

Approximate Inference for Continuous Time Markov Processes

Manfred Opper, Computer Science



collaboration with:

Cédric Archambeau (UCL)

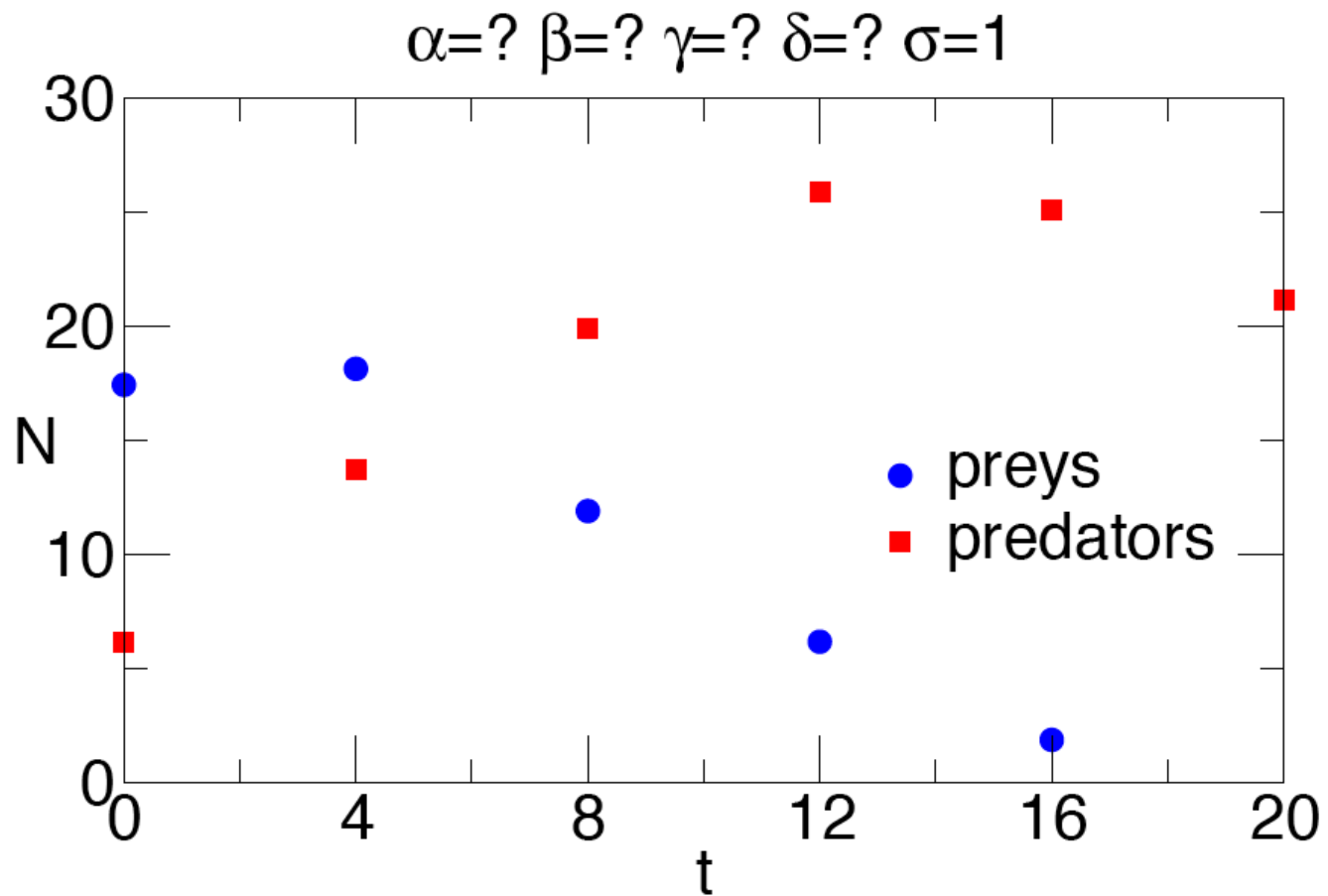
Dan Cornford (Aston)

Andreas Ruttor (TU Berlin)

John Shawe–Taylor (UCL)

Yuan Shen (Aston)

Guido Sanguinetti (U Sheffield)



Discrete observations from a continuous time series

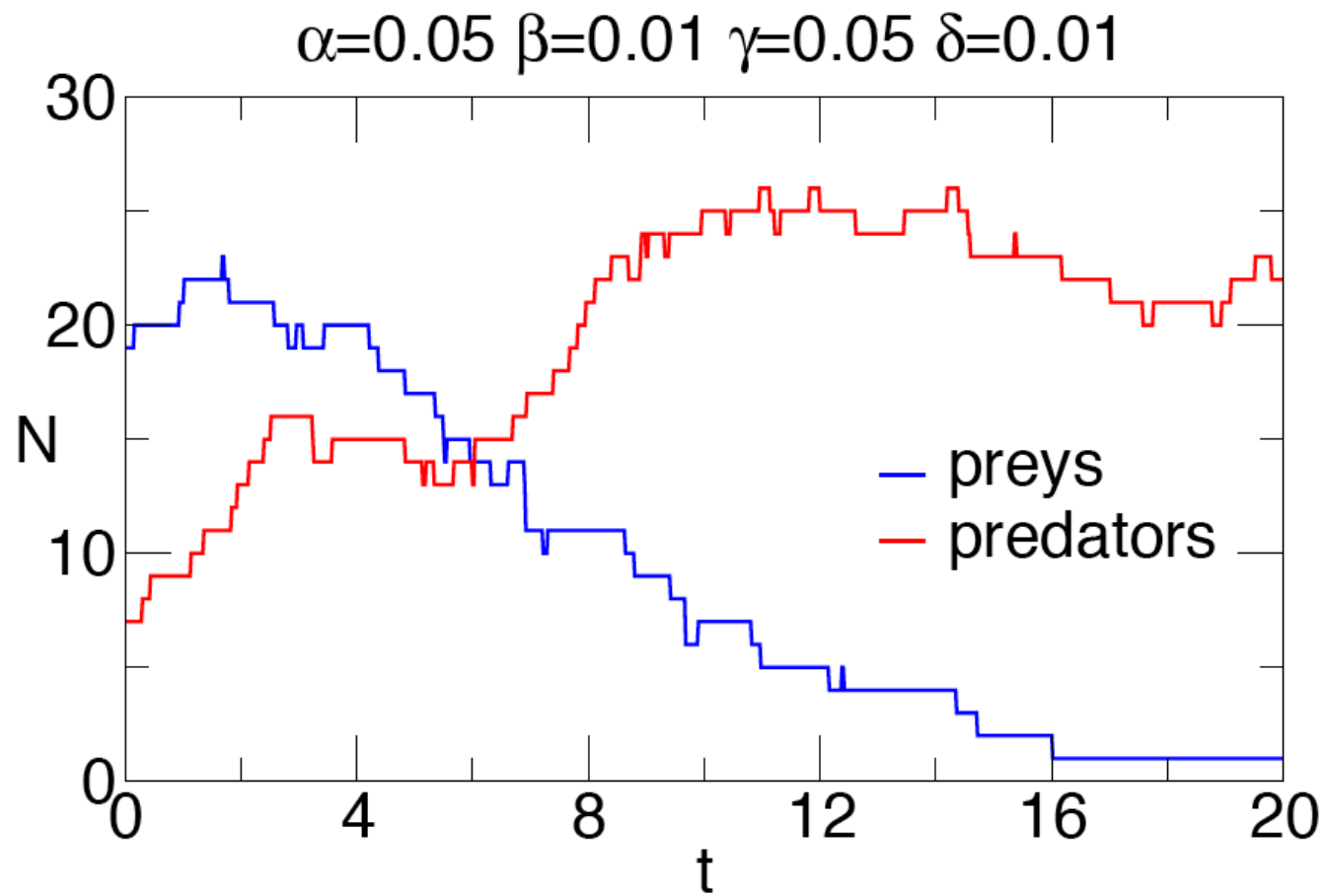
Stochastic Lotka Volterra Model

Prey \rightarrow 2 Prey with Rate αX_{Prey}

Prey $\rightarrow \emptyset$ with Rate $\beta X_{\text{Prey}} X_{\text{Pred}}$

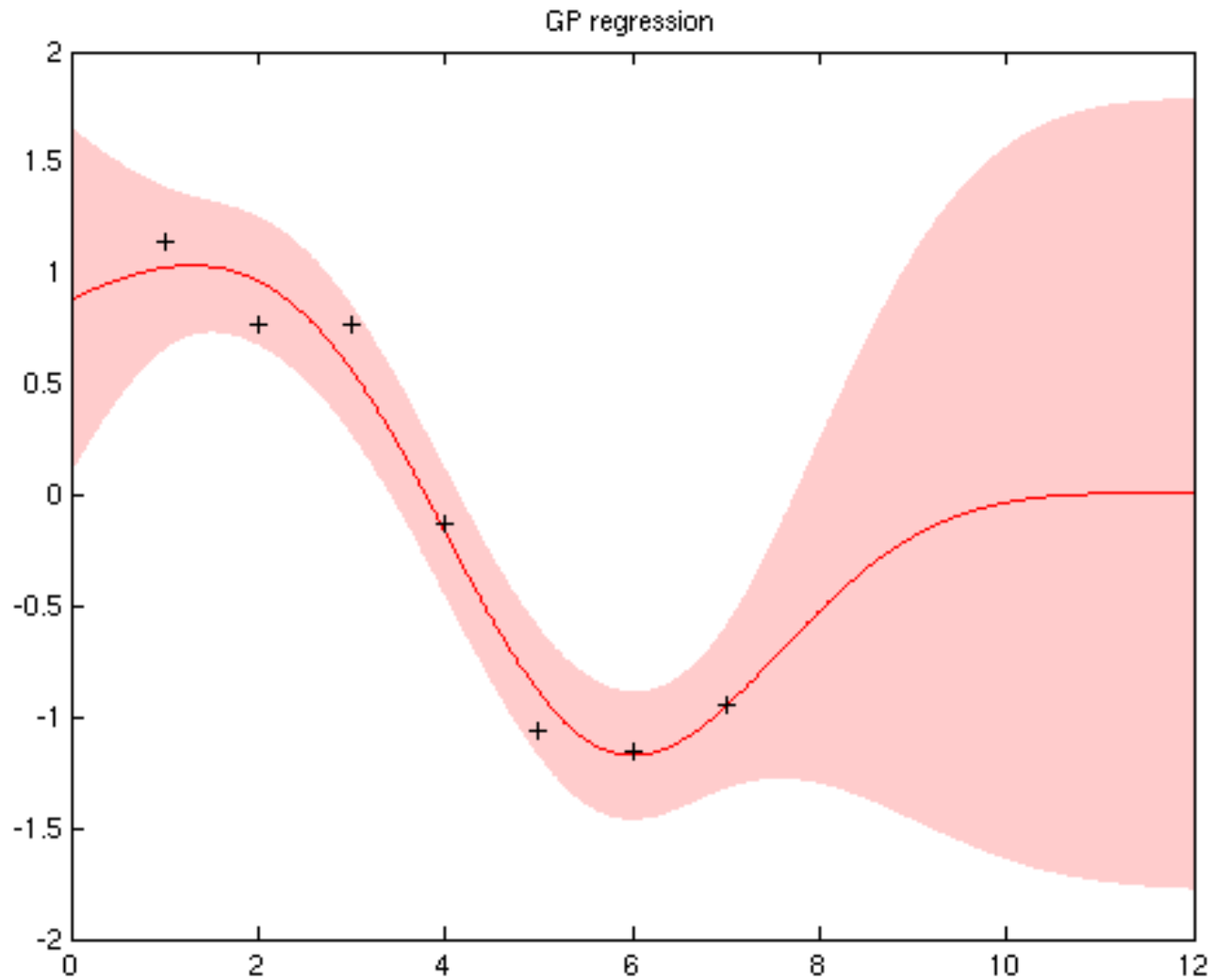
Predator \rightarrow 2 Predator with Rate $\delta X_{\text{Prey}} X_{\text{Pred}}$

