Nonparametric estimation of the drift in stochastic differential equations

Manfred Opper, Andreas Ruttor, Philip Batz,

May 22, 2015

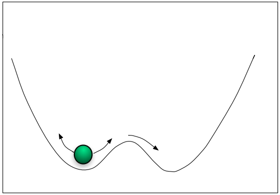
The problem

- Dynamics defined by set of ODEs

$$\frac{dX}{dt} = f(X)$$

  with $X \in R^d$.
- Learn the function $f(\cdot)$ from a set of noise free observations $x(t_1), x(t_2), \ldots, x(t_n)$.

Example with $f(x) \approx x - x^3$



The problem with noise

- In order to explore the 'phase space' add white noise noise $\rightarrow$ SDE

$$dX_t = \underbrace{f(X_t)}_{\text{Drift}} dt + \underbrace{D^{1/2}(X_t)}_{\text{Diffusion}} \times \underbrace{dW_t}_{\text{Wiener process}}$$

  Limit of discrete time process

$$X_{t+\Delta} - X_t = f(X_t)\Delta + D^{1/2}(X_t)\sqrt{\Delta}\,\epsilon_t \ .$$

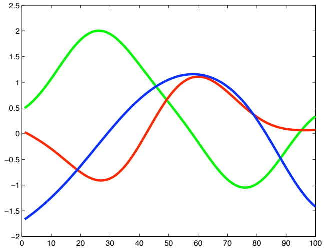
  with ϵ_t i.i.d. Gaussian for $\Delta \rightarrow 0$.
- 'Learn' the function $f(\cdot)$ from a set of noise free observations $x_{t_1}, \ldots, x_{t_n}$.

Outline

- Nonparametric estimates using an EM algorithm
- Nonparametric estimates using a pseudo–Bayesian approach

Nonparametric (Gaussian process) approach

- Learn the function $f(\cdot)$ under smoothness assumptions !
- **Idea** (Papaspilioupoulis, Pokern, Roberts & Stuart (2012), Pokern, Stuart & van Zanten (2013):
  Use a Gaussian Process prior distribution $p(f)$ with covariance kernel $K(x, x')$ over functions $f(\cdot)$.



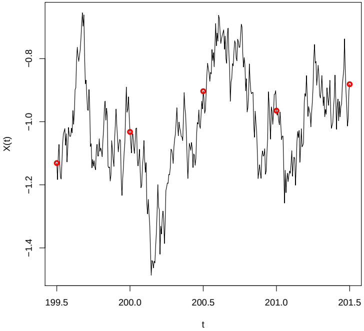
Densely observed path

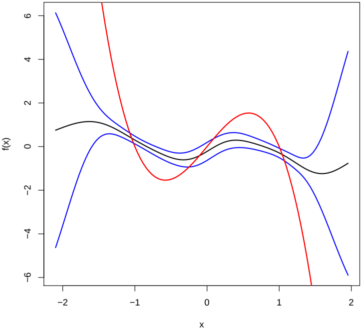
- In Euler discretization the SDE looks like this $X_{t+\Delta t} - X_t = f(X_t)\Delta + \sqrt{\Delta}\,\epsilon_t$, for $\Delta \to 0$.
- Hence the likelihood for the drift is (with a **densely observed** path $X_{0:T}$) is

$$p(X_{0:T}|f) \propto \exp\left[-\frac{1}{2\Delta}\sum_t ||X_{t+\Delta} - X_t||^2\right] \times$$
$$\exp\left[-\frac{1}{2}\sum_t ||f(X_t)||^2\,\Delta + \sum_t f(X_t)\cdot(X_{t+\Delta} - X_t)\right].$$

  allows for simple GP based estimation of the function $f(\cdot)$.
- This essentially leads to the estimate $f(x) \approx E\left[\frac{X_{t+\Delta}-X_t}{\Delta}|X_t = x\right]$. OK for $\Delta \to 0$.

For not so small Δ it does not work well ! Data from
$dx = (x - x^3)dt + dW$.





EM algorithm

Treat X_t for times t between observations as a **latent** random variables (Batz, Ruttor & Opper NIPS 2013).

1. **E-step**: Compute expected complete data likelihood

$$\mathcal{L}(f, f_{old}) = -E_{p_{old}} \left[\ln L(X_{0:T}|f)\right]$$

   where p_{old} = posterior $p(X_{0:T}|\mathbf{y})$ computed with the previous estimate f_{old} of the drift.

