Consistent nonparametric Bayesian inference for discretely observed diffusions

Harry van Zanten

Eindhoven University of Technology

Joint work with Frank van der Meulen (TU Delft)

Warwick, January 22, 2010

Outline

- The diffusion model: assumptions, basic properties
- Approaches to nonparametric inference
- Nonparametric Bayesian inference

- Bayesian data augmentation
- Bayesian inference based on continuous-time data
- Sampling diffusion bridges
- Simulated data example

- Consistency result
- Concrete priors
- Sketch of proof

- Wish list

The model

Model: stationary, ergodic SDE

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t,$$

with $\sigma > 0$ *known*, b the infinite-dimensional parameter of interest. (For simplicity, we take $\sigma \equiv 1$.)

Observations:

$$X_0, X_\Delta, X_{2\Delta}, \ldots, X_{n\Delta},$$

with $\Delta > 0$ *fixed*.

Aim: make inference about b on basis of these observations.

Some possible approaches

- For small Δ,

$$X_{i\Delta} - X_{(i-1)\Delta} \approx b(X_{(i-1)\Delta})\Delta + Z_i,$$

  with Z_i independent $N(0, \Delta)$. Use favorite nonparametric regression method. Problem: bias.

Some possible approaches

- For small Δ,

  $$X_{i\Delta} - X_{(i-1)\Delta} \approx b(X_{(i-1)\Delta})\Delta + Z_i,$$

  with Z_i independent $N(0, \Delta)$. Use favorite nonparametric regression method. Problem: bias.
- Spectral approach (Hansen et al. (1998), Gobet et al. (2004)): relate drift to eigenvalues/eigenfunctions of generator/transition operator.
  Restriction: works if generator has discrete spectrum (e.g. reflecting boundaries). Computation?

Some possible approaches

- For small Δ,

  $$X_{i\Delta} - X_{(i-1)\Delta} \approx b(X_{(i-1)\Delta})\Delta + Z_i,$$

  with Z_i independent $N(0, \Delta)$. Use favorite nonparametric regression method. Problem: bias.
- Spectral approach (Hansen et al. (1998), Gobet et al. (2004)): relate drift to eigenvalues/eigenfunctions of generator/transition operator.
  Restriction: works if generator has discrete spectrum (e.g. reflecting boundaries). Computation?
- What about a nonparametric version of the data augmentation approach? [REFS]

Model assumptions more precise - 1

We assume that $\sigma \equiv 1$ and $b : \mathbb{R} \to \mathbb{R}$ is locally integrable.

The SDE generates a one-dimensional diffusion with scale function

$$s_b(x) = \int_0^x \exp\Big(-2 \int_0^y b(z)\, dz \Big)\, dy.$$

We assume that

$$\lim_{x \downarrow -\infty} s_b(x) = -\infty, \qquad \lim_{x \uparrow \infty} s_b(x) = \infty.$$

Then the SDE generates a recurrent diffusion on $\mathbb{R}$.

Model assumptions more precise - 2

The speed measure of the diffusion is given by

$$m_b(dx) = \exp\left(2\int_0^x b(z)\,dz\right)dx.$$

We assume that $m_b(\mathbb{R}) < \infty$.

This ensures the diffusion is positive recurrent/ergodic and that $\mu_b = m_b/m_b(\mathbb{R})$ is the unique invariant probability measure.

We assume that the diffusion is started in μ_b, so that it is stationary.

Some basic properties of the model - 1

Define the transition operators of the diffusion by

$$P_t^b f(x) = \mathbb{E}_x^b f(X_t).$$

Under our assumptions

- Feller property:

$$P_t^b : C_b(\mathbb{R}) \to C_b(\mathbb{R}),$$

- we have transition densities p_b relative to Lebesgue measure:

$$P_t^b f(x) = \int f(y) p_b(t, x, y)\, dy.$$

Some basic properties of the model - 2

In particular, we have that the data

$$X_0, X_\Delta, \dots$$

form a (discrete-time) stationary, ergodic Markov chain, with unique invariant distribution μ_b and transition densities $(x, y) \mapsto p_b(\Delta, x, y)$.

