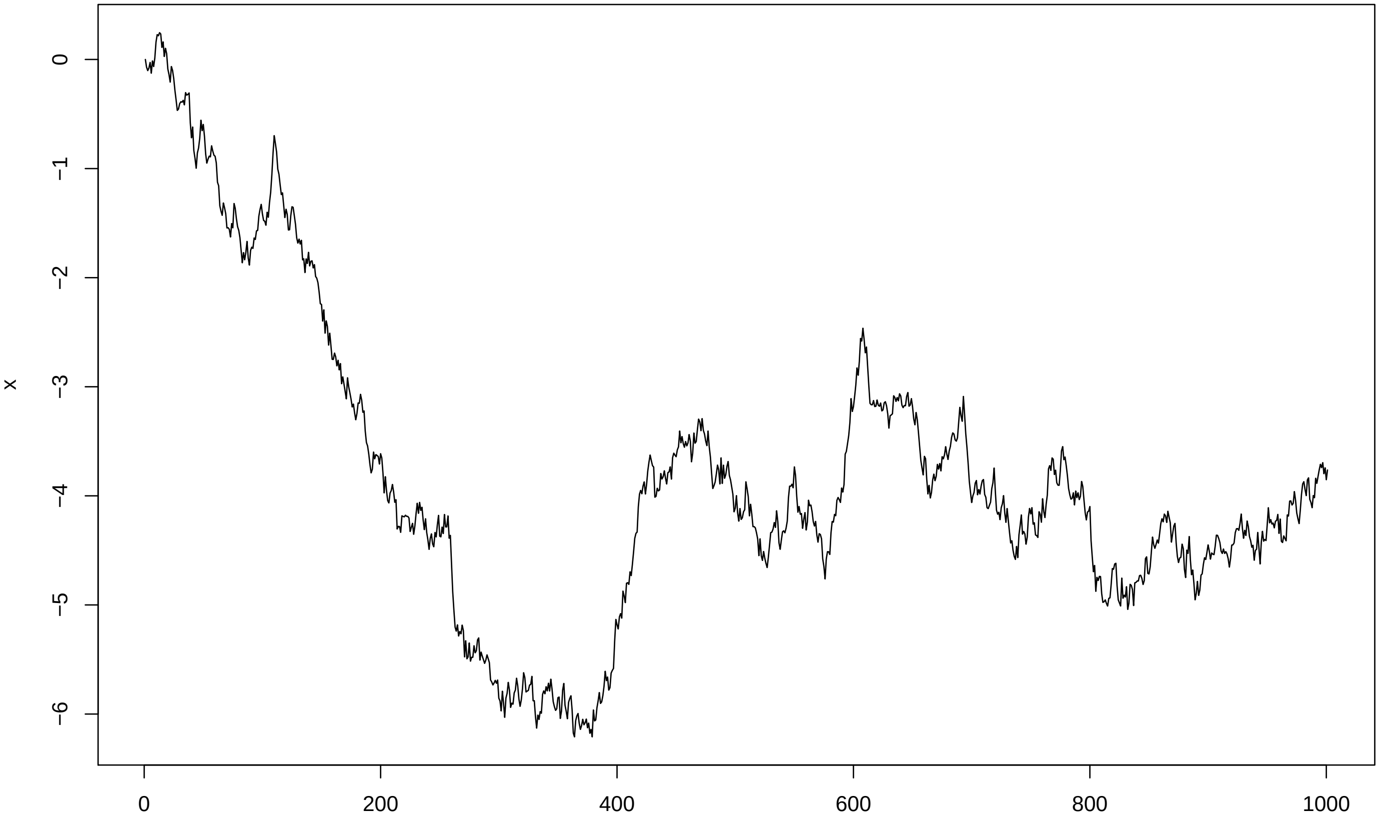
$$\mathrm{d}X_t = \mu(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t.$$

where B denotes d-dimensional Brownian motion, σ is a $d \times d$ matrix and μ is a d-vector.

Often understood intuitively and constructively via its dynamics over small time intervals. Approximately for small h:

$$X_{t+h}|X_t = x_t \quad \sim \quad x_t + h\mu(x_t) + h^{1/2}\sigma(x_t)Z$$

where Z is a d-dimensional standard normal random variable.


x
0
−1
−2
−3
−4
−5
−6
0
200
400
600
800
1000

Transition Densities

We will denote the transition density of the diffusion by

$$p(y|x,h)dy = p(X_{t+h} \in dy | X_t = x)$$

.

It satisfies Kolmogorov's forward equation:

$$\frac{\partial}{\partial t} p(y|x,t) = \mathcal{K}_y p(y|x,t),$$

for some forward-operator $\mathcal{K}_y$ which acts on y.

Generally the transition density is intractable with the usual exceptions: constant or linear drifts and a few others ...

Rejection sampling

Let f be a density of interest, and g be a density from which we can simulate. f/g bounded by K say.

1. Sample X from g.
2. Compute $p(X) = f(X)/(Kg(X))$.
3. Simulate $U \sim U(0,1)$.
4. Accept X if $p(X) > U$. Otherwise return to 1.

Blue steps are often unnecessary!

Retrospective rejection sampling

1. Sample $V \sim U(0,1)$.
2. Identify a function $h(V,X)$ and a set $A(V)$ such that
$$\mathbf{P}_V\{h(V,X) \in A(V)\} = p(X)$$
3. Simulate $h(X,V)$.
4. If $h(X,V) \in A(V)$ the accept. Otherwise return to 1.
5. Fill in missing bits of X from distribution of $X|h(X,V)$ as required.

Simulation of stochastic processes

Suppose that $X : [0,1] \to \mathbf{R}^d$ is a stochastic process with associated probabiltiy measure $\mathbf{P}_0$.

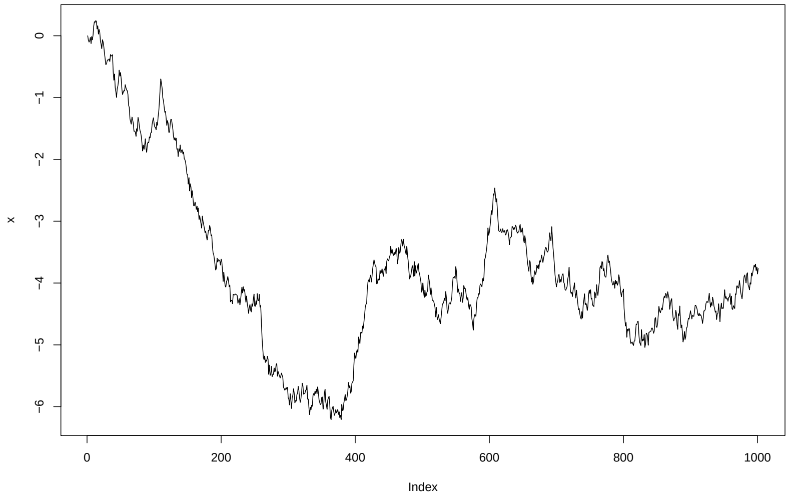
Suppose we are able to simulate from $\mathbf{P}_0$.

