Variational approximations to inference for stochastic differential equations

Manfred Opper

TU Berlin, Dept of Computer Science

Collaborators:

Cédric Archambeau (UCL)

Phillip Batz (TU Berlin)

Remi Barillec (Aston)

Dan Cornford (Aston)

Renata Retkute (Reading)

Ian Roulstone (Reading)

Andreas Ruttor (TU Berlin)

Guido Sanguinetti (Edinburgh)

John Shawe–Taylor (UCL)

Yuan Shen (Aston)

Michail Vrettas (Aston)

Overview

- Inference for stochastic differential equations

- Variational approach $\neq$ 4D–Var

- Variational approximations for path probabilities

- Experiments

- Outlook

Ito stochastic differential equations

for state $X_t \in R^d$

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

Limit of discrete time process X_k

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

ϵ_k i.i.d. Gaussian.

Inference Problems

Given **noisy observations** $\{y_i\}_{i=1}^{N} \equiv y_1, \ldots, y_N$ of **hidden process** X_{t_i} at times $t_i \leq T$ for $i = 1, \ldots, N$.

- Estimate X_t for $0 \leq t \leq T$ (**smoothing**).

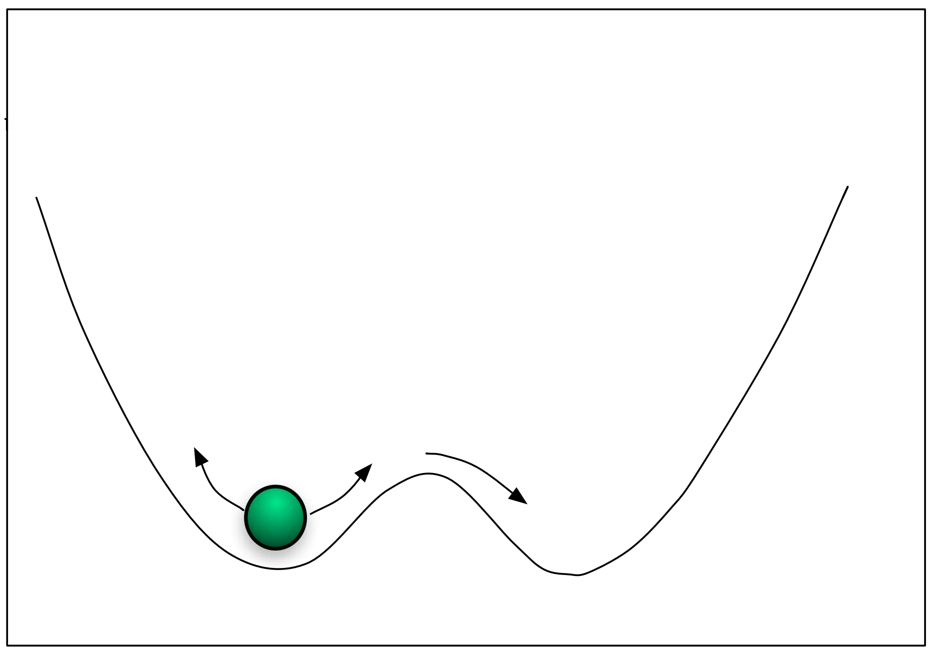
- Estimate system parameters θ contained in drift f and diffusion Σ.

Motion in double-well potential

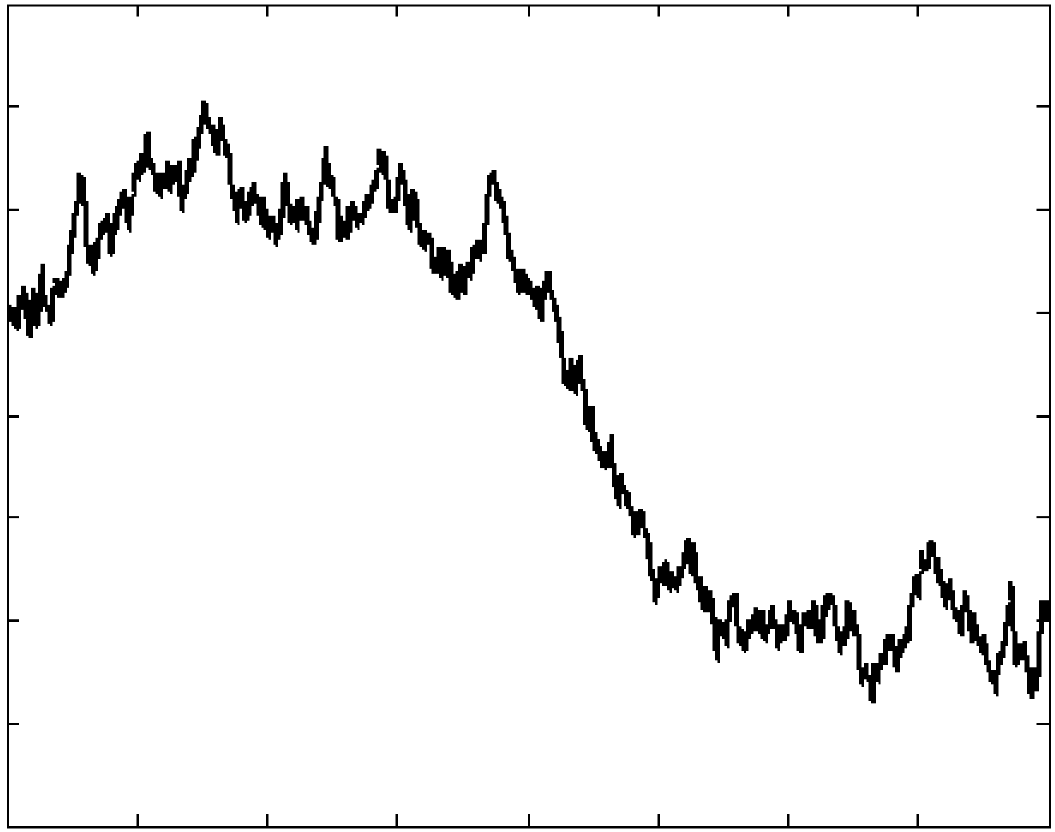
$$dX = f(X)dt + \sigma dW.$$

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

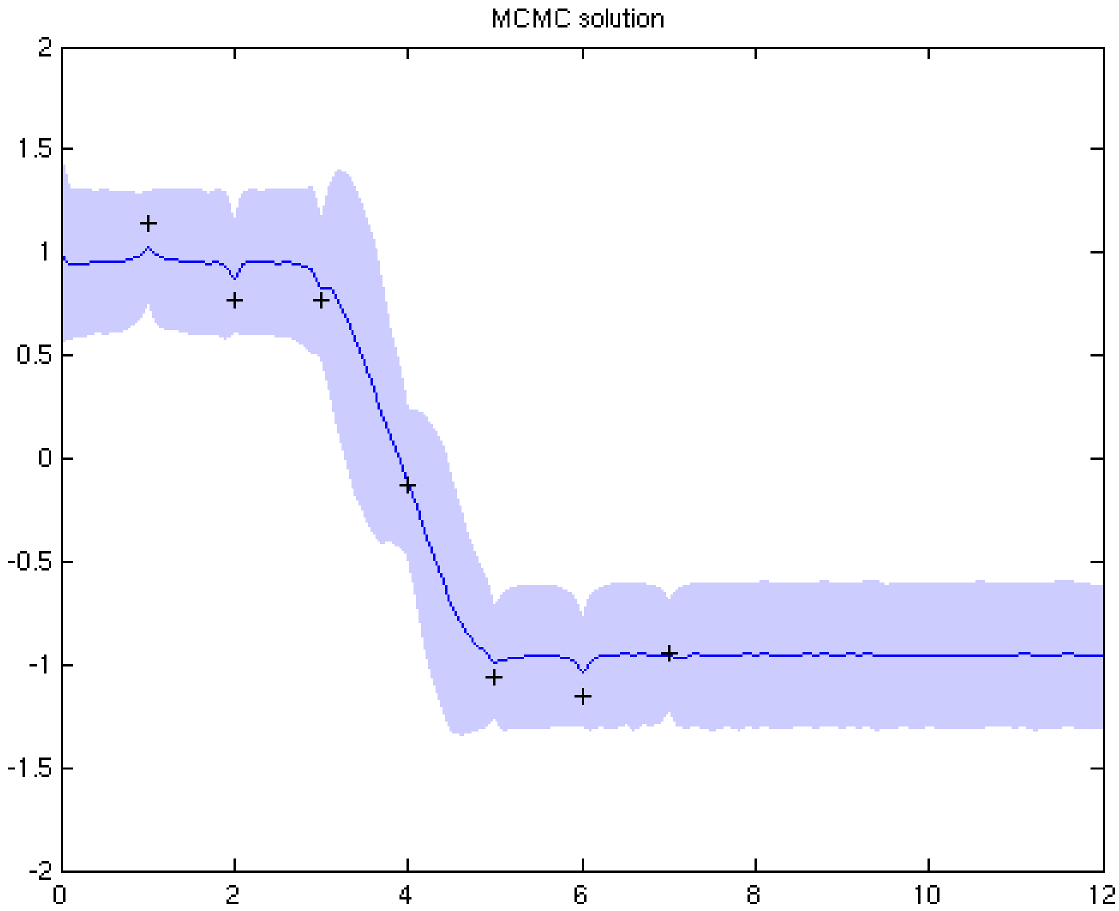
and $V(x)$ is a double well potential



A sample path might look like this



Observations & optimal prediction



What we would like to do

- **State estimation:** Use conditional (posterior) distribution over **paths** $X_{0:T}$ (an ∞ dimensional object)

$$\frac{dP(X_{0:T}|\{y_i\}_{i=1}^N,\theta)}{dP_{prior}(X_{0:T}|\theta)} = \frac{1}{p(\{y_i\}_{i=1}^N|\theta))} \times \prod_{n=1}^N p(y_n|X_{t_n},\theta),$$

  to compute prediction $E[X_t|\{y_i\}_{i=1}^N,\theta]$

- **Parameter estimation**: Maximise $p(\{y_i\}_{i=1}^N|\theta)$ with respect to θ (**Max Likelihood**) or use a prior $p(\theta)$ to compute $p(\theta|\{y_i\}_{i=1}^N) \propto p(\{y_i\}_{i=1}^N|\theta)p(\theta)$ (**Bayes**).

The conditional distribution and likelihood $p(\{y_i\}_{i=1}^N|\theta)$ are not easily tractable !

