Bayesian nonparametric inference for diffusion models with discrete sampling

Delft University of Technology

Jakob Söhl

joint work with Richard Nickl

Van Dantzig Seminar, Leiden, 26 October 2016

Outline

Diffusion Markov Processes

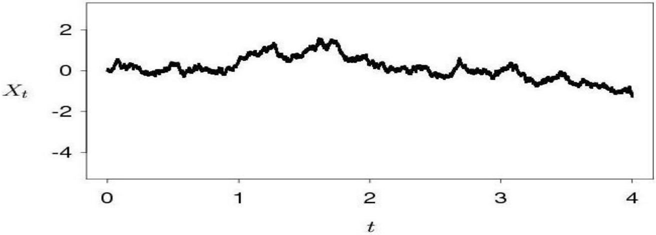
Consider a process $(X_t : t \geqslant 0)$ that solves the stochastic differential equation

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad t \geqslant 0.$$

Here b is a drift coefficient, σ the diffusion coefficient, $(W_t)_{t\geqslant 0}$ Brownian motion

Under mild assumptions on (σ, b), $(X_t : t \geqslant 0)$ is a unique Markov process with transition densities $p_{t,\sigma b}(x, y)$ describing the operator

$$\mathbb{E}_{\sigma b}[f(X_{t+s})|X_s = x] = \int_{\mathcal{Y}} f(y) p_{t,\sigma b}(x, y)\,\mathrm{d}y =: P_t f(x), \quad f \in C_b(\mathcal{Y}), \;\; s \geqslant 0.$$



Applications

→ Diffusion models are ubiquitous in modern science: They serve as fundamental building blocks in the modelling of dynamic phenomena in

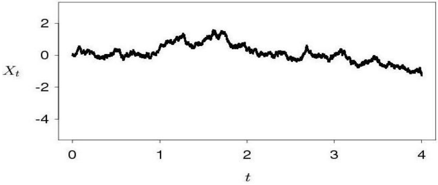
- physics, biology, geosciences
- evolutionary dynamics and life sciences
- engineering
- economics & finance

They are closely related to stochastic models that model a dynamical system by some differential operator L that propagates the system state perturbed with statistical noise.

Buzzwords: 'data assimilation, uncertainty quantification, filtering problems, Hidden Markov Models'.

→ Often the parameters (σ, b) are unknown and one wants to infer their values from some form of sample of the diffusion.

Statistical Inference & Observation Schemes



- An idealised assumption would be to observe an entire trajectory $(X_t : 0 \leqslant t \leqslant T)$, up to time T. Inference on b becomes possible as $T \to \infty$. (Note that σ is known in this case.)
- More realistic: discrete observations $X_0, X_\Delta, X_{2\Delta}, \ldots, X_{n\Delta}$ of the continuous process, where Δ is the 'observation distance'.
  - high-frequency observations: $\Delta \to 0$ and $n\Delta = T \to \infty$
  - low-frequency observations: $\Delta > 0$ fixed as $n \to \infty$.
- The high-frequency regime asymptotically reflects the 'continuous data' setting. Low-frequency is harder.

TU Delft

Some Spectral Theory

When the diffusion is restricted to a regular compact space by reflection, say $[0,1]$ for simplicity, the transition operator P_t coincides with the action of the semigroup $(e^{tL} : t \geqslant 0)$ on $L^2(\mu)$ where the infinitesimal generator

$$L = L_{\sigma b} = b(x)\frac{\mathrm{d}}{\mathrm{d}x} + \frac{\sigma(x)^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}$$

admits (subject to suitable boundary conditions) a discrete spectrum of eigenfunctions $u_k : k = 0, 1, 2, \dots$ with eigenvalues $\lambda_k \in [-Ck^2, -C'k^2]$, $k \geqslant 1$. Here μ is the invariant density of the Markov process. We deduce the expansion

$$p_{t,\sigma b}(x,y) = \sum_k e^{\lambda_k t} u_k(x) u_k(y) \mu(y), \quad x, y \in [0,1].$$

$\to$ In the case of a scalar diffusion reflected at $\{0,1\}$ the boundary conditions are of von Neumann type ($u_k'(0) = u_k'(1) = 0$). If $b = 0$ and $\sigma = 1$ we have reflected Brownian motion. Dirichlet conditions correspond to killed Brownian motion.

