Nonparametric Estimation for Stochastic Dynamical Systems

Harish S. Bhat

Applied Mathematics Unit
University of California, Merced

April 16, 2018

Basic Question

How can we estimate an SDE from vector-valued time series?

- Consider stochastic differential equation (SDE) in $\mathbb{R}^d$:

$$dX_t = f(X_t)dt + \Gamma dW_t.$$

- Here W_t is Brownian motion in $\mathbb{R}^d$.
- Suppose we have observations of X_t at times $t_n = nh$.
- How can we estimate both f and Γ?
- Here f is a vector field in $\mathbb{R}^d$ and Γ is a constant, $d \times d$ matrix.

Motivation

Why do we want to do this?

- Abundance of time series in new problem domains without canonical equations of motion.
- Automated model discovery complements traditional modeling efforts.
- SINDy (Brunton, Proctor, Kutz [PNAS '16]) is elegant, scalable procedure for discovering systems of ODEs from time series.
- However, as signal/noise ratio drops, SINDy becomes less effective.

Automated Discovery of Stochastic Models

This topic has attracted significant interest:

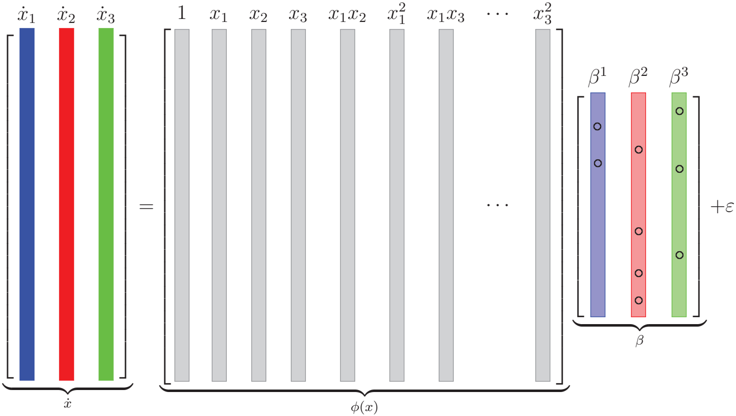
- Numerical optimization of likelihood function using adjoint method to compute gradients: Bhat and Madushani [DSAA '16]
- Variational mean field Gaussian process approximations: Archambeau, Opper, Shen, Cornford, Shawe-Taylor [NIPS '07]; Ruttor, Batz, Opper [NIPS '13]; Vrettas, Opper, Cornford [PRE '15]
- Diffusion bridge with reversible jump MCMC: van der Meulen, Schauer, van Zanten [CSDA '14]; van der Meulen, Schauer, van Waaij [SISP '17]
- Transforming SDE into random ODE: Bauer, Gorbach, Miladinovic, Buhmann [NIPS '17]

Goals

What I will do here:

- Provide probabilistic derivation of SINDy.
- Derive SDE nonparametric estimation method that combines SINDy with:
  - EM (expectation maximization) and
  - MCMC (Markov Chain Monte Carlo) diffusion bridge sampling.

SINDy (Sparse Identification of Nonlinear Dynamics)



$$\frac{dx}{dt} = \phi(x)\beta + \varepsilon$$

Idea: solve for β via sparsity-promoting (least squares) regression.

SINDy (Sparse Identification of Nonlinear Dynamics)

More Details

- Let $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$.
- Suppose there are $N + 1$ observations of x and p columns of $\phi(x)$.
- The j-th column of $\phi(x)$ is a function

$$\phi_j : \mathbb{R}^d \to \mathbb{R}.$$

- The user prescribes the ϕ_j's. In the example, we used polynomial functions of the coordinates; other choices are fine.
- Reduce problem to finding "expansion coefficients" β:

$$\frac{dx_i}{dt} = f_i(x) = \sum_{j=1}^{p} \beta_j^i \phi_j(x) + \varepsilon_i$$

Notation

Let $\mathcal{N}(\mu, \Sigma)$ denote a multivariate Gaussian/normal in $\mathbb{R}^d$ with:

- mean vector $\mu \in \mathbb{R}^d$, and
- $d \times d$ covariance matrix Σ.

Probabilistic Derivation of SINDy

Step 1: Discretize in Time

We return to the SDE

$$dX_t = f(X_t)dt + \Gamma dW_t.$$

Assume Γ is diagonal. Discretize in time (for now, using Euler):

$$\widetilde{X}_{n+1} = \widetilde{X}_n + f(\widetilde{X}_n)h + \Gamma h^{1/2} Z_{n+1}$$

where $Z_{n+1} \sim \mathcal{N}(0, I)$, independent from $\widetilde{X}_n$. Note that

$$[\widetilde{X}_{n+1} | \widetilde{X}_n = v] \sim \mathcal{N}(v + f(v)h, h\Gamma^2).$$

Now use expansion of f from earlier and assume coefficients are given:

$$[\widetilde{X}_{n+1} | \widetilde{X}_n = v, \beta] \sim \mathcal{N}(v + \phi(v)\beta h, h\Gamma^2).$$