The variational approximation

Approximate intractable posterior

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

by a $q(x)$ which belongs to a family of **simpler tractable** distributions (e.g. factorising = mean field, or Gaussian densities).

Optimise q by minimising the **relative entropy**

$$D[q\|p(\cdot|y)] = \int q(x) \ln \frac{q(x)}{p(x|y)}\, dx =$$

$$\int q(x) \ln \frac{q(x)}{p(x)}\, dx - \int q(x) \ln p(y|x)\, dx + \ln p(y)$$

The statistical physics version

Set $p(x|y) = \frac{1}{Z}\, e^{-H^y(x)}$ and $q(x) = \frac{1}{Z_0}\, e^{-H_0(x)}$

The variational bound on the free energy is

$$-\ln Z \leq -\ln Z_0 + \langle H^y(x) \rangle_0 - \langle H_0(x) \rangle_0$$

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

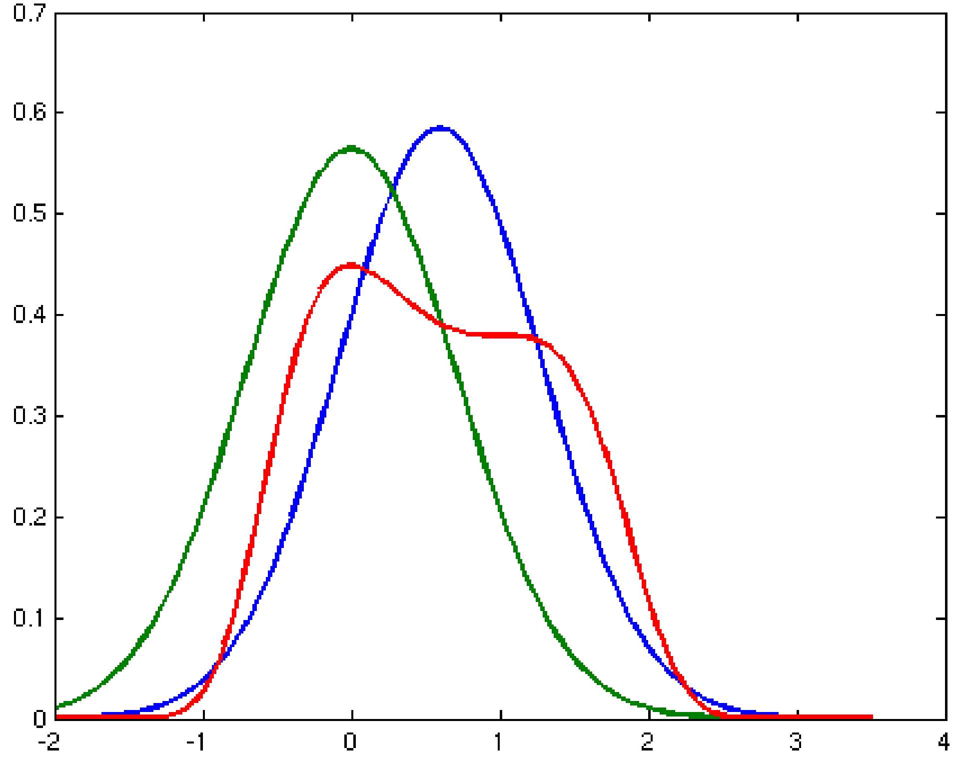
Equivalent to first order perturbation theory around H_0

Approximation for free energies is often better than the quality of q.

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

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

The Gaussian Variational Approximation



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

Variational free energy

$$\begin{aligned}
\mathcal{F}(q) &= D[q\|p(\cdot|y)] - \ln p(y) \\
&= D[q\|p] - \int q(x) \ln p(y|x)\, dx \\
&\geq -\ln p(y)
\end{aligned}$$

Approximate maximum likelihood estimate

Assume model depends on parameter θ. The free energy inherits the dependency.

Let $q^*(\theta) = \operatorname{argmin} \mathcal{F}_\theta(q)$. Since

$$- \ln p(y|\theta) \leq \mathcal{F}_\theta(q^*(\theta))$$

we can minimise $\mathcal{F}_\theta(q^*)$ wrt θ to get an approximate maximum likelihood estimate.

Approximate Bayesian parameter inference

Approximate posterior of parameters (Lappalainen, 2000):

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

How to choose the measure q for stochastic differential equations ?

- Process conditioned on data is Markovian!

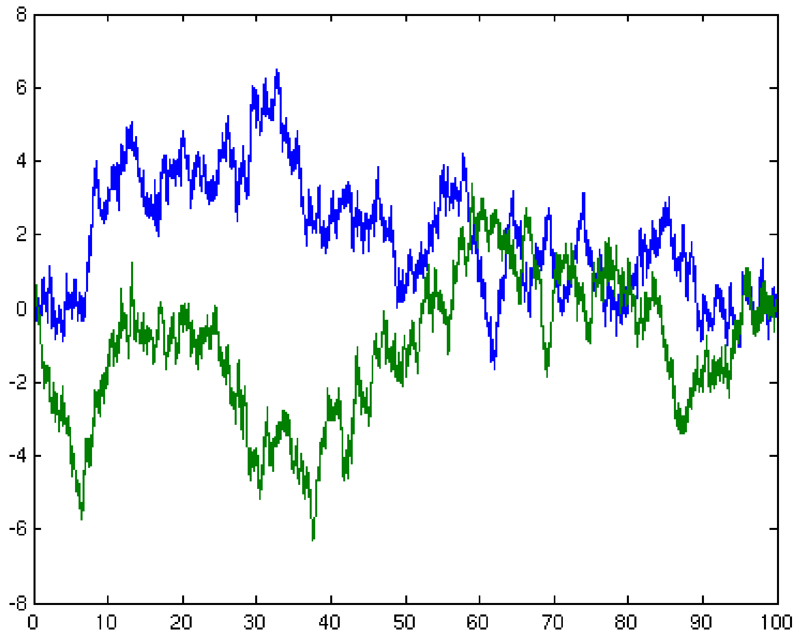
- It fulfils SDE

$$dX_t = g(X_t, t)dt + \Sigma^{1/2}(X_t)\, dW_t$$

with a new time dependent drift $g(X_t, t)$ but the **same diffusion** Σ.

Example

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



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

Relative entropy for path probabilities: A physics style derivation

Use representation of joint density in term of conditionals and the Markov property (assuming $q_0(x) = p_0(x)$) and work with time discretization $t_{k+1} - t_k = \Delta t$.

$$
\begin{aligned}
D\left[q \| p\right] &= \int dx_{0:T}\, q(x_{0:T}) \ln \frac{q(x_{0:T})}{p(x_{0:T})} \\
&\approx \sum_{k=0}^{K-1} \int dx\, q_{t_k}(x) \int dx'\, q_{t_{k+1},t_k}(x'|x) \ln \frac{q_{t_{k+1},t_k}(x'|x)}{p_{t_{k+1},t_k}(x'|x)} \\
&= \sum_{k=0}^{K-1} \int dx\, q_{t_k}(x) D\left[q_{t_{k+1},t_k}(\cdot|x) \| p_{t_{k+1},t_k}(\cdot|x)\right]
\end{aligned}
$$

in terms of transition and marginal probabilities.

We know that short time transition probability

is approximately Gaussian

$$p_{t+\Delta t,t}(x'|x) \propto \exp\left[-\frac{1}{2\Delta t}\left\|x' - x - f(x)\Delta t\right\|_\Sigma^2\right]$$

as $\Delta t \to 0$,

with $\|F\|_\Sigma^2 = F^\top \Sigma^{-1} F$.

Then for small Δt

$$D\left[q_{t_{k+1},t_k}(\cdot|x)\|p_{t_{k+1},t_k}(\cdot|x)\right] \approx \frac{1}{2}\|g(x,t) - f(x)\|_\Sigma^2 \Delta t$$

The relative entropy for Stochastic Differential Equations

Let q and p be measures over paths for SDEs with drifts $g(X,t)$ and $f(X,t)$ with **same diffusion** $\Sigma(X)$. Then

$$D\left[q\|p\right] = \frac{1}{2}\int_0^T dt \left\{\int dx \; q_t(x) \;\; \|g(x,t) - f_\theta(x)\|^2_\Sigma\right\}$$

$q_t(x)$ is the marginal density of X_t.

The variational problem (exact inference !)

Minimise variational free energy

$$\mathcal{F}_\theta(q) = \frac{1}{2}\int_0^T \int q(x,t)\ \|g(x,t) - f_\theta(x)\|_\Sigma^2\ dx\ dt\ - \sum_i E_q[\ln p(y_i|X_{t_i})]$$

with respect to the marginal density $q(x,t)$.

The marginal density q and the drift $g(x,t)$ are coupled through the Fokker - Planck equation

$$\frac{\partial q}{\partial t} = \left\{-\nabla g + \frac{1}{2}\mathrm{Tr}(\nabla\nabla^T\Sigma)\right\} q$$

Variation leads to forward backward PDEs.

The Variational Gaussian Approximation for SDEs

(Archambeau, Cornford, Opper & Shawe - Taylor, 2007)

- Approximate (Gaussian) process over paths $X_{0:T}$ induced by linear SDE:

$$dX_t = \{A(t)X_t + b(t)\}\, dt + \Sigma^{1/2} dW$$

- Diffusion Σ must be independent of X !

- Relative entropy is of the form $\mathcal{F}_\theta[m, S, A, b]$.

- Constraints are evolution eqs. for marginal **mean** $m(t)$ and **covariance** $S(t)$

$$\begin{aligned} \frac{dm}{dt} &= Am + b \\ \frac{dS}{dt} &= AS + SA^\top + \Sigma. \end{aligned}$$

$\rightarrow$ **nonlinear ODEs** instead of PDEs !