Pred $\rightarrow \emptyset$ with Rate γX_{Pred}



The actual time series and the reaction constants

Another set of data & estimated path with uncertainty

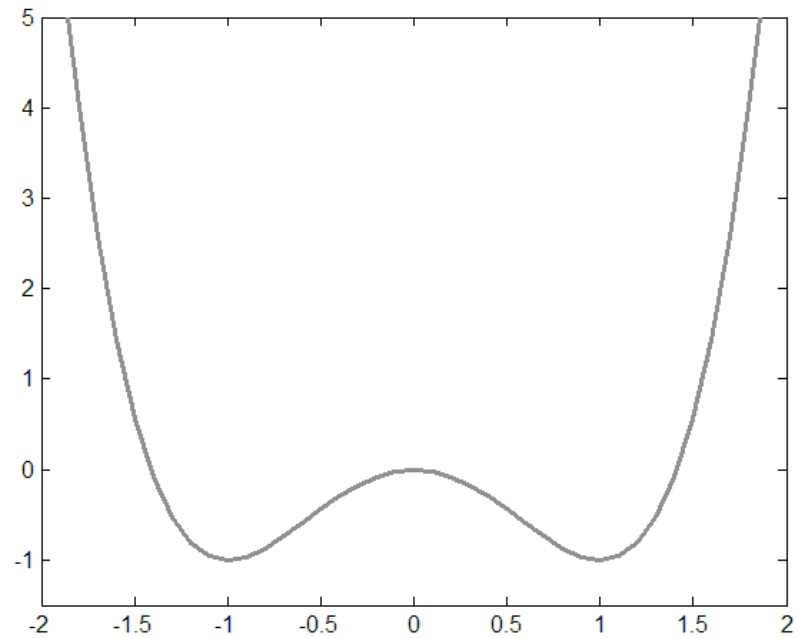


Motion in double-well potential

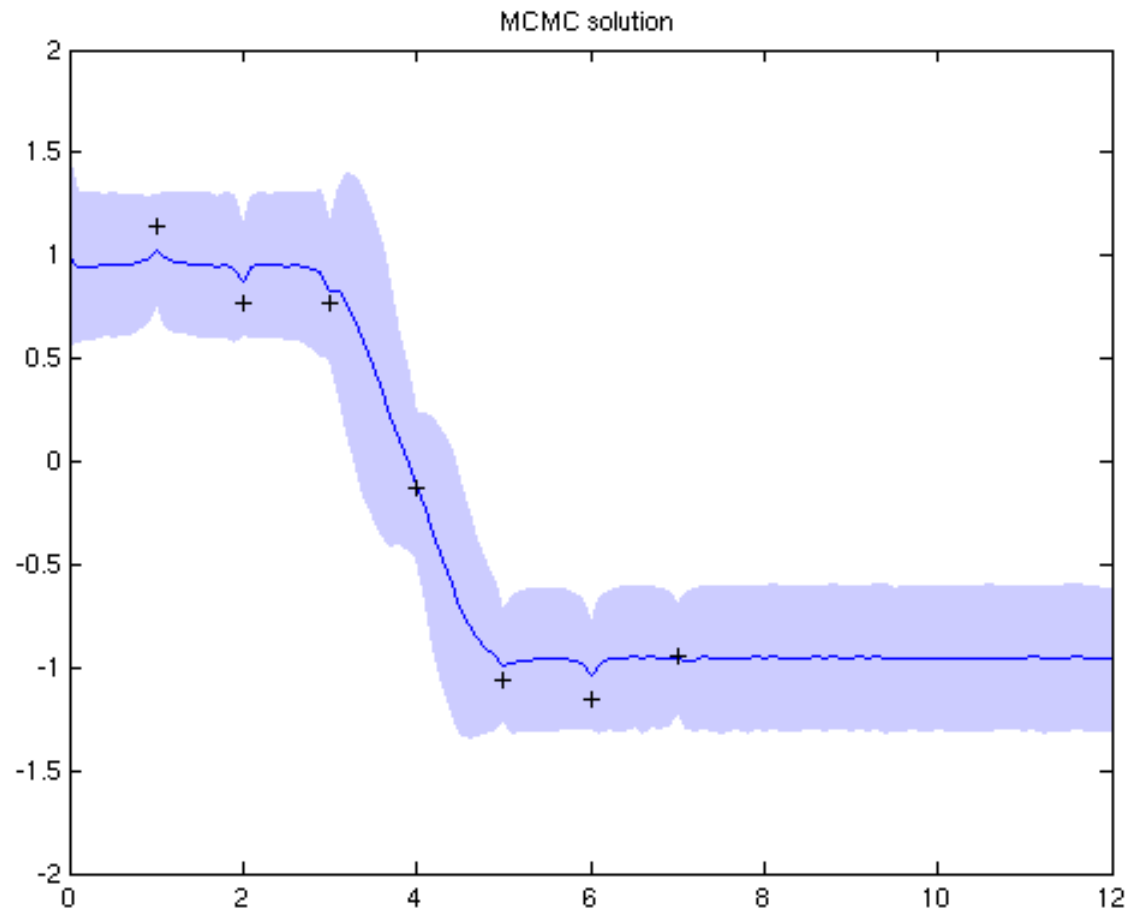
$$dX = f(X)dt + \sigma^2 dW.$$

with $f(x) = -\frac{dV(x)}{dx}$

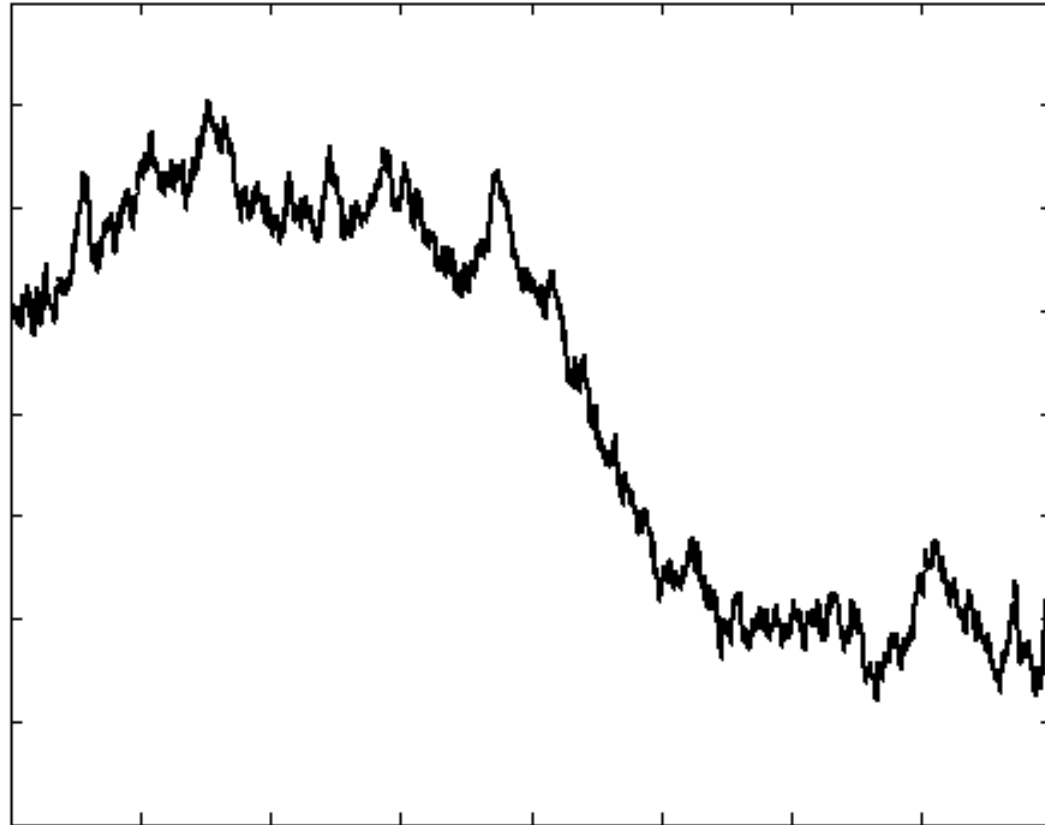
and $V(x)$ is a double well potential



The optimal prediction



The sample path might have looked like this



Overview

- Partly observed Diffusion and simple Markov jump processes
- Variational approach to inference & parameter estimation
- Exact inference
- Gaussian and mean field approaches
- weak noise approximation

Continuous time Markov processes

- **Markov jump processes:** Discrete states X . Probability of transition $X \rightarrow X'$ for small times $\Delta t \rightarrow 0$

$$p(X', t + \Delta t | X, t) \simeq \delta_{X'X} + \Delta t f(X'|X)$$

- **Diffusion processes:** (Ito) stochastic differential equation (SDE) for state $X(t) \in R^d$

$$dX(t) = \underbrace{f(X(t))}_{\text{Drift}} dt + \underbrace{D^{1/2}(X(t))}_{\text{Diffusion}} dW(t)$$

$W(t)$ is vector of independent *Wiener processes*.

limit of discrete time process X_k

$$\Delta X_k \equiv X_{k+1} - X_k = f(X_k)\Delta t + D^{1/2}(X_k)\sqrt{\Delta t} \epsilon_k .$$

ϵ_k i.i.d. Gaussian.

Inference Problems

Given **noisy observations** $Y \equiv Y_1, \dots, Y_n$ of
hidden process $X(t)$ at times t_i for $i = 1, \dots, n$.

- Estimate $X(t)$ for $0 \leq t \leq T$ and give uncertainty of prediction.
- Estimate unknown system parameters θ contained in rates $f(X'|X)$ (MJP) or drift f and diffusion D (Diff).

Solution to estimation problems

- **Optimal prediction** of $X(t)$: Conditional expectation $E[X(t)|Y]$.

The conditional (posterior) distribution over **paths** $X \equiv X_{0:T}$ is

$$\frac{dP(X|Y)}{dP(X)} = \frac{1}{Z} \times \prod_{n=1}^N p(Y_n|X(t_n)),$$

with **partition function** $Z = p(Y|\theta)$.

- **Parameter estimation**: Minimise $-\ln Z = -\ln p(Y|\theta)$ with respect to θ (**Max Likelihood**) or use a prior $p(\theta)$ to compute $p(\theta|Y) \propto p(Y|\theta)p(\theta)$ (**Bayes**).

Why is this not so easy ?

- Analytical solution leads to PDEs
- Monte Carlo requires path sampling (often slow convergence)
- Computation of free energy needs many 'temperatures'
- Gibbs sampler $p(\text{path}|\theta) \leftrightarrow p(\theta|\text{path})$ is nontrivial!

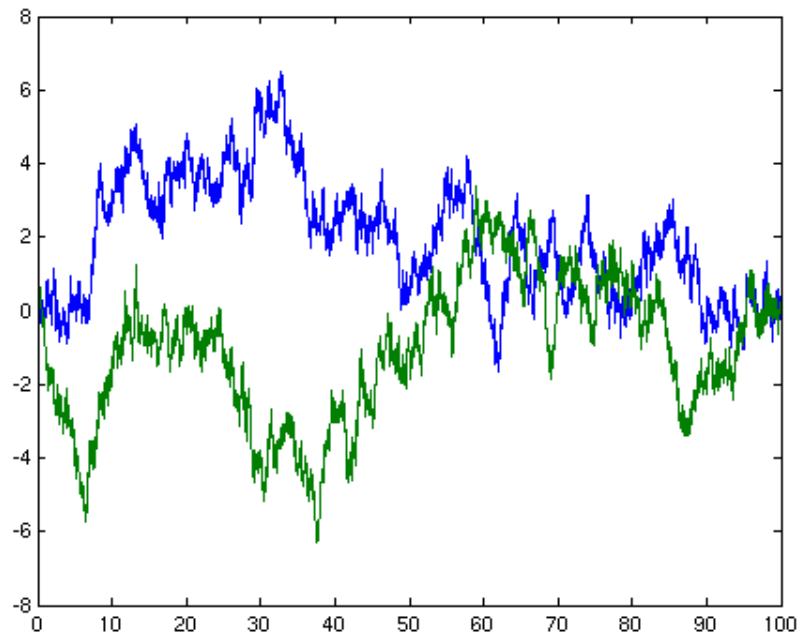
Comments

- Posterior is Markovian!
- MJP: Posterior has new rate function $g(X'|X, t)$
- Diffusion: Posterior process fulfils SDE

$$dX(t) = g(X(t), t)dt + D^{1/2}(X(t)) dW(t)$$

Example

Wiener process with single, noise free observation $y = x(t = T) = 0$



Posterior drift $g(x, t) = -\frac{x}{T-t}$ for $0 < t < T$.

