Variational Inference for Diffusion Processes

Cédric Archambeau

Xerox Research Centre Europe
cedric.archambeau@xerox.com

Joint work with Manfred Opper.

Statlearn '11
Grenoble, March 2011

Stochastic differential systems

Many real dynamical systems are continuous in time:

- Data assimilation (e.g. numerical weather prediction)
- Systems biology (e.g. cellular stress response, transcription factors)
- fMRI brain image data (e.g. voxel based activity)

Modelled by stochastic differential equations (SDEs):

$$d\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), t)dt + \mathbf{D}^{1/2}(\mathbf{x}(t), t)d\mathbf{w}(t),$$

where $d\mathbf{w}(t)$ is a Wiener process (Brownian motion):

$$d\mathbf{w}(t) = \lim_{\Delta t \to 0} \boldsymbol{\epsilon}_t \sqrt{\Delta t}, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

- Deterministic drift $\mathbf{f}$ and stochastic diffusion component $\mathbf{D}$
- Continuous-time limit of discrete-time state-space model

Stochastic differential systems

Why should we bother?

- A lot of theory, few (effective) data driven approaches
- Time discretisation is unavoidable in practice
- Physics models enforce continuity constraints, such that the number of observations can be relatively small
- High frequency fluctuations can be incorporated into the diffusion
- Any discrete representation can be chosen a posteriori
- Easy to handle irregular sampling/missing data

Bayesian approaches are natural:

- The SDE induces a non-Gaussian prior over sample paths
- Define a noise model (or likelihood) and simulate posterior process over trajectories via MCMC (Beskos et al., 2009)
- Or develop fast deterministic approximations

Overview

- Setting, notations and variational inference
- Partially observed diffusion processes
- Gaussian variational approximation
- Experiments and conclusion

Bayesian inference (framework and notations)

- Predictions are made by averaging over all possible models:

$$p(\mathbf{y}_*|\mathbf{y}) = \int p(\mathbf{y}_*|\mathbf{x})\; p(\mathbf{x}|\mathbf{y})\; d\mathbf{x}.$$

- The latent variables are inferred using Bayes' rule:

$$\underbrace{p(\mathbf{x}|\mathbf{y})}_{\textit{posterior}} = \frac{\overbrace{p(\mathbf{y}|\mathbf{x})}^{\textit{likelihood}}\;\overbrace{p(\mathbf{x})}^{\textit{prior}}}{\underbrace{p(\mathbf{y})}_{\textit{marginal likelihood}}}, \qquad p(\mathbf{y}) = \int p(\mathbf{y},\mathbf{x})\; d\mathbf{x}.$$

- Type II maximum likelihood estimation of the (hyper)parameters $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_{\mathrm{ML2}} = \operatorname*{argmax}_{\boldsymbol{\theta}}\; \ln p(\mathbf{y}|\boldsymbol{\theta}),$$

- The marginals are in general analytically intractable:
  1. We can use Markov chain Monte Carlo to simulate the integrals; potentially exact, but often slow.
  2. Or we can focus on fast(er) approximate inference schemes, such as variational inference.

Approximate Bayesian inference (variational inference)

- For any distribution $q(\mathbf{x}) \approx p(\mathbf{x}|\mathbf{y})$, we optimise a lower bound to the log-marginal likelihood:

$$\ln p(\mathbf{y}|\boldsymbol{\theta}) = \ln \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) \, d\mathbf{x} \geqslant \int q(\mathbf{x}) \, \ln \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})}{q(\mathbf{x})} \, d\mathbf{x} \doteq -\mathcal{F}(q, \boldsymbol{\theta}).$$

- (Variational) EM minimises the variational free energy iteratively and monotonically (Beal, 2003):

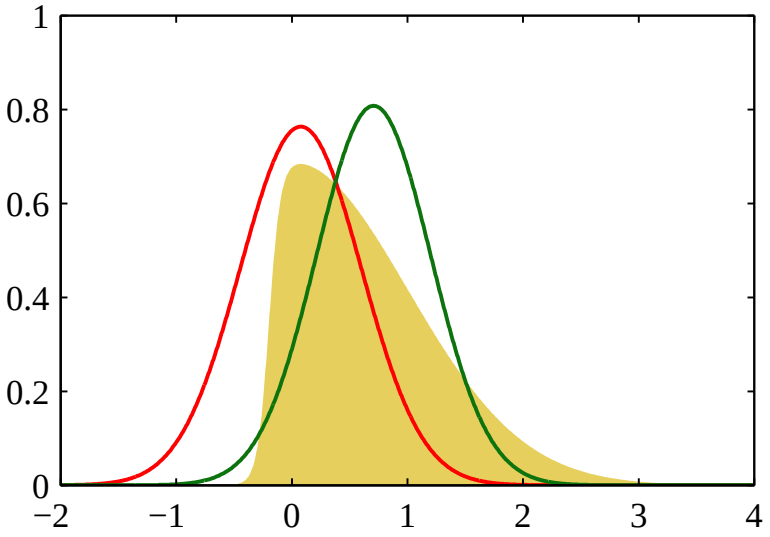
$$\begin{aligned} \mathcal{F}(q, \boldsymbol{\theta}) &= -\ln p(\mathbf{y}|\boldsymbol{\theta}) + \mathrm{KL}[q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})], \\ \mathcal{F}(q, \boldsymbol{\theta}) &= -\left\langle \ln p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) \right\rangle_{q(\mathbf{x})} - \mathrm{H}[q(\mathbf{x})]. \end{aligned}$$