Numerical approach

Introduce Lagrange multipliers

$$\begin{aligned}\mathcal{L} &= \mathcal{F}_\theta[m,S,A,b] - \mathrm{tr}\left\{\Psi^\top(t)\left(\frac{dS}{dt} - AS - SA^\top - \Sigma\right)\right.\\ &\left. -\lambda^\top(t)\left(\frac{dm}{dt} - Am - b\right)\right\}\, dt\end{aligned}$$

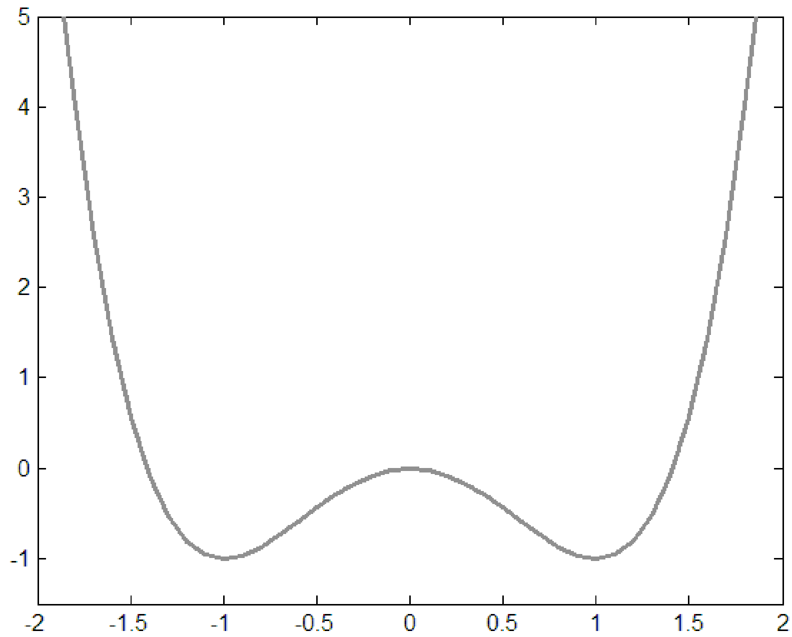
1. For given A and b run moment equations forward in time.

2. Derivatives wrt m and S lead to backward equation for Ψ and λ.

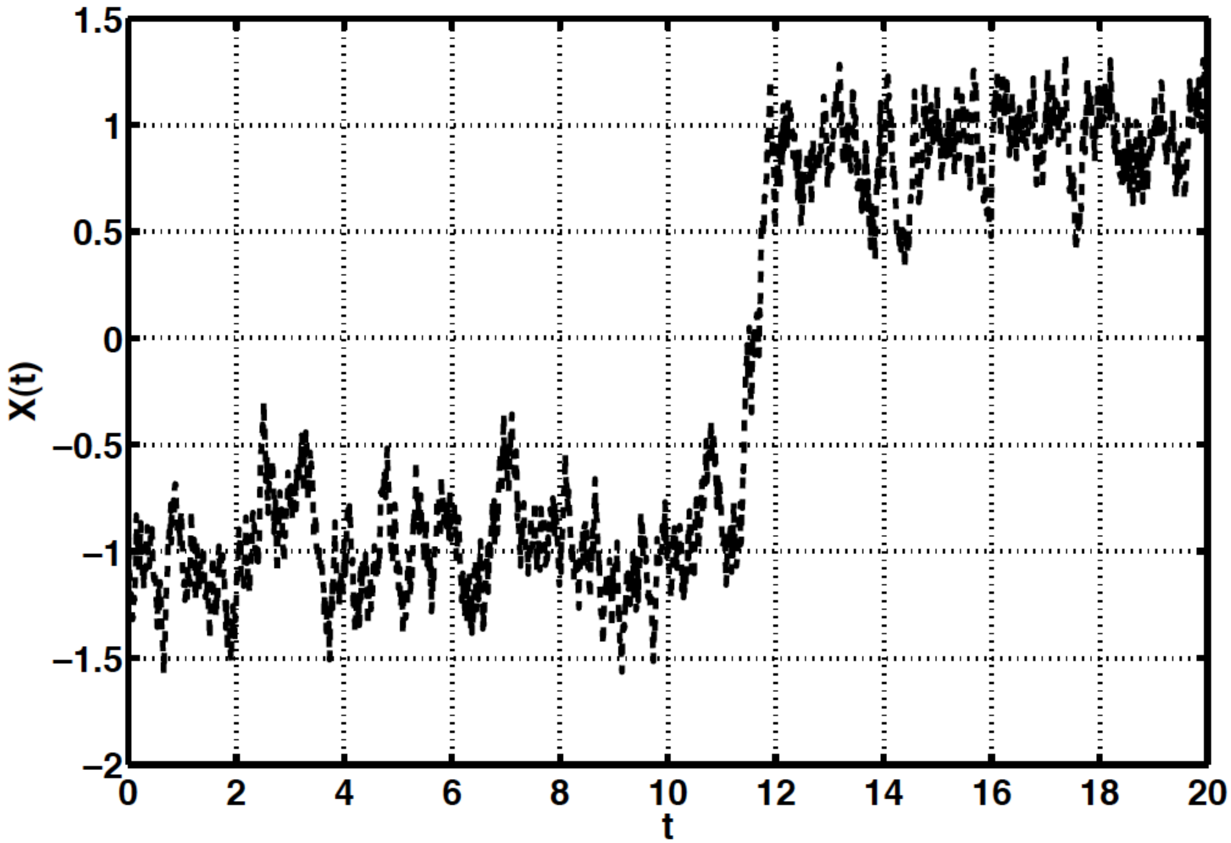
3. Compute gradient with respect to A and b.