Variational approach to inference

The posterior measure is the minimiser of

$$KL[Q\|P(\cdot|Y)] = \int dQ(X) \ln \frac{dQ(X)}{dP(X|Y)} \geq 0$$

Equivalent formulation: Minimise **variational free energy**

$$\begin{aligned} \mathcal{F}_\theta(Q, Y) &= KL[Q\|P(\cdot|Y)] - \ln p(Y|\theta) \\ &= KL[Q\|P_{prior}] - \sum_n E_Q[\ln P(Y_n|X(t_n))] \\ &\geq -\ln p(Y|\theta) \end{aligned}$$

The statistical physics version

Set $P(X|Y) = \frac{1}{Z} e^{-H(X)}$ and $Q(x) = \frac{1}{Z_0} e^{-H_0(X)}$

The **Variational free energy** is

$$\mathcal{F}_\theta(Q) = -\ln Z_0 + \langle H(X) \rangle_0 - \langle H_0(X) \rangle_0 \geq -\ln Z$$

(Feynman, Peierls, Bogolubov, Kleinert...)

The Euclidian path integral (for diffusion processes) would be something like this ...

$$Z = \int \mathcal{D}[X(t)] \exp \left[-\frac{1}{2\sigma^2} \int_0^T dt \left\{ \left(\frac{dx}{dt} \right)^2 + \frac{1}{2} f \cdot \frac{dx}{dt} - \|f\|^2 - \frac{1}{2} \sigma^2 \nabla f \right\} \right]$$

Full (approximate) Bayesian Inference

Approximate posterior of parameters (Lappalainen, 2000):

$$q(\theta|Y) \approx \frac{e^{-\mathcal{F}_\theta(Q,Y)} p(\theta)}{\int e^{-\mathcal{F}_\theta(Q,Y)} p(\theta) d\theta}.$$

The KL divergence for Markov processes

Probabilities $p(x_{1:K})$, $q(x_{1:K})$ over discrete time paths

$$\begin{aligned} KL [Q||P] &= \int dx_{1:K} p(x_{1:K}) \ln \frac{q(x_{1:K})}{p(x_{1:K})} \\ &= \sum_{k=1}^{K-1} \int dx_k q(x_k) \int dx_{k+1} q(x_{k+1}|x_k) \ln \frac{q(x_{k+1}|x_k)}{p(x_{k+1}|x_k)} \\ &= \sum_{k=1}^{K-1} \int dx_k q(x_k) KL [q(\cdot|x_k)||p(\cdot|x_k)] \end{aligned}$$

Depends on short - time limit of transition probabilities.

Short time transition probability

$$p(X', t + \Delta t | X, t) =$$

$$= \begin{cases} c \cdot \exp \left[-\frac{1}{2\Delta t} (\Delta X - f(X)\Delta t)^\top D^{-1}(X) (\Delta X - f(X)\Delta t) \right] & \text{for SDE} \\ \delta_{X'X} + \Delta t f(X'|X) & \text{for MJP} \end{cases}$$

as $\Delta t \rightarrow 0$.

The limit I : Stochastic Differential Equations

Let Q and P be measures over paths for SDEs with drifts $g(X, t)$ and $f(X, t)$ having the **same diffusion** $D(X)$. Then

$$KL [Q \| P] = E_Q \left[\ln \frac{dQ}{dP} \right] =$$
$$\frac{1}{2} \int_0^T dt \left\{ \int dx q(x, t) \left[(g(x, t) - f(x, t))^{\top} D^{-1}(x) (g(x, t) - f(x, t)) \right] \right\}$$

$q(x, t)$ is the (marginal) density of $X(t)$ with respect to Q .

The limit II : Markov jump processes

Assume transition rates $g(X'|X, t)$ and $f(X'|X, t)$

$$KL [Q||P] =$$

$$\int_0^T dt \sum_x q(x, t) \sum_{x':x' \neq x} \left\{ g(x'|x, t) \ln \frac{g(x'|x, t)}{f(x'|x, t)} + f(x'|x, t) - g(x'|x, t) \right\}$$

The variational problem (no approximation)

Minimise variational free energy

$$\mathcal{F}_\theta(Q, Y) = KL(Q, P) - \sum_i E_Q[\ln p(Y_i|X(t_i))]$$

and

$$KL(Q, P) = \frac{1}{2} \int_0^T dt \int dx (g - f)^\top D^{-1} (g - f) q(x, t)$$

The marginal density q fulfils the Fokker - Planck equation

$$\frac{\partial q}{\partial t} = \left\{ -\nabla g + \frac{1}{2} \text{Tr}(\nabla \nabla^T) D \right\} q \equiv L_g q$$

Variation of Lagrange function

$$L = \frac{1}{2} \int_0^T dt \int dx (g - f)^\top D^{-1} (g - f) q(x, t) - \sum_i E_Q[\ln p(Y_i|X(t_i))] \\ - \int_0^T dt \int dx \lambda(x, t) \left(\frac{\partial q}{\partial t} - L_g q \right)$$

with respect to q and g results in

$$g(x, t) = f(x) - D(x, t) \nabla \lambda(x, t)$$

where the 'potential' $\lambda(x, t)$ includes the effect of all data in the future ($> t$).

Setting $\lambda(x, t) = -\ln \psi(x, t)$ we find that ψ fulfils the **Kolmogorov Backward equation**

$$\left\{ \frac{\partial}{\partial t} + f^\top \nabla + \frac{1}{2} \text{Tr}(D(x, t) \nabla^\top \nabla) \right\} \psi(x, t) = 0$$

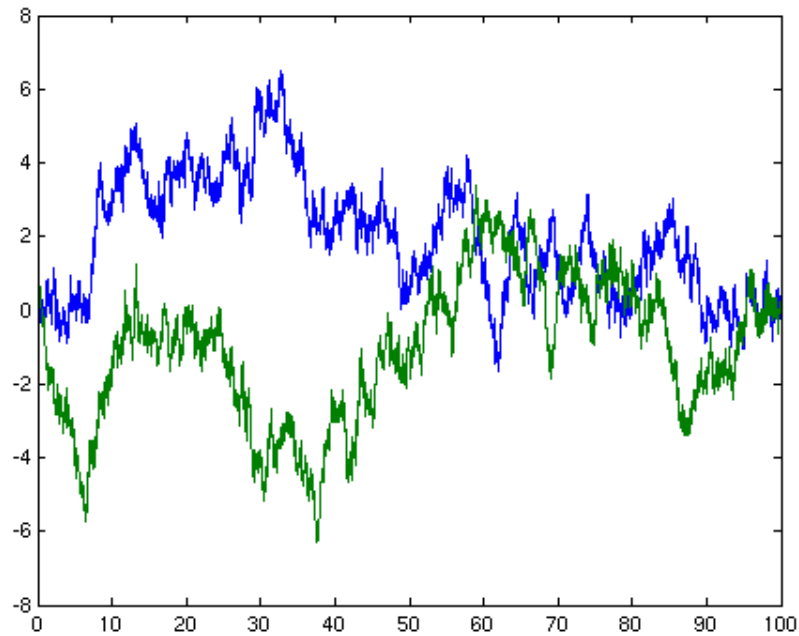
End - & jump conditions

$$\psi(x, t_n) = p(y_N | x) \quad \psi_k^+(x) = \psi_k^-(x) p(y_k | x) \quad k = 1, \dots, N - 1$$

Equivalent to KSP equations (Kushner '62, Stratonovich '60 & Pardoux '82).

Example

Wiener process with single, noise free observation $y = x(t = T) = 0$



Solution

$$g(x, t) = \frac{\partial \ln \psi(x, t)}{\partial x}$$

$$\frac{\partial \psi(x, t)}{\partial t} + \frac{\partial^2 \psi(x, t)}{\partial x^2} = 0$$

$$\psi(x, T) = \delta(x)$$

is solved by

$$\psi(x, t) \propto e^{-\frac{x^2}{2(T-t)}}$$

and leads to

$$g(x, t) = -\frac{x}{T-t}$$

for $0 < t < T$.

The Gaussian Variational Approximation

For previous applications in machine learning (see e.g. Barber & Bishop (1998), Seeger (2000), Honkela & Valpola (2005)).

- Diffusion must be independent of X .
- Linear approximate posterior SDE:

$$dX(t) = \{-A(t)X + b(t)\} dt + D^{1/2}dW$$

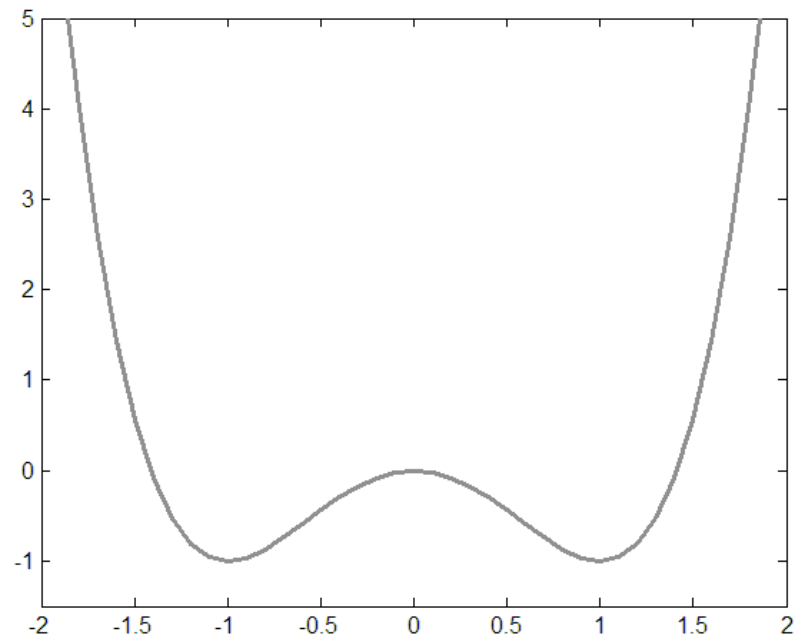
Constraints are evolution eqs. for marginal **mean** and **covariance**

$$\begin{aligned}\frac{dm}{dt} &= -Am + b(t) \\ \frac{dS}{dt} &= -AS - SA^\top + D.\end{aligned}$$

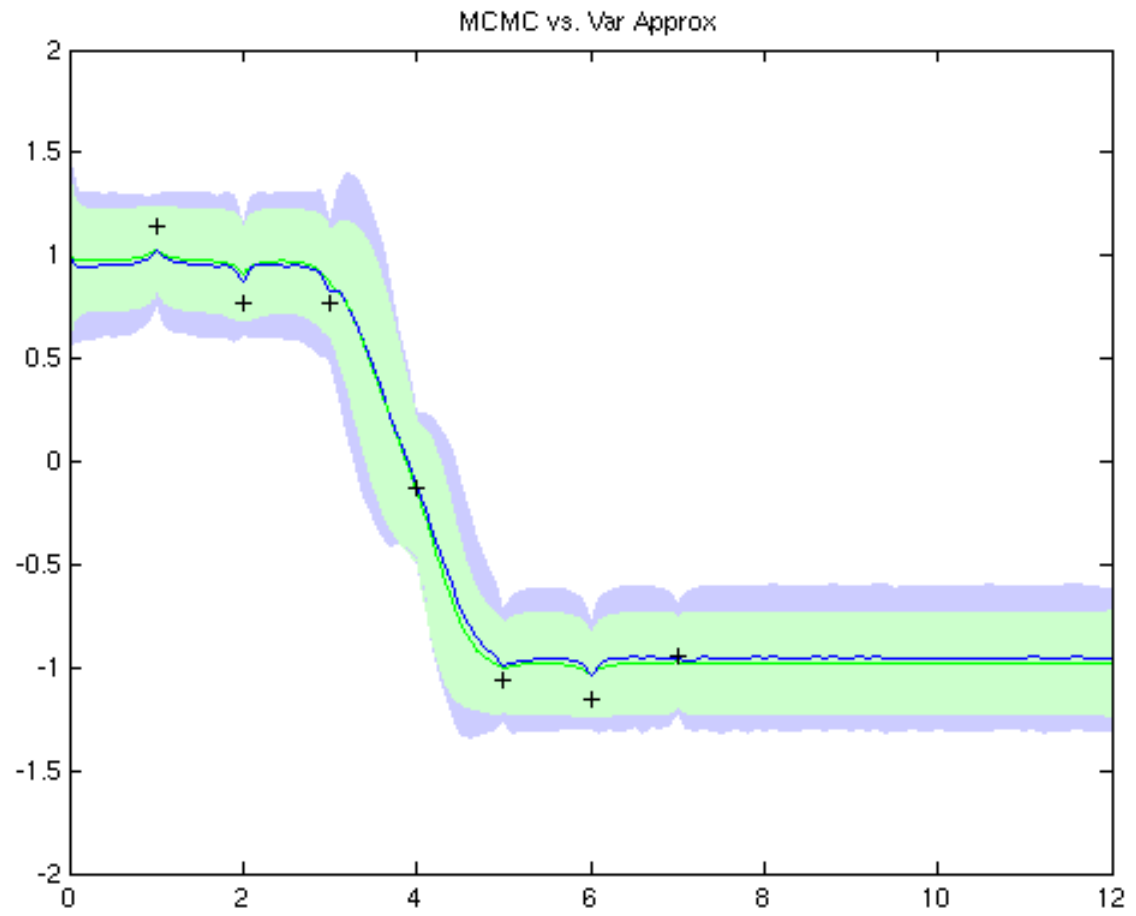
→ **nonlinear ODEs** instead of PDEs !