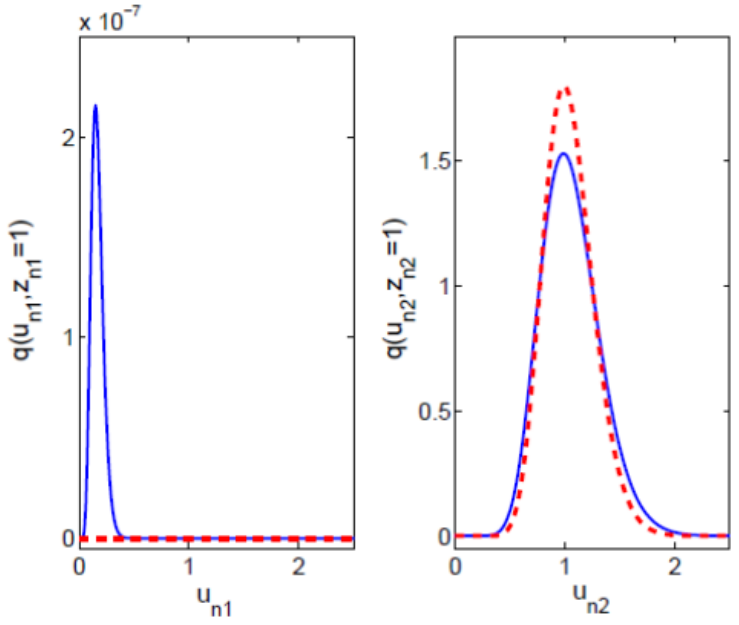
  where $\mathrm{KL}[q\|p] = \mathbb{E}_q\{\ln \frac{q}{p}\}$ is the Kullback-Leibler divergence and $\mathrm{H}[q] = -\mathbb{E}\{\ln q)\}$ the entropy.

- An alternative approach is to minimise $\mathcal{F}(q, \boldsymbol{\theta})$ with your favourite optimisation algorithm:

$$\mathcal{F}(q, \boldsymbol{\theta}) = -\left\langle \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \right\rangle_{q(\mathbf{x})} + \mathrm{KL}[q(\mathbf{x}) \| p(\mathbf{x}|\theta)].$$

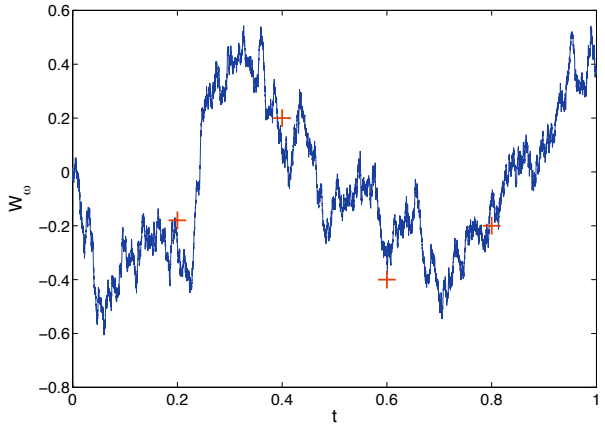
Variational inference (continued)





- Monotonic decrease of $\mathcal{F}$; convergence is easy to monitor (unlike MCMC)
- Deterministic, but different from Laplace approximation
- Usually q is assumed to have a factorised form ($q(\mathbf{x}) \approx p(\mathbf{x}|\mathbf{y})$)
- KL is wrt q; underestimation of correlations between latent variables
- Exmaple: variational treatment of Student-t mixtures

Partially observed diffusion process



- Model data by a latent diffusion process:

$$d\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), t)dt + \mathbf{D}^{1/2}(\mathbf{x}(t), t)d\mathbf{w}(t).$$

  where $\mathbf{f}$ and $\mathbf{D}$ have a known functional form.
- Discrete-time likelilhood observation operator:

$$\mathbf{y}_n = \mathbf{H}\mathbf{x}(t = t_n) + \boldsymbol{\eta}_n.$$

- Goal: infer the states $\mathbf{x}(t)$ and learn the parameters of $\mathbf{f}$ and $\mathbf{D}$ given the data.

Variational inference for diffusion processes

- We are interested in the posterior measure over the sample paths:

$$\frac{dP(\mathbf{x}(t)|\mathbf{y}_1, \ldots, \mathbf{y}_N)}{dP(\mathbf{x}(t))} = \frac{1}{Z} \prod_n P(\mathbf{y}_n|\mathbf{x}_{t=t_n}) \,.$$

  This quantity is non-Gaussian when $\mathbf{f}$ is nonlinear (and in general intractable).