Probabilistic Derivation of SINDy

Step 2: Markov property

The overall log likelihood breaks into pieces:

$$\log p(x_0, x_1, \ldots, x_N|\beta) = \sum_{n=1}^{N} \log p(x_n|x_{n-1}, \beta)$$

Identifying x_n with $\widetilde{X}_n$, we can use the Gaussian to write:

$$\log p(x_0, x_1, \ldots, x_N|\beta) = \sum_{n=1}^{N} \left[\sum_{i=1}^{d} -\frac{1}{2} \log(2\pi h \gamma_i^2) \right]$$
$$-\frac{1}{2h}(x_n - x_{n-1} - h \sum_{j=1}^{p} \phi_j(x_{n-1})\beta_j)^T \Gamma^{-2} (x_n - x_{n-1} - h \sum_{\ell=1}^{p} \phi_j(x_{n-1})\beta_\ell)$$

Probabilistic Derivation of SINDy

Step 3: Maximize Likelihood over β

Now we set $\Gamma = \sigma I$. The log likelihood becomes

$$\log p(x|\beta) = -\frac{Nd}{2}\log(2\pi h\sigma^2) - \frac{1}{2h\sigma^2}\sum_{n=1}^{N}\left\| x_n - x_{n-1} - h\sum_{j=1}^{p}\phi_j(x_{n-1})\beta_j \right\|_2^2.$$

Maximizing the log likelihood over β is equivalent to minimizing

$$\sum_{n=1}^{N}\left\| \frac{x_n - x_{n-1}}{h} - \sum_{j=1}^{p}\phi_j(x_{n-1})\beta_j \right\|_2^2.$$

Except that dx/dt has been discretized in time, this is precisely the same as the SINDy regression problem described above.

Lessons/Connections Learned

- Quality of derivative approximation tied to quality of time-discretization of SDE.
- SINDy appears to ignore the covariance Γ entirely. I view this as hidden price paid for SINDy's simplicity and scalability.
- Could return to end of Step 2 and maximize likelihood over both β and Γ.
- Iterative maximization over β and Γ in turn is iteratively reweighted least squares in Statistics.

Ongoing/future work: build in priors, sample from posterior $p(\beta|x)$, and thereby quantify uncertainty in β.

Extending this to SDE

Are we done?

- We just showed that if we change SINDy slightly and estimate both β and Γ, we've found an SDE model for the data!
- Caveats:
  1. Quality of approximation is controlled by h.
     As $h \downarrow 0$, the Euler transition density
     $$\widetilde{x}_{n+1} | \widetilde{x}_n = v$$
     will approach the true transition density of the SDE,
     $$X((n+1)h) | X(nh) = v.$$
  2. When h is small, above small modification to SINDy should work.
  3. What h is large, need additional ideas.

Large Interobservation Times

Idea

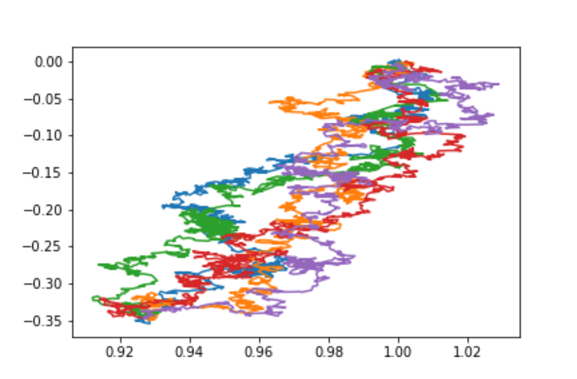
- Consider data points $\{x_n\}_{n=0}^N$ collected at times $t_n = n\Delta t$.
- When Δt is large, treat the fine scale data as missing data and use expectation maximization (EM).
- In the "E" step, we sample repeatedly from a diffusion bridge connecting x_n to x_{n+1} on a fine time scale h.
- In this way, we can compute *expected* log likelihoods and then maximize them in the "M" step.

Concrete Example

Noisy Pendulum

$$\dot{\mathbf{x}} = \begin{bmatrix} x_2 \\ -\sin x_1 \end{bmatrix} + \Gamma d\mathbf{W}_t$$

$\Gamma = \text{diag}(.075, .1)$. $\mathbf{x}(0) = (1, 0)$, $\mathbf{x}(T) = (0.927, -0.347)$, and $T = 0.418$. Bridges joining the two points with $h = T/1000$:



Expectation Maximization in Detail

Let z denote the missing fine scale data connecting each x_{n-1} to the next x_n. Then (x, z) is the *completed data* consisting of coarse-scale measurements augmented at a fine scale.

Algorithm to find parameters $\theta = (\beta, \gamma)$

1. Start with an initial guess $\theta^{(0)}$.
2. Define
$$Q(\theta, \theta^{(k)}) = E_{z|x,\theta^{(k)}}[\log p(x, z|\theta)].$$
3. Maximize Q over θ, i.e., set
$$\theta^{(k+1)} = \text{argmax}_\theta\, Q(\theta, \theta^{(k)}).$$

It will turn out that we can solve this maximization problem without numerical optimization. (Least squares!)