Remark: the transition densities can typically not be expressed expicitly in terms of the coefficients b and σ of the SDE!

Nonparametric Bayesian inference - 1

The likelihood for the data $X_0, X_\Delta, \ldots, X_{n\Delta}$ is given by

$$L_n(b) = \pi_b(X_0) \prod_{i=1}^{n} p_b(\Delta, X_{(i-1)\Delta}, X_{i\Delta}),$$

where $\mu_b(dx) = \pi_b(x)\,dx$. (Have $\pi_b(x) \propto \exp\left(2\int_0^x b(z)\,dz\right)$.)

Nonparametric Bayesian inference - 2

Bayesian approach: put prior Π on a collection of drift functions $\mathscr{B}$ and compute the posterior

$$\Pi(B \,|\, X_0, X_\Delta, \ldots, X_{n\Delta}) = \frac{\int_B L_n(b)\,\Pi(db)}{\int_{\mathscr{B}} L_n(b)\,\Pi(db)},$$

for $B \subset \mathscr{B}$.

Base further inference on the posterior.

Complication: $L_n(b)$ can not be computed explicitly!

Bayesian data augmentation - 1

- Likelihood for discrete data intractable.
- Likelihood for continuous data $(X_t : t \in [0, n\Delta])$ explicit.
- View continuous segments between the discrete observations as missing data.
- Use Gibbs sampling scheme to draw from the posterior.

Bayesian data augmentation - 2

Bayesian data augmentation in general (Tanner and Wong (1987)):

Suppose data y depend on parameter θ.

Likelihood $p(y \mid \theta)$ intractable. Artificial "data" z such that $p(y, z \mid \theta)$ is tractable.

Posterior:

$$p(\theta \mid y) = \int p(\theta \mid y, z) p(z \mid y)\, dz,$$

$$p(z \mid y) = \int p(z \mid y, \theta) p(\theta \mid y)\, d\theta.$$

Bayesian data augmentation - 3

Consider the Markov chain:

- start in (z_0, θ_0),
- draw z_1 from $p(z \mid y, \theta_0)$,
- draw θ_1 from $p(\theta \mid y, z_1)$
- etc.

This chain has an invariant probability distribution, who's marginals are $p(z \mid y)$ and $p(\theta \mid y)$.

Gibbs sampler

Bayesian data augmentation - 4

To draw from the discrete data posterior in our setting:

- Choose some initial estimator b.

Repeat a large number of times:

- Sample the missing continuous segments between the observations, given the observations and the current value of b, yielding a full path $(x_t : t \in [0, n\Delta])$.
- Draw new b from the continuous data posterior $\Pi(\cdot \mid x_t : t \in [0, n\Delta])$.

After "burn-in", this yields (approximately) draws from the discrete data posterior.

Bayesian inference based on continuous-time data - 1

The likelihood for the continuous data $(X_t : t \in [0, T])$ is given by

$$L_T^c(b) = \pi_b(X_0) \exp\Big(\int_0^T b(X_t)\, dX_t - \tfrac{1}{2} \int_0^T b^2(X_t)\, dt \Big),$$

and the continuous data posterior is given by

$$\Pi(B \,|\, X_t : t \in [0, n\Delta]) = \frac{\int_B L_{n\Delta}^c(b)\, \Pi(db)}{\int_{\mathscr{B}} L_{n\Delta}^c(b)\, \Pi(db)},$$

for $B \subset \mathscr{B}$.

Bayesian inference based on continuous-time data - 2

Clever choice for Π: Gaussian or conditionally Gaussian. Yields (partial) conjugacy.

