Likelihood representations for discretely observed stochastic processes

Frank van der Meulen (Vrije Universiteit Amsterdam)
Moritz Schauer (Chalmers University of Technology & University of Gothenburg)

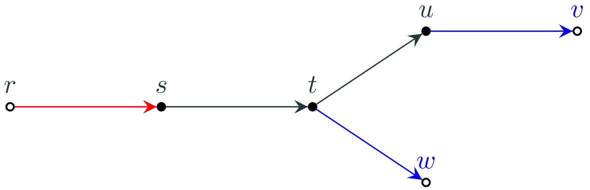
Bergamo-Waseda Workshop on Inference for Stochastic Processes and Applications
January 27, 2023

General problem setting

Problem setting

Setting

- A stochastic process X on a tree.
- At a branching vertex, the process evolves (conditionally) independent over the branches.
- Probability distribution depends on unknown θ.



- $\bullet$ latent vertex
- $\circ$ leaf/observation-vertex

Aims

General aims:

1. sampling values at $\bullet$, conditional on values at $\circ$;
2. estimating parameters in kernels;
3. not just on a tree, but on a general Directed Acyclic Graph (DAG).

Specific aim for this talk:
Derive a representation for the likelihood for a diffusion on a tree.

Likelihood can be expressed as expectation of a path functional:

$$L(\theta) = C_\theta \, \mathbb{E}_\theta \, \omega_\theta(X^\circ).$$

Motivational example from phylogenetics: Wright-Fisher diffusion on a tree

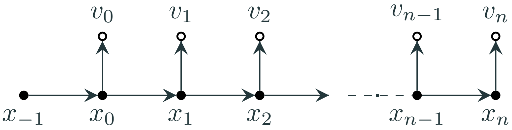
- Consider a population of diploid individuals. Alleles (**A**, **a**).
- X_t: probability of allele of type **A** at time t.
- Species may diverge... Consider a directed tree where along each branch

$$\mathrm{d}X_t = (\theta_1(1 - X_t) + \theta_2 X_t)\,\mathrm{d}t + \sqrt{X_t(1 - X_t)}\,\mathrm{d}W_t.$$

  Diffusion approximation to Wright-Fisher model with mutation for diploid individuals.

- Edge to observation $u \longrightarrow v$
- $x_v \mid x_u \sim \mathsf{Bin}(2n, x_u)$
  Observed number of alleles of type **A** among n individuals.

Related literature: state-space models – hidden Markov models



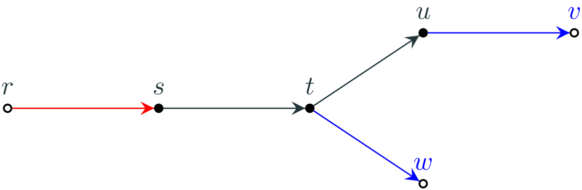
Well-known filtering, smoothing algorithms dating back to 1960-1970.

- Finite state space: Baum-Welch, Viterbi, forward-backward algorithm.
- Linear Gaussian models: Kalman filter, Rauch-Tung-Striebel smoother.
- Linear stochastic differential equations: Kalman-Bucy filter & smoother.

Recently much work on SMC (twisted particle samplers, controlled SMC).

Backward Filtering Forward Sampling

Working example



- Edge $r \to s$: $x_s \mid x_r \sim \Pi$ prior distribution on x_s.
- Edges $s \to t$ and $t \to u$: a diffusion process evolves for a fixed time-span:

$$\begin{aligned} \mathrm{d}X_t &= b_\theta(t, X_t)\,\mathrm{d}t + \sigma_\theta(t, X_t)\,\mathrm{d}W_t. \\ X_{t+h} &\approx X_t + hb_\theta(t, X_t) + \sigma_\theta(t, X_t)N(0, h). \end{aligned}$$

- Edges $u \to v$ and $t \to w$:

$$x_v \mid x_u \sim \Lambda_\theta(x_u, \cdot) \qquad x_w \mid x_t \sim \Lambda_\theta(x_t, \cdot).$$

Computing the likelihood

- Likelihood:

$$L(\theta) := p_\theta(x_v, x_w) = \int p_\theta(x_v, x_w \mid x_s)\Pi(\,\mathrm{d}x_s).$$

- Incorporate x_t

$$p_\theta(x_v, x_w \mid x_s) = \int p_\theta(x_t \mid x_s) p_\theta(x_v, x_w \mid x_t, \cancel{x_s})\,\mathrm{d}x_t.$$

- Branches evolve independently, conditional on x_t:

$$p_\theta(x_v, x_w \mid x_t) = p_\theta(x_v \mid x_t) p_\theta(x_w \mid x_t).$$

- Incorporate x_u

$$p_\theta(x_v \mid x_t) = \int p_\theta(x_u \mid x_t) p_\theta(x_v \mid x_u)\,\mathrm{d}x_u.$$

Backward Information Filter: recursive computation of the likelihood, starting from the leaves back to the root.