Motion in double-well potential

$$dX = 4X(1 - X^2)dt + \sigma^2 dW.$$

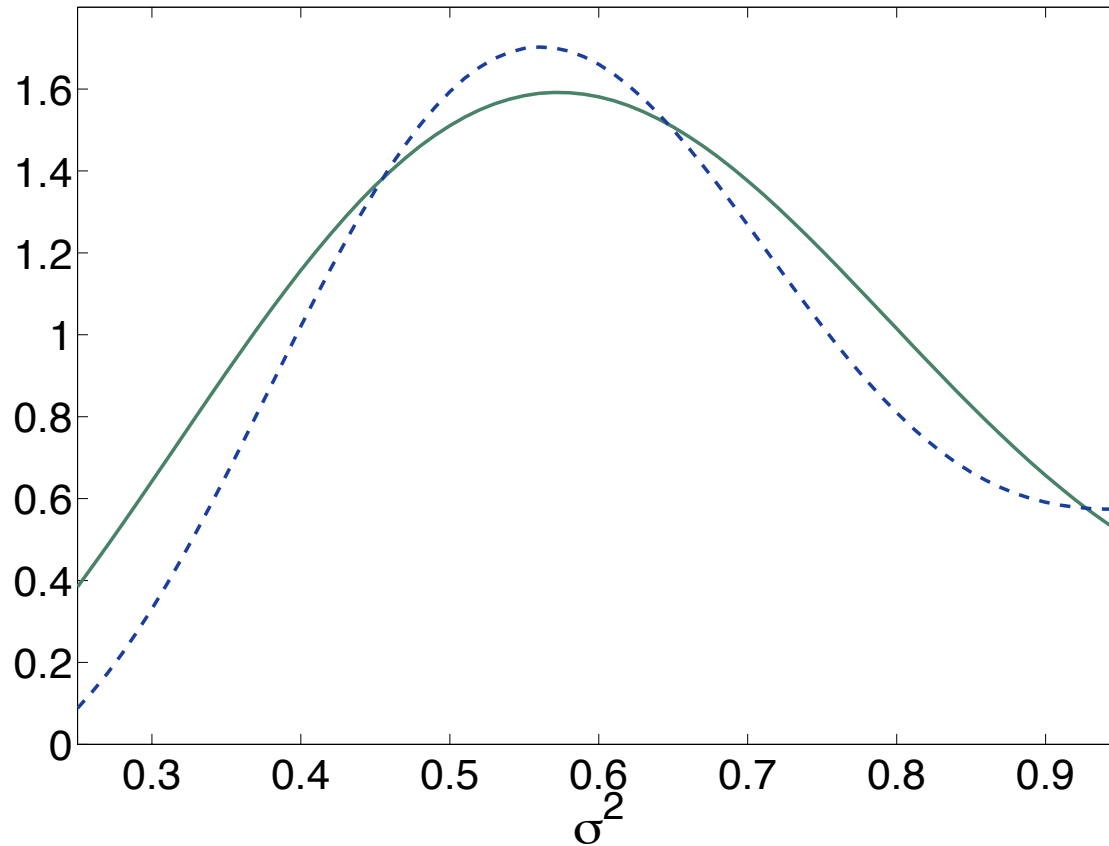


Variational result and comparison to MCMC

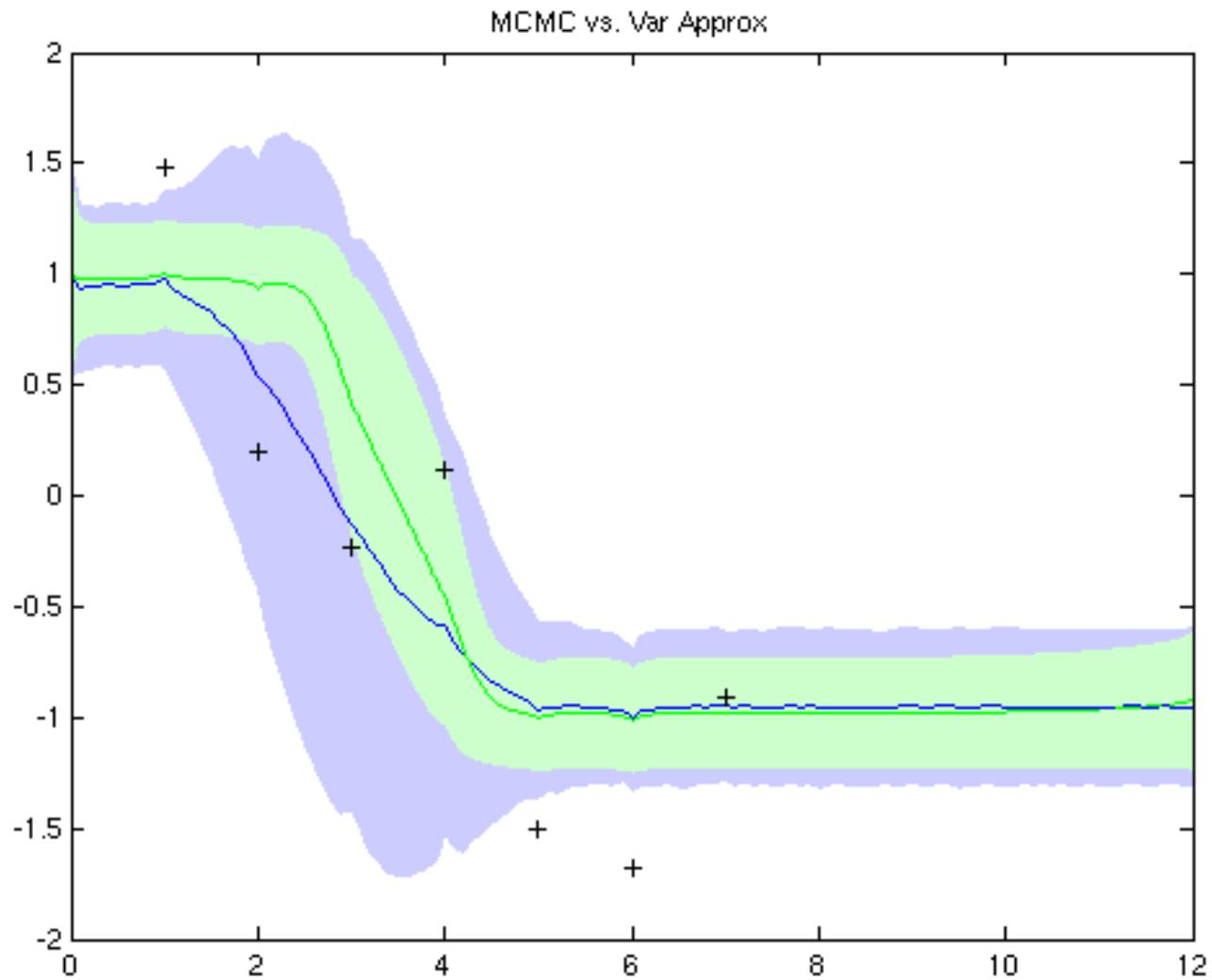


Estimation of Diffusion constant σ^2

$$q(\theta|Y) \approx \frac{e^{-\mathcal{F}_\theta(Q,Y)} p(\theta)}{\int e^{-\mathcal{F}_\theta(Q,Y)} p(\theta) d\theta}$$

