Exact Bayesian Inference for Markov Switching Diffusions

Timothée Stumpf Fétizon [tim.stumpf-fetizon@warwick.ac.uk]
June 14, 2021

Joint work with K. Łatuszyński, J. Palczewski, G. Roberts
Outline

Model

Inference Strategy

Methods

Designing 2-Coin Algorithms

Simulation Study

Outlook
Model
Many time series exhibits discrete regime shifts.

**Figure 1**: Pseudo-interest rate time series.
We model the regime as a Markov jump process.

![Figure 2: The corresponding trajectory of the regime.](image)

\[ Y(\tau_0) \xrightarrow{\text{Exp}(\lambda_{Y(\tau_0)})} Y(\tau_1) \xrightarrow{\text{Exp}(\lambda_{Y(\tau_1)})} Y(\tau_2) \xrightarrow{\text{Exp}(\lambda_{Y(\tau_2)})} \ldots \]

**Figure 3:** A 2-state Markov jump process.

Since the exponential distribution is memoryless, \( Y \) is Markovian. More generally, \( \lambda_{ij} \) gives the transition rate from state \( i \) to \( j \).
The diffusion process arises as a limit in discrete time.

Consider the process that evolves according to

\[
V_{t+\epsilon} - V_t = \mu(V_t, Y_t) \times \epsilon + \sigma(V_t, Y_t) \times (W_{t+\epsilon} - W_t)
\]  

(1)

Under [conditions], there is a limiting process as \( \epsilon \to 0 \). We write

\[
dV_t = \mu(V_t, Y_t) \, dt + \sigma(V_t, Y_t) \, dW_t
\]

(2)

We parameterize the *instantaneous drift* \( \mu_\theta \) and *volatility* \( \sigma_\theta \) in terms of a vector \( \theta \).

**Intractable likelihood problem!**

\[
\pi(v_{t+\epsilon} \mid v_t, y_t, \theta)
\]

typically not available!
We discretely observe the diffusion process.

Figure 4: What we see: Discrete observations $v_s$ on the path of $V$. Goal: Sample from posterior $\pi(\theta, y, \lambda|v_s)$ for some prior $\pi(\theta, y, \lambda)$!
Inference Strategy
We want to design an MCMC algorithm targeting the posterior on $(\theta, y, \lambda)$, s.t.:

- it targets the exact posterior, and the \textit{Markov chain central limit theorem} applies - estimates are unbiased, standard error decays according to $\mathcal{O}(\text{computational budget}^{-1/2})$.
- it is \textit{model agnostic} in principle - plug in $\mu_{\theta}$ and $\sigma_{\theta}$ and you’re good to go.
- it is an “\textit{algorithm for the people}” - no supercomputers required!
1. transform $V$ to a process $X$ with a tractable *dominating measure*.
2. augment with the *missing data* - the bridges between observations $v_s$ and $Y$ - such that conditional updates are “easy”!
3. devise an infinite-dimensional *Gibbs sampler* with updates (parameters|missing) and (missing|parameters).
4. carry out the updates based on *finite* information, using *Barker’s algorithm* in conjunction with *Bernoulli factories* and the *Exact algorithms*. 
Simplified setting and notation.

\[ \tau_0 = 0, \quad \tau_1, \quad \tau_2 = s \]

**Figure 5:** Event times; observations \( \tau_0 = 0 \) and \( \tau_1 = s \) and jumps \( \tau_1 \)

- **for simplicity:** assume that \( \sigma_\theta \) is constant in \( Y \) and we observe \( V \) at times 0 and \( s \).
- event times \( \tau \) consist of observation times \( \{0, s\} \) and intervening jump times. Denote consecutive jump times \( \hat{\tau} \sim \check{\tau} \).
- ancillary quantities are denoted by \( a \).
- random variables in upper case, realizations thereof in lower case.
- I skip inference for \( \lambda \), which is conditionally conjugate.
Event time augmentation.