TUDelft

Frequentist Estimation at Low Frequency

- In a seminal paper, Gobet, Hoffmann & Reiß (2004) studied the above model in the nonparametric setting. They started from the spectral identities

$$\sigma^2 = \frac{2\lambda_1 \int_0^{\cdot} u_1 \,\mathrm{d}\mu}{u_1'\mu}, \quad b = \lambda_1 \frac{u_1 u_1' \mu - u_1'' \int_0^{\cdot} u_1 \,\mathrm{d}\mu}{(u_1')^2 \mu}.$$

Frequentist Estimation at Low Frequency

- In a seminal paper, Gobet, Hoffmann & Reiß (2004) studied the above model in the nonparametric setting. They started from the spectral identities

$$\sigma^2 = \frac{2\lambda_1 \int_0^{\cdot} u_1 \, \mathrm{d}\mu}{u_1' \mu}, \quad b = \lambda_1 \frac{u_1 u_1' \mu - u_1'' \int_0^{\cdot} u_1 \, \mathrm{d}\mu}{(u_1')^2 \mu}.$$

- While estimation of μ is straightforward, recovery of the first eigen-pair (u_1, λ_1) requires estimation of the entire transition operator P_Δ. GHR show that this can be done empirically in a minimax optimal way, with resulting L^2-convergence rates

$$n^{-s/(2s+3)} \text{ for } \sigma^2 \text{ and } n^{-(s-1)/(2s+3)} \text{ for b}$$

whenever, for C^s a s-Hölder or Sobolev space,

$$(\sigma, b) \in \Theta_s = \{\|\sigma\|_{C^s} + \|b\|_{C^{s-1}} \leqslant B, \sigma \geqslant c > 0\}.$$

These rates reveal an ill-posed nonlinear inverse problem of order 1 and 2.

Bayesian Methods

From a Bayesian perspective it is natural to put a prior Π on the pair (σ, b). The resulting posterior distribution is obtained from Bayes' formula. For instance if the process is started in equilibrium, $X_0 \sim \mu_{\sigma b}$, then

$$\mathrm{d}\Pi((\sigma, b)|X_0, X_\Delta, \dots X_{n\Delta}) = \frac{\mu_{\sigma b}(X_0) \prod_{i=1}^n p_{\Delta,\sigma b}(X_{(i-1)\Delta}, X_{i\Delta})\, \mathrm{d}\Pi(\sigma, b)}{\int \mu_{\sigma b}(X_0) \prod_{i=1}^n p_{\Delta,\sigma b}(X_{(i-1)\Delta}, X_{i\Delta})\, \mathrm{d}\Pi(\sigma, b)}.$$

Direct evaluation is out of reach, since the transition probabilities depend in an analytically intractable, non-linear way on σ, b.

TU Delft

Sampling from the Posterior Distribution

Papaspiliopoulos, Pokern, Roberts & Stuart (2012) showed how one can sample from the posterior distribution when $\sigma = 1$ (or parametric) and the prior on b comes from a Gaussian process. One uses conjugacy under continuous sampling, combined with a 'latent' variables sampling idea.

Can this 'work', particularly if the prior only models the regularity of σ, b – so is ignorant of the 'inverse problem'? The same question can be asked about many similar Bayesian 'solutions' of inverse problems (Stuart (2010)).

Frequentist Posterior Contraction Rates for Inverse Problems

- Following the program of van der Vaart, Ghosal et al., one can ask whether the posterior distribution contracts about the 'true value' (σ_0, b_0) at the right rate. Do we have, for large enough $M > 0$ that

$$\Pi\Big((\sigma, b) : n^{s/(2s+3)}\|\sigma - \sigma_0\| + n^{(s-1)/(2s+3)}\|b - b_0\| > M|X_0, \dots, X_{n\Delta}\Big) \to 0$$

in $\mathbb{P}_{\sigma_0 b_0}$-probability as $n \to \infty$?