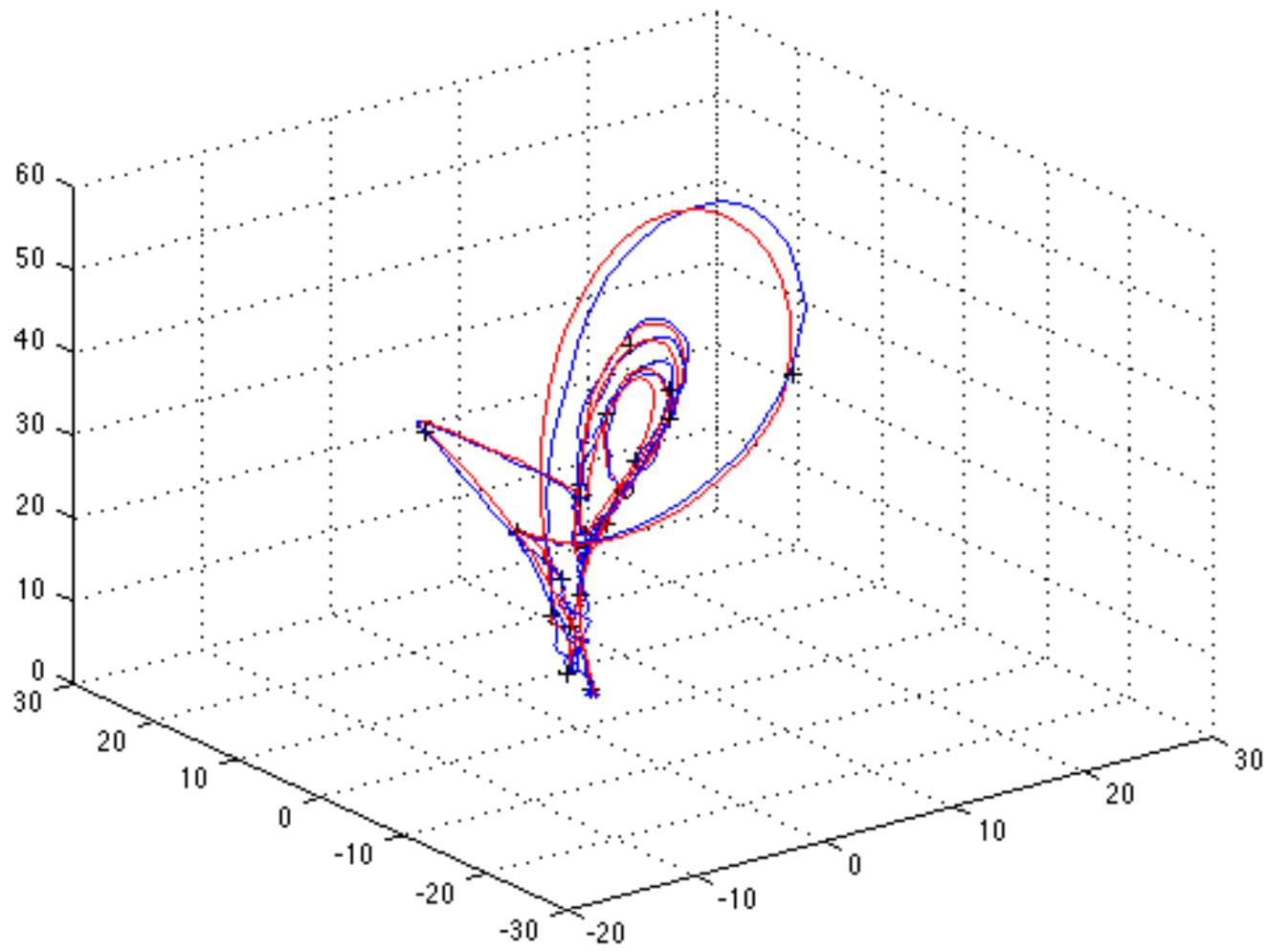


Large observation noise

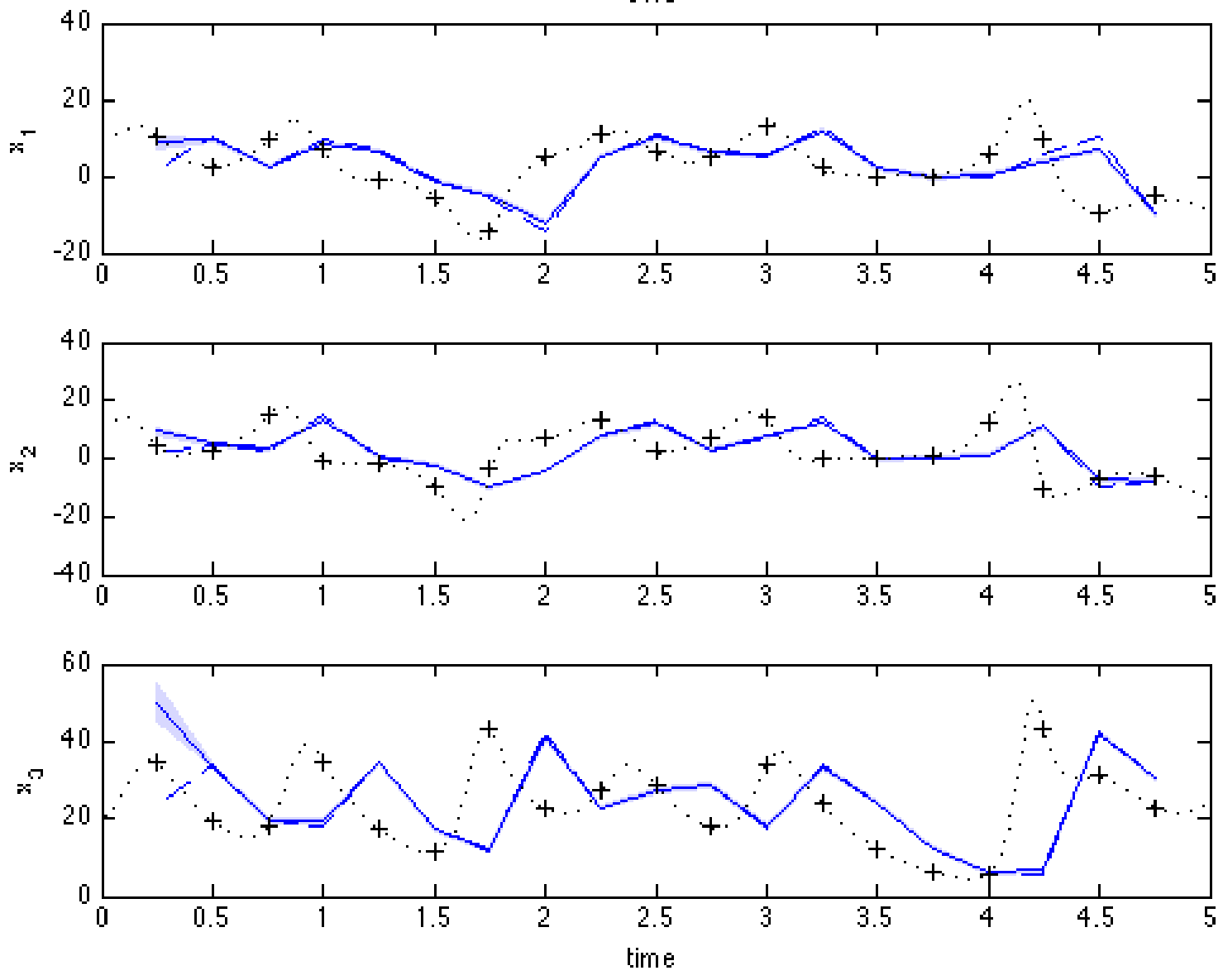


Double well with observation noise $\sigma_o = 0.6$

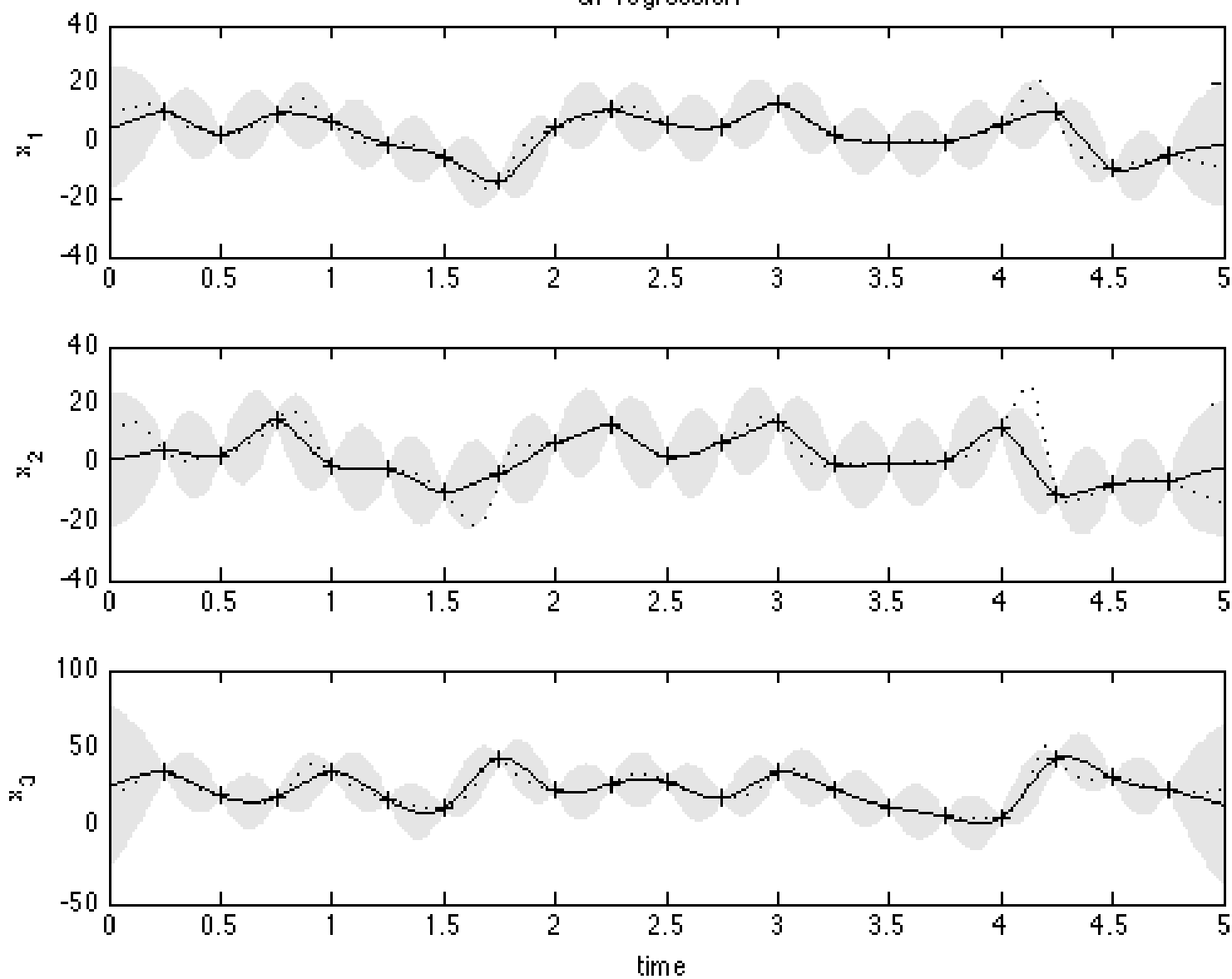
Noisy Lorenz system



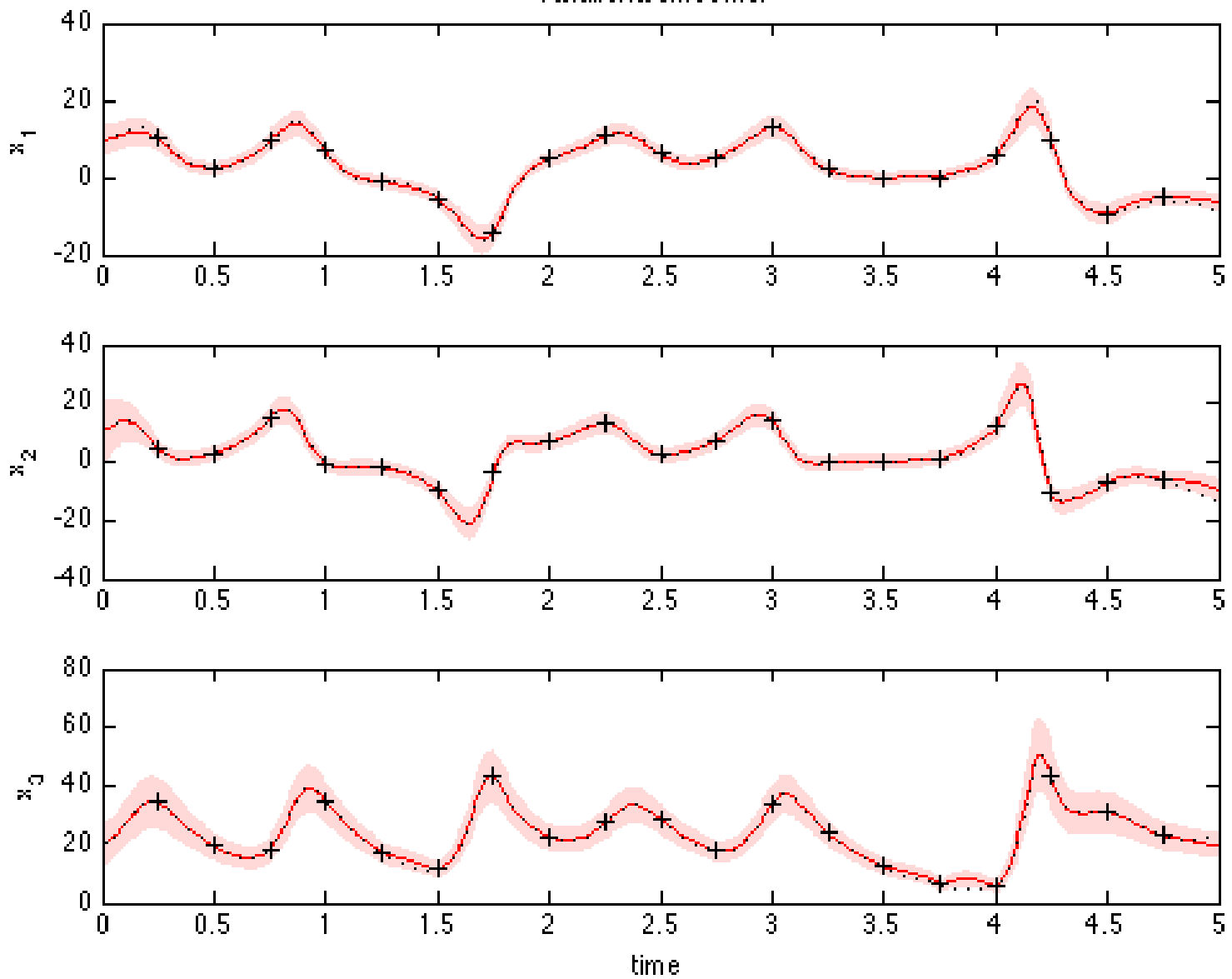
UKS



GP regression



Variational smoother



MJPs: Mean field approximation

Multivariate states $X = (X(1), \dots, X(d))$

Exact inference: Linear ODEs in S^d variables

Variational approximation: Optimise in family of factorising measures, i.e. of the type

$$Q[X_{0:T}] = \prod_{i=1}^d Q_i[X_{0:T}(i)]$$

Linear ODEs in Sd variables.