Suppose we observe $V$ at times $\tau$. Then by the Markov property

$$
\pi(v_{\tau \setminus 0} | v_0, y_{[0,s]}, \theta) = \prod_{(\check{\tau} \sim \ddot{\tau}) \in \mathcal{T}} \pi(v_\check{\tau} | v_\ddot{\tau}, y_\ddot{\tau}, \theta)
$$

(3)

i.e. we can apply tools from ordinary diffusion inference to address the terms $\pi(v_\check{\tau} | v_\ddot{\tau}, y_\ddot{\tau}, \theta)$. 

law of ordinary diffusion!
Define the *Lamperti transform*

\[
\eta_\theta(v_t) = \int_0^{v_t} \frac{da}{\sigma_\theta(a)}
\]

and the *reduced process* \(X_t = \eta_\theta(V_t)\) with induced measure \(\mathbb{X}_{x_0,y,\theta}\) and SDE

\[
dX_t = \delta_\theta(X_t,Y_t) \, dt + dW_t
\]

Then, by the *Girsanov theorem* and under [conditions],

\[
\frac{d\mathbb{X}_{x_\tau,y_\tau,\theta}(x_{\tilde{\tau}},\tilde{\tau})}{dW_{x_\tau}} = a \exp \left[ - \int_{\tilde{\tau}}^{\tilde{\tau}} \phi_\theta(x_t,y_t) \, dt \right]
\]

Wiener measure

\[
2^{-1} \left( \delta^2_\theta(x_t,y_t) + \partial_{x_t} \delta_\theta(x_t,y_t) \right)
\]

\[
\partial_{x_t} \delta_\theta(x_t,y_t)
\]
Diffusion path augmentation.

Changing the dominating measure to \( \text{Leb} \times \mathbb{W}_{x_\tau, x_\tau} \), obtain augmented transition density

\[
\pi(x(\hat{\tau}, \bar{\tau}) | x_\tau, y_\tau, \theta) = a \frac{d\mathbb{X}_{x_\tau, y_\tau, \theta}}{d\mathbb{W}_{x_\tau}}(x(\hat{\tau}, \bar{\tau}))
\]  

w.r.t. \( \text{Leb} \times \mathbb{W}_{x_\tau, x_\tau} \) \hspace{1cm} (7)

Switch to non-centered parameterization to ensure irreducibility:

\[
\omega_\theta(x_t) = x_t - \eta_\theta(v_\tau) - \frac{t - \tau}{\bar{\tau} - \hat{\tau}} (\eta_\theta(v_\tau) - \eta_\theta(v_\tau)), \quad t \in [\hat{\tau}, \bar{\tau})
\]  

(8)

Such that \( Z(x_\tau, x_\tau) = \omega_\theta(X(x_\tau, x_\tau)) \) is a standard Brownian bridge under \( \mathbb{W}_{x_\tau, x_\tau} \circ \omega_\theta^{-1} = \mathbb{B} \). Now,

\[
\pi(v_\tau, z(\hat{\tau}, \bar{\tau}) | v_\tau, y_\tau, \theta) = a \frac{d\mathbb{X}_{x_\tau, y_\tau, \theta}}{d\mathbb{W}_{x_\tau}}(\omega_\theta^{-1}(z(\hat{\tau}, \bar{\tau})), \eta_\theta(v_\tau))
\]  

w.r.t. \( \text{Leb} \times \mathbb{B} \) \hspace{1cm} (9)
Infinite dimensional Gibbs sampler.

Put it all together:

\[
\pi(\theta, \lambda, h, y|v_0) \propto \pi(v_s, h|v_0, y, \theta) \pi(y|\lambda) \pi(\theta)\pi(\lambda) \tag{10}
\]

(10) augmented posterior \quad \text{aug trans density} \quad \text{regime prior} \quad \text{param prior}

\[
\pi(v_s, h|v_0, y, \theta) = \prod_{\tau \sim \tau \in \tau} \pi(v_{\tau^-}, z_{(\tau^-)}, v_{\tau^+}, y_{\tau^-}, \theta) \tag{11}
\]

(11) w.r.t. \((\text{Leb} \times \mathcal{B})|\tau|^{-1}\)

\[
H = V_{\tau\setminus\{0,s\}} \cup Z_{[0,s]\setminus\tau} \tag{12}
\]

(12) augmentation set

We can now define an *ergodic* Gibbs sampler:

\[
\begin{align*}
\text{(missing|param)}: \quad & \pi(h, y|v_0, v_s, \theta, \lambda) \propto \pi(h, v_s|v_0, y, \theta)\pi(y|\lambda) \tag{13} \\
\text{(param|missing)}: \quad & \pi(\theta|v_0, v_s, h, y) \propto \pi(h, v_s|v_0, y, \theta)\pi(\theta) \tag{14}
\end{align*}
\]

(13) (14)

The second update is of particular interest!
What to do about the path integral?