- For an approximate measure $Q(\cdot)$, we minimise the variational free energy over a certain time interval:

$$\mathcal{F}(Q, \boldsymbol{\theta}) = -\left\langle \ln P(\mathbf{y}_1, \ldots, \mathbf{y}_N|\mathbf{x}(t), \boldsymbol{\theta}) \right\rangle_{Q(\mathbf{x}(t))} + \mathrm{KL}[dQ(\mathbf{x}(t))\|dP(\mathbf{x}(t))],$$

  where $t \in [0, T]$.

- What is a suitable $Q(\cdot)$?

Gaussian variational approximation

- We restrict ourselves to a state independent diffusion matrix D.
- Consider the following linear, but time-dependent SDE:

$$d\mathbf{x}(t) = \mathbf{g}(\mathbf{x}(t), t)dt + \mathbf{D}^{-1/2}(t)d\mathbf{w}(t),$$

where

$$\mathbf{g}(\mathbf{x}(t), t) \doteq -\mathbf{A}(t)\mathbf{x}(t) + \mathbf{b}(t).$$

- It induces a non-stationary Gaussian measure, with marginal mean and marginal covariance satisfying a set of ODEs:

$$\begin{aligned}\dot{\mathbf{m}}(t) &= -\mathbf{A}(t)\mathbf{m}(t) + \mathbf{b}(t),\\ \dot{\mathbf{S}}(t) &= -\mathbf{A}(t)\mathbf{S}(t) - \mathbf{S}(t)\mathbf{A}^{\top}(t) + \mathbf{D}(t).\end{aligned}$$

- We view $\mathbf{A}(t)$ and $\mathbf{b}(t)$ as variational parameters and approximate the posterior process by this non-stationary Gaussian process.

Gaussian process

Multivariate Gaussian:

- Probability density over D random variables (based on correlations).
- Characterized by a mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$\mathbf{f} \equiv (f_1, \ldots, f_D)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Gaussian process (GP):

- Probability measure over random functions ($\approx$ infinitely long vector).
- Marginal over any finite subset of variables is a consistent finite dimensional Gaussian!
- Characterized by a mean function and a covariance function (kernel):

$$f(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot)).$$

- Gaussian processes for ML (Rasmussen and Williams, 2006)
- **A** and **b** specify the kernel (in general no closed form solution)

Consistency constraints and smoothing algorithm

- The objective function is of the form

$$\mathcal{F}(Q,\theta) = \int E_{obs}(t)dt + \int E_{sde}(t)dt + \mathrm{KL}[q(\mathbf{x}_0)\|p(\mathbf{x}_0)],$$

where

$$E_{sde}(t) = -\frac{1}{2}\langle(\mathbf{f}_t - \mathbf{g}_t)^\top \mathbf{D}^{-1}(\mathbf{f}_t - \mathbf{g}_t)\rangle_{Q(\mathbf{x}_t)}.$$

- The diffusion matrix of the linear SDE is by construction equal to the diffusion matrix of the original SDE (so that $\mathcal{F}$ is finite).
- We enforce consistent Gaussian marginals by using the following ODEs as constraints (forward propagation):

$$\begin{aligned}\dot{\mathbf{m}}(t) &= -\mathbf{A}(t)\mathbf{m}(t) + \mathbf{b}(t),\\ \dot{\mathbf{S}}(t) &= -\mathbf{A}(t)\mathbf{S}(t) - \mathbf{S}(t)\mathbf{A}^\top(t) + \mathbf{D}(t).\end{aligned}$$

- Differentiating the Lagrangian leads to a set of ODEs for the Lagrange multipliers (backward propagation):

$$\dot{\boldsymbol{\lambda}}(t) = -\nabla_{\mathbf{m}} E_{sde}(t) + \mathbf{A}^\top(t)\boldsymbol{\lambda}(t), \quad \boldsymbol{\lambda}_n^+ = \boldsymbol{\lambda}_n^- - \nabla_{\mathbf{m}} E_{obs}(t)|_{t=t_n},$$

$$\dot{\boldsymbol{\Psi}}(t) = -\nabla_{\mathbf{S}} E_{sde}(t) + 2\boldsymbol{\Psi}(t)\mathbf{A}(t), \quad \boldsymbol{\Psi}_n^+ = \boldsymbol{\Psi}_n^- - \nabla_{\mathbf{S}} E_{obs}(t)|_{t=t_n}.$$

Optimal Gaussian variational approximation

- The non-linear SDE is reduced to a set of linear ODEs describing the evolution of the means, covariances and Lagrange multipliers.
- The smoothing algorithm consists of a forward and a backward integration for fixed $\mathbf{A}(t)$ and $\mathbf{b}(t)$.
- The observation are incorporated in the backward pass (cf. jump conditions).
- The optimal Gaussian variational approximation is obtained by optimising $\mathcal{F}$ wrt the variational parameters $\mathbf{A}(t)$ and $\mathbf{b}(t)$.
- At equilibrium, the variational parameters satisfy the following conditions:

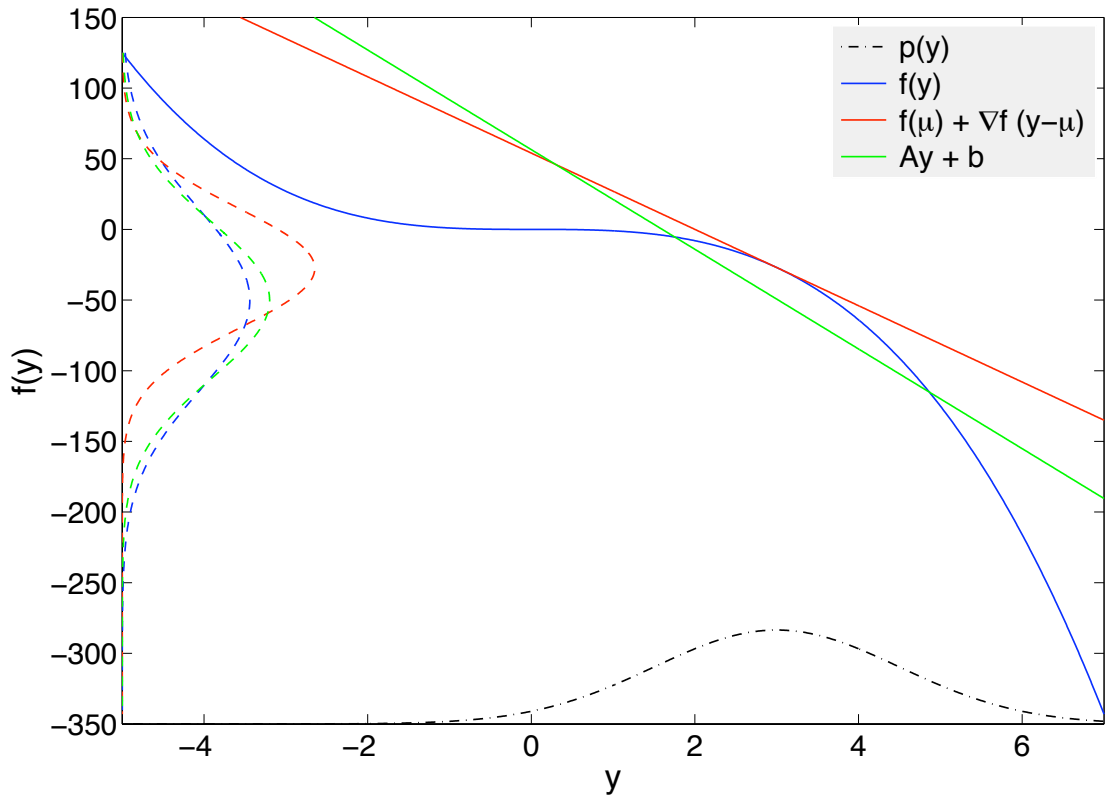
$$\mathbf{A} = -\left\langle \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right\rangle + 2\mathbf{D}\boldsymbol{\Psi},$$

$$\mathbf{b} = \langle \mathbf{f}(\mathbf{x}) \rangle + \mathbf{A}\mathbf{m} - \mathbf{D}\boldsymbol{\lambda}.$$

- The variational solution is closely related to statistical linearisation:

$$\{\mathbf{A}, \mathbf{b}\} \leftarrow \underset{\mathbf{A},\mathbf{b}}{\operatorname{argmin}} \left\langle \|\mathbf{f}(\mathbf{x}) + \mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\rangle.$$

Illustration of the statistical linearisation principle



Related approaches

Continuous-time sigma point Kalman smoothers (KS; Sarkka and Sottinen, 2008):

- Unscented KS and central difference KS.
- Gaussian approximation of the transition density.
- No feedback loop to adjust the sigma points.

Perfect simulation approaches (Beskos et al., 2009):

- No discrete time approximation of the transition density.
- Transition density is non-Gaussian.
- Drift is restricted to derive from a potential.
- Convergence is difficult to monitor, potentially slower.

Other approaches include Particle smoothers, Hybrid MCMC (Eyinck et al., 2004) etc.

Diffusions with multiplicative noise

- Apply explicit transformation to obtain a diffusion process with constant diffusion matrix; such a transformation does not always exist in the multivariate case.
- Construct Gaussian variational approximation based on the following ODEs, which hold for any non-linear SDE:

$$\dot{\mathbf{m}}(t) = -\mathbf{A}(t)\mathbf{m}(t) + \mathbf{b}(t),$$
$$\dot{\mathbf{S}}(t) = -\mathbf{A}(t)\mathbf{S}(t) - \mathbf{S}(t)\mathbf{A}^{\top}(t) + \langle \mathbf{D}(\mathbf{x}(t), t) \rangle_{Q(\mathbf{x}_t)}.$$

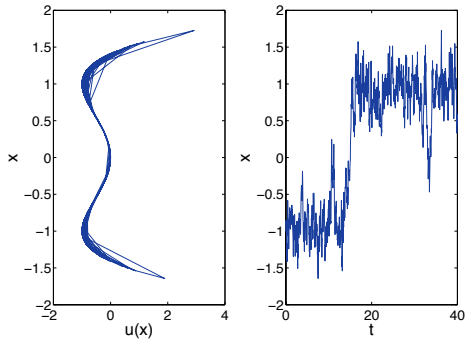
The smoothing algorithm is analogue; the expression of $\mathbf{A}(t)$ and $\mathbf{b}(t)$ is more involved.