Simple gene autoregulatory network

- Proteins (y) produced (*translation*) with rate proportional to mRNA (x) abundance

$$f_p(y + 1|x, y) = \gamma x$$

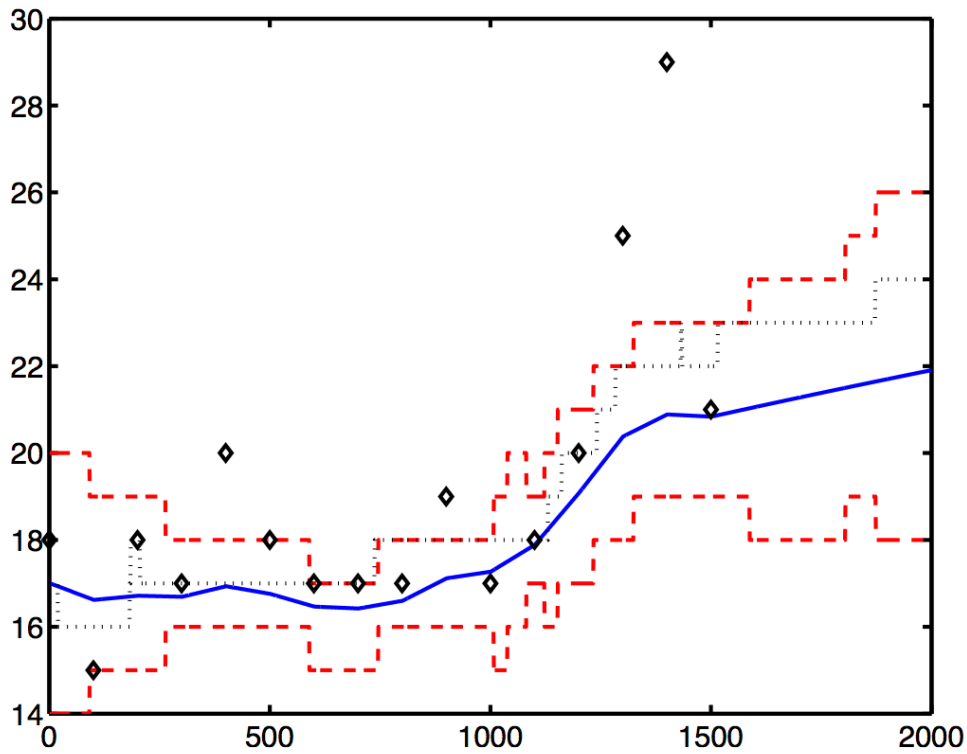
- If protein numbers $y > y_c$, mRNA production drops dramatically by factor 100.

$$f_{RNA}(x + 1|x, y) = \alpha (1 - 0.99 \times \Theta(y - y_c))$$

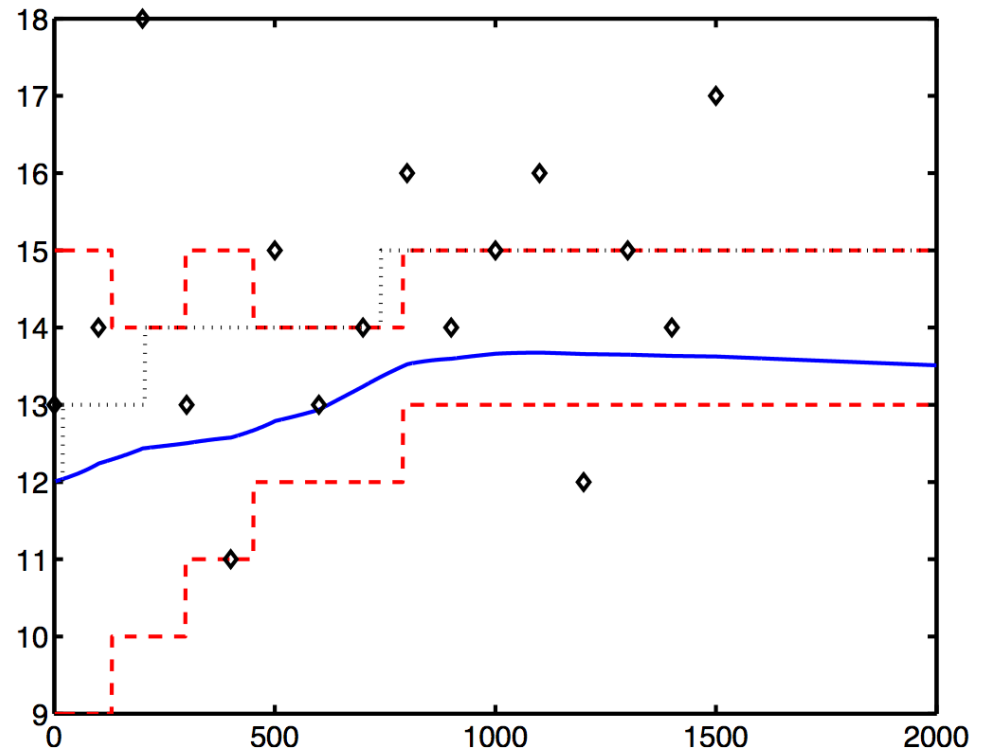
- Decay of proteins and mRNA

$$\begin{aligned} f_p(y - 1|x, y) &= \delta y \\ f_{RNA}(x - 1|x, y) &= \beta x \end{aligned}$$

Protein



mRNA

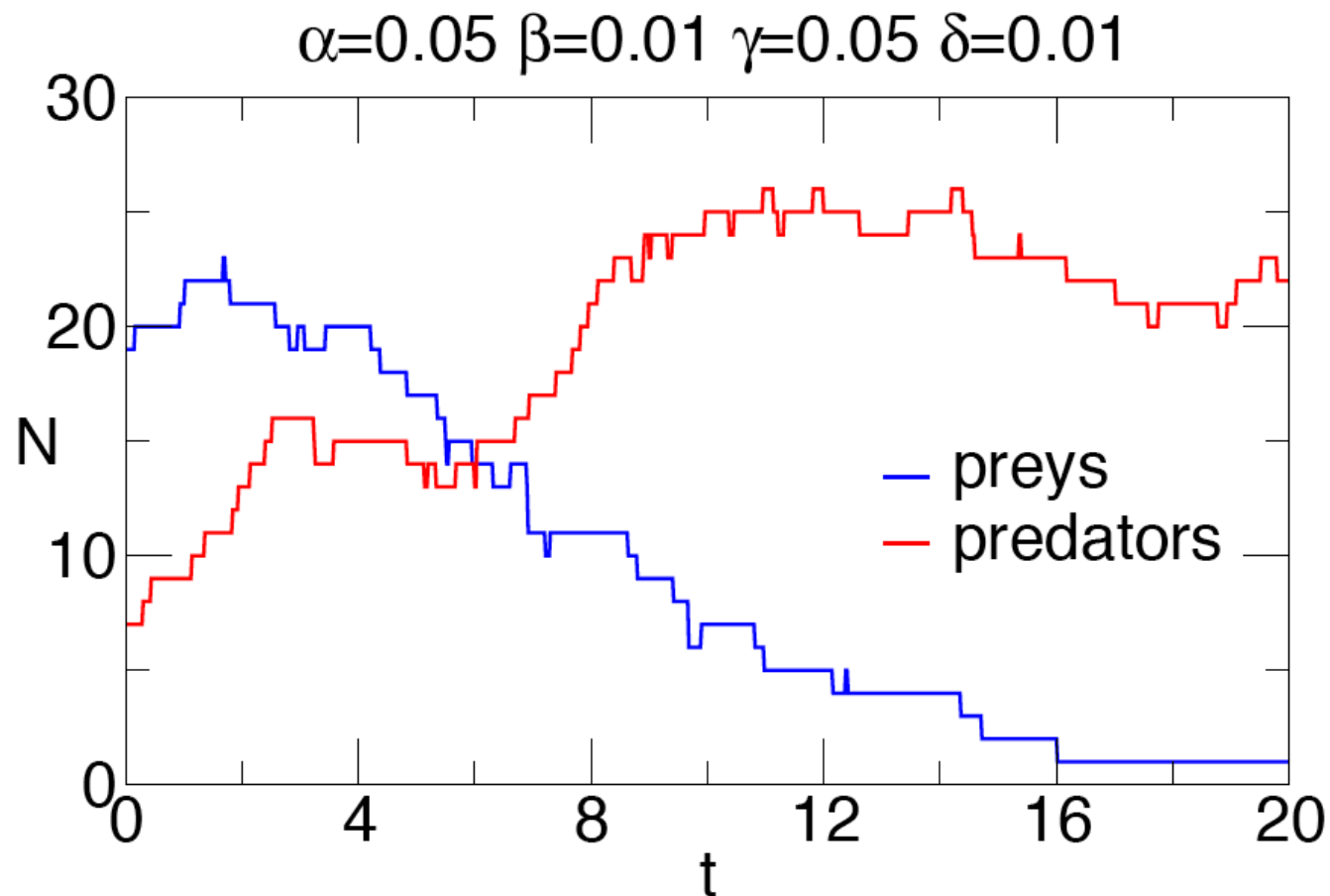


(M. Opper & G. Sanguinetti, NIPS07)

Critical parameter $y_c = 20$ (exact) and $y_c = 19$ (ML estimate).

Diffusion Limit of MJP

Stochastic Lotka Volterra Model: Jumps small compared to number of species \rightarrow **SDE with multiplicative noise**



Back to SDEs: Problems with Gaussian approx.

- Gaussian SDEs must have state independent diffusion terms $D \neq D(X)$ (additive noise).
- $KL(Q\|P) = \infty$ if Q and P have different diffusion terms $D(X)$.
- Gaussian approximation impossible for $D = D(X)$ (multiplicative noise) !
- Transformations $Z = \phi(X)$ to additive noise usually impossible in multivariate case.

Weak noise solution to KSP equation

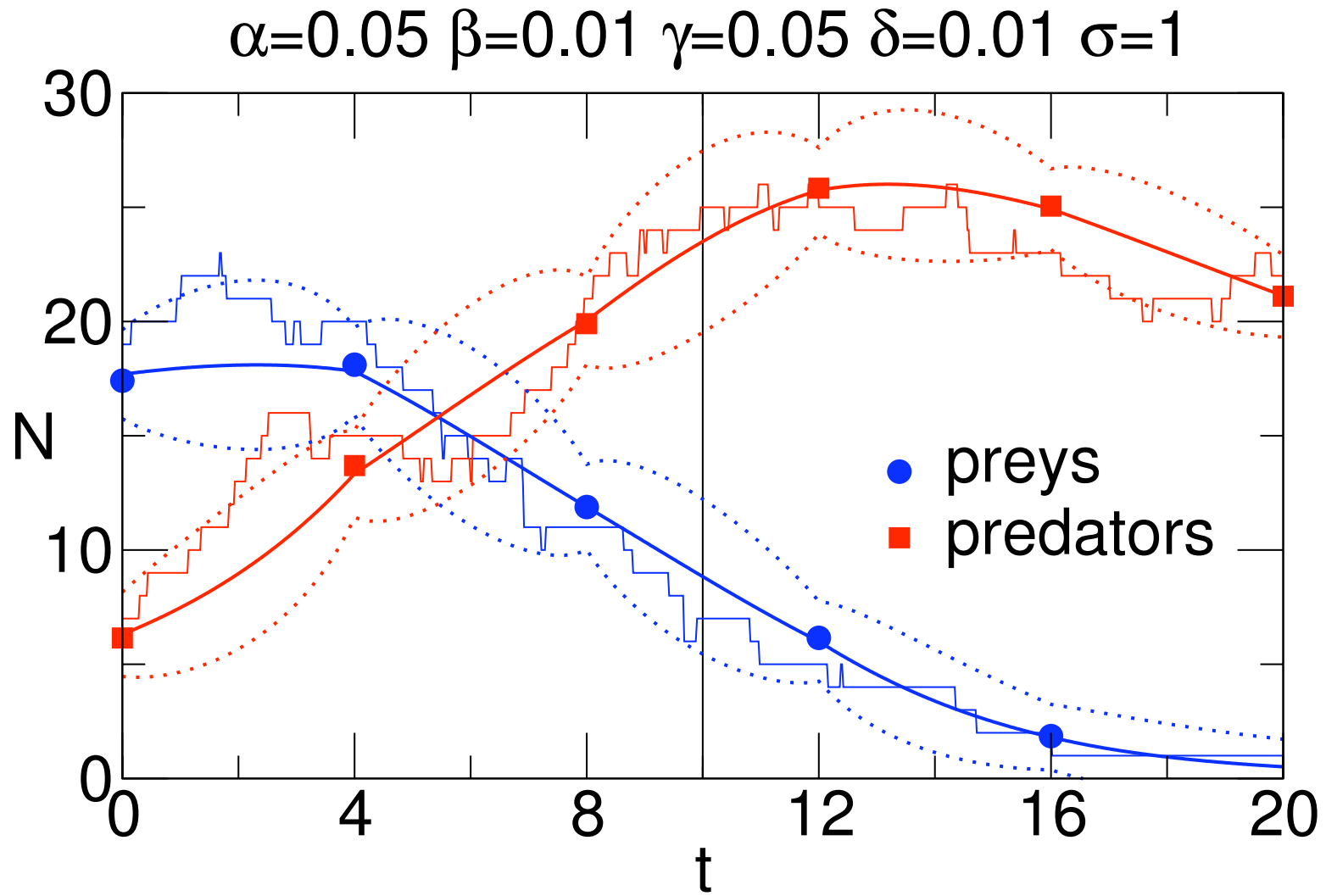
$$\left\{ \frac{\partial}{\partial t} + f^\top \nabla + \frac{1}{2} \text{Tr}(D(x, t) \nabla^\top \nabla) \right\} \psi(x, t) = 0$$