Frequentist Posterior Contraction Rates for Inverse Problems

- Following the program of van der Vaart, Ghosal et al., one can ask whether the posterior distribution contracts about the 'true value' (σ_0, b_0) at the right rate. Do we have, for large enough $M > 0$ that

$$\Pi\left((\sigma, b) : n^{s/(2s+3)}\|\sigma - \sigma_0\| + n^{(s-1)/(2s+3)}\|b - b_0\| > M | X_0, \dots, X_{n\Delta}\right) \to 0$$

  in $\mathbb{P}_{\sigma_0 b_0}$-probability as $n \to \infty$?
- For general linear inverse problems

$$Y = Af + \epsilon; \quad A : \mathbb{H}_1 \to \mathbb{H}_2 \text{ linear, compact,}$$

  with Gaussian white noise ϵ, results are available: see Knapik, van der Vaart & van Zanten (2011), Agapiou, Larsson & Stuart (2013) for the Gaussian conjugate setting, and Ray (2013) for a general approach.

Bayesian Estimation for Low-Frequency Observations

For nonlinear settings, very little is known. Particularly in the diffusion model with low-frequency observations only consistency in a weak topology (with $\sigma = 1$ known) has been proved so far (van der Meulen & van Zanten, 2013).

There are extensions to multidimensional diffusions (Gugushvili & Spreij, 2014) and to jump diffusions (Koskela, Spano & Jenkins, 2015).

All three papers assume $\sigma = 1$ known and show consistency in a weak topology.

Wavelet Series Priors I

ψ_{lk} boundary corrected Daubechies wavelets, $0 < \alpha < \beta < 1$,
$\mathcal{I} = \{(l,k) : \psi_{lk} \text{ supported in } [\alpha,\beta]\}$

Model diffusion coefficient σ by

$$\log(\sigma^{-2}(x)) = \sum_{(l,k)\in\mathcal{I}} \frac{2^{-l(s+1/2)}}{l^2} u_{lk}\psi_{lk}(x), \qquad u_{lk} \sim^{iid} U(-B,B).$$

Comments:

- Could replace uniform distributions $U(-B,B)$ by any distribution with bouded support and density bounded away from zero.
- Could truncate sum in l at $L_n \to \infty$ sufficiently fast.
- By connection between Hölder norms and wavelet series $\log(\sigma^{-2})$ is modelled as typical s-Hölder smooth function (with a 'convenient' log-factor).

TU Delft

Wavelet Series Priors II

Model invariant density μ through

$$H(x) = \sum_{(l,k)\in\mathcal{I}} \frac{2^{-l(s+3/2)}}{l^2} \bar{u}_{lk}\psi_{lk}(x), \qquad \bar{u}_{lk} \sim^{iid} U(-B,B),$$

$$\mu = e^H / \int e^H.$$

Drift coefficient b indirectly given by

$$2b = (\sigma^2)' + \sigma^2(\log \mu)'.$$

Overall Prior is given by $\Pi = \mathcal{L}(\sigma^2, ((\sigma^2)' + \sigma^2 H')/2)$.

Comments:

- Priors on b, σ^2 are not independent.
- Invariant density is modelled explicitly.

TU Delft

Assumptions on σ_0 and μ_0

We define the Hölder-type space

$$\mathcal{C}^t([0,1]) := \{f \in C([0,1]) : \|f\|_{\mathcal{C}^t} < \infty\}, \quad \text{where}$$

$$\|f\|_{\mathcal{C}^t} := \sum_{k=0}^{\lfloor t \rfloor} \|D^k f\|_\infty + \sup_{h>0} \sup_{x\in[0,1]} \frac{|D^{\lfloor t \rfloor} f(x+h) - D^{\lfloor t \rfloor} f(x)|}{h^{t-\lfloor t \rfloor} \log(1/h)^{-2}}.$$

Assume diffusion coefficient $\sigma_0 \in \mathcal{C}^s$ is of form

$$\log \sigma_0^{-2}(x) = \sum_{(l,k)\in\mathcal{I}} \tau_{lk}\psi_{lk}(x), \quad x \in [0,1], \qquad \text{with } 2^{l(s+1/2)} l^2 |\tau_{lk}| \leqslant B.$$

Assume invariant density $\mu_0 \in \mathcal{C}^{s+1}$ is of form

$$\log \mu_0(x) = \sum_{(l,k)\in\mathcal{I}} \nu_{lk}\psi_{lk}(x), \quad x \in [0,1], \qquad \text{with } 2^{l(s+3/2)} l^2 |\nu_{lk}| \leqslant B.$$