Simple example: finite Gaussian series prior

Take for Π the law of $\sum Z_k h_k$, for deterministic functions h_k and i.i.d. standard normals Z_k. Then under the posterior

$$b \sim \sum c_k h_k,$$

where $c \sim N((I+\Sigma)^{-1}\mu, (I+\Sigma)^{-1})$, for

$$\mu_k = \int_0^{n\Delta} h_k(X_t)dX_t, \qquad \Sigma_{ij} = \int_0^{n\Delta} h_i(X_t)h_j(X_t)\,dt.$$

Sampling diffusion bridges - 1

Basic problem: Suppose X solves the SDE

$$dX_t = b(X_t)\,dt + dW_t, \qquad t \in [0,1].$$

Sample from the conditional distribution of X given that $X_0 = u$, $X_1 = v$.

Recall Girsanov:

$$\frac{d\mathbb{P}_X}{d\mathbb{P}_W}(X) = \exp\left(\int_0^T b(X_t)\,dX_t - \tfrac{1}{2}\int_0^T b^2(X_t)\,dt \right).$$

Sampling diffusion bridges - 2

General observation: if X and W are conditioned on the same functional, the resulting conditional distributions remain equivalent, and the Radon-Nikodym derivative remains the same up to a constant.

In particular: let $\mathbb{P}_X^{u,v}$ be the law of the diffusion bridge we are after, let $\mathbb{P}_W^{u,v}$ be the law of the Brownian bridge that goes from u to v. Then

$$\frac{d\mathbb{P}_X^{u,v}}{d\mathbb{P}_W^{u,v}}(X) \propto \exp\left(\int_0^1 b(X_t)\, dX_t - \tfrac{1}{2} \int_0^1 b^2(X_t)\, dt \right).$$

Hence: can do importance sampling to get exact simulations (Beskos and Roberts (2005), Beskos et al. (2006)).

Simulation example - 1

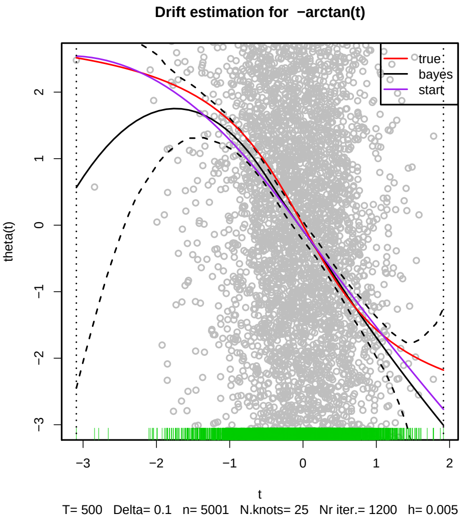

Figure: $\Delta = 0.1$

Simulation example - 2

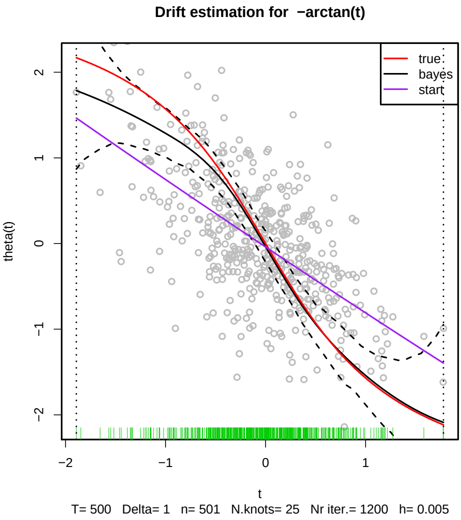

Figure: $\Delta = 1$

Consistency - 1

Assume there exists a “true value” b_0. Does the posterior contract around b_0 as $n \to \infty$?

General remarks:

- For nonparametric models, intuitively reasonable priors may lead to inconsistent procedures (Freedman, Diaconis).
- Existing consistency results for Markov chain data (Tang and Ghosal (2007), Ghosal and Van der Vaart (2007)) are not applicable.
- More generally: usual “testing approach” can not be followed (yet).

We prove consistency for a weak topology, avoiding testing arguments.