Set $D \rightarrow \epsilon D$ and assume **scaling form**

$$\psi(x, t) \approx z(t) \exp \left[-\frac{1}{2\epsilon} (x - m(t))^\top S^{-1}(t) (x - m(t)) \right]$$

To leading order, obtain the ODEs

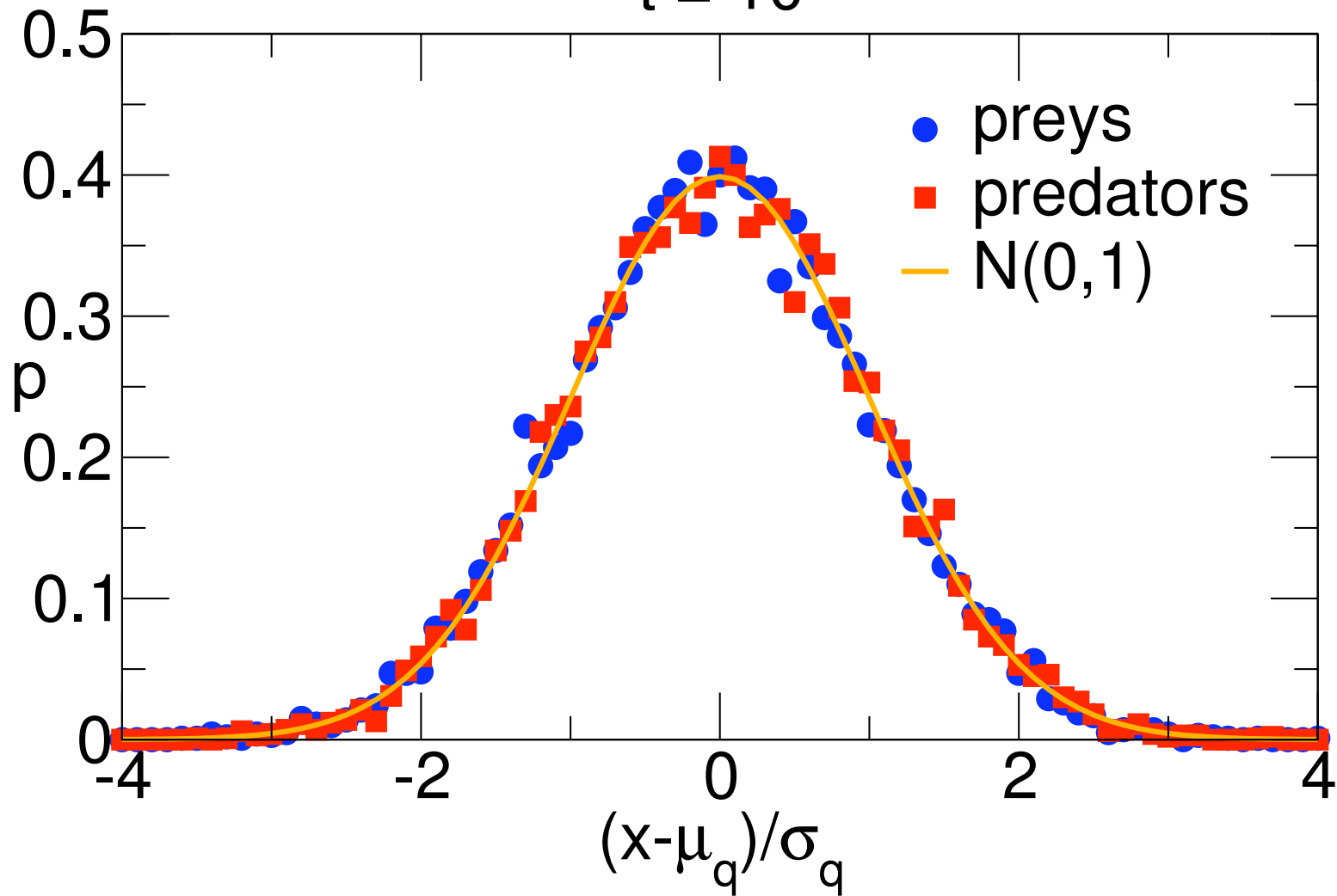
$$\frac{dm}{dt} = f(m) \quad \text{and} \quad \frac{dS}{dt} = AS + SA^\top - D(m) \quad \text{with} \quad A_{ij}(t) = \left. \frac{\partial f_i}{\partial x_j} \right|_{x(t)=m(t)} .$$



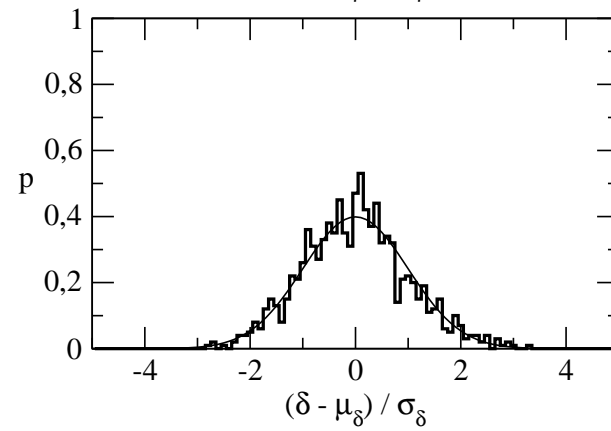
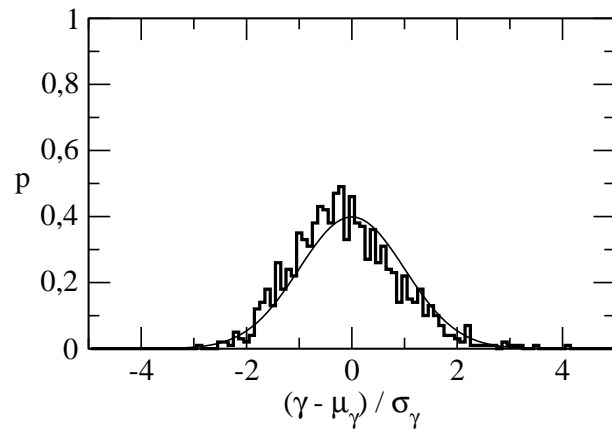
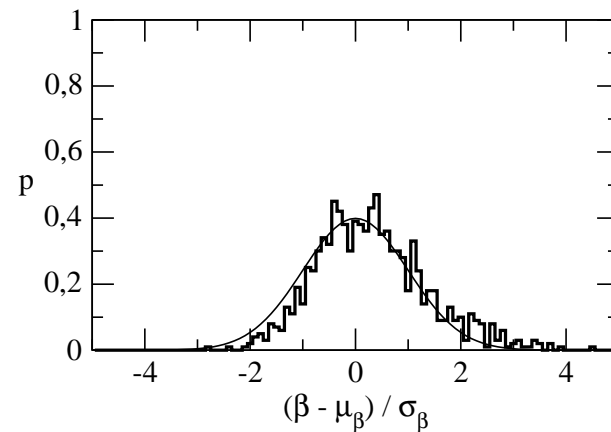
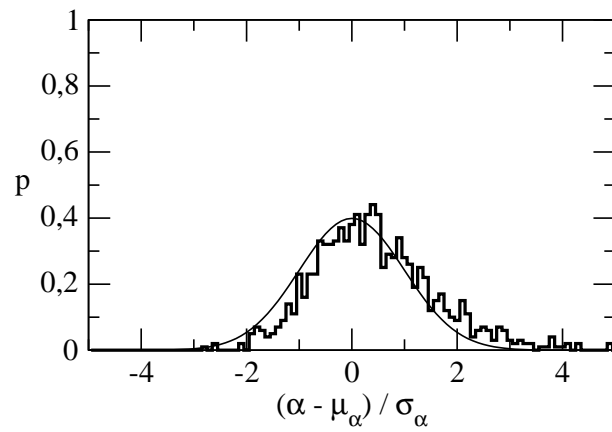
(A. Ruttor & M. Opper)