2. **M-Step**: Recompute the MAP drift function as

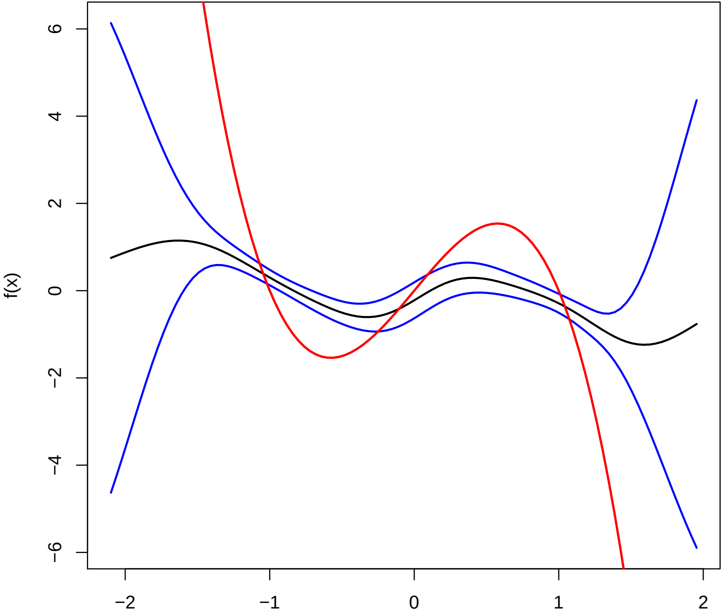
$$f_{new} = \arg\min_f \left(\mathcal{L}(f, f_{old}) - \ln P_0(f)\right) \qquad (1)$$

Problems

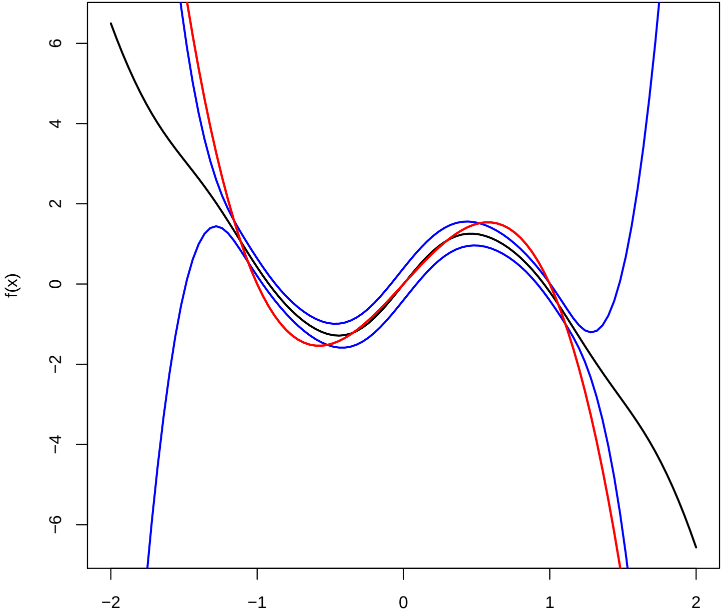
1. E-step requires posterior marginal densities q_t for diffusion processes with *arbitrary* prior drift functions $f(x)$.
2. GP has to deal with an infinite amount of densely imputed data.