Suppose that we wish to simulate from a different distribution $\mathbf{P}$ which cannot be directly simulated, but for which we can write:

$$\frac{d\mathbf{P}}{d\mathbf{P}_0}(X) \propto \exp\{-r \int_0^1 \phi(X_s)ds\} = a(X)$$

for some function ϕ taking values in $[0,1]$.

This applies to very wide range of stochastic processes, eg point processes in space and time, diffusions, jump diffusions, processes used in Bayesian non-parametrics.



For example, given this trajectory, $a(X)$ describes the Radon-Nikodym derivative between $\mathbf{P}$ and $\mathbf{P}_0$ for this particular trajectory.

Rejection for sample paths

Would like to just propose a sample path fom $\mathbf{P}_0$ and use rejection sampling. However

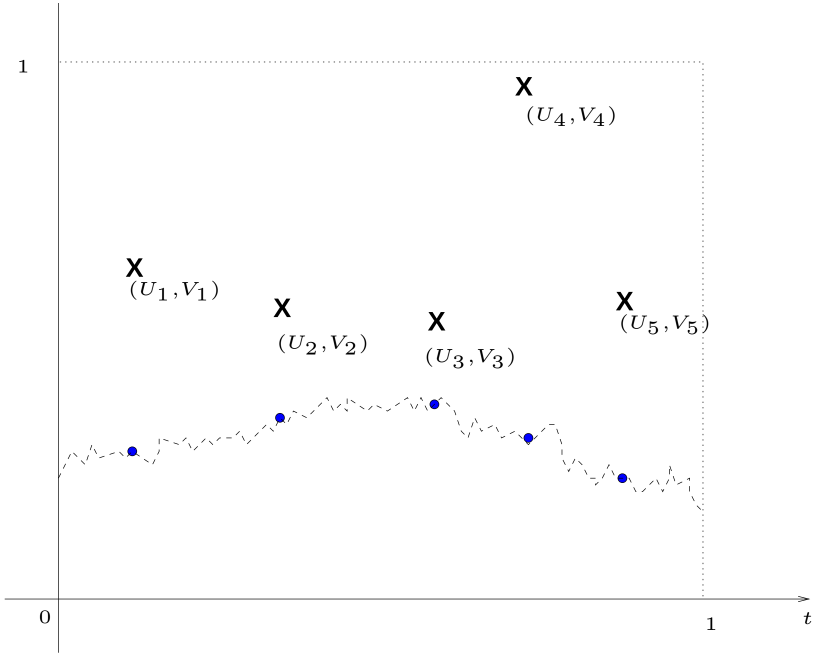
- Just storing all of X could require infinite storage capacity.
- Calculating $\int_0^1 \phi(X_s)ds$ is likely to require infinite computation

We could approximate in some way, but this seems unsatisfactory, and it would typically be very difficult to quantify the resulting approximation error.

Retrospective rejection simulation

Key observation: $a(x)$ is the probability of a Poisson random variable of parameter $r\int_0^1 \phi(X_s)ds$ taking value 0.

Or ... the probability that a Poisson process of rate r on the unit square has no points on the epi graph $\{(u,v) \in [0,1]^2;\ v \le \phi(u)\}$.



Simulation of diffusions

Continuous, strong Markov processes described by stochastic differential equation:

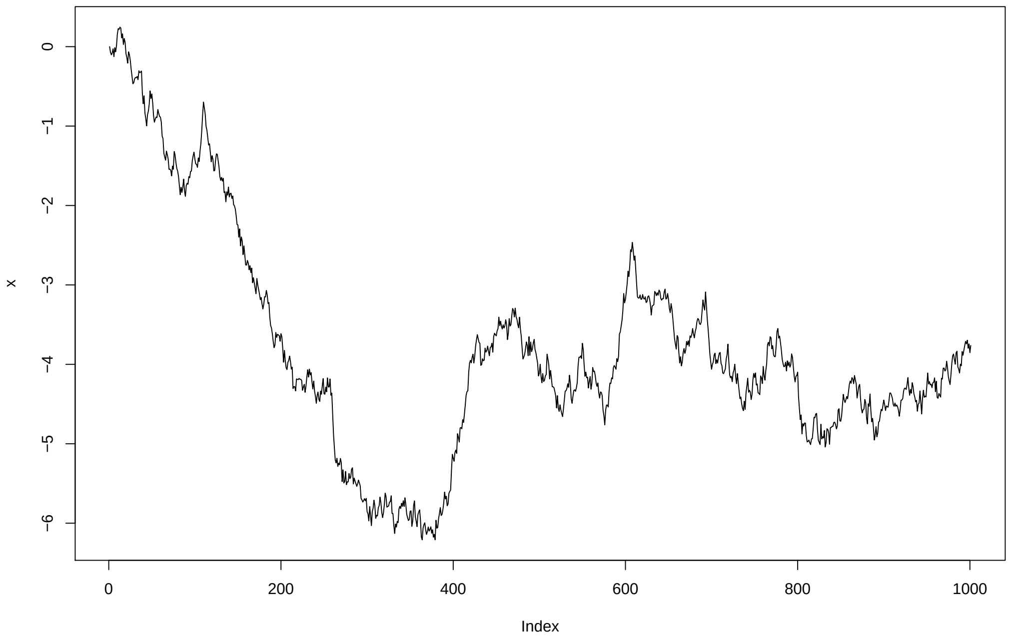
$$dX_t = \alpha(X_t)dt + \sigma(X_t)dB_t$$

where B is standard Brownian motion.

This can be interpreted constructively as

$$X_{t+\epsilon} = X_t + \epsilon\alpha(X_t) + \sigma(X_t)N(0, \epsilon)$$

approximately for ‘small’ ϵ (the **Euler approximation**) written as



Interested in simulating without discretisation error and obtaining a realisation of the whole path in some sense.

Diffusion densities

Consider simplest case, σ constant and drift α which is bounded with bounded derivative.

$$dX_t = \alpha(X_t)dt + dB_t$$

and let the law of this diffusion on $[0,1]$ be denoted $\mathbf{P}$, with $\mathbf{W}_0$ being that of the Brownian motion (Wiener measure) .