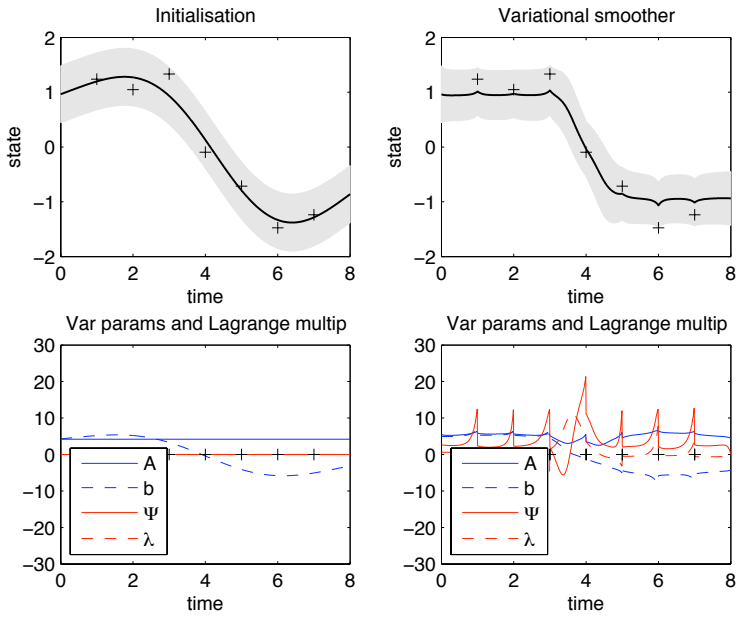
Bi-stable dynamical system

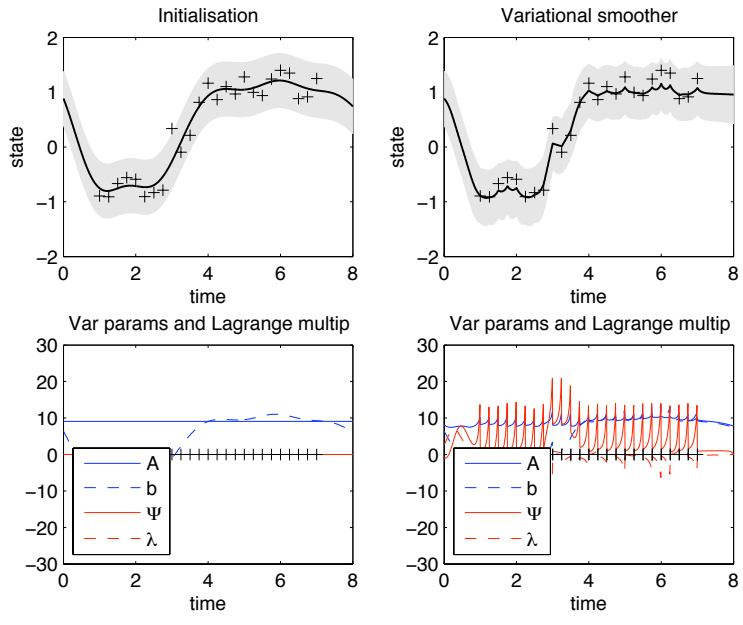
The deterministic drift is defined as

$$f(t, x) = 4x(\theta - x^2), \qquad \theta > 0.$$

The system is driven by the stochastic noise.

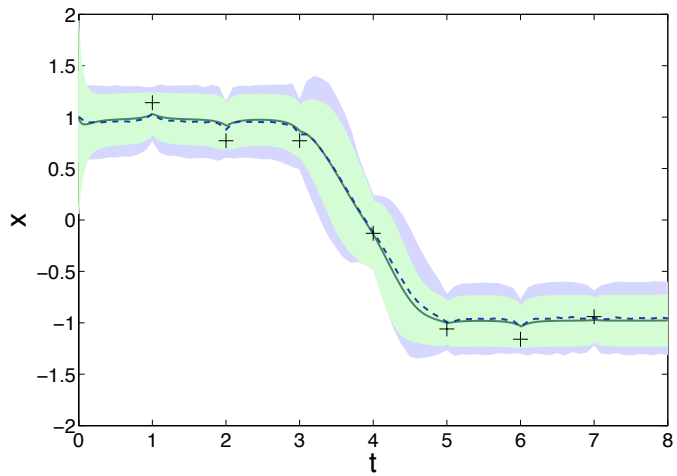




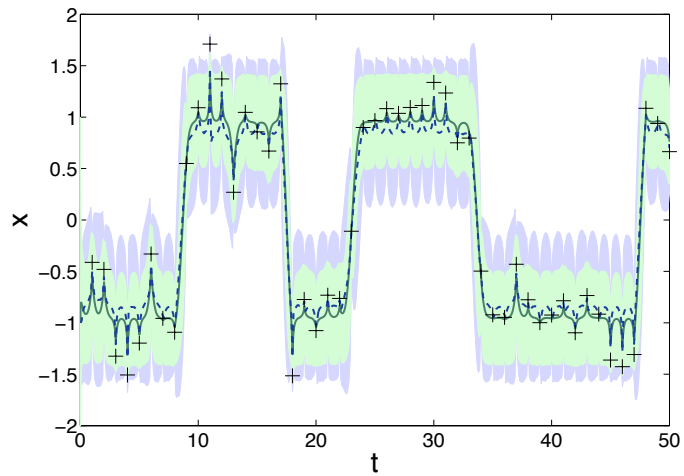


Comparison to hybrid Markov Chain Monte Carlo (Eyinck et al., 2004)