Example: Motion in double-well potential

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

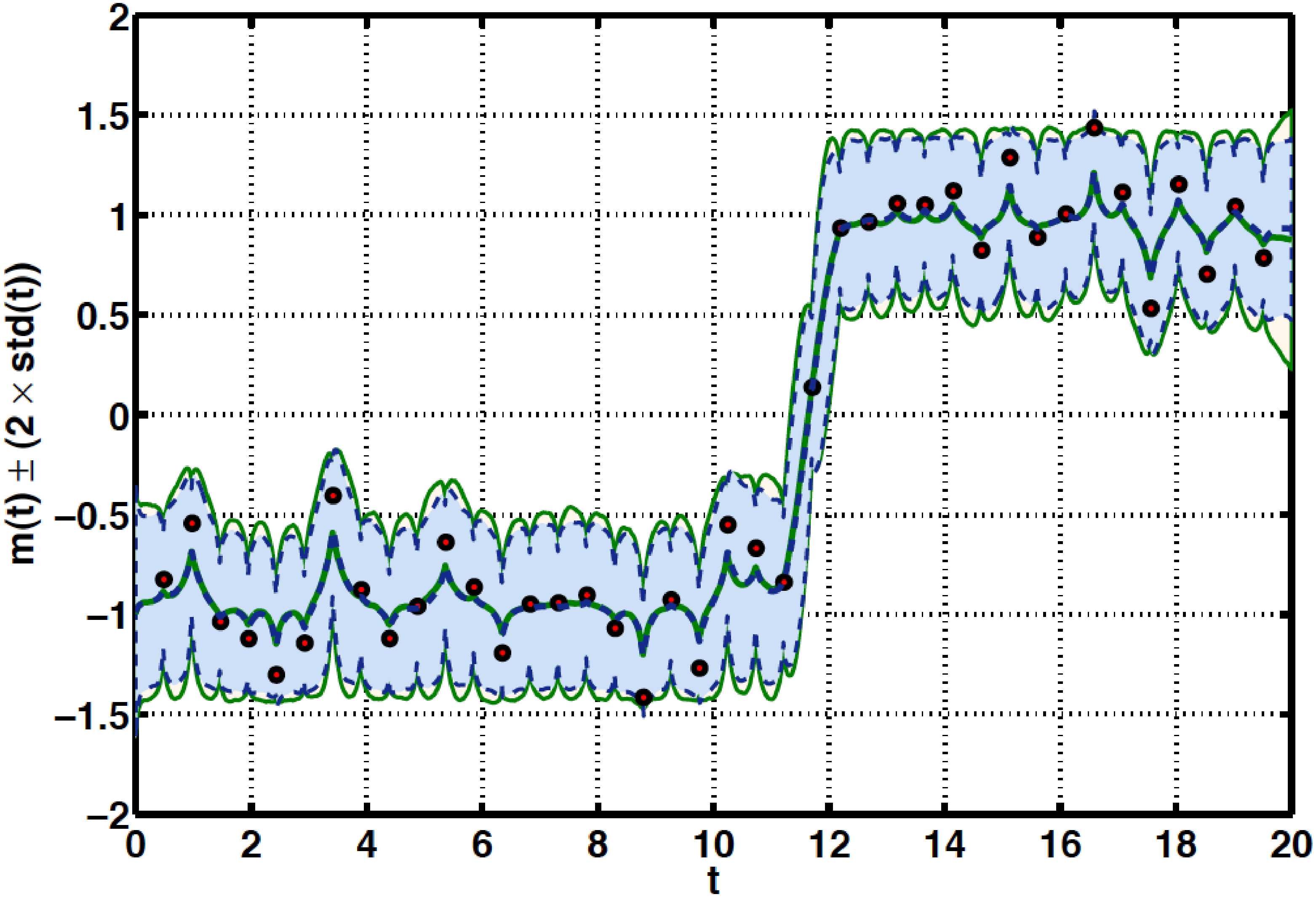


A trajectory

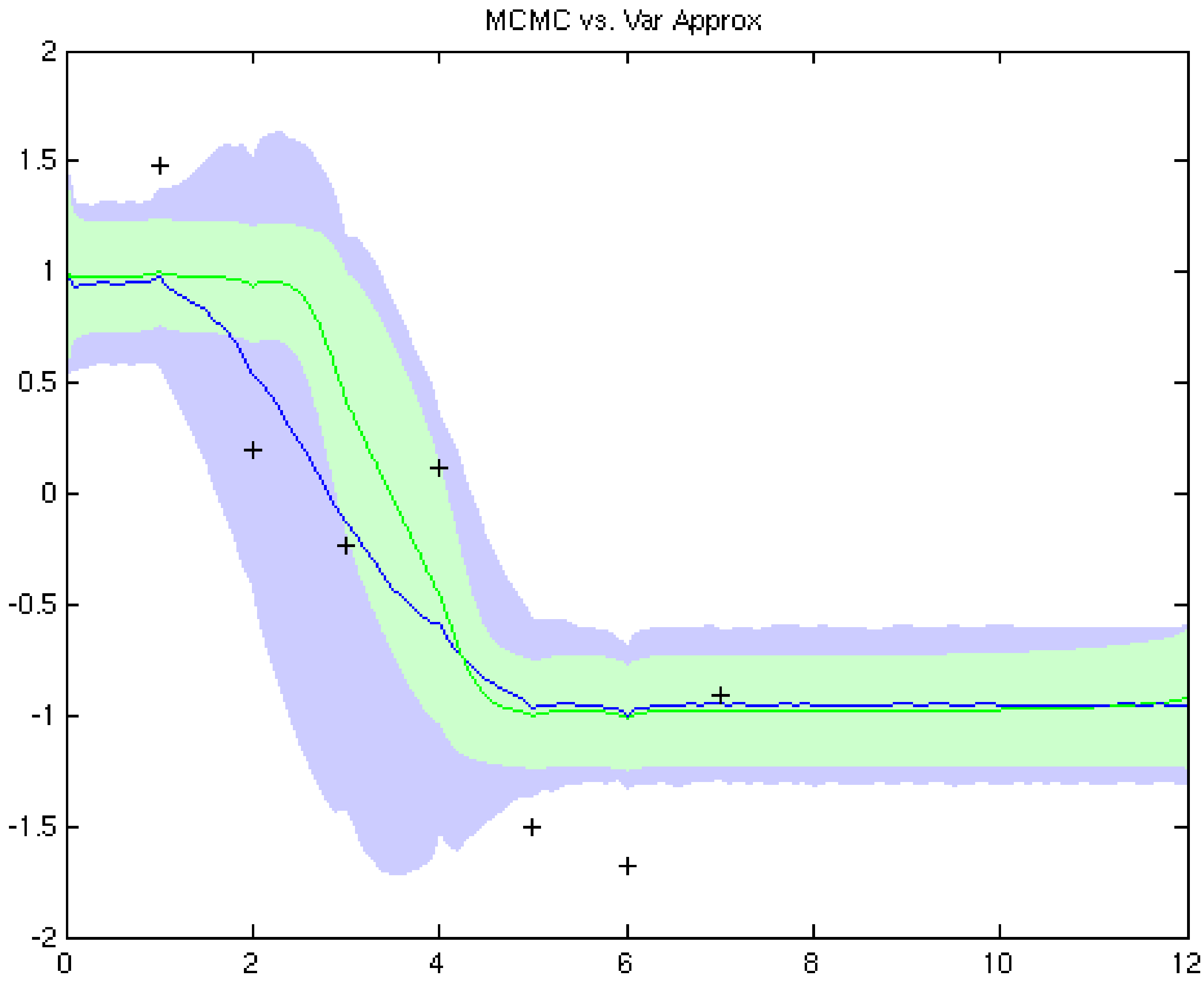


Prediction & comparison with hybrid Monte Carlo

$T = 20$, $\theta = 1$, $\sigma^2 = 0.8$ with $N = 40$ observations with noise $\sigma_o^2 = 0.04$. Fixed initial conditions.