Then under very weak regularity conditions

$$\frac{d\mathbf{P}}{d\mathbf{W}}(X) = G(X)$$

where G is given by the **Cameron-Martin-Girsanov** formula:

$$\log G(X) = \int_0^1 \big(\alpha(X_s)dX_s - \alpha^2(X_s)/2\big)\, ds$$

Towards a simulation algorithm: simplifying G

By a suitable rearrangement we can rewrite

$$\frac{d\mathbf{P}}{d\mathbf{W}}(X) = G(X) \propto \exp\left\{ A(X_1) - r\int_0^1 \phi(X_s)ds \right\} := a(X)$$

where ϕ always always takes values in the interval $[0, 1]$.

This is almost in the exponential form required for the Poisson process idea above.

So we consider biased Brownian motion proposals for rejection sampling:

$$\mathbf{P}_0(X_1 \in dx) \propto \exp\{A(x) - x^2/2\}\ dx \qquad (*)$$

with $\mathbf{X}|X_1$ ~Brownian bridge, so that

$$\frac{d\mathbf{P}}{d\mathbf{P}_0} \propto \exp\left\{-r\int_{s=0}^{1} \phi(X_s)ds\right\}.$$

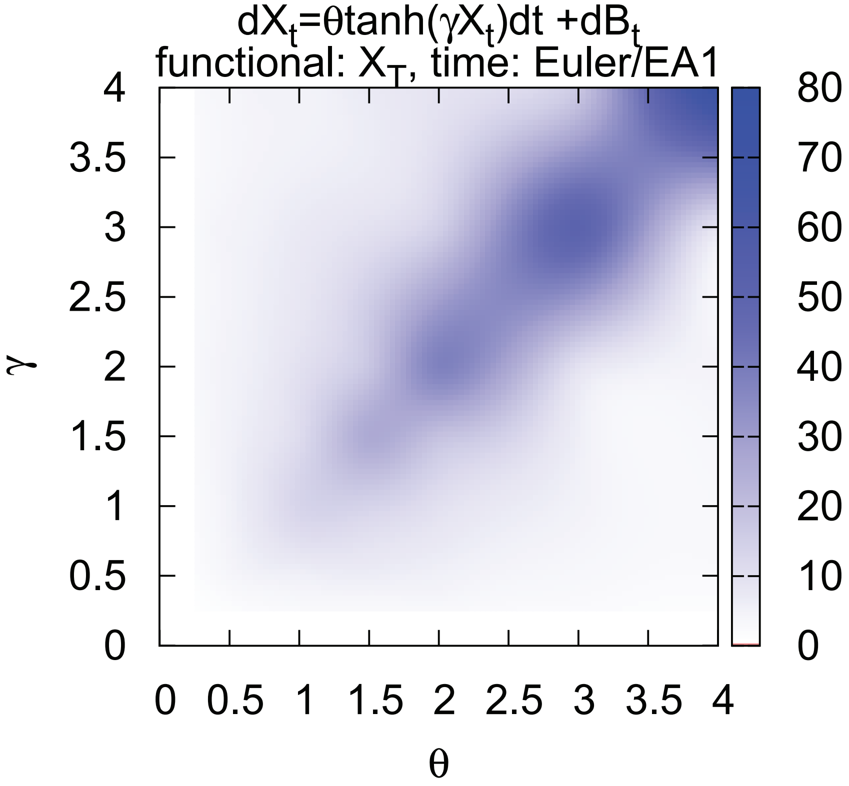
Let Φ be a Poisson process of rate r on $\{0 \leq y \leq \phi(X_s),\ 0 \leq s \leq 1\}$. Then

$$\mathbf{P}\left(\Phi \text{ is the empty configeration} = \exp\left\{-r\int_0^1 \phi(X_s)ds\right\}\right).$$

The basic algorithm (EA1)

1. Set $B_0 = 0$. Simulate B_1 from (*)
2. Generate Poisson process of rate r on $[0,1]\times[0,1]$: $\Phi = \{(U_1, V_1), \ldots (U_n,$
3. For each U_i, draw B_{U_i} from its appropriate Brownian bridge probabilit
4. If $\phi(B_{U_i}) > V_i$ for ANY i, erase skeleton and go to (1).
5. Output the currently stored skeleton $\{(0, B_0), (1, B_1), (U_i, B_{U_i}),\ 1 \leq i \leq n\}$.

Part of a simulation study



The Exact Algorithm for multi-dimensional diffusions

Generally simulation and inference for diffusions is performed by approximating the diffusions by a discrete-time Markov process.

For multi-dimensional diffusions, we can adopt the exact algorithm if:

- The volatility can be transformed to be constant via the Lamperti transform: ie we can find a 1-1 function η satisfying the matrix valued differential equation

$$\nabla \eta \sigma = I_d$$

- The drift of the transformed diffusion is the gradient of a potential: $\mu(x) = \nabla A(x)$.

This can be applied to almost all 1-d diffusions for which CMG theorem holds, but only certain classes of d-dimensional ones.

Why?

The exact Algorithm is a Rejection Sampler based on proposing paths from a driftless version of the diffusion (with same volatility).

The acceptance probability for the path is (for $\sigma(x) = I_d$) proportional to:

$$\begin{aligned} &\exp\left\{\int_0^T \mu(X_t)\mathrm{d}X_t - \frac{1}{2}\int_0^T |\mu(X_t)|^2\mathrm{d}t\right\} \\ &= \exp\left\{A(X_T) - A(X_0) - \frac{1}{2}\int_0^T \left(|\mu(X_t)|^2 + \nabla\mu(X_t)\right)\mathrm{d}t\right\}. \end{aligned}$$

Whilst this cannot be evaluated, events with this probability can be simulated.

The condition $\mu(x) = \nabla A(x)$ is required to replace the stochastic integral by a Lebesgue one. It is a necessary and sufficient condition for Girsanov's formula to be bounded for bounded sample paths.

The condition $\sigma(x)$ is constant is so that we can simulate from the driftless diffusion.

- Importance sampling seems doomed if we cannot sample from an distribution wrt which target is absolutely continuous.

Consider two diffusions with different diffusion coefficients, σ_1 and σ_2, then their laws as NOT mutually absolutely continuous ...

even though their finite-dimensional distributions typically are.

Avoiding time-discretisation Errors: Why?

Beskos, Papaspiliopoulos, Roberts and Fearnhead (2006) extend the rejection sampler to an importance sampler, and show how this can used to perform inference for diffusions which avoids time-discretisation approximations.

Why may these methods be useful?