The augmented transition density contains an integral over a rough path:

$$\pi(h, v_s|v_0, y, \theta) = a \exp \left[ - \int_0^s \varphi_\theta(\omega^{-1}_\theta(z_t), y_t) \, dt \right]$$  \hspace{1cm} (15)

Can’t evaluate in finite time! Multiple possible approaches...

- *Pseudo-marginal* method, using unbiased estimators of the exponentiated path integral.
- even more augmentation...
- **Here:** Combine *Barker’s algorithm* with *Bernoulli factories!* Keeps the state space as is.
Methods
Let $\pi(a)$ be a target density. Propose update $a^\dagger$ according to $\kappa(a^\dagger | a)$. The *Metropolis algorithm* accepts with probability

$$\min \left[ 1, \frac{\kappa(a | a^\dagger) \pi(a^\dagger)}{\kappa(a^\dagger | a) \pi(a)} \right]$$

(16)

But there are other options! Barker’s algorithm accepts with probability

$$\frac{\kappa(a | a^\dagger) \pi(a^\dagger)}{\kappa(a^\dagger | a) \pi(a) + \kappa(a | a^\dagger) \pi(a^\dagger)}$$

(17)

This results in higher asymptotic variance for a given proposal! So why bother?
Enter the 2-coin algorithm.

```
start

\[
\begin{align*}
C_0 & \xleftarrow{\frac{c_1}{c_1+c_2}} C_1 \\
& \xleftarrow{1-p_1} \quad 1-p_2 \\
& \xrightarrow{p_2} C_2 \\
\end{align*}
\]

\[\text{return } 1 \quad \text{return } 0\]
```

**Figure 6**: Probability flow diagram of the 2-coin algorithm.

Suppose we can generate coins with probability \(p_1\) and \(p_2\). Then, the 2-coin algorithm generates coins with odds

\[
\frac{c_1 p_1}{c_1 p_1 + c_2 p_2}
\]

This is an example of a *Bernoulli factory*.

**Notice!**

runtime \(\to \infty\) as \(p_1, p_2 \to 0\)!
Assume there exist $\varphi_{\theta}^\downarrow, \varphi_{\theta}^\uparrow$ such that

$$
\varphi_{\theta}^\downarrow(z_{(\tau, \tilde{\tau})}, y_{\tilde{\tau}}) \leq \varphi_{\theta}(\omega_{\theta}^{-1}(z_t), y_{\tilde{\tau}}) \leq \varphi_{\theta}^\uparrow(z_{(\tau, \tilde{\tau})}, y_{\tilde{\tau}}), \quad t \in [\tau, \tilde{\tau})
$$

(19)

Barker acceptance odds for parameter proposal $\theta^\dagger \sim \kappa(\theta^\dagger | \theta)$ update are

$$
\frac{\alpha}{1 - \alpha} = \frac{\pi(h, v_s | v_0, y, \theta^\dagger)}{\pi(h, v_s | v_0, y, \theta)} \times \frac{\pi(\theta^\dagger)}{\pi(\theta)} \times \frac{\kappa(\theta | \theta^\dagger)}{\kappa(\theta^\dagger | \theta)}
$$

(20)

$$
\alpha = \prod_{(\tau \sim \tilde{\tau}) \in \tau} \frac{c_{\tilde{\tau}}}{c_{\tau}} \exp \left[ \int_{\tau} \varphi_{\theta^\dagger}^\downarrow(z_{(\tau, \tilde{\tau})}, y_{\tilde{\tau}}) - \varphi_{\theta^\dagger}(\omega_{\theta^\dagger}^{-1}(z_t), y_{\tilde{\tau}}) \, dt \right]
\frac{\exp \left[ \int_{\tau} \varphi_{\theta}^\downarrow(z_{(\tau, \tilde{\tau})}, y_{\tilde{\tau}}) - \varphi_{\theta}(\omega_{\theta}^{-1}(z_t), y_{\tilde{\tau}}) \, dt \right]}{\in (0, 1)}
$$

(21)
Let $0 \leq f(t) \leq f^\uparrow$ for $t \in [0, s]$. Simulate a unit intensity Poisson process on $[0, s] \times [0, f^\uparrow]$. Then

$$\Pr \left[ \text{all points above the graph of } f \right] = \exp \left[ - \int_0^s f(t) \, dt \right]$$  \hspace{1cm} (22)

So we only need to interpolate $f$ at a finite set of times. Apply this within a 2-coin algorithm to simulate coins with probability

$$\exp \left[ \int_{\tau}^{\hat{\tau}} \varphi_{\theta}^\downarrow (\hat{z}(\tau, \hat{\tau}), y_{\hat{\tau}}) - \varphi_{\theta}(\omega_{\theta}^{-1}(z_t), y_{\tau}) \, dt \right]$$  \hspace{1cm} (23)
Designing 2-Coin Algorithms
Naive 2-coin algorithms don’t scale.