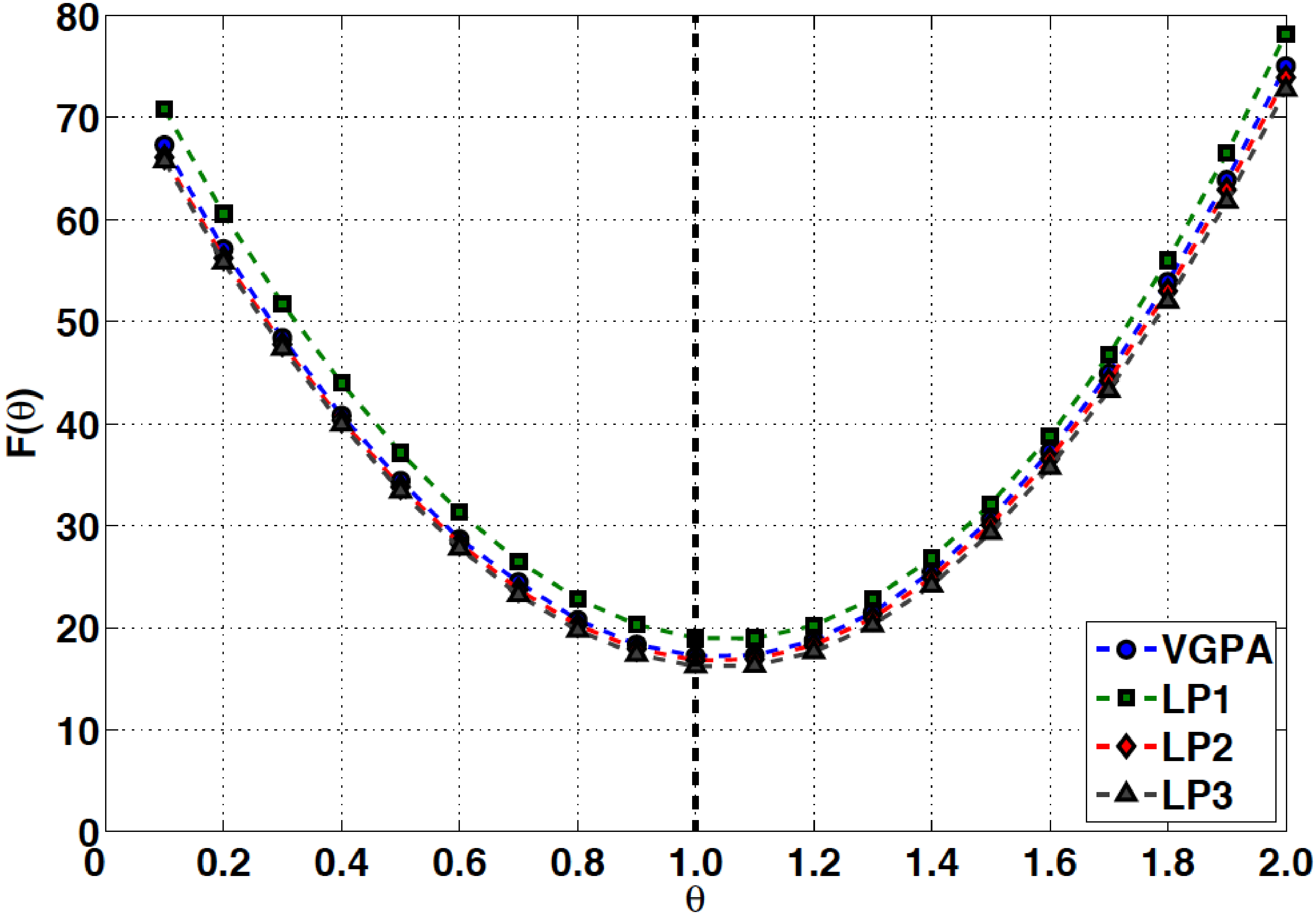


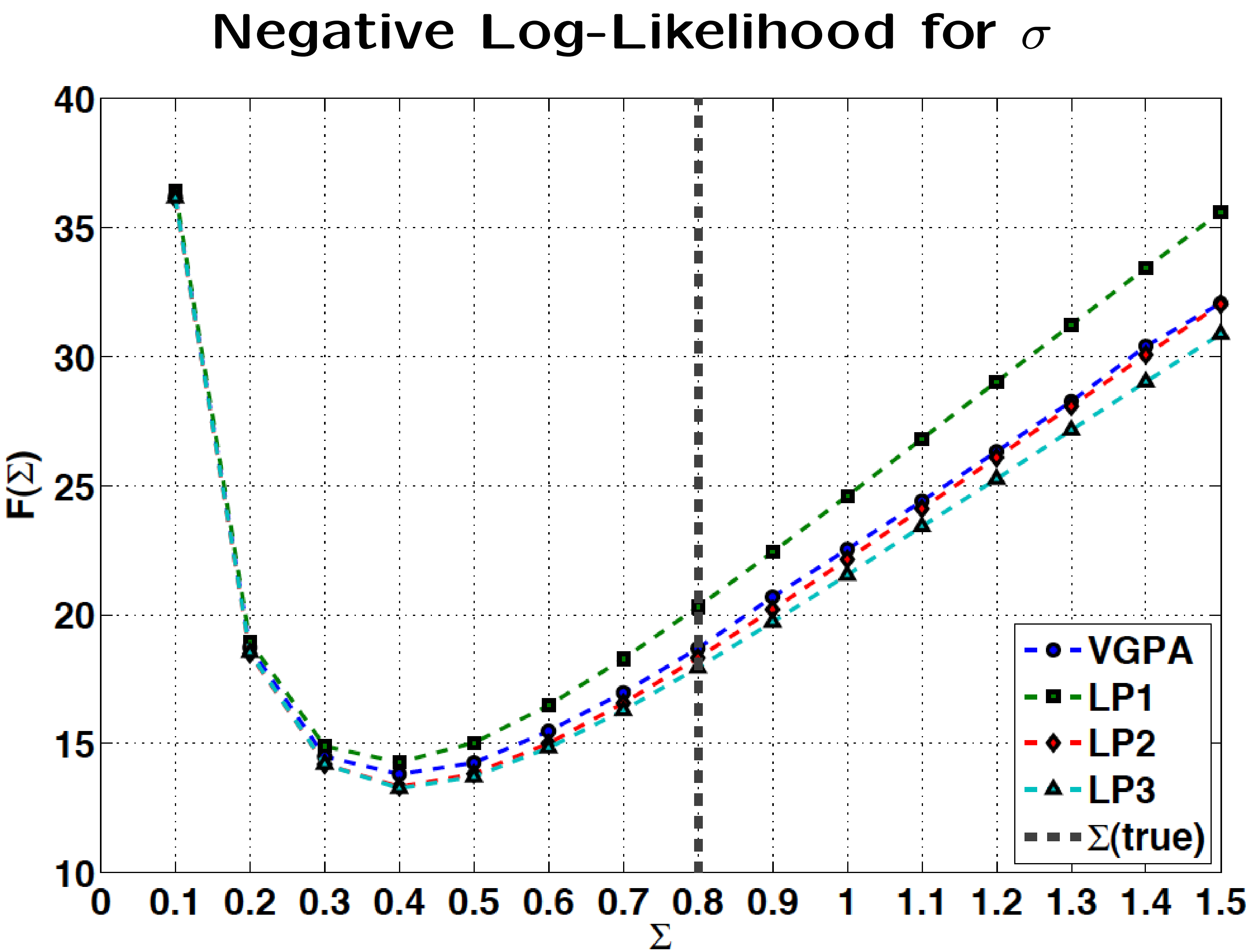
Large observation noise



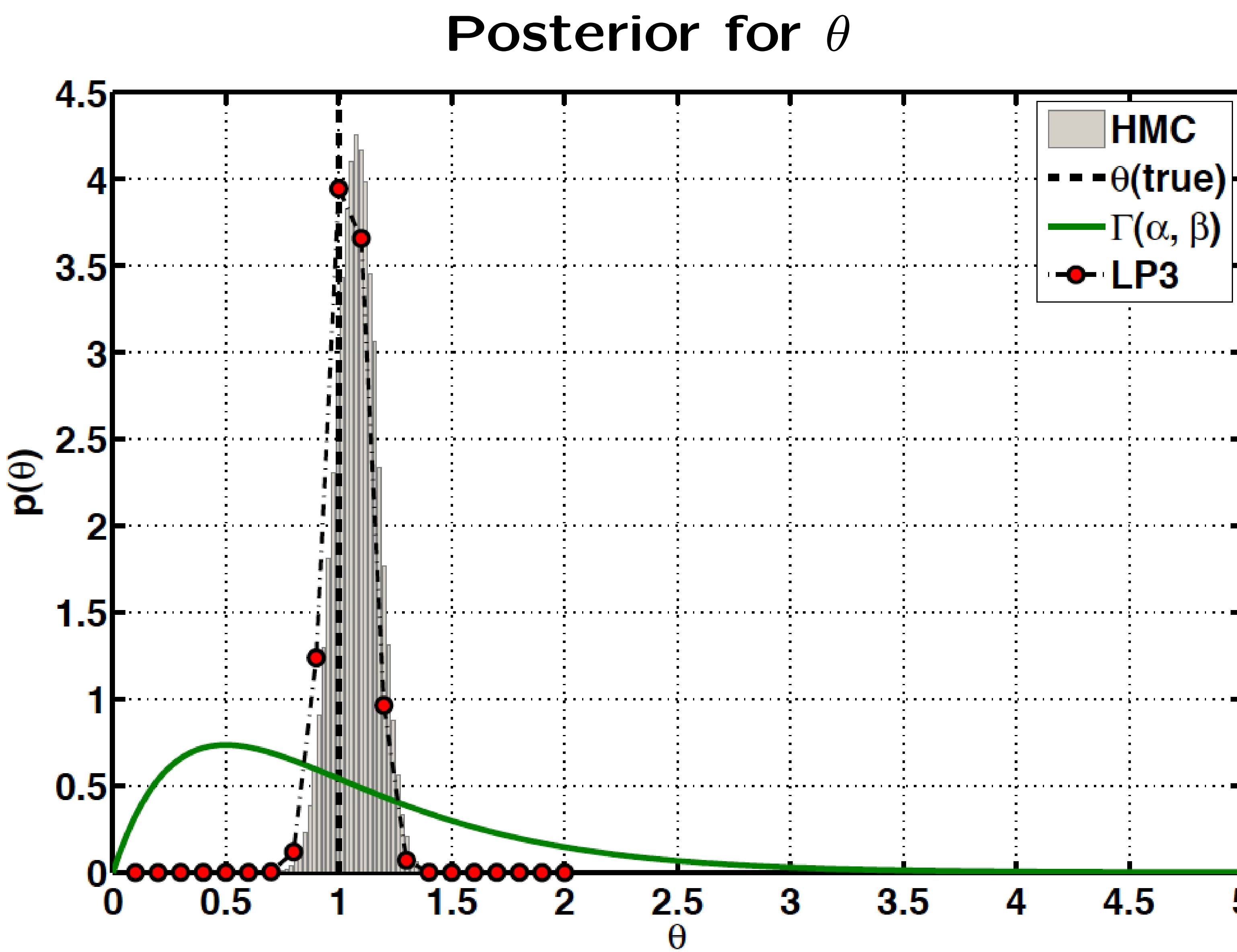
Double well with observation noise $\sigma_o = 0.6$

Negative Log-Likelihood for θ



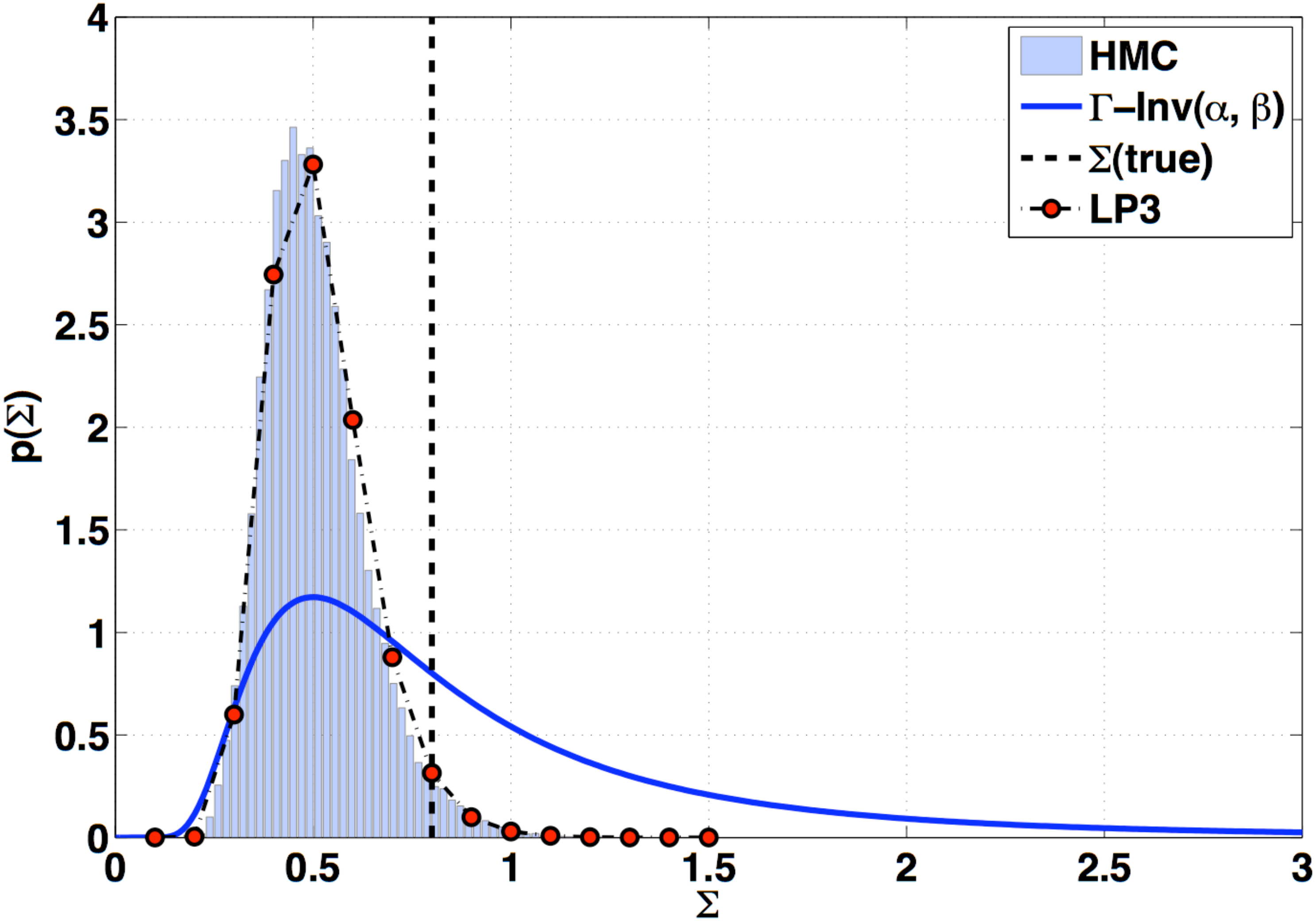


Negative Log-Likelihood for σ
F(Σ)
Σ
40
35
30
25
20
15
10
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
VGPA
LP1
LP2
LP3
Σ(true)



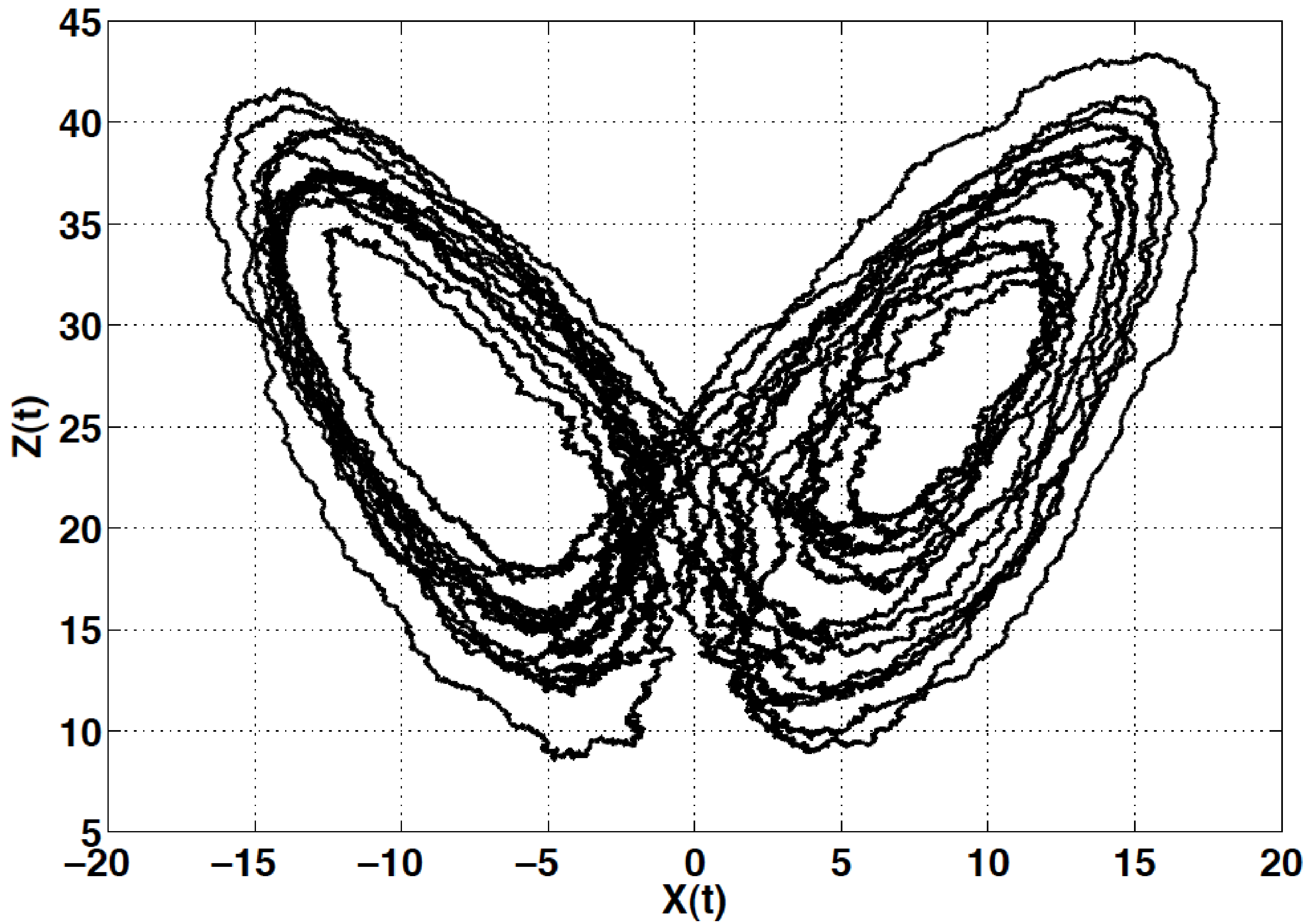
Posterior for θ
HMC
θ(true)
Γ(α, β)
LP3
p(θ)
θ
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Posterior for σ