- Error in estimates are purely Monte Carlo. Thus it is easier to quantify the error.
- Time-discretisation may tend to use substantially finer discretisations than are necessary: possible computational gains?
- Want methods which are robust as $h \to 0$
- Error is $O(C^{-1/2})$, where C is CPU cost. Alternative approaches have errors that can be e.g. $O(C^{-1/3})$ or worse (though see multigrid work by Giles).

Generalising the exact algorithm using importance sampling

Our aim was to try and extend the ability to perform simulation / estimation / inference without time-discretisation approximations to a wider class of diffusions.

The key is to be able to unbiasedly estimate expectations, such as $\mathrm{E}(f(X_t))$ or $\mathrm{E}(f(X_{t_1}, \ldots, X_{t_m}))$.

The approach we have developed can be applied to general continuous-time Markov processes, and is a continuous-time version of sequential importance sampling.

We construct a signed measure-valued stochastic processes (which is non-Markov) $\{\xi_t, t \geq 0\}$ with

$$\mathrm{E}(\xi_t(f)) = \mathrm{E}(f(X_t))$$

Unbiased estimation almost as good as simulation given recent advances in combining importance sampling and simulation algorithms such as MCMC.

Importance Sampling

Importance Sampling (**IS**) is a Monte Carlo integration technique. Consider the integral

$$I = \int f(x)p(x)\mathrm{d}x = \int \frac{h(x)}{q(x)}q(x)\mathrm{d}x,$$

where $p(x)$ and $q(x)$ are densities, $f(x)$ is arbitrary and $p(x) > 0 \Rightarrow q(x) > 0$. Here we are setting $h(x) = f(x)p(x)$.

We can view this as an expectation with respect to $q(x)$. Thus

1. Sample x_i, $i = 1, \ldots, N$, iid from $q(x)$;
2. Estimate the integral by the unbiased, consistent estimator:

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N}\frac{h(x_i)}{q(x_i)}.$$

Sequential Importance Sampling (SIS)

As this gives an estimate of the expectation of $f(X)$ for arbitrary functions f, we can think of the sample from $q(x)$, and the corresponding weights as giving an approximation to the distribution defined by $p(x)$.

This idea can be extended to Markov processes:

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i \mid x_{i-1}).$$

With a proposal process defined by $q(x_1)$ and $q(x_i \mid x_{i-1})$.

Sequential Importance Sampling (SIS)

To obtain one weighted sample:

1. Simulate $X_1^{(i)}$ from $q(x_1)$; assign a weight $\tilde{w}_1^{(i)} = p(x_1)/q(x_1)$.
2. For $t = 2, \ldots, n$; simulate $X_t^{(i)} | x_{t-1}^{(i)}$ from $q(x_t | x_{t-1}^{(i)})$, and set

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-1}^{(i)} \frac{p(x_t^{(i)} | x_{t-1}^{(i)})}{q(x_t^{(i)} | x_{t-1}^{(i)})}.$$

New Approach: CIS

We now derive a continuous-time importance sampling (CIS) procedure for unbiased inference for general continuous-time Markov models.

We will describe the CIS algorithm for generating a single realisation. So at any time t we will have x_t and w_t, realisations of random variables X_t, W_t such that

$$\mathrm{E}_p(f(X_t)) = \mathrm{E}_q(f(X_t)W_t).$$

The former expectation is wrt to the target diffusion, the latter wrt to CIS procedure.

We will use a proposal process with tractable transition density $q(x|y,t)$ (and forward-operator $\mathcal{K}_x^{(1)}$).

A discrete-time SIS procedure

First consider a discrete-time SIS method aimed at inference at times $h, 2h, 3h, \ldots,$.

(0) Fix x_0; set $w_0 = 1$, and $i = 1$.

(1) Simulate $X_{ih} = x_{ih}$ from $q(x_{ih}|x_{(i-1)h})$.

(2) Set

$$w_i = w_{i-1}\frac{p(x_{ih}|x_{(i-1)h}, h)}{q(x_{ih}|x_{(i-1)h}, h)}$$

(3) Let $i = i + 1$ and goto (1).

Problems: cannot calculate weights, and often the efficiency degenerates as $h \to 0$ for fixed T.

As $h \to 0$, where q and p are discetisations of absolutely continuous diffusions, the limit is given by Girsanov's formula.

We want it to work in the case where q and p are mutually singular also!

Random weight SIS

It is valid to replace the weight in the SIS procedure by a random variable whose expectation is equal to the weight.

A simple way to do this here is to define

$$r(y,x,h) = 1 + \left(\frac{p(y|x,h)}{q(y|x,h)} - 1 \right) \frac{1}{\lambda h},$$

and introduce a Bernoulli random variable U_i, with success probability λh.

Then

$$\frac{p(y|x,h)}{q(y|x,h)} = \mathrm{E}\left\{(1-U_i)\cdot 1 + U_i r(y,x,h)\right\}.$$

Random weight SIS

Now we can have a random weight SIS algorithm:

(0) Fix x_0; set $w_0 = 1$, and $i = 1$.

(1) Simulate $X_{ih} = x_{ih}$ from $q(x_{ih}|x_{(i-1)h})$.

(2) Simulate U_i. If $U_i = 1$ then set $w_i = w_{i-1} r(x_{ih}, x_{(i-1)h}, h)$, otherwise $w_i = w_{i-1}$.

(3) Let $i = i + 1$ and goto (1).

This is a less efficient algorithm than the previous one, but it enables us to now use two tricks: retrospective sampling and Rao-Blackwelisation.

Retrospective Sampling

We only need to update the weights at time-points where $U_i = 1$. At these points we need to simulate $X_{ih}, X_{(i-1)h}$ to calculate the new weights.

If j is the most recent time when $U_j = 1$, then the distribution of X_{ih} is given by $q(x_{ih}|x_{jh}, (i-j)h)$ (assuming time-homogeneity for simplicity).

Given x_{jh} and x_{ih} the conditional distribution of $X_{(i-1)h}$ is

$$q(x_{(i-1)h}|x_{jh}, x_{ih}) = \frac{q(x_{(i-1)h}|x_{jh}, (i-j-1)h)q(x_{ih}|x_{(i-1)h}, h)}{q(x_{ih}|x_{jh}, (i-j)h)}.$$

New SIS algorithm

Using these ideas we get:

(0) Fix x_0; set $w_0 = 1$, $j = 0$ and $i = 1$.

(1) Simulate U_i; if $U_i = 0$ goto (3).