Calibration of states

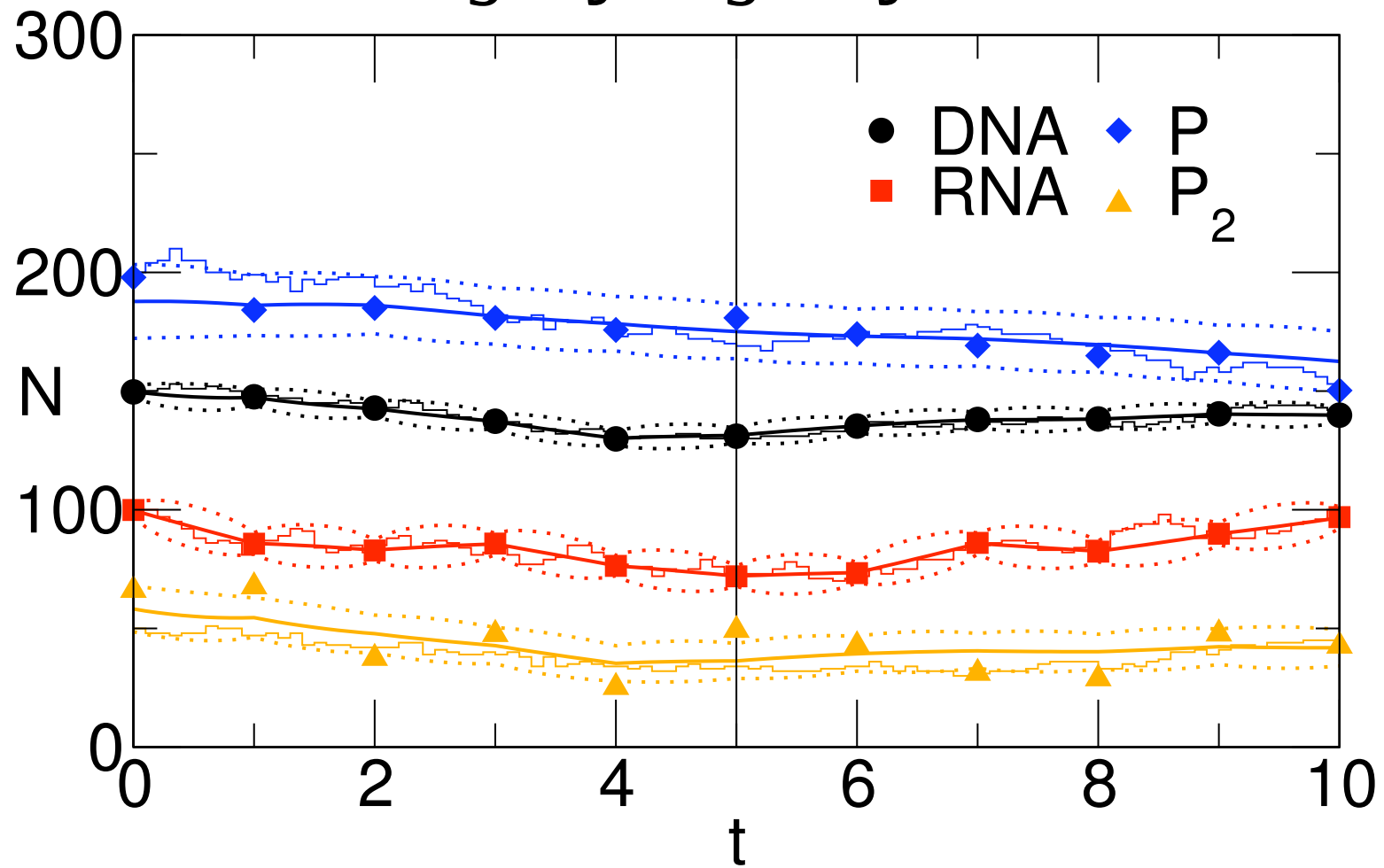
$t = 10$



Prediction error for parameters

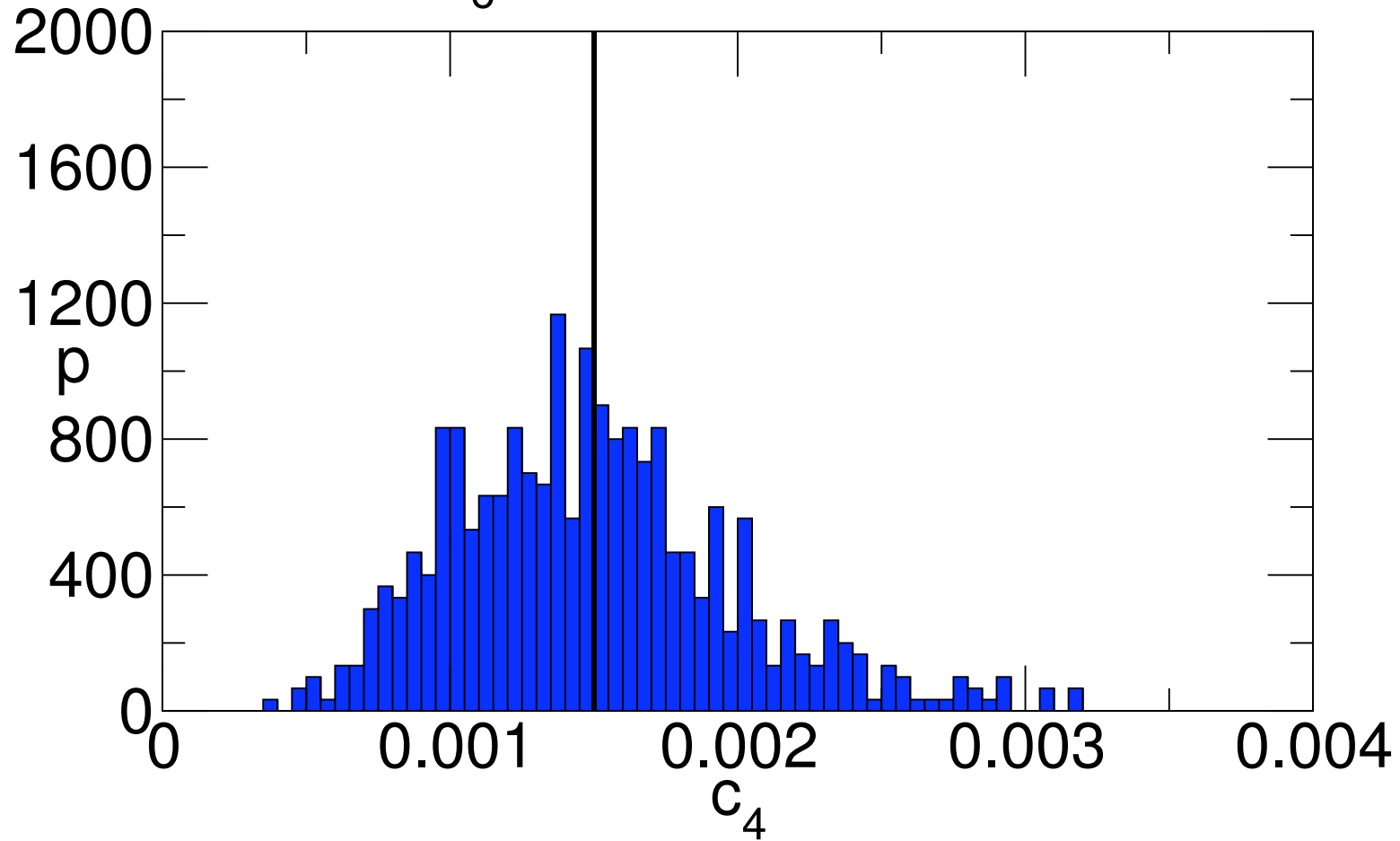


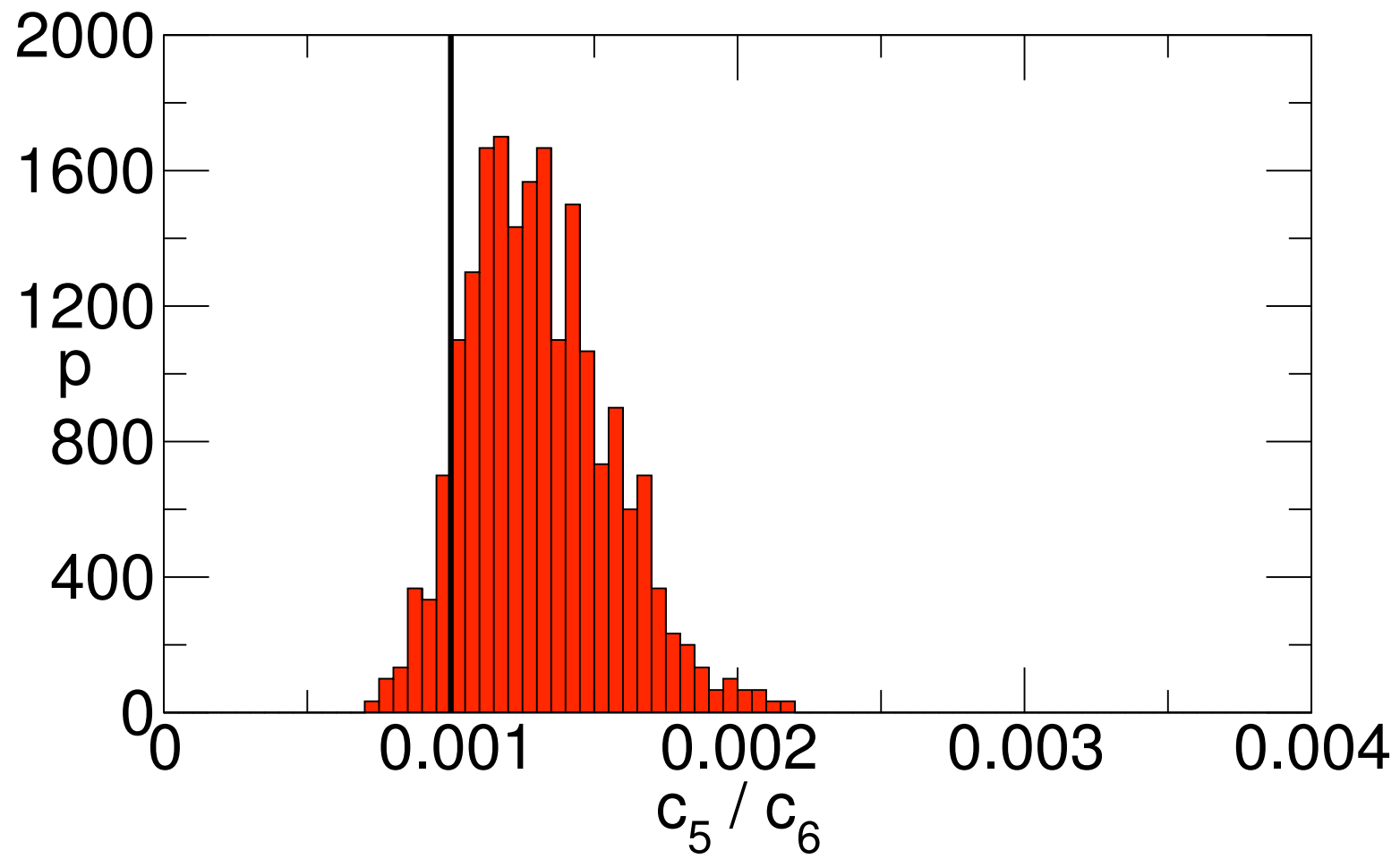
A slightly larger system



Hybrid MC for posterior density of parameters

$N_0=11$ $\sigma=(1, 1, 10, 10)$





Hybrid systems

$$\frac{dx}{dt} = -\lambda x(t) + Az(t) + b$$

$z(t) \in \{-1, 1\}$ and $z \rightarrow -z$ with rates γ_{\pm} .

Present & Future work

- Variational path densities as proposal for MCMC.
- Realistic higher dimensional systems: Need good suboptimal variational parametrisations.
- Variational approach to multiplicative noise !
- Other processes, coloured noise
- Perturbative corrections

A Hamiltonian approach for the 'potential' case

$$\mathcal{A}_{sde} = \int_0^T \mathcal{L}_{sde} dt \quad \text{with} \quad \mathcal{L}_{sde} = \frac{1}{2\sigma^2} \langle \|f + Ax - b\|^2 \rangle$$

Try to express A and b by S and m using

$$\frac{dm}{dt} = -Am + b \quad \frac{dS}{dt} = -AS - SA^\top + \sigma^2 I$$

Possible if $f(x) = -\nabla U(x)$ and after some calculations

$$\begin{aligned} 2\sigma^2 \mathcal{L} = & \left(\frac{dm}{dt} \right)^2 + \langle \|f\|^2 \rangle + \sigma^2 \langle \nabla f \rangle + 2 \frac{d\langle U \rangle}{dt} \\ & + \frac{\sigma^4}{4} \text{Tr} (S^{-1}) - \frac{\sigma^2}{2} \text{Tr} \left(S^{-1} \frac{dS}{dt} \right) + \frac{1}{2} \int_0^\infty dz \text{Tr} \left(e^{-zS} \frac{dS}{dt} \right)^2 \end{aligned}$$

The 1 - D case

Set $S \rightarrow s^2$ (variance). Then

$$\frac{ds}{dt} = -\alpha s + \frac{\sigma^2}{2s} \quad \rightarrow \quad \alpha = \frac{\sigma^2}{2s^2} - \frac{1}{s} \frac{ds}{dt}$$

Thus the action is

$$\mathcal{A} = \frac{1}{2\sigma^2} \int_0^T \left\{ \left(\frac{ds}{dt} \right)^2 + \left(\frac{dm}{dt} \right)^2 - V_{eff}(m, s) \right\} dt + \text{data \& surface terms}$$

with

$$V_{eff}(m, s) = - \left(\frac{\sigma^4}{4s^2} + \langle f^2(x) \rangle + \sigma^2 \langle f'(x) \rangle \right)$$

The Hamilton equations

Between observations we have the Hamiltonian flow

$$\begin{aligned} 2\frac{d^2m}{dt^2} &= -\frac{\partial V_{eff}(m, s)}{\partial m} \\ 2\frac{d^2s}{dt^2} &= -\frac{\partial V_{eff}(m, s)}{\partial s} \end{aligned}$$

Data and surface terms

$$\mathcal{A}_{data} = \frac{1}{2\sigma_0^2} \sum_i \left\{ s^2(t_i) + (y_i - m(t_i))^2 \right\} - \frac{1}{2} \ln \frac{s(T)}{s(0)} + \frac{1}{\sigma^2} (\langle U(T) \rangle - \langle U(0) \rangle)$$

lead to jump conditions for momenta

$$\begin{aligned} \frac{dm}{dt}(t_i^+) - \frac{dm}{dt}(t_i^-) &= \frac{m(t_i) - y_i}{2\sigma_0^2} \\ \frac{ds}{dt}(t_i^+) - \frac{ds}{dt}(t_i^-) &= \frac{1}{2\sigma_0^2} s(t_i) \end{aligned}$$

Ornstein - Uhlenbeck Process

$$f = -\gamma x \rightarrow V_{eff}(m, s) = -\frac{\sigma^4}{4s^2} - \gamma^2 s^2 - \gamma^2 x^2 + \sigma^2 \gamma$$

The potential separates in m and s !

Potential $V_{eff}(m)$ and **Prediction** $m(t) = y \frac{\sinh(\gamma t)}{\sinh(\gamma T)}$ for perfect observation $x(T = 5) = 3$.