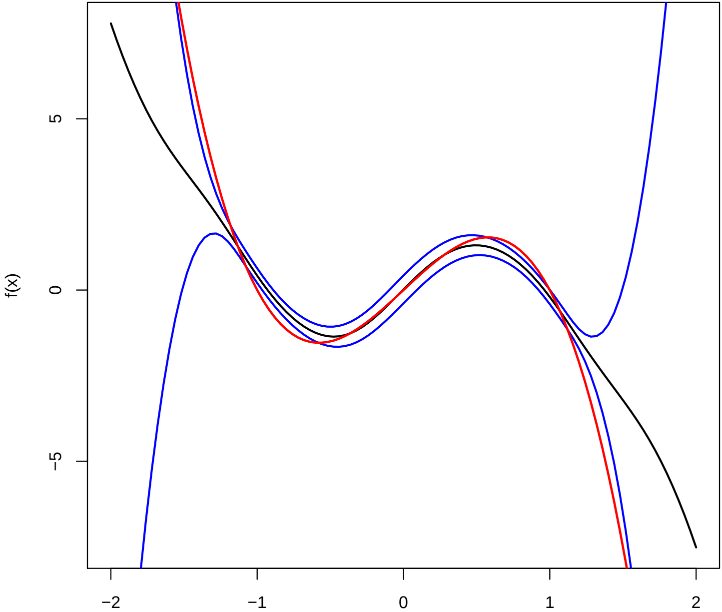
Approximate solutions

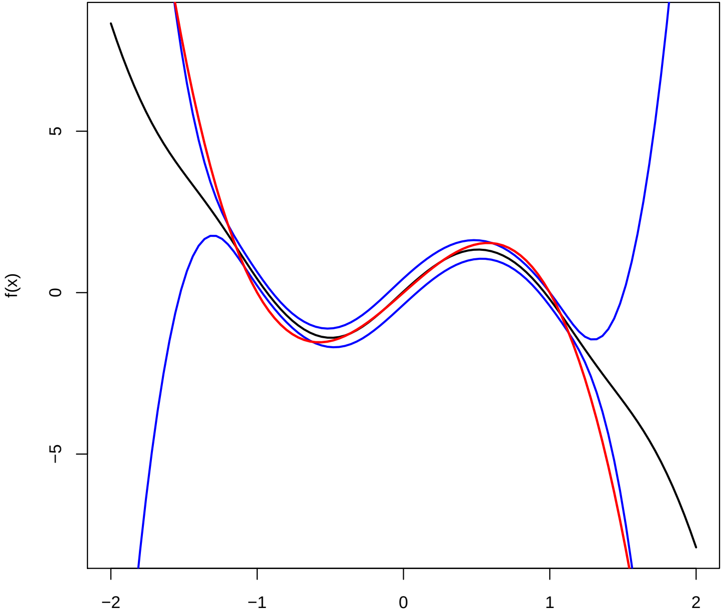
1. Linearize drift between consecutive observations (Ornstein Uhlenbeck bridge).
2. Work with sparse GP approximation.