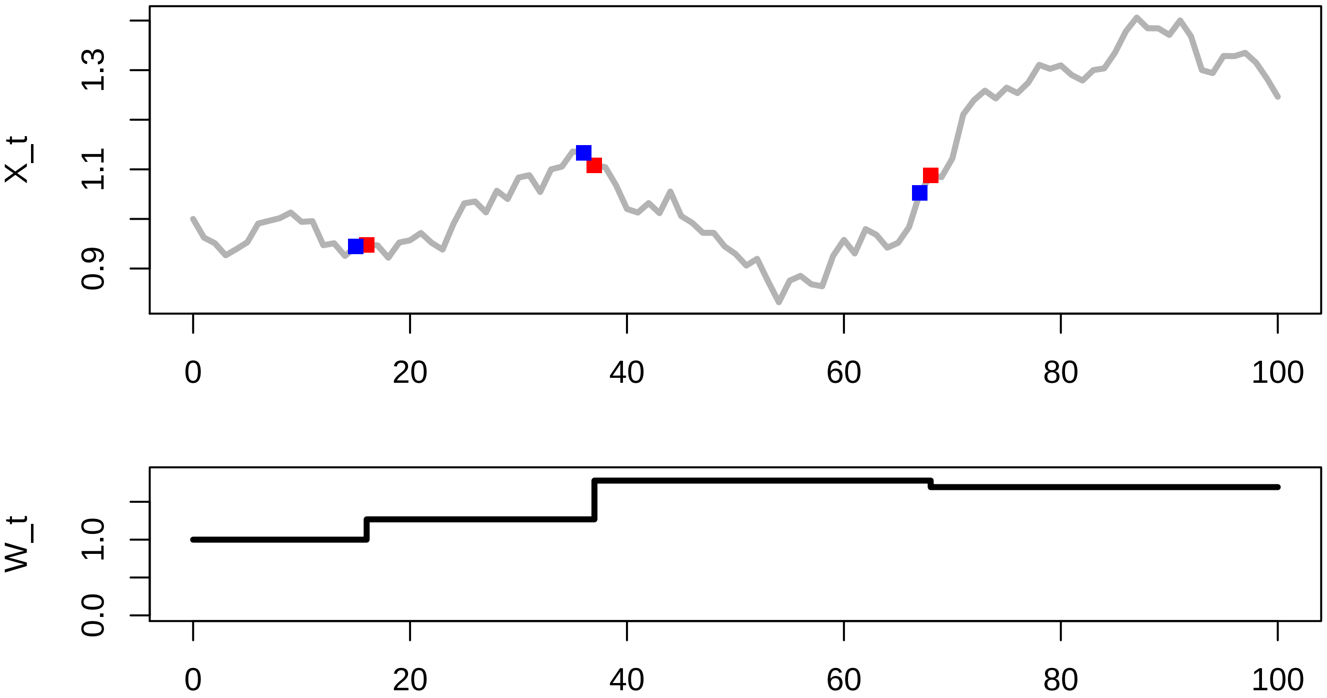
(2) $[U_i = 1]$ Simulate X_{ih} from $q(x_{ih}|x_{jh}, (i-j)h)$ and $X_{(i-1)h}$ from $q(x_{(i-1)h}|x_{jh}, x_{ih})$. Set

$$w_i = w_j r(x_{ih}, x_{(i-1)h}, h).$$

(3) Let $i = i + 1$ and goto (1).

If we stop the SIS at a time point t, then X_t can be drawn from $q(x_t|x_{jh}, t - jh)$; and the weight is w_j.

Example



Rao-Blackwellisation

At time ih, the incremental weight depends on x_{ih} and $x_{(i-1)h}$. Rather than simulating both we simulate x_{ih}, and use an expected incremental weight

$$\rho_h(x_{ih}, x_{jh}, (j-i)h) = \mathrm{E}\left(r(x_{ih}, X_{(i-1)h}, h) \mid x_{jh}\right),$$

with expectation with respect to the conditional distribution of $X_{(i-1)h}$ given x_{jh}, x_{ih} under the proposal:

$$\mathrm{E}\left(r(x_{ih}, X_{(i-1)h}, h) \mid x_{jh}\right) = \int r(x_{ih}, x_{(i-1)h}, h) q(x_{(i-1)h} | x_{jh}, x_{ih}) \mathrm{d}x_{(i-1)h}.$$

New SIS algorithm

Using these ideas we get:

(0) Fix x_0; set $w_0 = 1$, $j = 0$ and $i = 1$.

(1) Simulate U_i; if $U_i = 0$ goto (3).

(2) $[U_i = 1]$ Simulate X_{ih} from $q(x_{ih}|x_{jh},(i-j)h)$ and set

$$w_i = w_j \rho_h(x_{ih}, x_{jh}, (i-j)h).$$

(3) Let $i = i + 1$ and goto (1).

If we stop the SIS at a time point t, then X_t can be drawn from $q(x_t|x_{jh}, t - jh)$; and the weight is w_j.

Continuous-time SIS

The previous algorithm cannot be implemented as we do not know $p(\cdot|\cdot, h)$. However, if we consider $h \to 0$ we obtain a continuous-time algorithm that can be implemented.

The Bernoulli process converges to a Poisson-process.

In the limit as $h \to 0$, if we fix $t = ih$ and $s = jh$ we get

$$\rho(x_t, x_s, t-s) = \lim_{h \to 0} \rho_h(x_t, x_s, t-s) = 1 + \frac{1}{\lambda} \left(\frac{(\mathcal{K}_x - \mathcal{K}_x^{(1)}) q(x|x_s, t-s)}{q(x|x_s, t-s)} \right) \bigg|_{x=x_t}.$$

CIS Algorithm

(0) Fix x_0; set $w_0 = 1$ and $s = 0$.

(1) Simulate the time t of the next event after s in a Poisson process of rate λ.

