Continuous-time Importance Sampling for Multivariate Diffusion Processes (Avoiding time-discretisation approximation error)

Paul Fearnhead

Lancaster University

`www.maths.lancs.ac.uk/∼fearnhea`

Part of ongoing work with Krys Latuszynski, Gareth Roberts and Giorgos Sermaidis

Diffusions

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

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

Intuitively this defines the 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 standard normal random variable.

Transition Densities

We will denote the transition density of the diffusion by

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

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. Exceptions include models with both $\sigma(x) = \sigma$ and $\mu(x) = a + bx$. X_t is then a Gaussian process.

The Exact Algorithm

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

However, work by Beskos, Papaspiliopoulos and Roberts demonstrate how to simulate from a class of diffusion models where (possibly after transformation):

- The volatility is the identity: $\sigma(x) = \mathbf{I}$.
- The drift is the gradient of a potential: $\mu(x) = \nabla A(x)$.

This can be applied to almost all 1-d diffusions, and almost no others.

Current Approaches: The Exact Algorithm

The exact Algorithm is a Rejection Sampler based on proposing paths from Brownian motion.

The acceptance probability for the path is (for $\sigma(x) = 1$) proportional to:

$$\begin{aligned} &\exp\left\{-\int_0^T \mu(X_t)\mathrm{d}X_t + \frac{1}{2}\int_0^T \mu(X_t)^2\mathrm{d}t\right\} \\ &= \exp\left\{A(X_T) - A(X_0) - \frac{1}{2}\int_0^T \left(\mu(X_t)^2 + \mu'(X_t)\right)\mathrm{d}t\right\}. \end{aligned}$$

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

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?
- Error is $O(C^{-1/2})$, where C is CPU cost. Alternative approaches have errors that are e.g. $O(C^{-1/3})$ or worse.

Our Aim

Our aim was to try and extend the ability to perform 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 Exact Algorithm: Generalising Conditions

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 the identity as otherwise we do not have a proposal distribution that is tractable and absolutely continuous wrt to the target:

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.

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.

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

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).

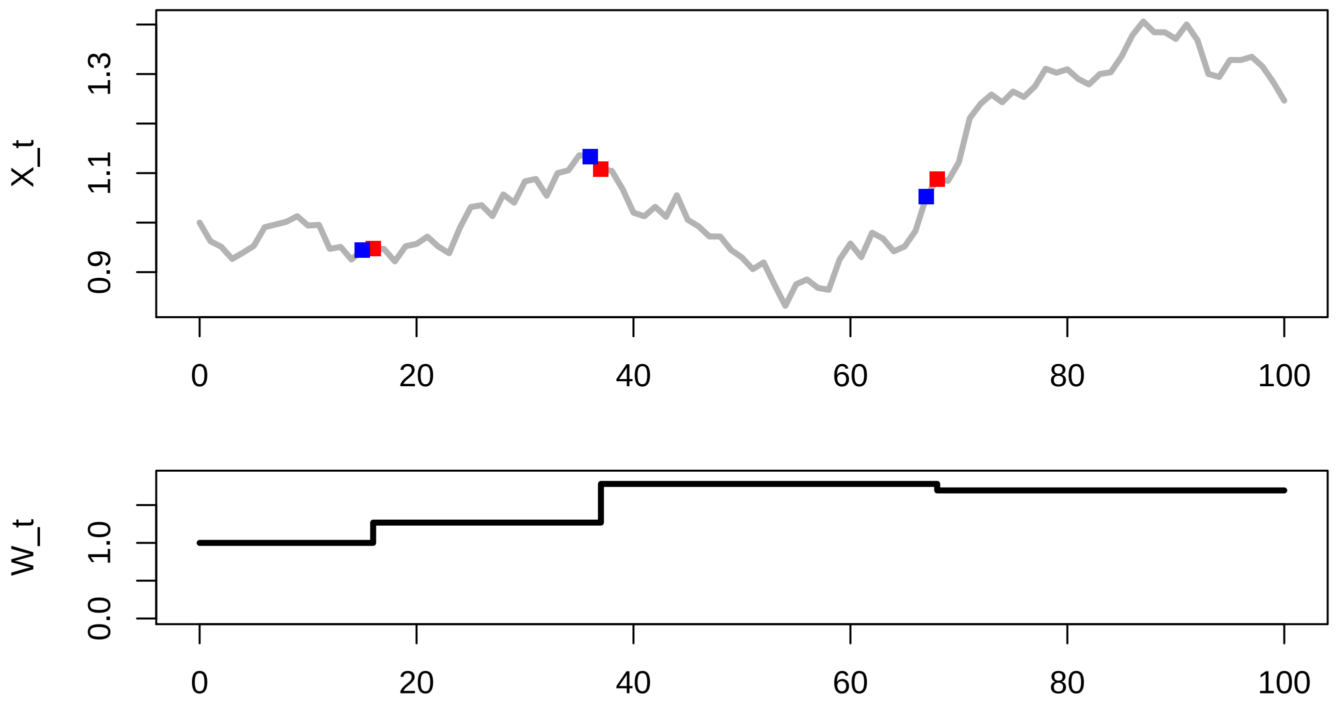
(2) $[U_i = 1]$ Simulate X_{ih} from $q(x_{ih}|x_{jh}, (i-j)h)$ and $X_{(i-1)h}$ from $q(x_{(i-1)h}|x_{jh}, x_{ih})$. Set