Lorenz 1963

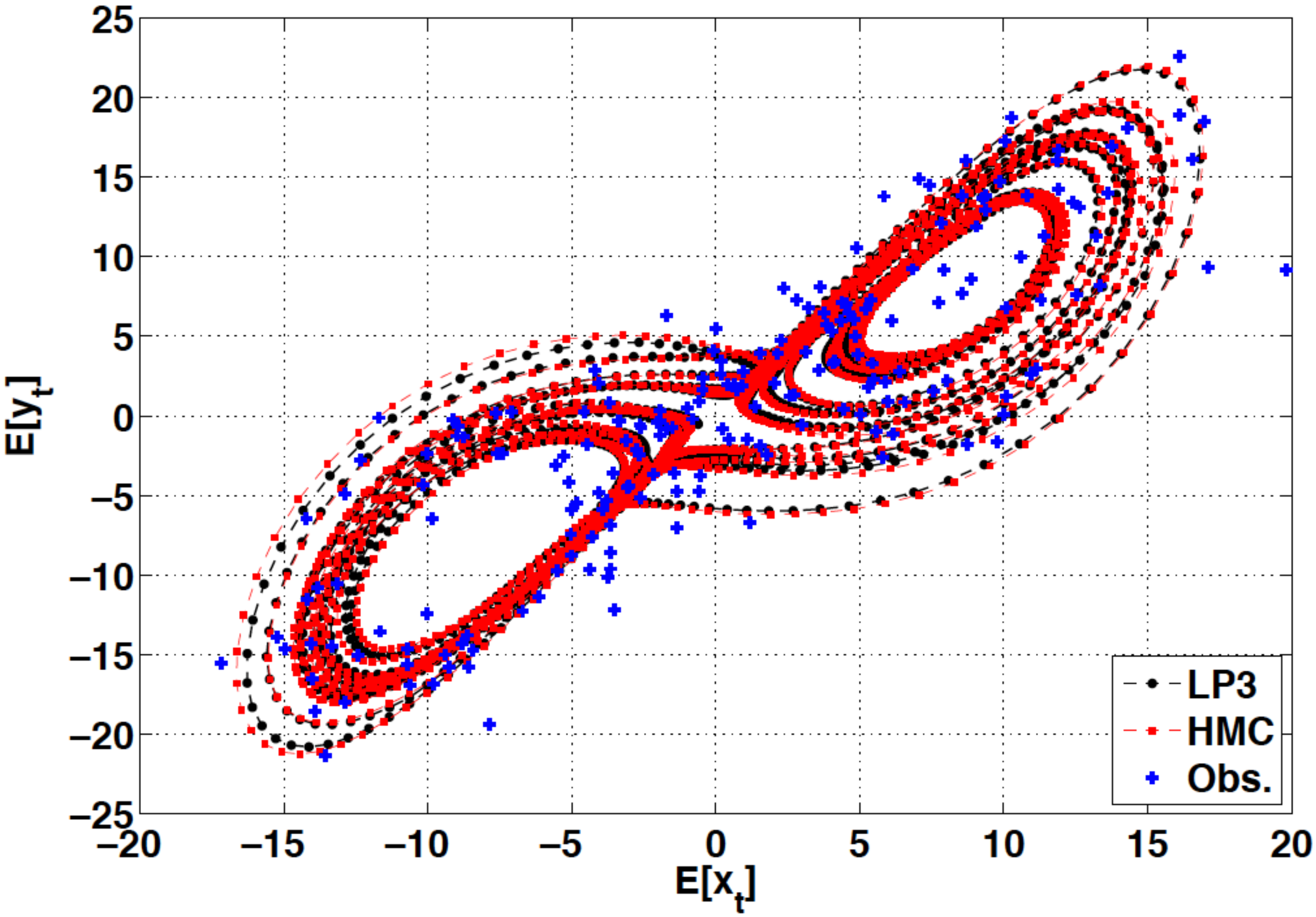
$$
\begin{aligned}
dx_t &= \sigma(y_t - x_t)dt + \sqrt{\Sigma^x}dW^x \\
dy_t &= (\rho x_t - y_t - x_t z_t)dt + \sqrt{\Sigma_y}dW^y \\
dz_t &= (x_t y_t - \beta z_t)dt + \sqrt{\Sigma_z}dW^z
\end{aligned}
$$

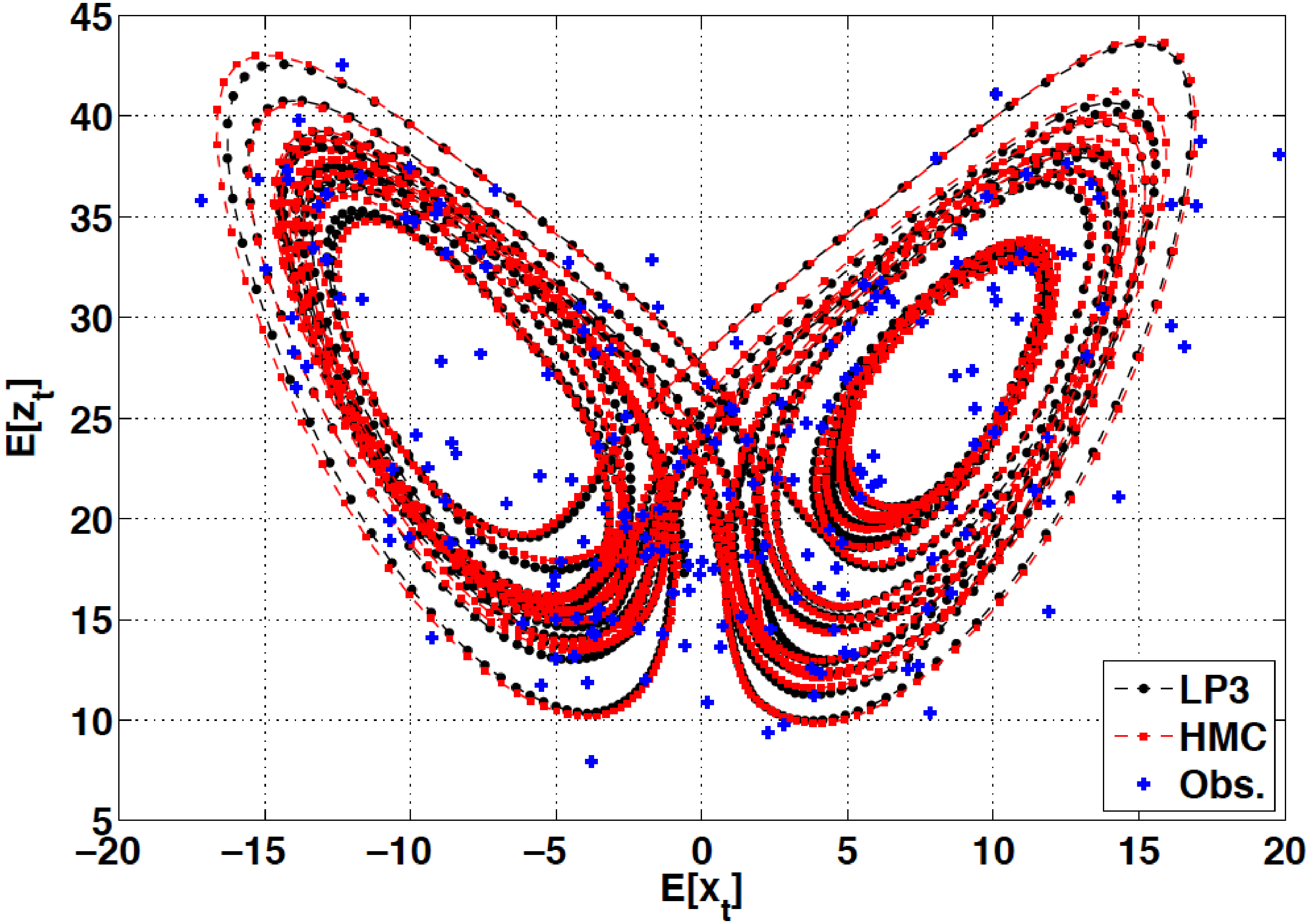


45
40
35
30
25
20
15
10
5
Z(t)
–20
–15
–10
–5
0
5
10
15
20
X(t)