Contraction Theorem

For $s \geqslant 2$ we define Θ_s by

$$\left\{(\sigma, b) : \|\sigma\|_{\mathcal{C}^s} \leqslant D, \|b\|_{\mathcal{C}^{s-1}} \leqslant D, \inf_x \sigma(x) \geqslant d, \text{ boundary conditions}\right\}$$

Theorem

$(X_t : t \geqslant 0)$ reflected diffusion with $(\sigma_0, b_0) \in \Theta_s$. σ_0 and μ_0 as above. Π wavelet series prior. Then for all $0 < \alpha < \beta < 1$ there exists $\gamma > 0$ such that in the $L^2([\alpha, \beta])$-norm

$$\Pi\left((\sigma, b) : \begin{array}{l} n^{s/(2s+3)}\|\sigma^2 - \sigma_0^2\|_{L^2} > \log^\gamma n \quad \text{or} \\ n^{(s-1)/(2s+3)}\|b - b_0\|_{L^2} > \log^\gamma n \end{array} \middle| X_0, \dots, X_{n\Delta}\right) \to 0$$

in $\mathbb{P}_{\sigma_0 b_0}$-probability for $\Delta > 0$ fixed and $n \to \infty$.

TU Delft

Comments on Contraction Theorem

- The contraction theorem shows that the posterior distribution contracts about the true parameters at the minimax rate within $\log n$ factors.
- Note that the above prior does not require knowledge of the 'inverse problem' at all, in particular not the singular value decomposition of the operator.
- Bayes formula gives a (near-) optimal solution of this ill-posed non-linear inverse problem. It illustrates the power of the Bayesian approach to inverse problems.

Comments on the Conditions

- The additional logarithmic factor in the definition of $\mathcal{C}^s$ might change the minimax rate by a logarithmic factor $(\log n)^\eta$, $\eta > 0$.

- The assumption $\mu_0 \in \mathcal{C}^{s+1}$ is restricting (σ_0, b_0) beyond having to lie in Θ_s. As the lower bounds by GHR are for $\mu_0 \equiv 1 \in \mathcal{C}^{s+1}$ this does not affect the minimax rates.

- μ_0 assumed to be in $\mathcal{C}^{s+1}$ and μ modelled explicitely since information theoretic distance involves the term $\|\mu - \mu_0\|_{L^2}$.

TU Delft

General Contraction Theorem

The basic strategy follows Ghosal, Ghosh & van der Vaart (2000)

Small ball probability condition: C, L, r constants so that

$$\Pi\left(B_{\varepsilon_n,r}\right) \geqslant e^{-Cn\varepsilon_n^2},$$

and $\Pi(\mathcal{B}\backslash\mathcal{B}_n) \leqslant Le^{-(C+4)n\varepsilon_n^2}$ for some sequence $\mathcal{B}_n \subseteq \mathcal{B}$

Tests: Sequence of tests $\Psi_n \equiv \Psi(X_0, \ldots, X_{n\Delta})$ and of metrics d_n such that for $M > 0$ large enough,

$$\mathbb{E}_{\sigma_0 b_0}[\Psi_n] \to_{n\to\infty} 0, \qquad \sup_{(\sigma,b)\in\mathcal{B}_n : d_n((\sigma,b),(\sigma_0,b_0))\geqslant M\varepsilon_n} \mathbb{E}_{\sigma b}[1-\Psi_n] \leqslant Le^{-(C+4)n\varepsilon_n^2}.$$

Give posterior contraction: Then the posterior $\Pi(\cdot|X_0, \ldots, X_{n\Delta})$ satisfies

$$\Pi((\sigma, b) : d_n((\sigma, b), (\sigma_0, b_0)) > M\varepsilon_n | X_0, \ldots, X_{n\Delta}) \to 0$$

in $\mathbb{P}_{\sigma_0 b_0}$-probability, as $n \to \infty$.