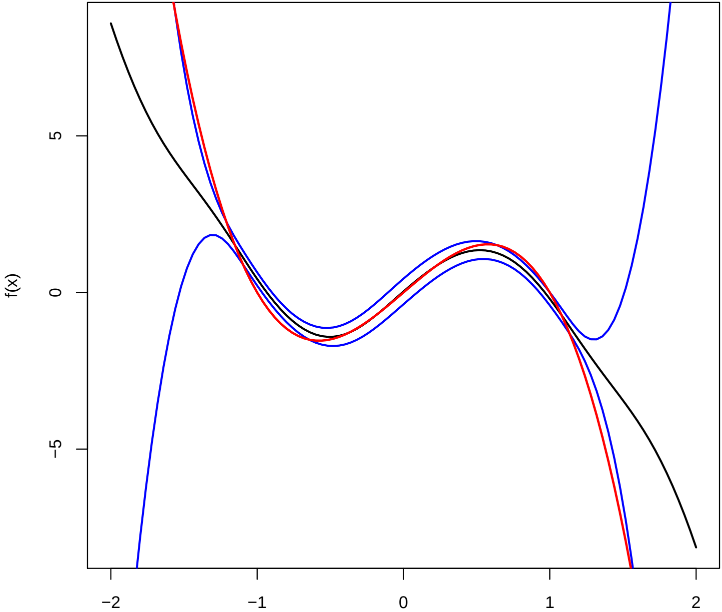
f(x)
6
4
2
0
−2
−4
−6
−2
−1
0
1
2

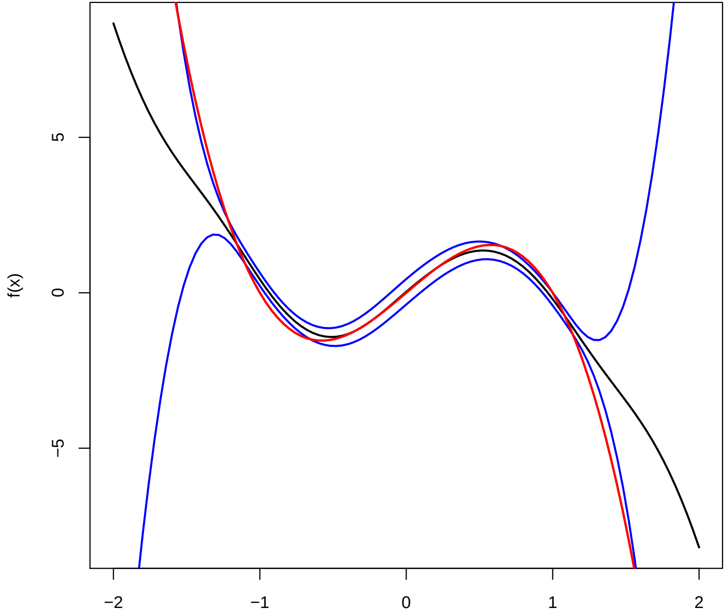

f(x)
6
4
2
0
−2
−4
−6
−2
−1
0
1
2

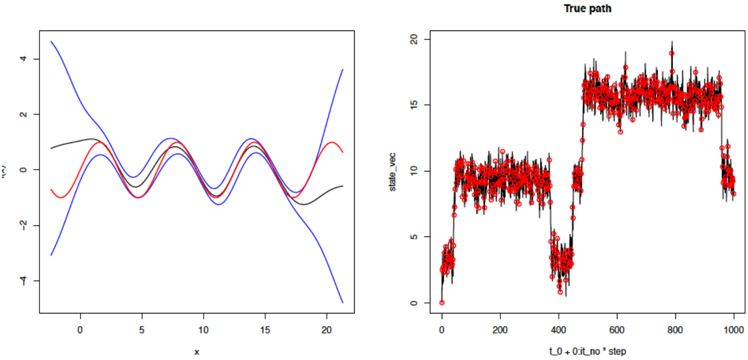

f(x)
5
0
−5
−2
−1
0
1
2



f(x)
5
0
−5
−2
−1
0
1
2


f(x)
5
0
−5
−2
−1
0
1
2


f(x)
5
0
−5
−2
−1
0
1
2



True path
x
state_vec
t_0 + 0:it_no * step

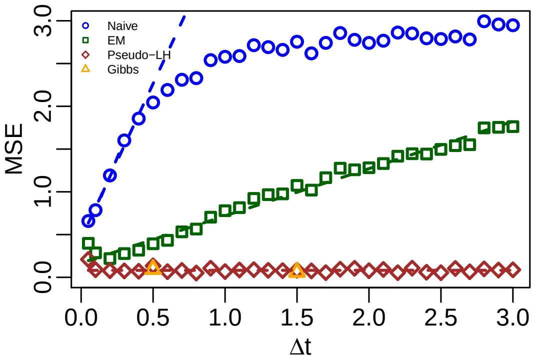


Figure: Comparison of the MSE for different methods for double well over different time intervals.

- Get rid of the Δt dependence.
- Find a method that works just with the empirical density of data !
- Of course this can't always work (potential conditions)

Basic idea: 1-D

- The stationary density for a SDE with drift $f(\cdot)$ fulfils the Fokker–Planck equation

$$-\frac{d}{dx}\left(f(x)p(x)\right) + \frac{\sigma^2}{2}\frac{d^2}{dx^2}p(x) = 0$$

Basic idea: 1-D

- The stationary density for a SDE with drift $f(\cdot)$ fulfils the Fokker–Planck equation

$$-\frac{d}{dx}\left(f(x)p(x)\right) + \frac{\sigma^2}{2}\frac{d^2}{dx^2}p(x) = 0$$

- Consider the functional

$$\varepsilon[f] = \int p(z)\left\{f^2(z) + \sigma^2 f'(z)\right\} dx,$$

- Minimisation wrt f yields

$$2f_*(z)p(z) - \sigma^2 p'(z) = 0, \tag{2}$$

- Replace $\varepsilon[f]$ by empirical estimate 'pseudo–likelihood'

$$\frac{1}{n}\sum_{i=1}^{n}\left\{f^2(z_i) + \sigma^2 f'(z_i)\right\}$$

- Use a GP prior over f for regularization.