Prediction and comparison with hybrid HMC

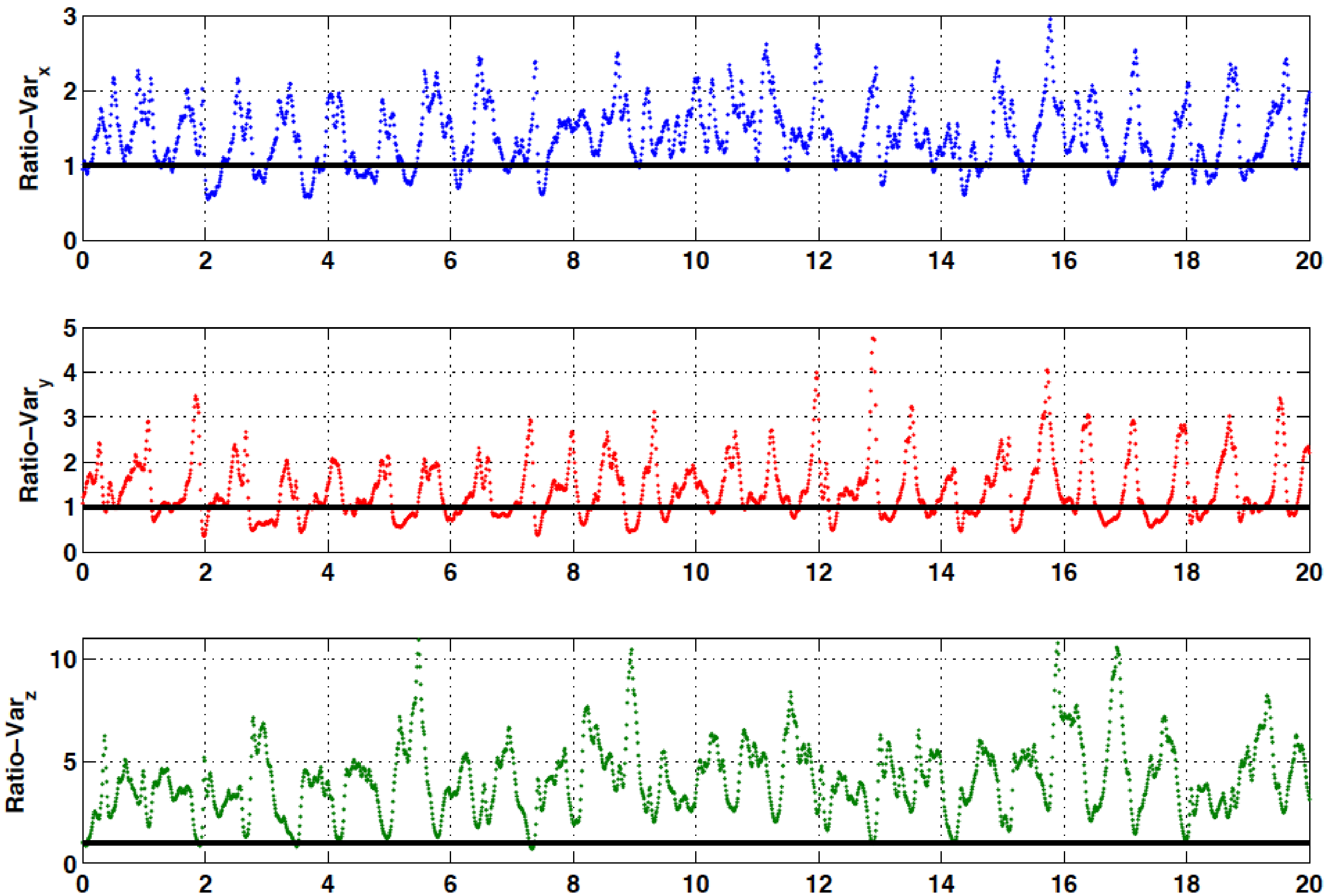
$(\sigma, \rho, \beta) = (10, 28, 2.6667)$, $T = 20$, $\Sigma = 6\mathbf{I}$ and $N = 200$ observations with noise $\Sigma_o = 2\mathbf{I}$.



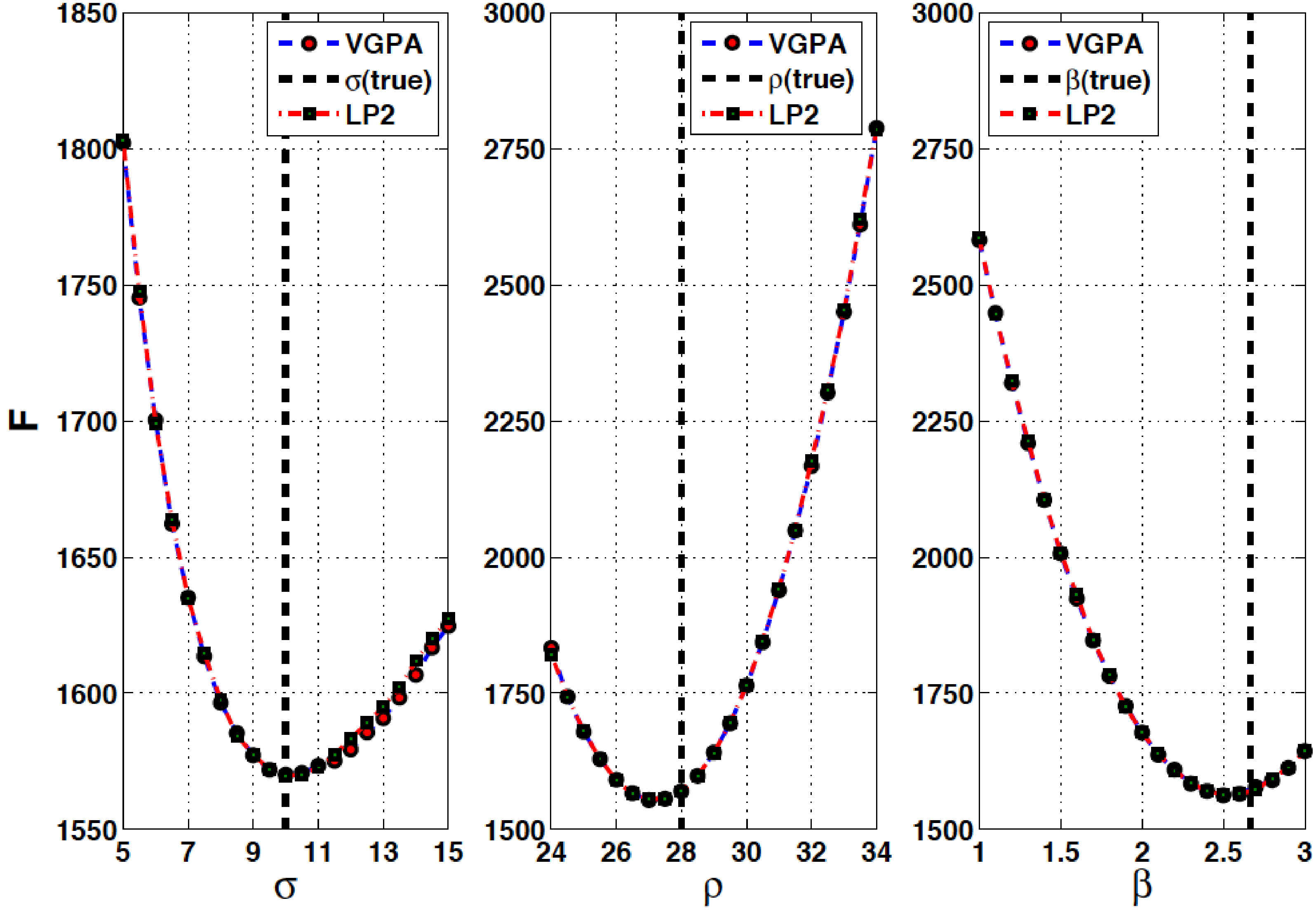


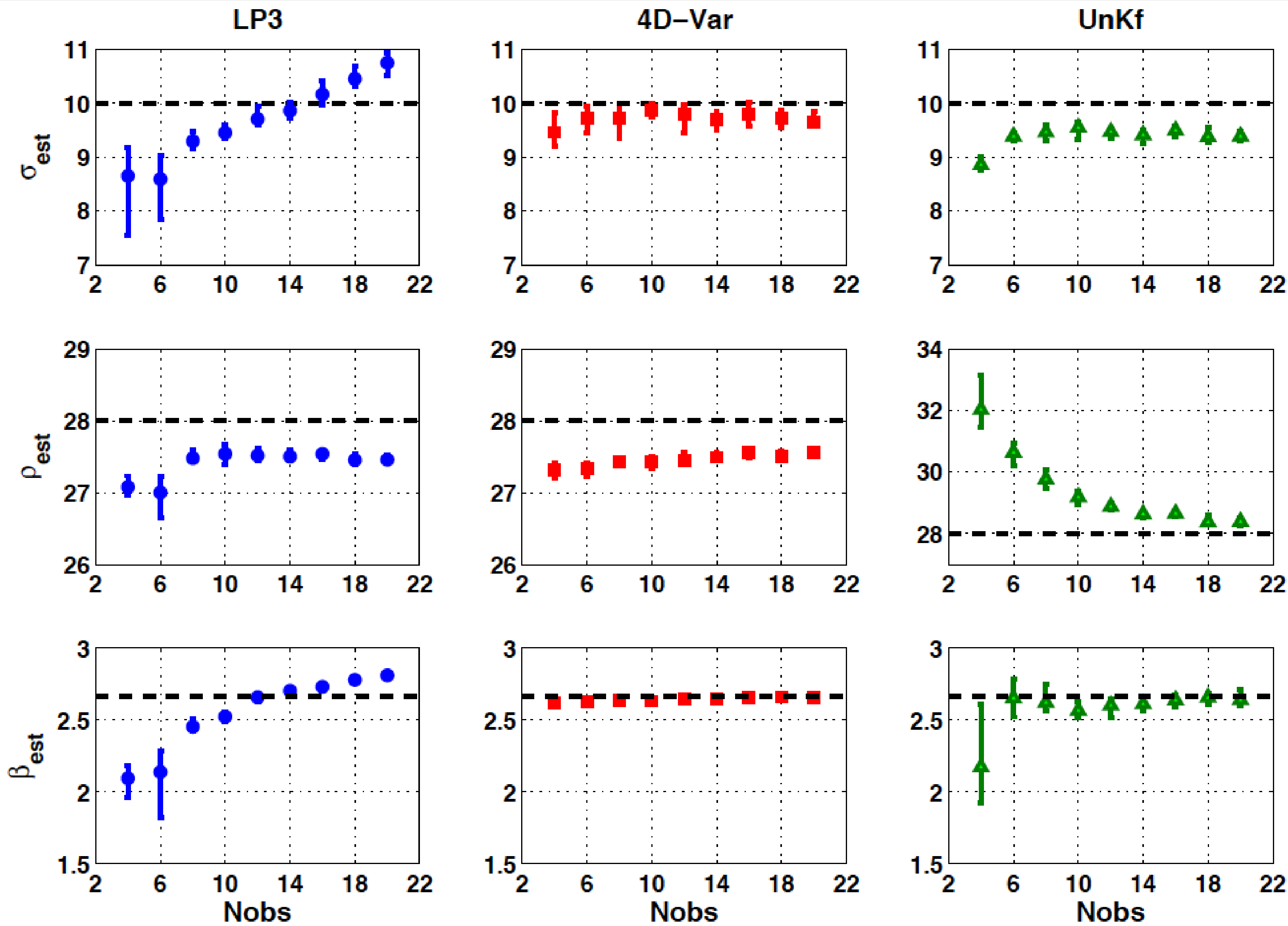
45
40
35
30
25
20
15
10
5
−20
−15
−10
−5
0
5
10
15
20
E[z_t]
E[x_t]
LP3
HMC
Obs.