Small Ball Probability Condition

$\mathcal{B} \subseteq \Theta$ with a σ-field $\mathcal{S}$, Π prior distribution on $\mathcal{S}$, $(\sigma_0, b_0) \in \Theta$, $\varepsilon_n \to 0$, $\sqrt{n}\varepsilon_n \to \infty$, and C, r fixed constants
Suppose Π satisfies

$$\Pi\left(B_{\varepsilon_n, r}\right) \geqslant e^{-Cn\varepsilon_n^2},$$

where

$$B_{\varepsilon,r} = \Bigg\{(\sigma, b) \in \mathcal{B} : \mathrm{KL}((\sigma_0, b_0), (\sigma, b)) \leqslant \varepsilon^2,$$
$$\mathrm{Var}_{\sigma_0 b_0}\left(\log \frac{p_{\sigma b}(\Delta, X_0, X_\Delta)}{p_{\sigma_0 b_0}(\Delta, X_0, X_\Delta)}\right) \leqslant 2\varepsilon^2,$$
$$\mathrm{KL}(\mu_{\sigma_0 b_0}, \mu_{\sigma b}) \leqslant r, \mathrm{Var}_{\sigma_0 b_0}\left(\log \frac{\mu_{\sigma b}(X_0)}{\mu_{\sigma_0 b_0}(X_0)}\right) \leqslant 2r\Bigg\}.$$

with transition density $p_{\sigma b}$ and invariant density $\mu_{\sigma b}$.

TU Delft

Bound on Information Theoretic Distance

Information theoretic distance

$$\mathsf{KL}((\sigma_0, b_0), (\sigma, b)) := \mathbb{E}_{\sigma_0 b_0}\left[\log\left(\frac{p_{\sigma_0 b_0}(\Delta, X_0, X_\Delta)}{p_{\sigma b}(\Delta, X_0, X_\Delta)}\right)\right],$$

$p_{\sigma b}$ transition density, expectation $\mathbb{E}_{\sigma_0 b_0}$ w.r.t. stationary distribution

Need good bound on KL:

$$\begin{aligned}\mathsf{KL}((\sigma_0, b_0), (\sigma, b)) &\lesssim \|p_{\sigma b} - p_{\sigma_0 b_0}\|_{L^2}^2 \\ &\lesssim \|P_\Delta^{\sigma b} - P_\Delta^{\sigma_0 b_0}\|_{HS}^2 \\ &\lesssim \|e^{\Delta/L_{\sigma b}^{-1}} - e^{\Delta/L_{\sigma_0 b_0}^{-1}}\|_{HS}^2 \\ &\lesssim \|L_{\sigma b}^{-1} - L_{\sigma_0 b_0}^{-1}\|_{HS}^2,\end{aligned}$$

where $P_\Delta^{\sigma b}$ transition operator and $L_{\sigma b}$ infinitesimal generator.

TUDelft

Bound on Information Theoretic Distance II

Inverse of infinitesimal generator

$$L_{\sigma b}^{-1} f(x) = \int K_{\sigma b}(x,z) f(z) \mu_0(z) \, \mathrm{d}z$$

Bound distance between integral kernels

$$\begin{aligned} \mathrm{KL}((\sigma_0, b_0), (\sigma, b)) &\lesssim \| L_{\sigma b}^{-1} - L_{\sigma_0 b_0}^{-1} \|_{HS}^2 \\ &\lesssim \int \int (K_{\sigma b} - K_{\sigma_0 b_0})^2 (x,z) \mu_0(x) \mu_0(z) \, \mathrm{d}x \, \mathrm{d}z \\ &\lesssim \| \mu_{\sigma b} - \mu_{\sigma_0 b_0} \|_{L^2([0,1])}^2 + \left\| \frac{1}{\sigma^2} - \frac{1}{\sigma_0^2} \right\|_{(B_{1\infty}^1)^*}^2 + \| b - b_0 \|_{(B_{1\infty}^2)^*}^2 , \end{aligned}$$

with dual spaces of Besov spaces $B_{1\infty}^1$ and $B_{1\infty}^2$.

Concentration of Frequentist Estimators and Tests

• A Birgé-Le Cam Hellinger testing theory like the one used in Ghosal, Ghosh, van der Vaart, is not available for (non-linear) inverse problems.