- Reference solution
- Based on a discrete approximation
- Generate complete sample paths from posterior
- Modified MCMC scheme to increase acceptance rate (Molecular Dynamics)
- Still requires to generate in order of 100,000 samples for good results
- Hard to check convergence

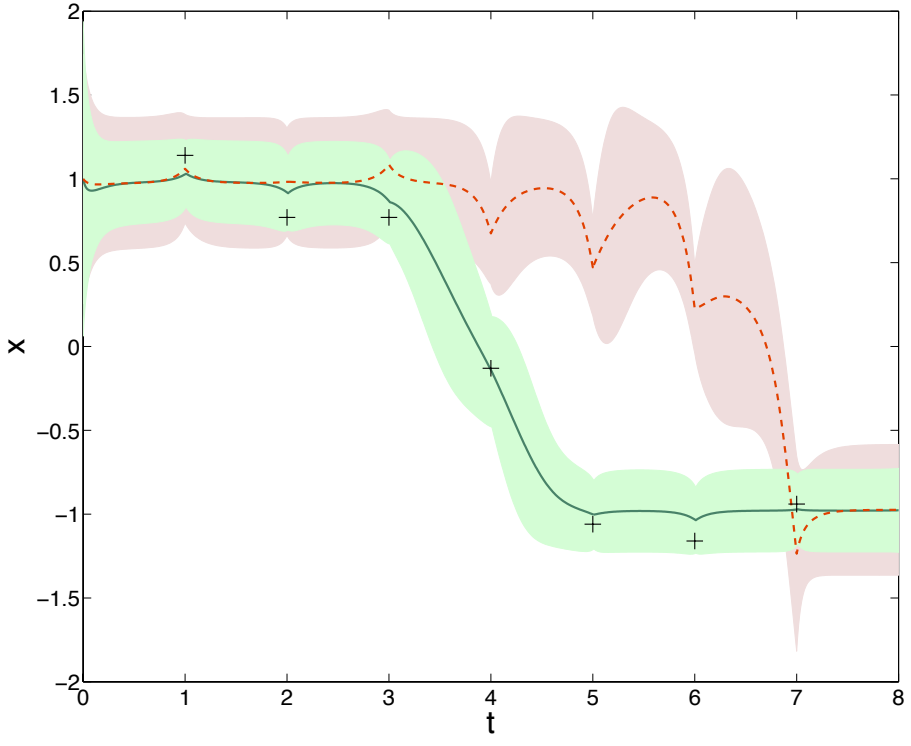


(a) $\theta = 1$, $\sigma = 0.5$.

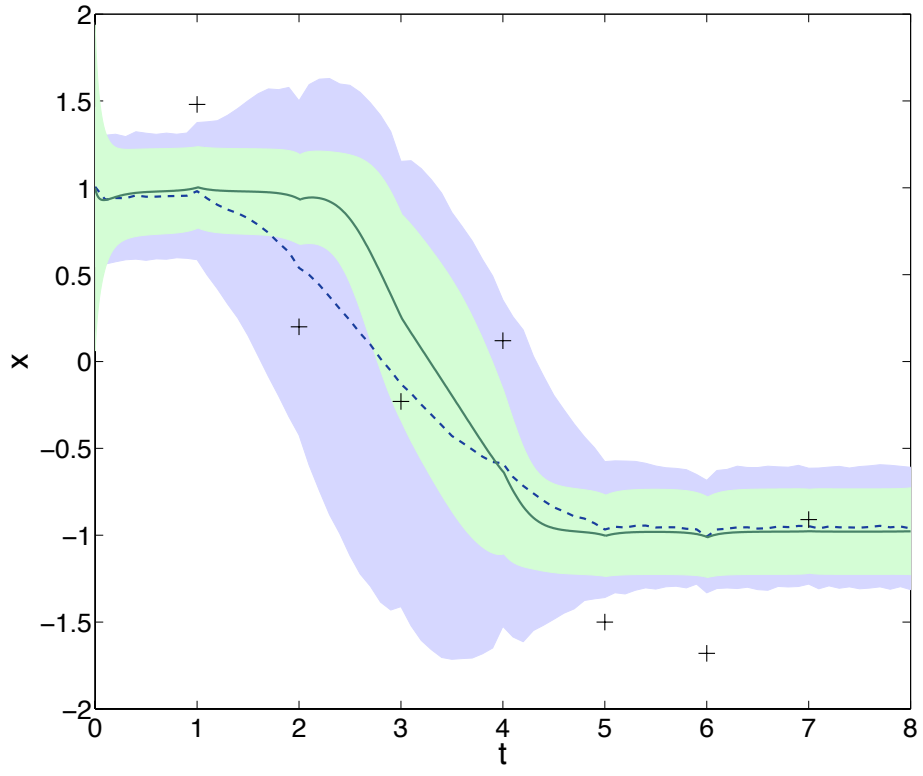


(b) Large noise.