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

(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 $s = 0$.

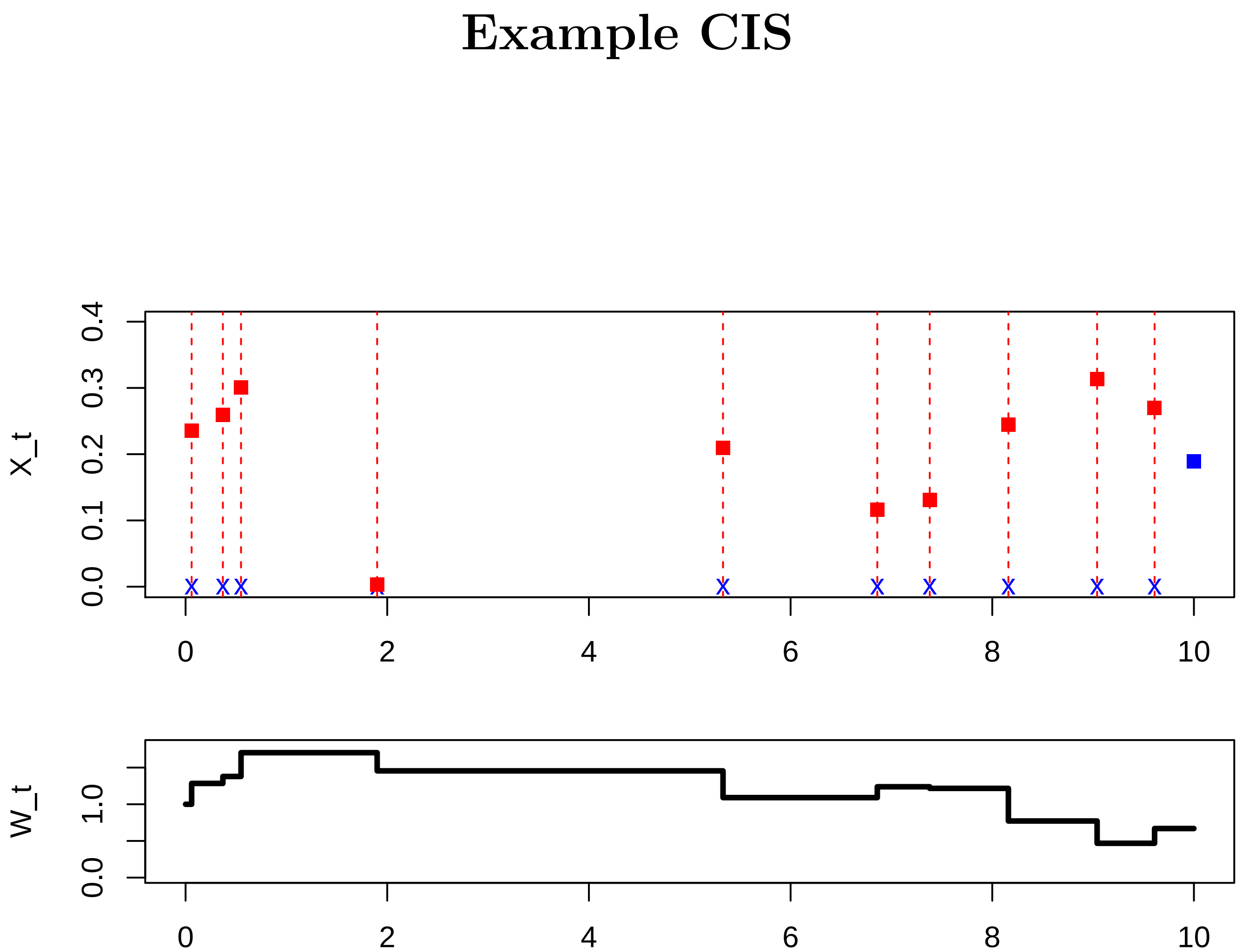
(1) Simulate the time t of the next event after s in a Poisson process of rate λ.