• Instead we use a 'concentration of measure approach' to such tests, put forward in Giné & Nickl (2011). In the present setting, for $\hat{\sigma}$ and $\hat{b}$ estimators by Gobet, Hoffmann & Reiß (2004) we can prove:

Theorem

There exists $R > 0$ such that for n large enough we have uniformly over Θ_s, $s \geqslant 2$,

$$\mathbb{P}\left(\begin{array}{c} \|\hat{\sigma}^2 - \sigma^2\|_{L^2([\alpha,\beta])} \geqslant Rn^{-s/(2s+3)} \quad \text{or} \\ \|\hat{b} - b\|_{L^2([\alpha,\beta])} \geqslant Rn^{-(s-1)/(2s+3)} \end{array} \right) \leqslant \exp\left(-Dn^{1/(2s+3)}\right).$$

This means exponential concentration of $\hat{\sigma}^2$ and $\hat{b}$ at minimax rates $n^{-s/(2s+3)}$ and $n^{-(s-1)/(2s+3)}$, respectively.

TUDelft

Concentration Inequality

Bernstein-type inequality

There exists $\kappa > 0$ such that for all reflected diffusions $\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t$, $t \in [0, \infty)$ with $(\sigma, b) \in \Theta := \Theta_2$ and arbitrary initial distribution, $\forall f : [0,1] \to \mathbb{R}$ bounded, $\forall x > 0$ and $\forall n \in \mathbb{N}$,

$$\mathbb{P}\left(\Big|\sum_{j=0}^{n-1}(f(X_{j\Delta}) - \mathbb{E}_\mu[f])\Big| > x\right)$$

$$\leqslant \kappa \exp\left(-\frac{1}{\kappa}\min\left(\frac{x^2}{n\|f\|_{L^2(\mu)}^2}, \frac{x}{\log(n)\|f\|_\infty}\right)\right).$$

TUDelft

Concentration Inequality for Suprema of Empirical Processes

Class of functions $\mathcal{F} = \{f_i : i \in I\}$ with $0 \in \mathcal{F}$ and $\dim I = d$

$V^2 = \kappa n \sup_{f \in \mathcal{F}} \|f\|^2_{L^2(\mu)}$ and $U = \kappa \log n \sup_{f \in \mathcal{F}} \|f\|_\infty$

Theorem

For $\tilde{\kappa} = 18$ and for all $x \geqslant 0$ we have

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{j=0}^{n-1} (f(X_{j\Delta}) - \mathbb{E}_\mu[f]) \right| \geqslant \tilde{\kappa} \left(V\sqrt{d+x} + U(d+x) \right) \right) \leqslant 2\kappa e^{-x}.$$

Follows from chaining and previous concentration inequality.

Using duality arguments from Giné & Nickl (2011) this gives deviation bounds for the estimation errors of frequentist estimators of σ^2, b.

Concentration inequality builds on results by Adamczak (2008) for Markov chains based on regeneration approach.

Lessons for General Non-Linear Inverse Problems $Y = Af + \epsilon$

Bayesian methods for inverse problems should work in principle.
Proving that may be quite difficult though!

Two key modifications of the standard Ghosal-Ghosh-van der Vaart approach are required:

• If A is the operator to invert (possibly after linearisation), one needs to show that the information distance is bounded above by $\|Af\|$ where $\|\cdot\|$ would be the information distance when $A = Id$. This allows to take 'faster' ϵ_n-sequences in the small ball computations. In our case the main contribution is to achieve this by considering negative Besov norms on (σ, b).

• In absence of robust Hellinger tests, one can show that for a large support set in the prior a frequentist estimator that solves the inverse problem admits tight sub-Gaussian exponential concentration bounds on its estimation error, which can be used in the construction of tests. In our case we had to derive new concentration inequalities for samples means of discretely sampled diffusions.

References

Ghosal S., Ghosh, J.K. and van der Vaart, A.W. (2000): Convergence rates of posterior distributions. *Ann. Statist.* 28, 500–531.

Giné, E., Nickl, R. (2011): Rates of contraction for posterior distributions in L^r-metrics. *Ann. Statist.* 39, 2883-2911.

Gobet, E., Hoffmann, M. and Reiß, M. (2004): Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Statist.* 32(5), 2223-2253.

Nickl, R. and Söhl, J. (2016): Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.*, to appear.

Papaspiliopoulos, O., Pokern, Y., Roberts, G. O. and Stuart, A. M. (2012): Nonparametric estimation of diffusions: a differential equations approach. *Biometrika* 99, 511-531.

Ray, K. (2013): Bayesian inverse problems with non-conjugate priors. *Elect. J. Stat.* 7, 2516-2549.

van der Meulen, F. and van Zanten, H. (2013): Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. *Bernoulli* 19, 44-63.

Thank you for your attention!