Comparison to the continuous-time Unscented Kalman Smoother



Failure mode

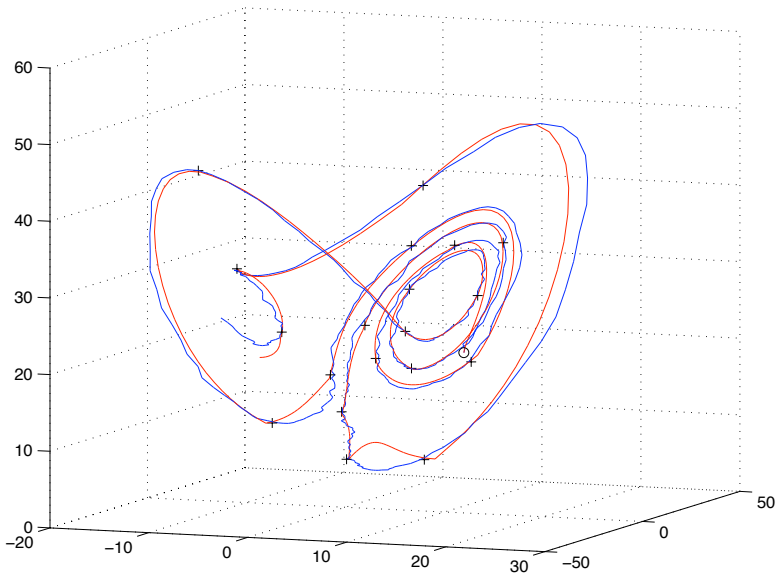


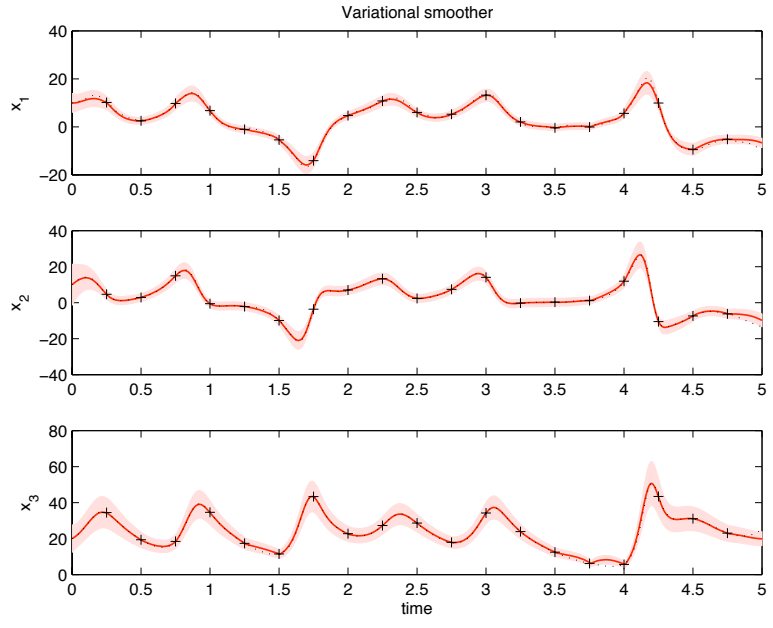
Stochastic Lorenz attractor

The Lorenz attractor:

$$
\begin{aligned}
f_x &= \sigma(y - z), & \sigma &> 0, \\
f_y &= \rho x - y - xz, & \rho &> 0, \\
f_z &= xy - \beta z, & \beta &> 0.
\end{aligned}
$$

When adding stochastic noise the system becomes chaotic.





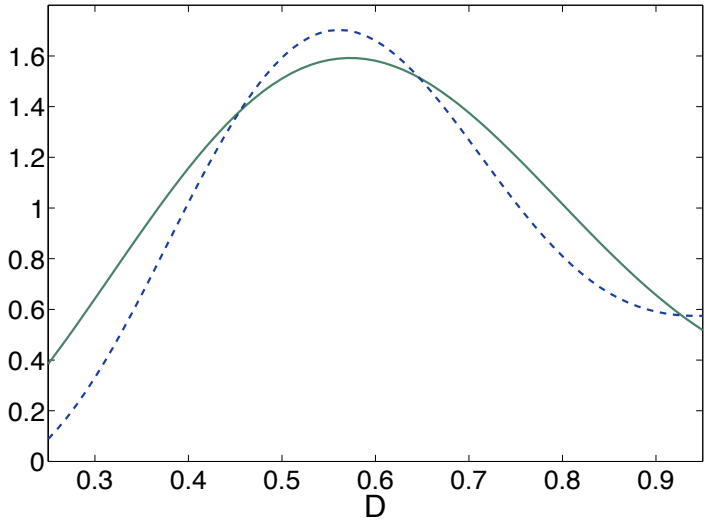
Parameter inference

- (Variational) EM fails for the diffusion coefficient:

$$\lim_{\delta \to 0} \sum_{i=1}^{T/\delta} (\mathbf{x}_{i\delta} - \mathbf{x}_{(i-1)\delta})(\mathbf{x}_{i\delta} - \mathbf{x}_{(i-1)\delta})^\top = \int_0^T \mathbf{D}(\mathbf{x}(t), t) dt \quad a.s.$$