(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
X_t
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
W_t
0.0
1.0
0
2
4
6
8
10

Does it work?

Not always! A necessary and sufficient condition for the method to be valid (ie unbiased) is that the weight process $\{w_s;\ s \geq 0\}$ is a martingale.

This does not automatically happen as W_t can be negative: thus $\mathrm{E}(|W_t|)$ can be infinite.

CIS: Implementation for diffusions

The target process is

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

Denote event times by $\tau_1, \tau_2, \ldots$, and $\tau(t)$ the time of the most-recent even prior to t.

Proposal process needs to be tractable: e.g. constant drift and volatility.

- Can allow rate of Poisson process to depend on time since last event: $\lambda = \lambda(t - \tau(t))$.
- At each renewal, can update the importance process:

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

Does it work?

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 Libshitz, and σ is bounded away from 0, then the copycat scheme is valid.
2. 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.

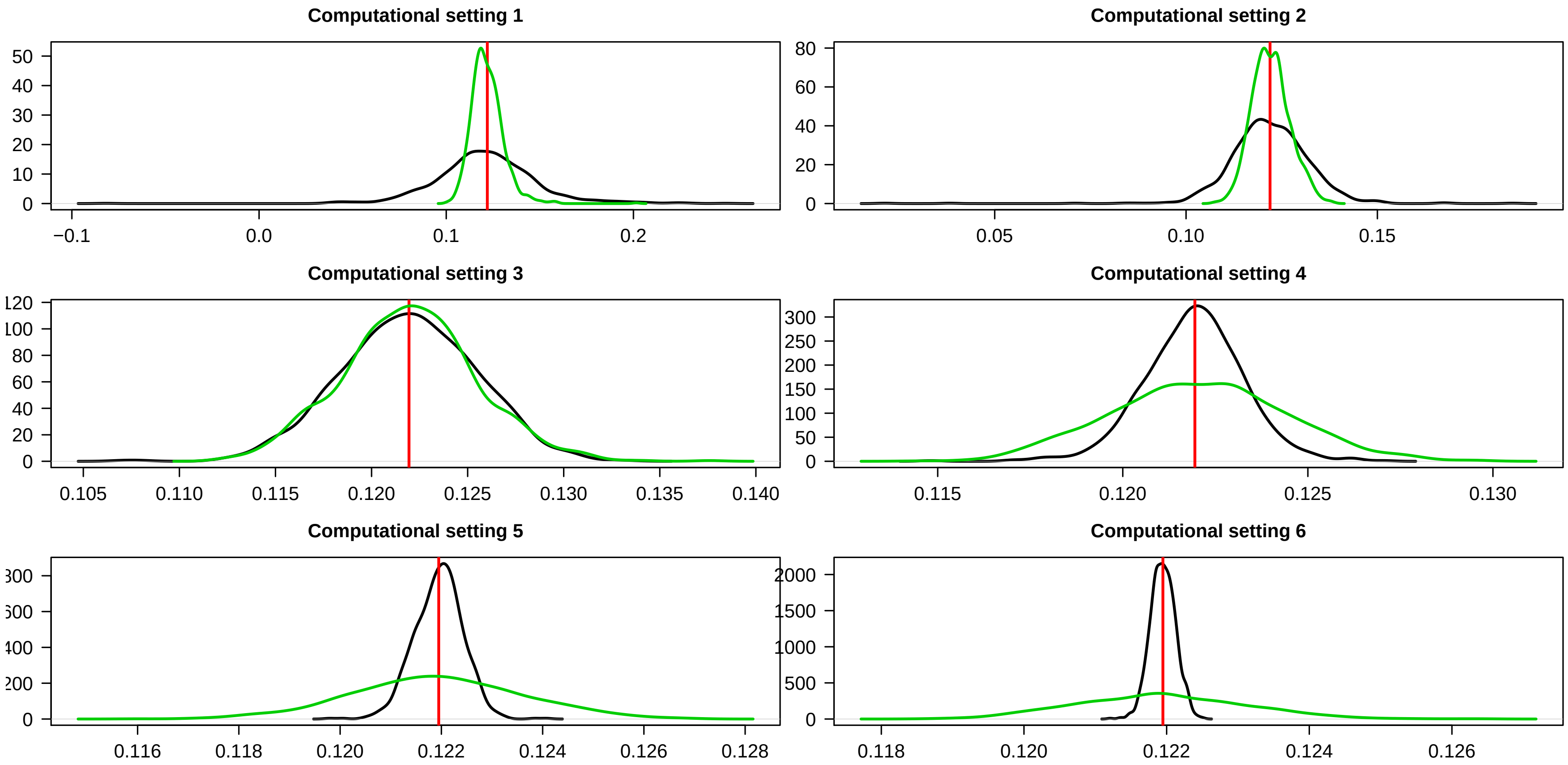
Example: CIR Diffusion

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.