2nd order stochastic differential equations

$$
\begin{aligned}
dX_t &= V_t dt \\
dV_t &= \{f(X_t, V_t) + r(X_t, V_t)\}dt + D_v^{1/2}(X_t, V_t)dW_t
\end{aligned}
$$

- Assume r is known and $f = \nabla_v \phi(x, v)$ (integrability condition).
- Let $p(x, v)$ be the stationary density of the SDE and $\mathcal{L}_0^\dagger$ the generator of the SDE for $f = 0$.
- The minimisation of the functional wrt ϕ

$$
\varepsilon[\phi] = \int p(x, v) \left\{ \mathcal{L}_0^\dagger \phi(x, v) + \frac{1}{2} |\nabla_v \phi(x, v)|^2 \right\} dxdv
$$

leads to the Fokker–Planck equation $\mathcal{L}_0 p - \nabla_v(fp) = 0$ for the stationary density, where $\mathcal{L}_0$ is the adjoint operator.

- Approximate $\varepsilon[\phi]$ by sample (x_i, v_i), $i = 1, \ldots, n$ drawn at random from $\phi(x, v)$ and minimize approximate functional

$$\varepsilon_{emp}[\phi] = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathcal{L}_0^\dagger \phi(x_i, v_i) + \frac{1}{2}|\nabla_v \phi(x_i, v_i)|^2\right\}$$

regularised by kernel method – equivalent to Gaussian process regression.

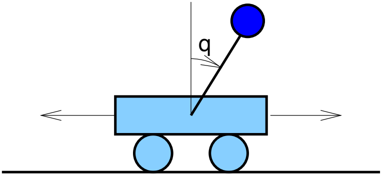
- Can be applied to models with $f(x, v) = f(x) - \Gamma v$.

- Approximate $\varepsilon[\phi]$ by sample (x_i, v_i), $i = 1, \ldots, n$ drawn at random from $\phi(x, v)$ and minimize approximate functional

$$\varepsilon_{emp}[\phi] = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathcal{L}_0^{\dagger}\phi(x_i, v_i) + \frac{1}{2}|\nabla_v \phi(x_i, v_i)|^2\right\}$$

regularised by kernel method – equivalent to Gaussian process regression.
- Can be applied to models with $f(x, v) = f(x) - \Gamma v$.
- If Γ is known, $\mathcal{L}_0^{\dagger}\phi(x, v)$ is independent of the diffusion $D(x, v)$!

Cart and pole



$$\begin{aligned} dX &= Vdt \\ dV &= \left(-\frac{g}{l}\sin X - \frac{\lambda}{ml^2}V\right)dt + \sigma \cos X dW, \end{aligned}$$

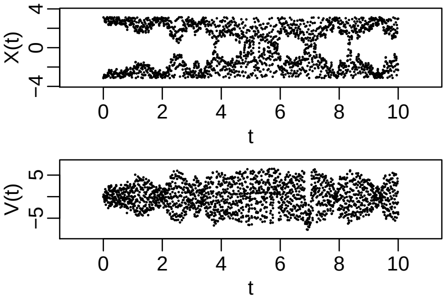


Figure: Full sample path of the Cart and pole model.

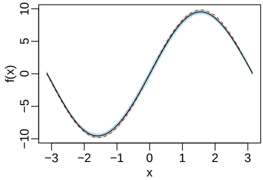


Figure: Estimated drift function for Cart and pole.

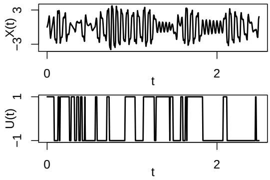
Other noise models

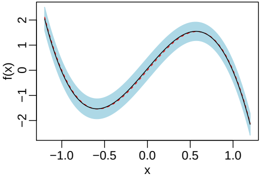
- Generalise to generators $\mathcal{L}_0^\dagger$ of other Markov processes
- Example: Model with Telegraph process $U_t = \pm 1$

$$\begin{aligned} dX &= V dt \\ dV &= (f(X) - \lambda V + U) dt \\ U &\sim \mathcal{TP} \end{aligned}$$

Equation for stationary density is derived from the functional

$$\varepsilon[f] = \frac{1}{2} \sum_u \int \left\{ f^2(x) + 2f'(x)v^2 + 2f(x)(u - \lambda v) \right\} p(x, v, u) dx dv$$





Naw, tha' ain't Bayes

- Can in some cases be reduced to a true likelihood in the limit of densely observed data
- But work well also for completely independent samples !
- Model selection: frequentist with hold out data
- GP error Bayes not useful.