Computing the likelihood: Markov kernels

- Initialise from the leaves: assume $\Lambda(x; \mathrm{d}y) = \lambda(x, y)\nu(\mathrm{d}y)$ and set

$$h_{u \to v}(x) = \lambda(x, x_v) \qquad h_{t \to w}(x) = \lambda(x, x_w).$$

- Collect incoming "message": $h_u(x) = h_{u \to v}(x)$.
- On edge $t \to u$ assume Markov kernel $\kappa_{t,u}$.

$$h_{t \to u}(x) = \int \kappa_{t,u}(x, \mathrm{d}y) h_u(y).$$

- Collect incoming "messages": $h_t(x) = h_{t \to u}(x) h_{t \to w}(x)$.
- On edge $s \to t$ assume Markov kernel $\kappa_{s,t}$.

$$h_{s \to t}(x) = \int \kappa_{s,t}(x, \mathrm{d}y) h_t(y).$$

- Collect incoming "message": $h_s(x) = h_{s \to t}(x)$.
-

$$L(\theta) = \int \Pi(\mathrm{d}x) h_s(x).$$

Backward Information Filter

Computing the likelihood consists of composing:

- $\int \kappa(x,\, \mathrm{d}y) h(y) \,\mathrm{d}y$;
- pointwise product $h_1(x) h_2(x)$ (collecting incoming messages).

⚠ **Difficulties**:

- $\kappa(x,\, \mathrm{d}y)$ not known on $[s,t]$ and $[t,u]$;
- $\int \kappa(x,\, \mathrm{d}y) h(y) \,\mathrm{d}y$ cannot be computed in closed-form;
- For computations, we need a finite-dimensional representation of h.

Sampling from the smoothing distribution

- Conditional process follows the same dependency structure as the unconditional process.
- Sample x_s from
$$\Pi^\star(\mathrm{d}y) := \frac{h_s(y)\Pi(\mathrm{d}y)}{\int h_s(y)\Pi(\mathrm{d}y)}.$$
- On edge $s \to t$ sample x_t from
$$\kappa^\star_{s,t}(x, \mathrm{d}y) := \frac{\kappa_{s,t}(x, \mathrm{d}y)h_t(y)}{\int \kappa_{s,t}(x, \mathrm{d}y)h_t(y)}.$$
- Etc.

Continuous-time transition over an edge

Rethinking the discrete-time case

- Edge $\overset{s}{\bullet} \longrightarrow \overset{t}{\bullet}$ with diffusion evolving.
  Suppose $x \mapsto h_t(x)$ is given; wish to find $x \mapsto h_s(x)$.
- We have just seen that

$$h_s(x) = \int h_t(y) \kappa_{s \to t}(x, \mathrm{d}y) = \mathbb{E}\left[h_t(X_t) \mid X_s = x\right].$$

- Define for $\tau \in [s, t]$

$$h(\tau, x) := \mathbb{E}\left[h_t(X_t) \mid X_\tau = x\right].$$

- Well-known that h solves the Cauchy problem

$$(\mathcal{A}h)(\tau, x) := (\mathcal{L}h + \partial_\tau h)(\tau, x) = 0, \quad \text{subject to} \quad h(t, \cdot) = h_t(\cdot),$$

  with $\mathcal{L}$ the infinitesimal generator of X.

Approximating h

- Solving Kolmogorov backward equation is usually very difficult.
- **Key idea**: solve Kolmogorov's backward equation for simplified dynamics: solve backwards in time

$$(\widetilde{\mathcal{A}}g)(\tau, x) = 0 \quad \text{subject to} \quad g(t, \cdot) = g_t(\cdot)$$

  to get $g_s(\cdot) = g(s, \cdot)$.
- **Is this valid?** If so, how to correct for the approximation made?
- ⚠ Not knowing transition probabilities remains a problem!