Consistency - 2

Topology on the class $\mathscr{B}$ of drift functions:

Let ν be a finite measure. For $b \in \mathscr{B}$, $f \in C_b(\mathbb{R})$ and $\varepsilon > 0$, let

$$U^b_{f,\varepsilon} = \{b' \in \mathscr{B} : \|P^{b'}_\Delta f - P^b_\Delta f\|_{1,\nu} < \varepsilon\}.$$

Topology on $\mathscr{B}$: such that the collection of sets

$$\{U^b_{f,\varepsilon} : f \in C_b(\mathbb{R}), \varepsilon > 0\}$$

forms sub-base for the neighborhood system at b.

(i.e.: any open neighborhood of $b \in \mathscr{B}$ is a union of finite intersections of the form $U^b_{f_1,\varepsilon_1} \cap \cdots \cap U^b_{f_m,\varepsilon_m}$.)

Consistency - 3

Remarks about the topology:

- If ν assigns positive mass to non-empty open intervals, then the topology is Hausdorff.
- Let $A_b f = bf' + f''/2$ for a C^2-function f. Then for small Δ,

$$P_\Delta^{b_1} f - P_\Delta^{b_2} f \approx \Delta(A_{b_1} f - A_{b_2} f) = \Delta(b_1 - b_2) f'.$$

Hence for small Δ, the constructed topology is close to the $L^1(\nu)$-topology.

Consistency - 4

Theorem.

Let Π be a prior on a class $\mathscr{B}$ that is locally uniformly equicontinuous and uniformly bounded. Then if

$$\Pi\Big(b \in \mathscr{B} : \|b - b_0\|_{2,\mu_0} < \varepsilon\Big) > 0 \quad \text{for all } \varepsilon > 0,$$

it holds for every open neighborhood U_{b_0} of b_0 that

$$\Pi(b \notin U_{b_0} \,|\, X_0, X_\Delta, \dots, X_{n\Delta}) \to 0 \qquad \mathbb{P}_{b_0}\text{-a.s.}$$

as $n \to \infty$.

Consistency - 5

In words:

If a prior gives mass 1 to a locally uniformly equicontinuous and uniformly bounded set of drift functions, then we have consistency for all b_0 in the $L^2(\mu_0)$-support of the prior.

Remark: The basic version of the exact algorithm for simulating diffusion bridges requires that

$$\sup_{b \in \mathscr{B}} \left(\|b\|_\infty + \|b'\|_\infty \right) < \infty.$$

Clearly this is stronger than what is required here for consistency.

Some concrete priors - 1

Discrete net prior for arbitrary $\mathscr{B}$ satisfying the requirements of the theorem:

Fix probability distributions (p_n) and (q_n) on $\mathbb{N}$ and $0 < \varepsilon_n \downarrow 0$. Let $\mathscr{B}_m = \{b|_{[-m,m]} : b \in \mathscr{B}\}$. By Arzelà-Ascoli, $\mathscr{B}_m$ is totally bounded for the uniform norm. For every n we fix a finite ε_n-net $\mathscr{B}_{m,\varepsilon_n} = \{b_1^{m,n}, \dots, b_{k_{m,n}}^{m,n}\}$ for $\mathscr{B}_m$. We extend every function in the net to the whole real line by making it constant on $(-\infty, -m]$ and $[m, \infty)$. Define

$$\Pi = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{k_{m,n}} \frac{p_m q_n}{k_{m,n}} \delta_{b_k^{m,n}}.$$

Some concrete priors - 2

General construction:

- Fix distribution (p_m) on $\mathbb{N}$. For every m, let Π_m be a prior on $C_L^\alpha[-m,m]$ ($\alpha \in (0,1)$, $L > 0$).
- $\bar{C}_L^\alpha[-m,m]$: extend functions in $C_L^\alpha[-m,m]$ by making them constant on $(-\infty,-m]$ and $[m,\infty)$. Let $\bar{\Pi}_m$ be the obvious extension of Π_m.
- Define

$$\Pi = \sum_{m=1}^{\infty} p_m \bar{\Pi}_m.$$

If every Π_m has full support relative to the uniform norm and $b_0 \in C_L^\alpha$, then we have consistency.

Some concrete priors - 3

Integrated Brownian motion:

Fix $m, L > 0$, B a Brownian motion with standard normal initial law.