- Type II ML based on gradient techniques is ok as we change the sample paths together with the diffusion coefficient.
- Cheap estimate of the posterior (sanity check; Lappalainen and Miskin, 2000):

$$q(\theta) = \frac{e^{-\mathcal{F}(Q,\theta)} p(\theta)}{\int e^{-\mathcal{F}(Q,\theta)} p(\theta) d\theta}$$



Conclusion

- Stochastic process models are very powerful when the number of observations is small compared to the complexity of the dynamics.
- Gaussian variational approximation for non-linear SDEs boils down to solving a set of ODEs.
- Preferred integration scheme can be used, no discrete time approximation of the transition density.
- Can be viewed as generalisation of sigma-point Kalman smoother for certain instantiations of the statistical linearisation principle.
- Considerably faster than (most) MCMC schemes.
- Diffusion matrix can be estimated; multiplicative noise is ok (in principle).
- Error bars are underestimated.

References

- C. Archambeau, M. Opper: Approximate inference for continuous time Markov processes. Inference and Estimation in Probabilistic Time-Series Models. Cambridge University Presse, 2011.
- C. Archambeau, M. Opper, Y. Shen, D. Cornford, J. Shawe-Taylor: Variational Inference for Diffusion Processes. NIPS 20, pp.17-24, 2008.
- A. Beskos, et al. : Monte-Carlo maximum likelihood estimation for discretely observed diffusion processes. Annals of Statistics, 37:1, pp 223-245, 2009.
- G. L. Eyink, J. L. Restrepo and F. J. Alexander: A mean field approximation in data assimilation for nonlinear dynamics. Physica D, 194:347368, 2004.
- I. Karatzas and S. E. Schreve. Brownian Motion and Stochastic Calculus. Springer, 1998.
- H. Lappalainen and J.W. Miskin: Ensemble learning. In M. Girolami, editor, Advances in Independent Component Analysis, pp 7692. Springer-Verlag, 2000.
- C. E. Rasmussen and C. K.I. Williams: Gaussian Processes for Machine Learning. MIT Press, 2006.
- S. Särkkä and T. Sottinen: Application of Girsanov Theorem to Particle Filtering of Discretely Observed Continuous-Time Non-Linear Systems. Bayesian Analysis, 3:3, pp 555-584, 2008.

Informal proof for $\mathrm{KL}[Q(\mathbf{x}(t))\|P(\mathbf{x}(t))]$

Consider the Euler-Muryama discrete approximation of the SDEs:

$$\Delta \mathbf{x}_k = \mathbf{f}_k \Delta t + \mathbf{D}^{1/2} \Delta \mathbf{w}_k, \quad \mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \Delta t \mathbf{I}),$$
$$\Delta \mathbf{x}_k = \mathbf{g}_k \Delta t + \mathbf{D}^{1/2} \Delta \hat{\mathbf{w}}_k, \quad \hat{\mathbf{w}}_k \sim \mathcal{N}(\mathbf{0}, \Delta t \mathbf{I}),$$

where $\Delta \mathbf{x}_k \equiv \mathbf{x}_{k+1} - \mathbf{x}_k$.

The joint distributions of discrete sample paths $\{\mathbf{x}_k\}_{k \geq 0}$ for the true process and its approximation follow from the Markov property:

$$p(\mathbf{x}_0, \ldots, \mathbf{x}_K | \mathbf{D}) = p(\mathbf{x}_0) \prod_{k>0} \mathcal{N}(\mathbf{x}_{k+1} | \mathbf{x}_k + \mathbf{f}_k \Delta t, \mathbf{D} \Delta t),$$
$$q(\mathbf{x}_0, \ldots, \mathbf{x}_K | \mathbf{D}) = q(\mathbf{x}_0) \prod_{k>0} \mathcal{N}(\mathbf{x}_{k+1} | \mathbf{x}_k + \mathbf{g}_k \Delta t, \mathbf{D} \Delta t).$$

The KL between the two discretised processes is then given by

$$\mathrm{KL}[q\|p] = \mathrm{KL}[q(\mathbf{x}_0)\|p(\mathbf{x}_0)] - \sum_{k>0} \int q(\mathbf{x}_k) \left\langle \ln \frac{p(\mathbf{x}_{k+1}|\mathbf{x}_k)}{q(\mathbf{x}_{k+1}|\mathbf{x}_k)} \right\rangle_{q(\mathbf{x}_{k+1}|\mathbf{x}_k)} d\mathbf{x}_k$$
$$= \mathrm{KL}[q(\mathbf{x}_0)\|p(\mathbf{x}_0)] + \frac{1}{2} \sum_{k>0} \langle (\mathbf{f}_k - \mathbf{g}_k)^\top \mathbf{D}^{-1} (\mathbf{f}_k - \mathbf{g}_k) \rangle_{q(\mathbf{x}_k)} \Delta t,$$

Passing to the limit is ok! (Formal proof based on the Girsanov theorem.)