The more technical part

Change of measure (1/3)

Characterisation [1]: for $\tau \in [s,t]$

$$f(\tau, X_\tau) - \int_s^\tau (\mathcal{A}f)(z, X_z)\,\mathrm{d}z$$

is a martingale iff

$$M_\tau = \frac{f(\tau, X_\tau)}{f(s, X_s)} \exp\left(- \int_s^\tau \frac{\mathcal{A}f}{f}(z, X_z)\,\mathrm{d}z \right)$$

is a martingale (which then satisfies $\mathbb{E}\, M_t = \mathbb{E}\, M_s = 1$).

[1] Cf. Palmowski & Rolski Bernoulli **8**, 2002.

Change of measure (2/3)

- Take $f \equiv h$ (recall $\mathcal{A}h = 0$). Define measure

$$\mathbb{P}^\star(B) = \mathbb{E}\left[\frac{h(t, X_t)}{h(s, x_s)}; B\right]$$

This is the law of the conditioned process.
Doob's h-transform.

- Take $f \equiv g$ (recall $\widetilde{\mathcal{A}}g = 0$). Define measure

$$\mathbb{P}^\circ(B) = \mathbb{E}\left[\frac{g(t, X_t)}{g(s, X_s)} \exp\left(-\int_s^t \frac{\mathcal{A}g}{g}(\tau, X_\tau)\,\mathrm{d}z\right); B\right]$$

This is the law of the guided process.

Change of measure (3/3)

What process is X under $\mathbb{P}^\circ$?

$$\mathcal{L}^\circ f = g^{-1} \left(\mathcal{L} f g - f \mathcal{L} g \right).$$

Then under $\mathbb{P}^\circ$, with $r(\tau, x) = \nabla_x \log g(\tau, x)$

$$\mathrm{d}X^\circ_\tau = b(\tau, X^\circ_\tau)\,\mathrm{d}\tau + \sigma\sigma'(\tau, X^\circ_\tau) r(\tau, X^\circ_\tau)\,\mathrm{d}\tau + \sigma(\tau, X^\circ_\tau)\,\mathrm{d}W_\tau.$$

Likelihood ratio (correction term):

$$\frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\mathbb{P}^\circ} = \frac{h(t, X^\circ_t)}{g(t, X^\circ_t)} \frac{g(s, X^\circ_s)}{h(s, X^\circ_s)} \underbrace{\exp\left(\int_s^t \frac{\mathcal{A}g}{g}(\tau, X^\circ_\tau)\,\mathrm{d}\tau \right)}_{\omega(X^\circ_{[s,t]})}.$$

Note that

$$\mathcal{A}g = \partial_t g + \mathcal{L}g = \cancel{\partial_t g + \widetilde{\mathcal{L}}g} + \left(\mathcal{L} - \widetilde{\mathcal{L}} \right) g.$$

Backward Filtering Forward Guiding

Backward Filtering

Backward Information Filter for simplified dynamics.

- Initialise $g_{u \to v}(x) = \lambda(x, x_v)$ and $g_{t \to w}(x) = \lambda(x, x_w)$.
- $g_u(x) = g_{u \to v}(x)$
- On $t \to u$ solve backwards $\widetilde{\mathcal{A}} g = 0$ subject to $g(u, \cdot) = g_u(\cdot)$.
  This gives $g_{t \to u}(x) = g(t, x)$.
- On $s \to t$ solve backwards $\widetilde{\mathcal{A}} g = 0$ subject to $g(u, \cdot) = g_t(\cdot)$.
  This gives $g_{s \to t}(x) = g(s, x)$.
- $g_s(x) = g_{s \to t}(x)$.
- $g_r(x) = \int g_s(x) \Pi(\,\mathrm{d}x)$.

Forward Guiding

Forward sample guided process.

- Sample X°_s from

$$\Pi^\circ(\mathrm{d}y) = \frac{\Pi(\mathrm{d}y) g_s(y)}{\int \Pi(\mathrm{d}y) g_s(y)}.$$

- sample X° on $[s, t]$
- sample X° on $[t, u]$
- Sample $X^\circ_v \sim \Lambda(X^\circ_u, \cdot)$ and $X^\circ_w \sim \Lambda(X^\circ_t, \cdot)$.

Likelihood representation

- $\mathbb{P}^\star$: law of true conditioned process (using h).
- $\mathbb{P}^\circ$: law of guided process (using g).