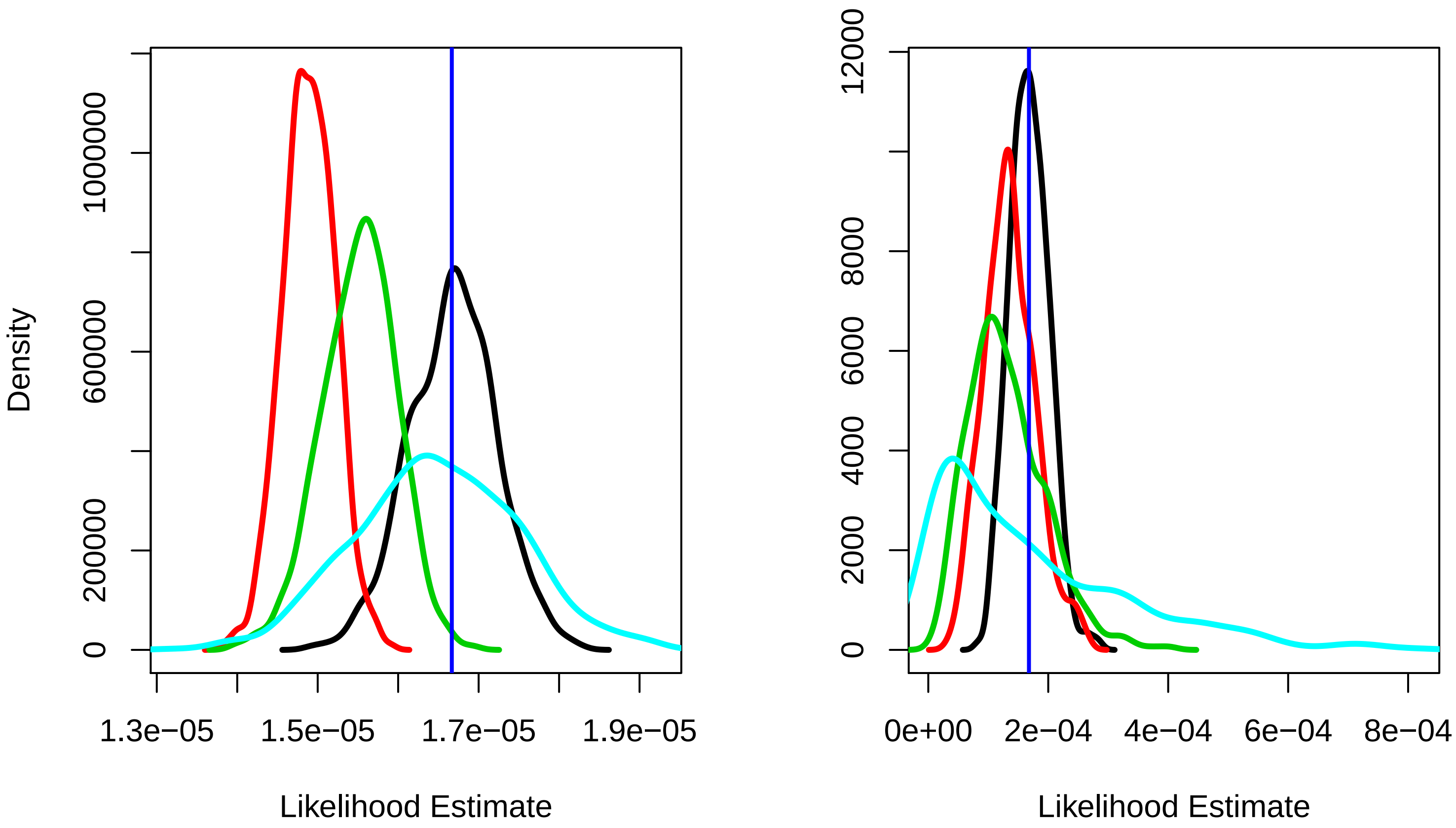
Auto-Regulatory System

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.