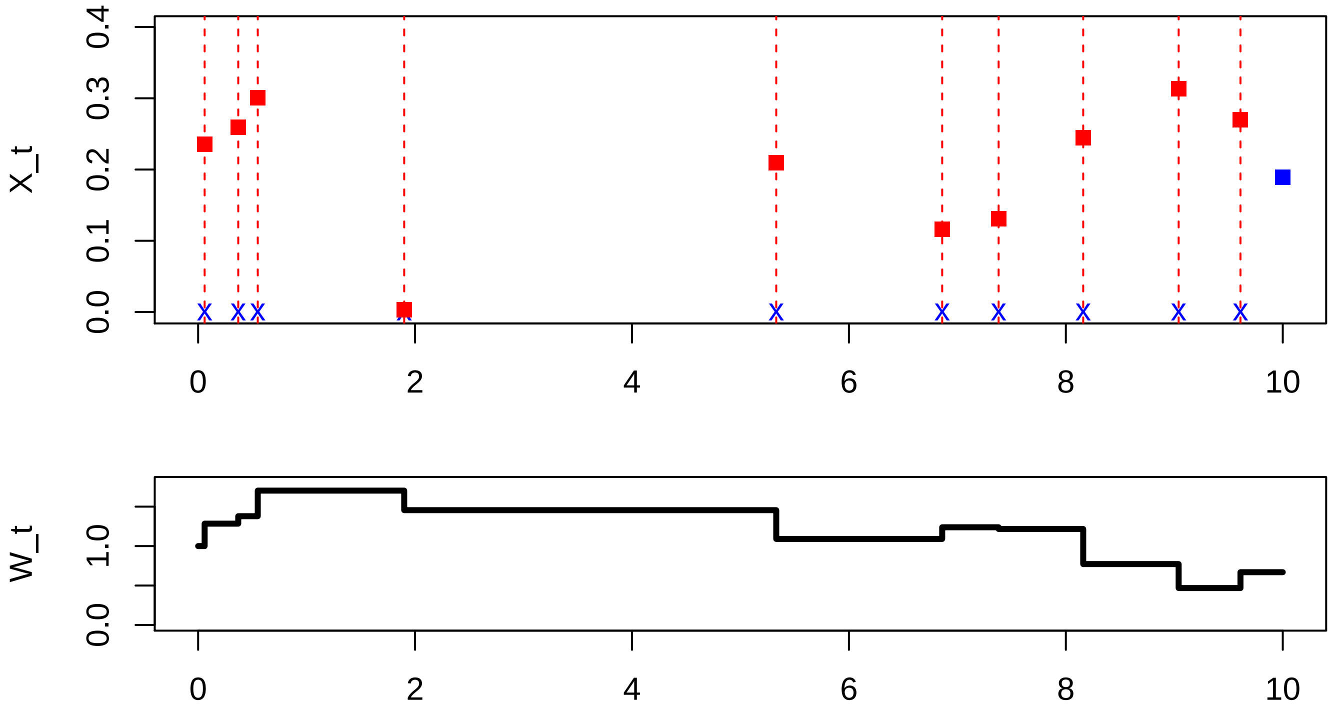
(2) Simulate X_t from $q(x_t|x_s, t-s)$; and set

$$w_t = w_s \times \rho(x_t, x_s, t-s).$$

(3) Goto (1).

If we stop the SIS at a time point T, then X_T can be drawn from $q(x_T|x_s, T-s)$; and the weight is w_j.

Example CIS



CIS for diffusions

The target process is

$$\mathrm{d}X_t = \mu(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t.$$

- Define an exogenous renewal process $\{\tau_1, \tau_2 \ldots\}$ with inter-arrival rate $\lambda = \lambda(t - \tau(t))$.
- Update weights at each renewal according to above formula.
- At each renewal, update the importance process:

$$\mathrm{d}X_t = b(\tau_i)\mathrm{d}t + v(X_{\tau_i})\mathrm{d}B_t.$$

Does it work?

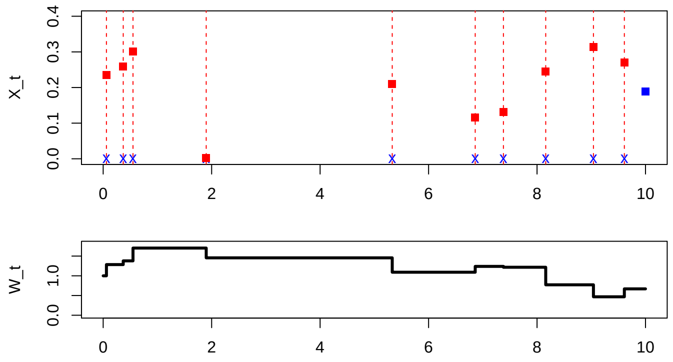
Not always! A necessary (and it turns out sufficient) condition for the method to be valid (ie unbiased) is that the weight process $\{w_s;\ s \geq 0\}$ is a martingale. Then the CIS algorithm provides unbiased estimates of the diffusion marginal distributions (and by iterations its FDDs).

In almost all cases where the proposal is not chosen to have $v(\tau_i) = \sigma(X_{\tau_i})$ then the weight process turns out to NOT be in L^1!

What about the copycat scheme? $v(\tau_i) = \sigma(X_{\tau_i})$, $b(\tau_i) = \mu(X_{\tau_i})$

Theorem:

1. If σ and μ are globally Libschitz, and σ is bounded away from 0, then the copycat scheme is valid.
2. If σ and μ are also bounded above, then for all $p > 1$, there exists $\epsilon > 0$ such that chosing $\lambda(u) \propto u^{-1+\epsilon}$ ensures that $\{w_s, s \geq 0\}$ is an L^p martingale.



$$w_T = \prod_{i=1}^{N_T} \rho_i$$

where

$$\rho_i = 1 + \frac{1}{\lambda}\left(\frac{(\mathcal{K}_x - \mathcal{K}_x^{(1)})q(x|x_s, t-s)}{q(x|x_s, t-s)}\right)\Bigg|_{x=x_t}.$$

Comments and Extensions

For general diffusions care is needed to ensure these conditions are satisfied – we have results which give rules for implementing the procedure in these cases.

There is substantial extra flexibility – such as letting the Poisson rate depend on the time since the last event, or coupling the Poisson rate with the proposal process.

There are numerous variance reduction methods that can be used (antithetic sampling, and extra importance sampling and different proposal distribution for the process at event times).