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\mathbb{P}^\circ}(X^\circ) \quad &= \quad \frac{h_s(X_s^\circ)/\int h_s(x)\Pi(\,\mathrm{d}x)}{g_s(X_s^\circ)/\int g_s(x)\Pi(\,\mathrm{d}x)} \\ &\quad \times \frac{h_t(X_t^\circ)}{g_t(X_t^\circ)}\frac{g_s(X_s^\circ)}{h_s(X_s^\circ)}\omega(X_{[s,t]}^\circ) \\ &\quad \times \frac{h_u(X_u^\circ)}{g_u(X_u^\circ)}\frac{g_t(X_t^\circ)}{h_t(X_t^\circ)}\omega(X_{[t,u]}^\circ)\end{aligned}$$

$$\frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\mathbb{P}^\circ}(X^\circ) = \frac{\int g_s(x)\Pi(\,\mathrm{d}x)}{\int h_s(x)\Pi(\,\mathrm{d}x)}\omega(X_{[s,u]}^\circ)$$

Main result

Bidirectional scheme:

1. Backwards Filtering (for simplified process).
2. Forward simulate Guided process.

Likelihood can be expressed as expectation of a path functional.

$$L(\theta) = \int g_s(x)\Pi(\mathrm{d}x) \times \mathbb{E}\,\omega(X^\circ_{[s,u]}).$$

How to solve the Kolmogorov Backward Equation?

For $\tau \in (s, t]$,

$$(\widetilde{\mathcal{L}} g + \partial_\tau g)(\tau, x) = 0, \qquad g(t, \cdot) = g_t(\cdot).$$

Examples/strategies:

1. If $\widetilde{\mathcal{L}}$ is the infinitesimal generator of a linear diffusion process, then

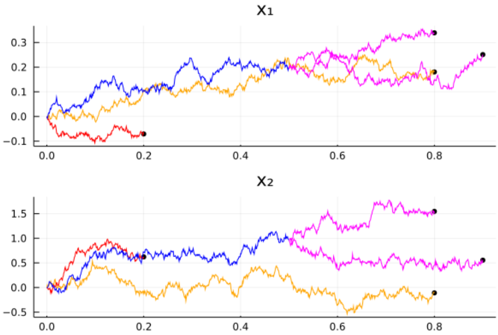
$$\log g(\tau, x) = c(\tau) + F(\tau)' x + x' H(\tau) x$$

   with ODE-system for $(H(\tau), F(\tau), c(\tau))$.

2. Ansatz $g(\tau, x) = \sum_j c(\tau) \psi_j(t)$. Derive ODE for $c(\tau)$.

Numerical example

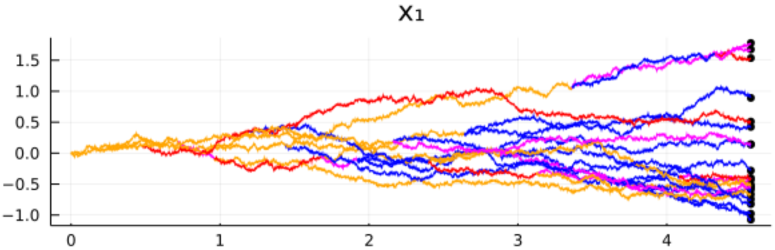
Example: branching diffusion

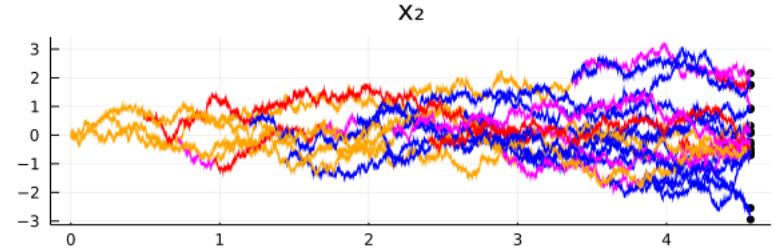


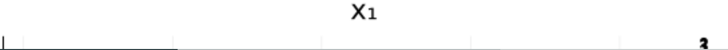
SDE on a tree where on each branch

$$\mathrm{d}X_t = \tanh.\left(\begin{bmatrix} -\theta_1 & \theta_1 \\ \theta_2 & -\theta_2 \end{bmatrix} X_t\right) \mathrm{d}t + \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \mathrm{d}W_t.$$