E-Step

Let $z^{(r)}$ denote the r-th diffusion bridge sample path,

$$z^{(r)} \sim z|x, \theta^(k).$$

Let $y = (x, z)$ denote the completed data. With R bridge samples, we approximate the expected value as follows:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \mathbb{E}_{z|x,\boldsymbol{\theta}^{(k)}}[\log p(x, z \mid \boldsymbol{\theta})] \approx \frac{1}{R} \sum_{r=1}^{R} \Bigg[\sum_{j=1}^{N} \left[\sum_{i=1}^{d} -\frac{1}{2} \log(2\pi h \gamma_i^2) \right]$$

$$-\frac{1}{2h}(y_j^{(r)} - y_{j-1}^{(r)} - h \sum_{k=1}^{M} \phi_k(y_{j-1}^{(r)}) \beta_k)^T \Gamma^{-2} (y_j^{(r)} - y_{j-1}^{(r)} - h \sum_{\ell=1}^{M} \phi_\ell(y_{j-1}^{(r)}) \beta_\ell) \Bigg]$$

M-Step

To maximize Q over θ, we first assume Γ is known and maximize over β. The solution is given by forming the matrix

$$\mathcal{M}_{k,\ell} = \frac{1}{R}\sum_{r=1}^{R}\sum_{j=1}^{N} h\phi_k^T(y_{j-1}^{(r)})\Gamma^{-2}\phi_\ell^T(y_{j-1}^{(r)})$$

and the vector

$$\rho_k = \frac{1}{R}\sum_{r=1}^{R}\sum_{j=1}^{N} \phi_k^T(y_{j-1}^{(r)})\Gamma^{-2}(y_j^{(r)} - y_{j-1}^{(r)}).$$

We then solve the system

$$\mathcal{M}\beta = \rho$$

for β. Now that we have β, we maximize Q over γ and get:

$$\gamma_i^2 = \frac{1}{RNh}\sum_{r=1}^{R}\sum_{j=1}^{N}((y_j^{(r)} - y_{j-1}^{(r)} - h\sum_{\ell=1}^{M}\beta_\ell\phi_\ell(y_{j-1}^{(r)}))\cdot e_i)^2$$

where e_i is the i^{th} canonical basis vector in $\mathbb{R}^d$.

The Above in Plain English

- Given coarse data, bridge the data R times.
- Each time you bridge the data, calculate the regression matrices/vectors that you would've calculated in SINDy.
- Now average the R regression matrices/vectors and solve the resulting system for β.
- Do the same for γ.

Upshot

Properties of EM

- Can prove that from one step to the next, the algorithm monotonically increases both Q and the true (incomplete) log likelihood.
- Hence, starting from guess $\theta^{(0)}$, we use EM to evolve

$$\theta^{(k+1)} = \mathrm{argmax}_{\theta}\, Q(\theta, \theta^{(k)})$$

Then, as $k \to \infty$,

$$\theta^{(k)} \to \theta^*,$$

a local maximizer of the likelihood function.

Key Technical Idea

MCMC Diffusion Bridge Sampling

We implement diffusion bridge using MCMC approach that goes back at least to Roberts and Stramer [2001]. Here is the idea:

- Given Γ and two points (x_0, x_1) at times $(0, \Delta t)$, sample from the Brownian bridge process

$$dY_t = \frac{x_1 - Y_t}{\Delta t - t} dt + \Gamma dW_t$$

  with initial condition $Y_0 = x_0$. Then $Y_{\Delta t} = x_1$.

- We treat this sample as a proposal, Z_t^*. Suppose the previous MCMC bridge was Z_t^{old}.

Key Technical Idea

MCMC Diffusion Bridge Sampling

- By Girsanov's theorem, the Metropolis accept probability is

$$\min\left\{1, \frac{\psi(Z_t^*)}{\psi(Z_t^{\text{old}})}\right\}$$

with

$$\log\psi(z) = \Gamma^{-2}\underbrace{\int_0^{\Delta t} f(Z_s)dZ_s}_{\text{stochastic}} - \frac{1}{2}\Gamma^{-2}\underbrace{\int_0^{\Delta t} f(Z_s)^2\, ds}_{\text{ordinary}}.$$

- Both the stochastic and ordinary integrals can be easily evaluated numerically.
- In practice, we obtain reasonable acceptance rates in this way.

Ongoing/Future Work

Preliminary results are promising! We plan to submit to NIPS '18.

- Testing the EM+bridge approach on stochastic versions of nonlinear dynamical systems in $d = 3$.
- Quantifying relationship between magnitude of noise and quality of reconstruction.
- Applying these techniques to real data.
- Comparing against competing approaches.

Email me at hbhat@ucmerced.edu if you'd like any notes or code on this.

Support from NSF DMS-1723272 gratefully acknowledged.