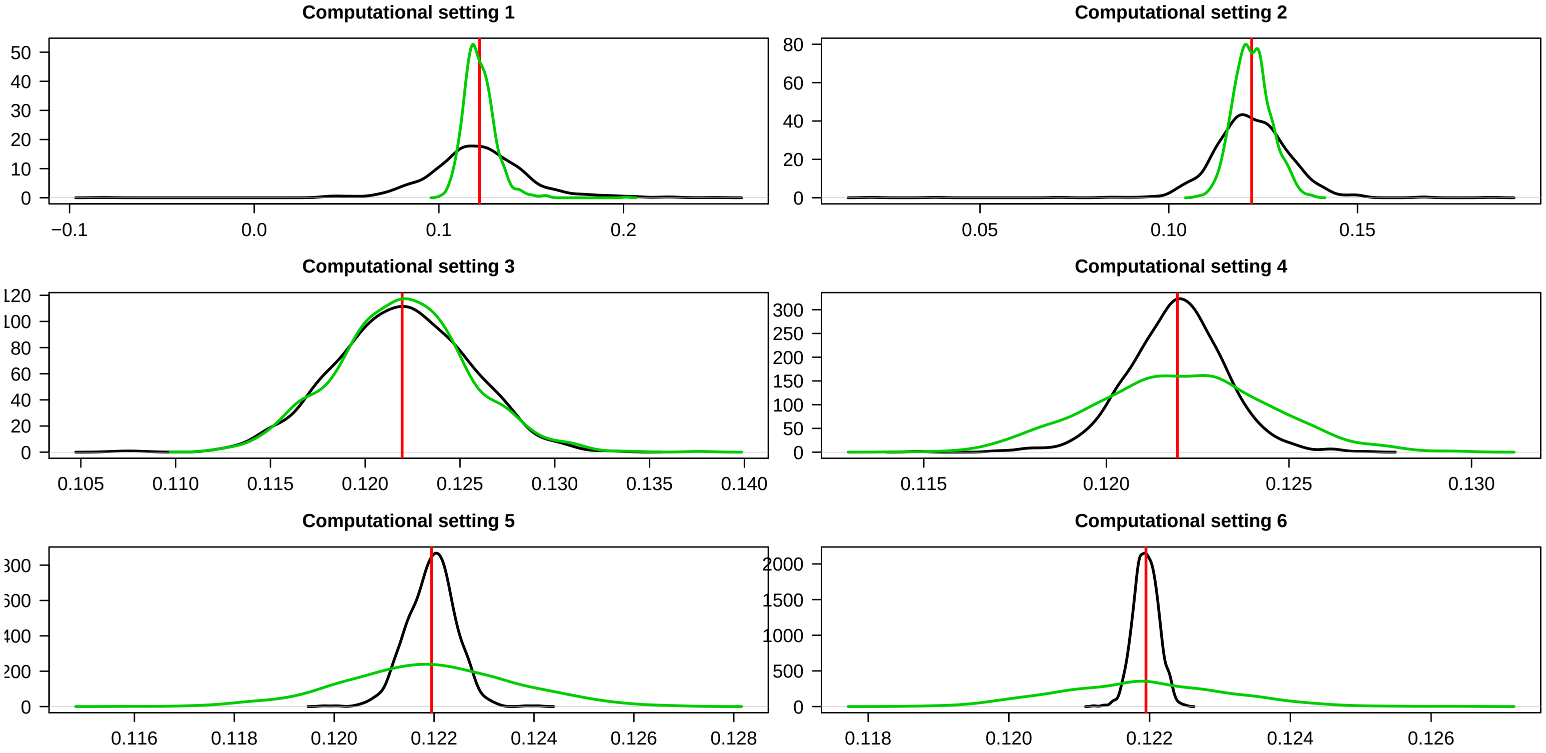
Example: CIR Diffusion

We consider estimating the transition density for a 2-d CIR model:

$$\begin{bmatrix} \mathrm{d}X_t^{(1)} \\ \mathrm{d}X_t^{(2)} \end{bmatrix} = \begin{bmatrix} -\rho_1(X_t^{(1)} - \mu_1) \\ -\rho_2(X_t^{(2)} - \mu_2) \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \sigma_1\sqrt{X_t^{(1)}} & 0 \\ \rho\sigma_2\sqrt{X_t^{(2)}} & \sigma_2\sqrt{(1-\rho^2)X_t^{(2)}} \end{bmatrix} \begin{bmatrix} \mathrm{d}B_t^{(1)} \\ \mathrm{d}B_t^{(2)} \end{bmatrix}$$

We compare the CIS with a time-discretisation approach based on the ideas in Durham and Gallant (2002), for varying CPU cost.

Example: CIR Diffusion



Example: Hybrid Systems

CIS can be applied to other continuous-time Markov processes.

One example is a hybrid linear diffusion/Markov-jump process:

$$\mathrm{d}X_t = (a(t, Y_t) + b(t, Y_t)X_t)\,\mathrm{d}t + \sigma(t, Y_t)\mathrm{d}B_t,$$

and Y_t is a Markov-jump process with generator (rate-matrix) $Q(X_t)$.

Such processes arise in systems biology and epidemic models

Example: Hybrid Systems

If we can bound the rate, $\lambda(X_t, y_t)$ of leaving a state y_t by $\bar{\lambda}$, then we can simulate from this process using thinning:

- Simulate the next time, τ from a Poisson Process with rate $\bar{\lambda}$.
- Simulate X_τ.
- With probability $\lambda(X_\tau, y_t)/\bar{\lambda}$ simulate an event in the Y_t process.

CIS can be implemented in a way similar to thinning, but does not require a bound, $\bar{\lambda}$. Instead if $\lambda(X_\tau, y_t) > \bar{\lambda}$ we get an Importance Sampling Correction.