Numerical illustration: SDE on a tree







Numerical illustration: SDE on a tree

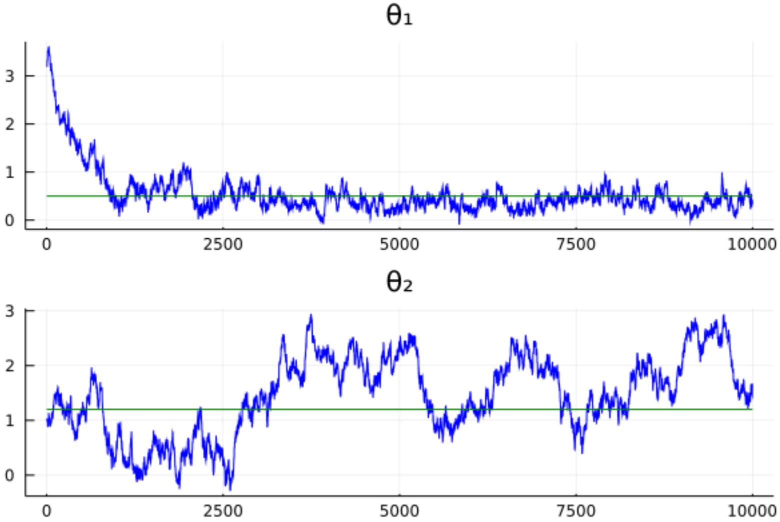
On each branch

$$\mathrm{d}X_t = \tanh.\left(\begin{bmatrix} -\theta_1 & \theta_1 \\ \theta_2 & -\theta_2 \end{bmatrix} X_t\right) \mathrm{d}t + \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \mathrm{d}W_t.$$

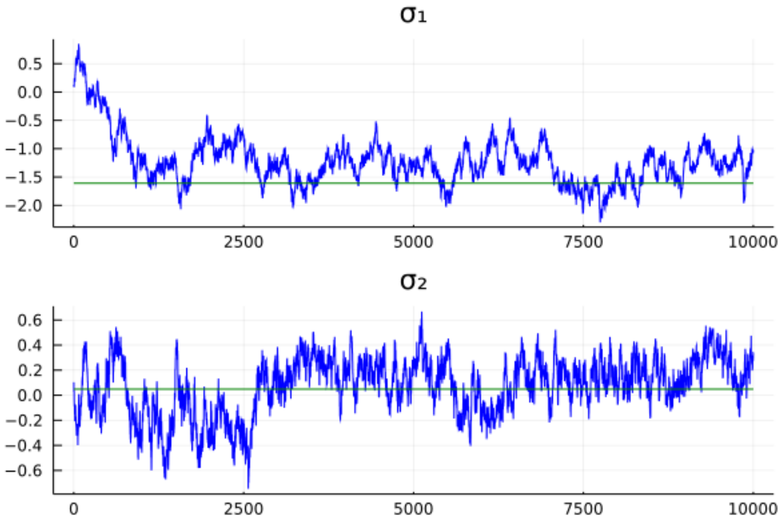
- Backward filter a linear process.
- Write X° as pushforward of (ξ, Z), with $\xi = (\theta_1, \theta_2, \sigma_1, \sigma_2)$
- MCMC on (ξ, Z)

Implementation in `MitosisStochasticDiffEq.jl` by Frank Schäfer (MIT).

Numerical illustration: SDE on a tree



Numerical illustration: SDE on a tree



Wrap-up / conclusions

Wrap-up

Backward Filtering Forward Guiding: framework for doing likelihood based inference in directed acyclic graphs, where transitions over edges may correspond to the evolution of a stochastic process for a certain time span.

- Defining guided processes on graphical models (for "non-tree"-case: see preprint).
- Both discrete-time and continuous-time transitions incorporated.
- **Approach is general and not restricted to diffusions on a tree.**
- Not covered: compositionality results (some category theory, see earlier versions on arXiv).

Ongoing: SPDEs, SDEs on manifolds, chemical reaction networks.

Open postdoc position at VU Amsterdam.

References

- Continuous-discrete smoothing of diffusions
  MIDER, SCHAUER, VDM, Electronic Journal of Statistics

  Bayesian inference for partially observed diffusions in SSM.
- Automatic Backward Filtering Forward Guiding for Markov processes and graphical models, VDM AND SCHAUER, arXiv, submitted.

  A generalisation to Markov processes on graphical models including ideas on compositionality from category theory.
- Introduction to Automatic Backward Filtering Forward Guiding, VDM, arXiv, submitted.

  Gentle introduction to the more advanced paper.
- Conditioning continuous-time Markov processes by guiding, CORSTANJE, VDM AND SCHAUER, to appear in Stochastics.

  Derivation of conditions for $\mathbb{P}^\star \ll \mathbb{P}^\circ$ for general continuous-time Markov processes.
- Inference in Hidden Markov Models, CAPPÉ, MOULINES AND RYDÉN

  Good source on filtering, smoothing, parameter estimation in HMM.