Regardless of $|\theta^\dagger - \theta|$, under standard conditions

$$
\lim_{s \to \infty} \exp \left[ \int_0^s \varphi_{\theta}(z_{(\hat{\tau},\zeta)}, y_{\hat{\tau}}) - \varphi_{\theta}(\omega_{\theta}^{-1}(z_t), y_{\hat{\tau}}) \, dt \right] = 0 \quad (24)
$$

almost surely, so the 2-coin algorithm slows down as the time series extends. But there are various two-coin algorithms resulting in the same coin probability...
An alternative 2-coin algorithm.

Rearrange terms...

\[
\frac{c_{\theta}^{\dagger}}{c_{\theta}} \exp \left[ \int_{\tau_{\theta}}^{\hat{\tau}} \varphi_{\theta}(z(t, \tau), y(t)) - \varphi_{\theta}(\omega_{\theta}(z(t), y(t))) \, dt \right]
\]

By the mean value theorem and the Cauchy-Schwarz inequality,

\[
\varphi_{\theta}(\omega_{\theta}(z(t), y(t))) - \varphi_{\theta}(\omega_{\theta}(z(t), y(t)))
\leq \sup_{\text{convhull}[\theta, \theta], \tau} |\nabla_{\theta} \varphi_{\theta}(\omega_{\theta}(z(t), y(t)))| |\theta - \theta|
\]

\[
\rightarrow 0 \quad \text{as} \quad |\theta - \theta| \rightarrow 0
\]
Bounding the new path integral.

We have to find

$$\sup_{\text{convhull}[\theta^\dagger,\theta], t} |\nabla_\theta \varphi_\theta(\omega_\theta^{-1}(z_t), y_{\tau})|$$

(29)

$\nabla_\theta \varphi_\theta(\omega_\theta^{-1}(z_t), y_{\tau})$ is usually not concave, so we take a symbolic approach. To find $\sup f(a)$, solve for

$$\sup \{f(a) : f'(a) = 0\}$$

(30)

If $f'(a) = 0$ doesn’t have an analytical solution, expand to $f = g + h$, and bound

$$\sup f \leq \sup g + \sup h$$

(31)

by finding roots of $g'$ and $h'$. Expressions are complicated even for simple models - use computer algebra systems to do the heavy lifting!
Simulation Study
Consider a generalized CIR model with SDE

\[ dV_t = \beta_{Y_t} (\mu_{Y_t} - V_t) \, dV_t + V_t^{3/4} \, dW_t \]  \hspace{1cm} (32)

Where \( V_t > 0 \) almost surely. A priori

\[ \beta_1, \beta_2, \mu_1, \mu_2 \sim \log N [0, 1] \]  \hspace{1cm} (33)

Notice!

Posterior is invariant to label inversions!
Figure 7: Parameter traces for $\theta$, colored by regime state.
Figure 8: Posterior density of jump times in $Y$. Red lines correspond to the ground truth.
In conclusion.

Our exact algorithm is slowed down...

- by dependence between $Y$ and $\theta$ due to Gibbs sampling.
- by large or variable drift, slowing down the 2-coin algorithm.

But other methods have the same downsides!

- integration of the posterior wrt $Y$ is intractable even for tractable diffusions, so some form of conditional updating is unavoidable.
- accuracy of approximate methods degrades when drift is variable.
Outlook
Open questions.

Work in progress...

- finish algorithm for general $\sigma_\theta(V_t, Y_t)$.
- apply to real data (misspecification!).
- benchmark against pseudo-marginal implementation.
- which rate of posterior contraction gives a scalable algorithm?
- MAP estimation for $Y$.
- try more than 2 states.

Important, but probably intractable...

- optimal scaling. Tradeoff between 2-coin and MCMC efficiency!
Stay tuned for the pre-print!