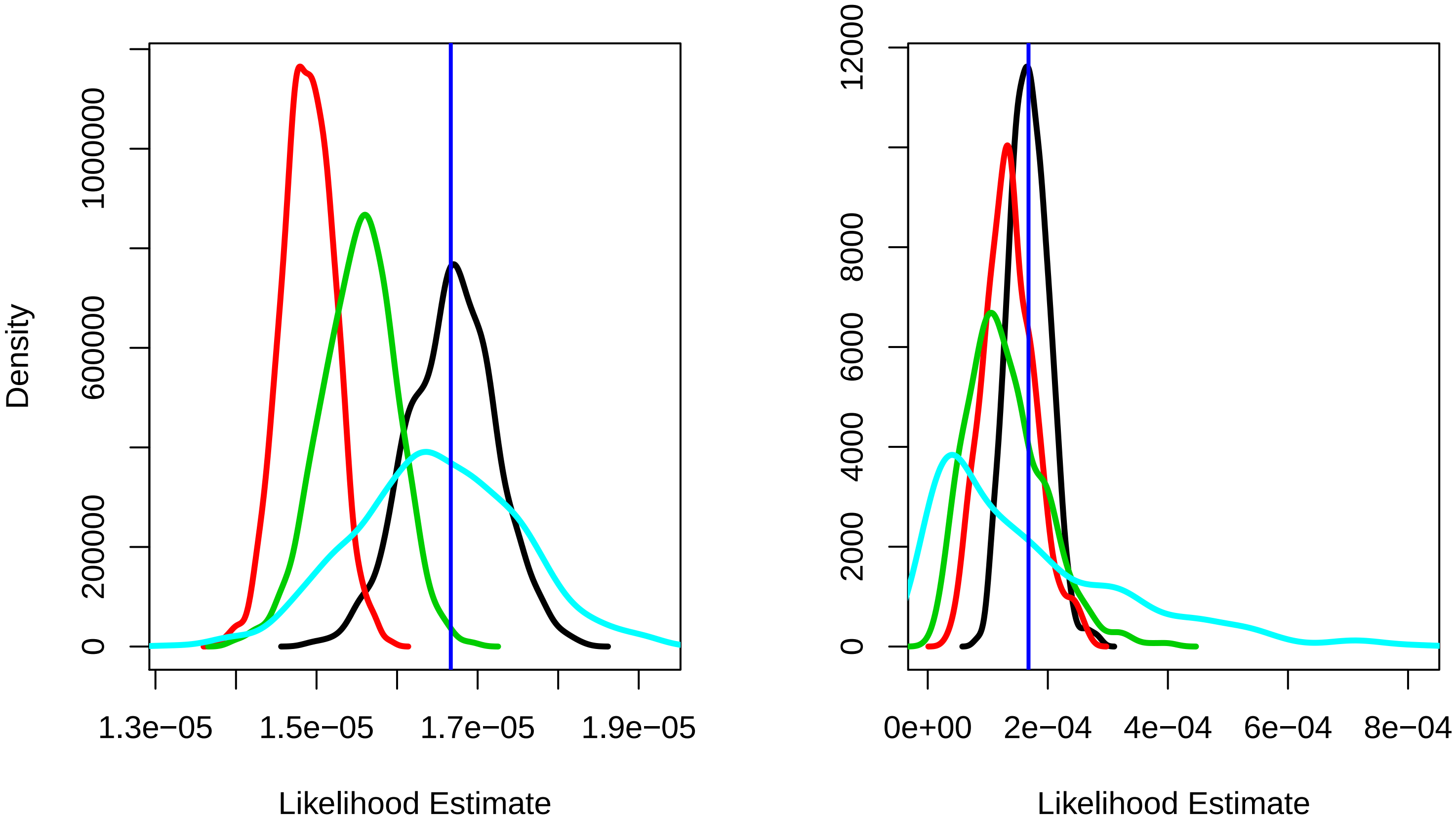
Auto-Regulatory System

We applied this to a hybrid system based on a 4-dimensional model of an autoregulatory system.

We looked at the accuracy of estimating the likelihood of data at a single time-point.

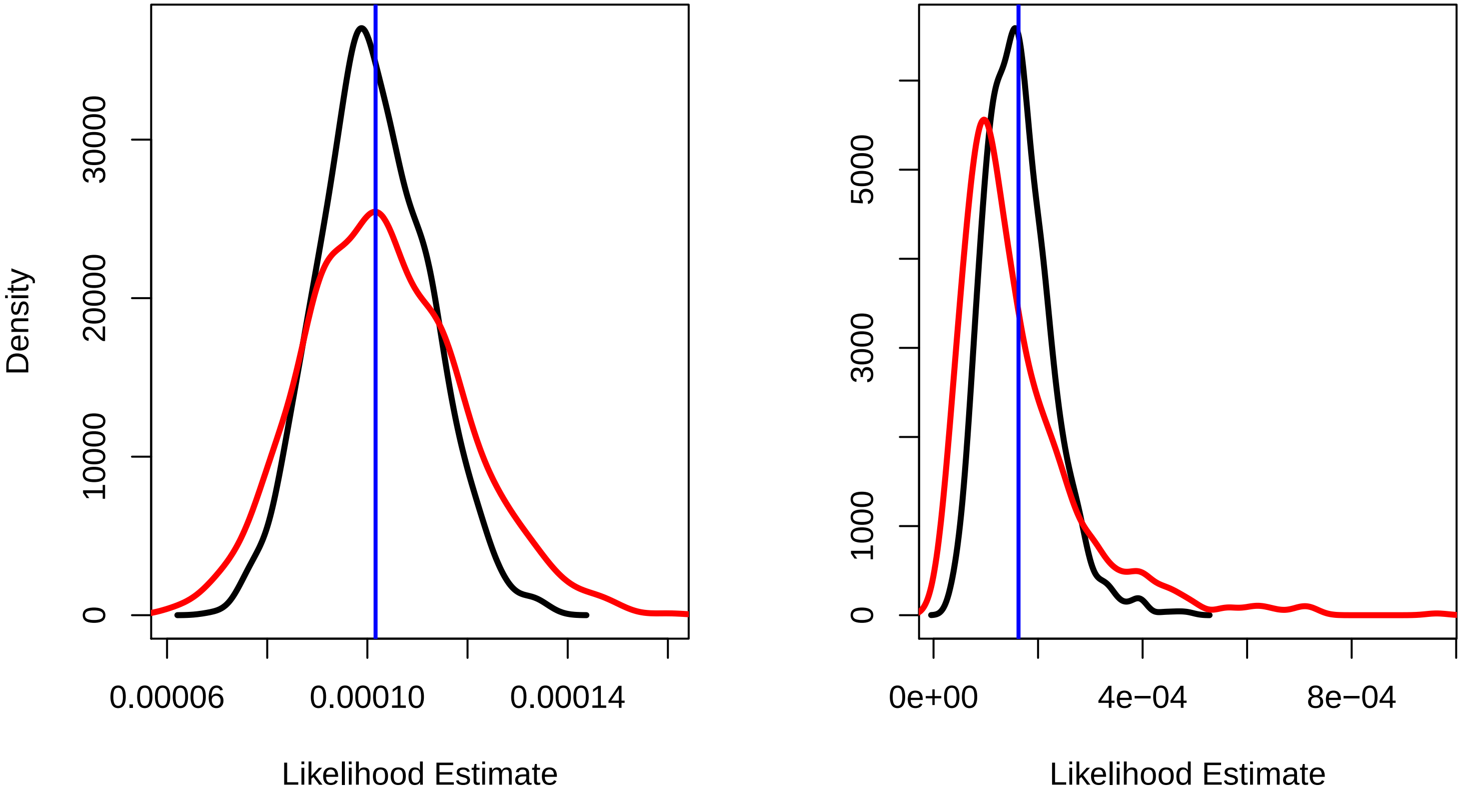
We utilised the tractability of the X_t process after the last event-time at which we (potentially) updated the Y_t process to improve the accuracy of our estimate – this advantages methods with fewer event times.

Auto-Regulatory System: Comparison with Euler



Comparison with (approximate) Thinning

Thinning with bound on rates chosen so that $\Pr(\lambda(X_\tau, y_t) < \bar{\lambda}) \approx 1$



Discussion

Exact simulation of diffusions is possible for pretty much all one-dimensional diffusions and some multi-dimensional ones.

However it is intrinsically limited in the multi-dimensional case.

This is a very flexible and potentially powerful method.

Can be used to unbiasedly estimate density (likelihood), expectations, etc

Theory established for diffusions, and very recently for jump diffusions

In diffusion case, links to importance sampling approach of Wagner. Our approach has the usual advantages of sequential importance sampling: resampling, adapting proposals etc. So SIS is more widely applicable.

Dealing with the negative weights is an important issue.