- For $|t| \leq m$, define $Y_t = (B_{t+m})_L$, where

$$(a)_L = \begin{cases} L & \text{if } a \geq L, \\ a & \text{if } |a| \leq L, \\ -L & \text{if } a \leq -L. \end{cases}$$

- Let U be uniform on $[-L, L]$, independent of B. Let Π_m be the law of

$$Z_t = \left(U + \int_{-m}^{t} Y_s \, ds \right)_L, \qquad |t| \leq m.$$

Some concrete priors - 4

Wavelets:

For an appropriate wavelet basis $\psi_{j,k}$, a Hölder ball of functions on $[-m, m]$ can be represented as

$$\Big\{ \sum_{j=1}^{\infty} \sum_{k=1}^{2^j} b_{j,k}\psi_{j,k} : \|b\|_{\infty,\infty}^{\alpha} \leq L \Big\},$$

for some $L > 0$, where $\|b\|_{\infty,\infty}^{\alpha} = \sup_j 2^{\alpha j} 2^{j/2} \max_k |b_{j,k}|$.
Let

$$\Pi_m \sim \sum_{j=1}^{J} \sum_{k=1}^{2^j} \delta_j U_{j,k} \psi_{j,k},$$

for J, $U_{j,k}$ independent, $U_{j,k}$ uniformly distributed variables in $[-L, L]$, and $\delta_j = 2^{-\alpha j} 2^{-j/2}$.

Sketch of proof - 1

- Enough to show that $\Pi(B \mid X_0, \ldots, X_{n\Delta}) \to 0$, $\mathbb{P}_{b_0}$-a.s., where
$$B = \{b \in \mathscr{B} : \|P_\Delta^b f - P_\Delta^{b_0} f\|_{1,\nu} > \varepsilon\},$$
with $\varepsilon > 0$ and f continuous $\|f\|_\infty \leq 1$.
- Equicontiuity assumption on $\mathscr{B}$ implies $\{P_\Delta^b f : b \in \mathscr{B}\}$ is locally uniformly equicontinuous as well.
- Suffices to show that $\Pi(B' \mid X_0, \ldots, X_{n\Delta}) \to 0$, $\mathbb{P}_{b_0}$-a.s., where
$$B' = \Big\{b \in \mathscr{B} : P_\Delta^b f(x) - P_\Delta^{b_0} f(x) > \eta \quad \forall x \in I\Big\},$$
for $\eta > 0$ and I a bounded open interval.

Sketch of proof - 2

- We have

$$\Pi(B' \mid X_0, X_\Delta, \ldots, X_{n\Delta}) = \frac{\int_{B'} L_n(b)\,\Pi(db)}{\int_{\mathscr{B}} L_n(b)\,\Pi(db)},$$

- By ergodicity,

$$\frac{1}{n}\log L_n(b) \overset{\text{as}}{\to} -\mathrm{KL}(b_0, b) \gtrsim -\|b - b_0\|^2_{2,\mu_0}.$$

- Hence, for n large and some $c > 0$, the denominator is bounded from below by

$$e^{-nc\delta}\Pi(b : \|b - b_0\|^2_{2,\mu_0} < \delta)$$

for every $\delta > 0$.

Sketch of proof - 3

- Square root of numerator of expression for posterior:

$$D_n = \sqrt{\int_{B'} L_n(b)\,\Pi(db)}.$$

- The process

$$M_n = D_n\Big(1 - \frac{\eta^2}{8}\Big)^{-\sum_{i=1}^{n-1} 1_{X_i \in I}}$$

  is a supermartingale.
- By Doob's convergence theorem $M_n \overset{\text{as}}{\to} M_\infty$, by ergodicity $n^{-1}\sum_{i=1}^{n-1} 1_{X_i \in I} \overset{\text{as}}{\to} \mu_0(I)$.
- Hence $D_n \overset{\text{as}}{\to} 0$, exponentially fast.

Wish list

- estimating σ^2 as well
- consistency in stronger topologies
- contraction rates
- optimal priors
- adaptation/model selection
- diffusions in $\mathbb{R}^d$