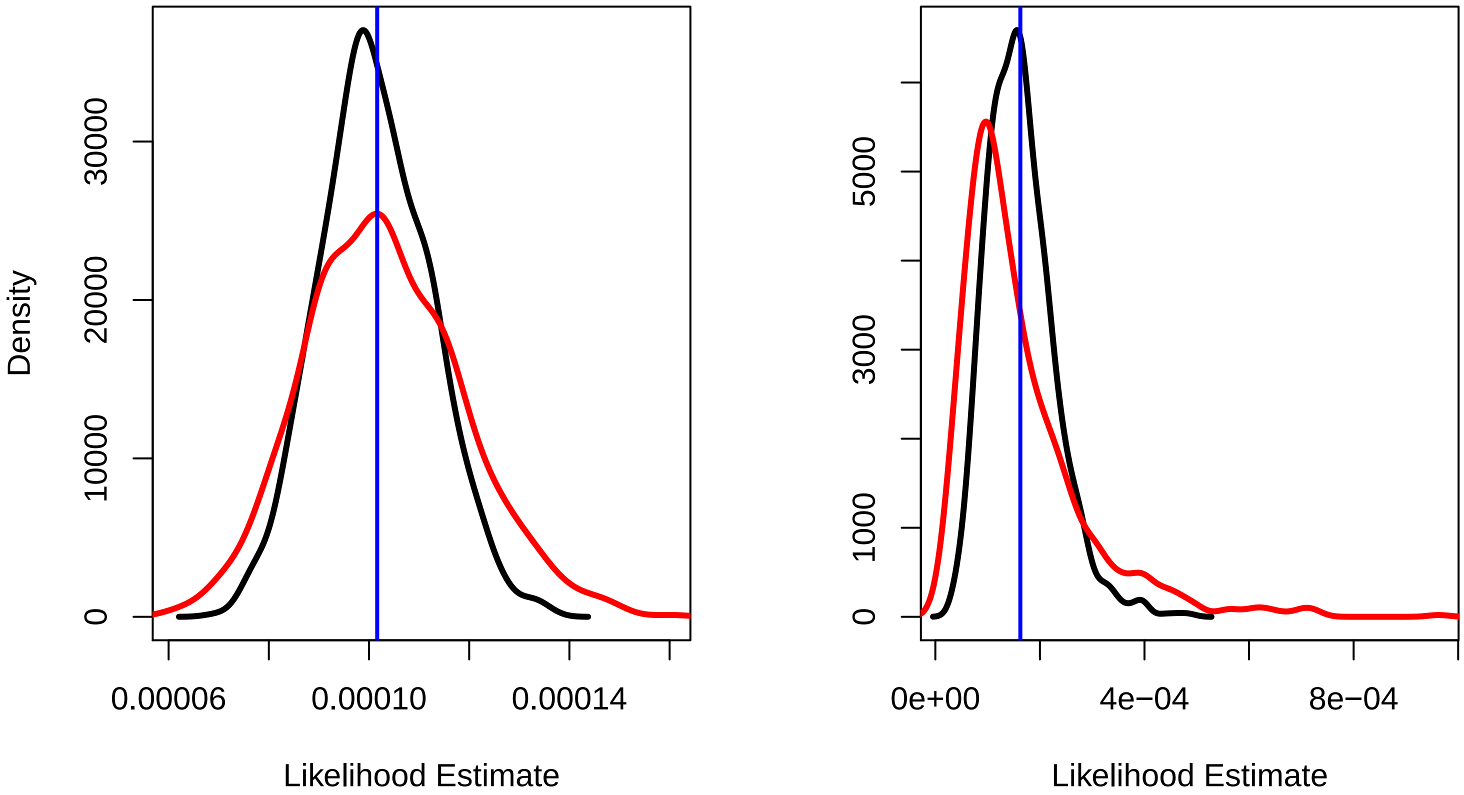
We utilised the tractability of the X_t process after the last event-time at which we (potentially) updated the Y_t process to improve the accuracy of our estimate – this advantages methods with fewer event times.

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

This is a very flexible and potentially powerful method. Can be used to unbiasedly estimate density (likelihood), expectations, etc.

There are numerous variance reduction methods that can be used

There is a related approach for diffusions by Wolfgang Wagner. His approach can be viewed as Importance Sampling, whereas ours is most similar to Sequential Importance Sampling. This has advantages in terms of using ideas (resampling, adapting proposals) from SIS to improve accuracy.

There are links of our method with Thinning of Jump-Markov processes.

Dealing with the negative weights is an important issue.