Predicted marginal variance/ HMC prediction



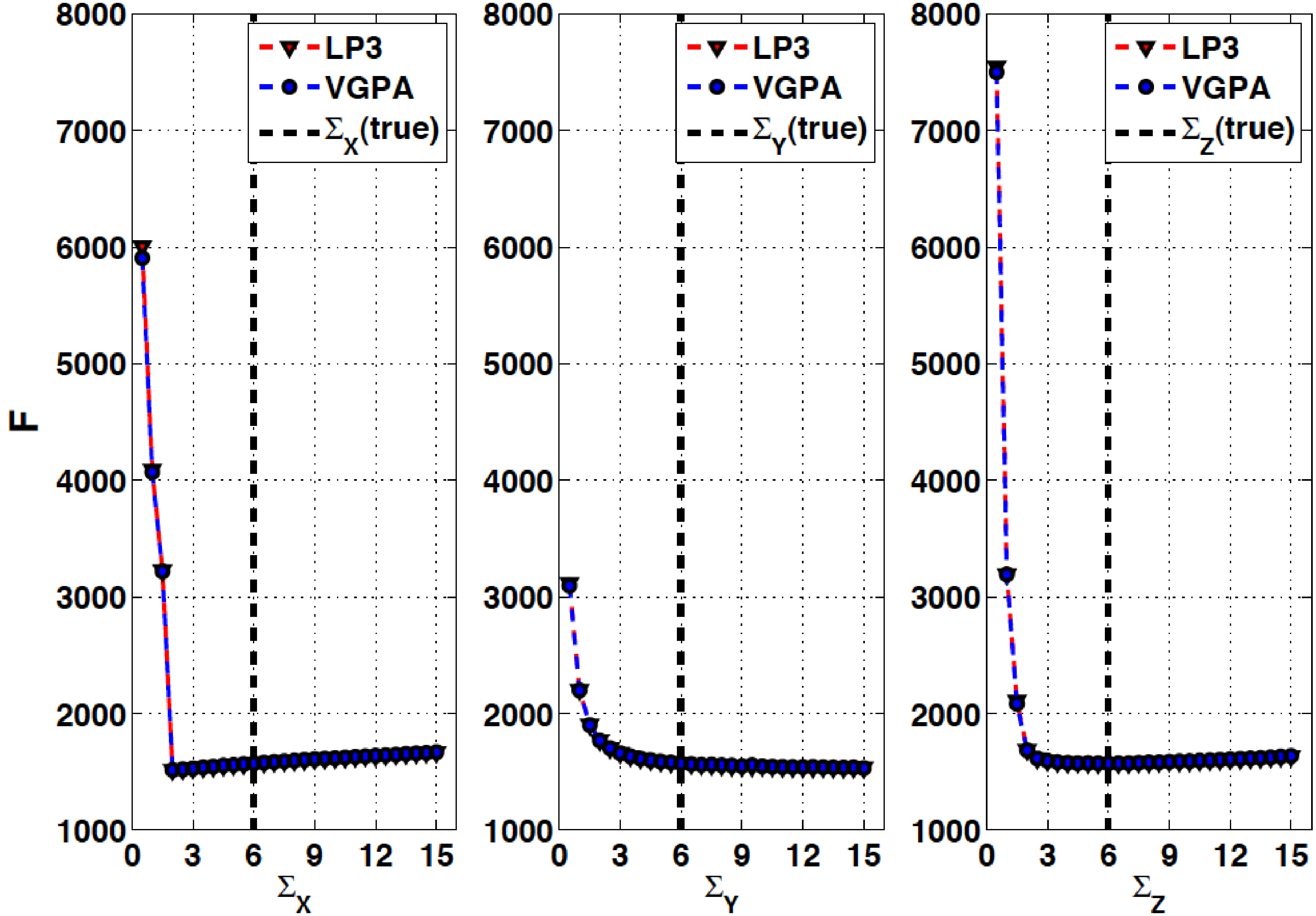
Negative Log-Likelihood for drift parameters

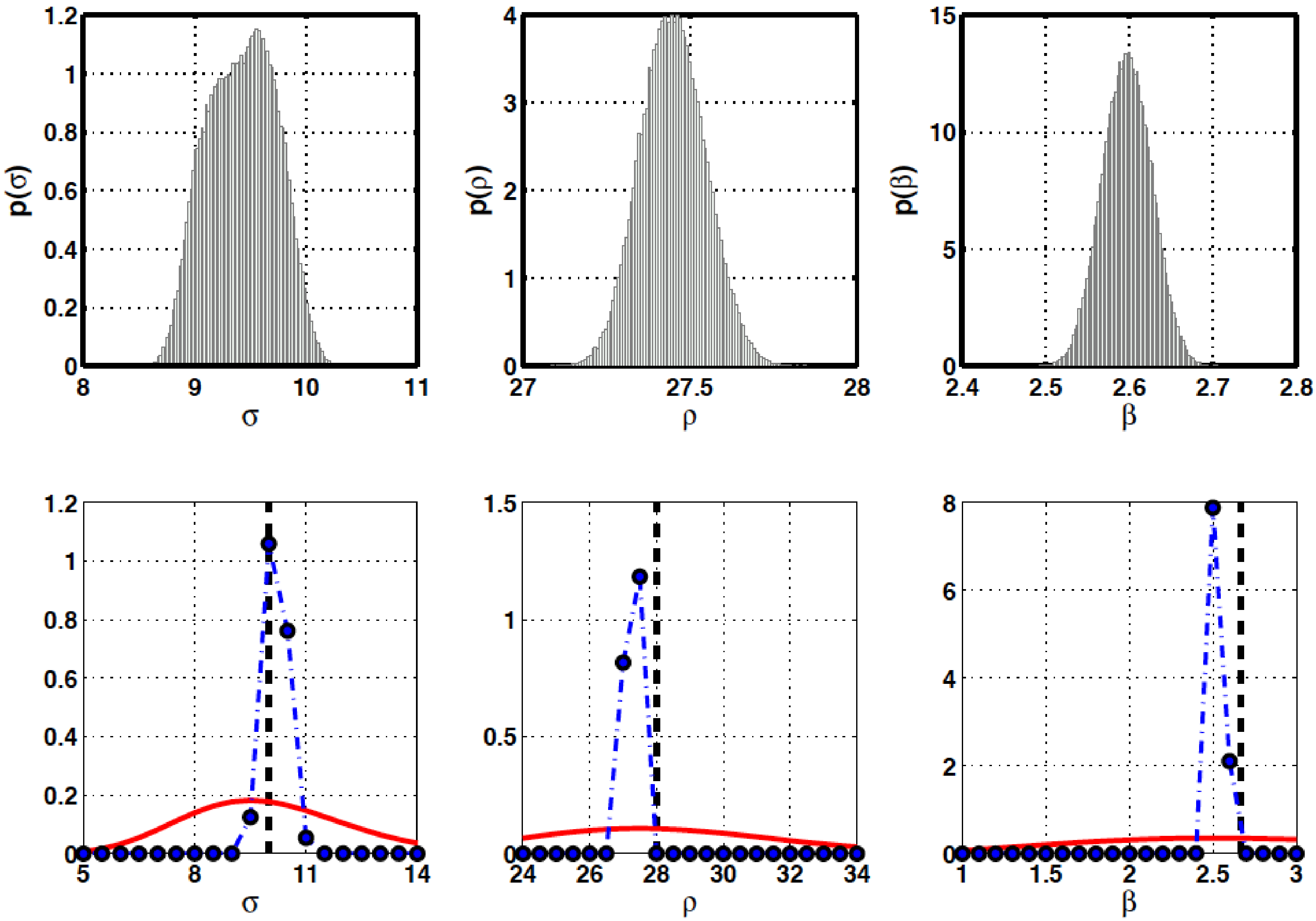




LP3
4D−Var
UnKf
σest
ρest
βest
Nobs
Nobs
Nobs

Negative Log-Likelihood for diffusion parameters




p(σ)
σ
p(ρ)
ρ
p(β)
β
σ
ρ
β

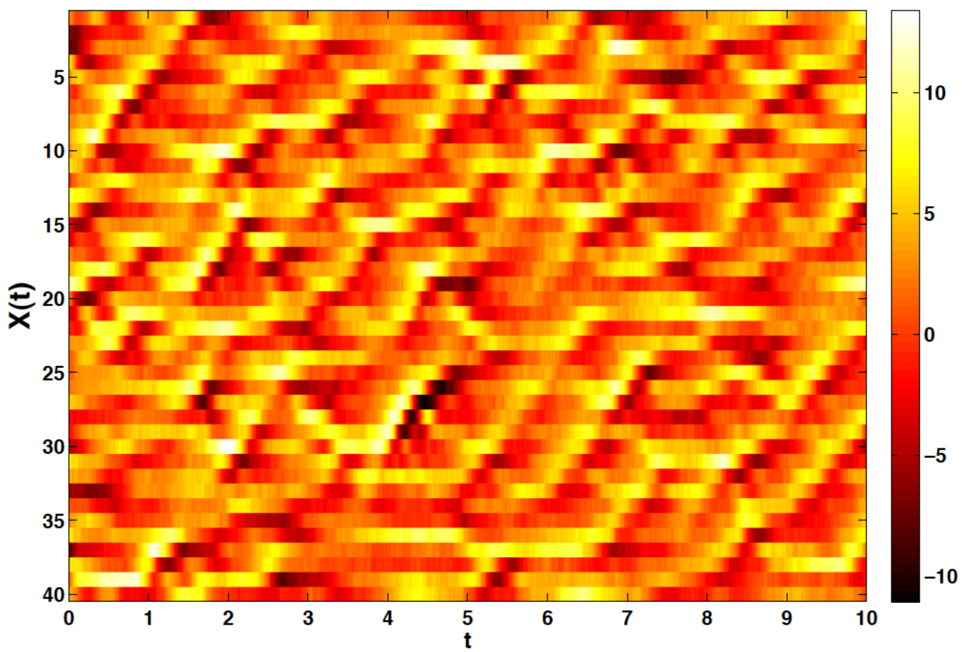
More dimensions

Lorenz 1998 model:

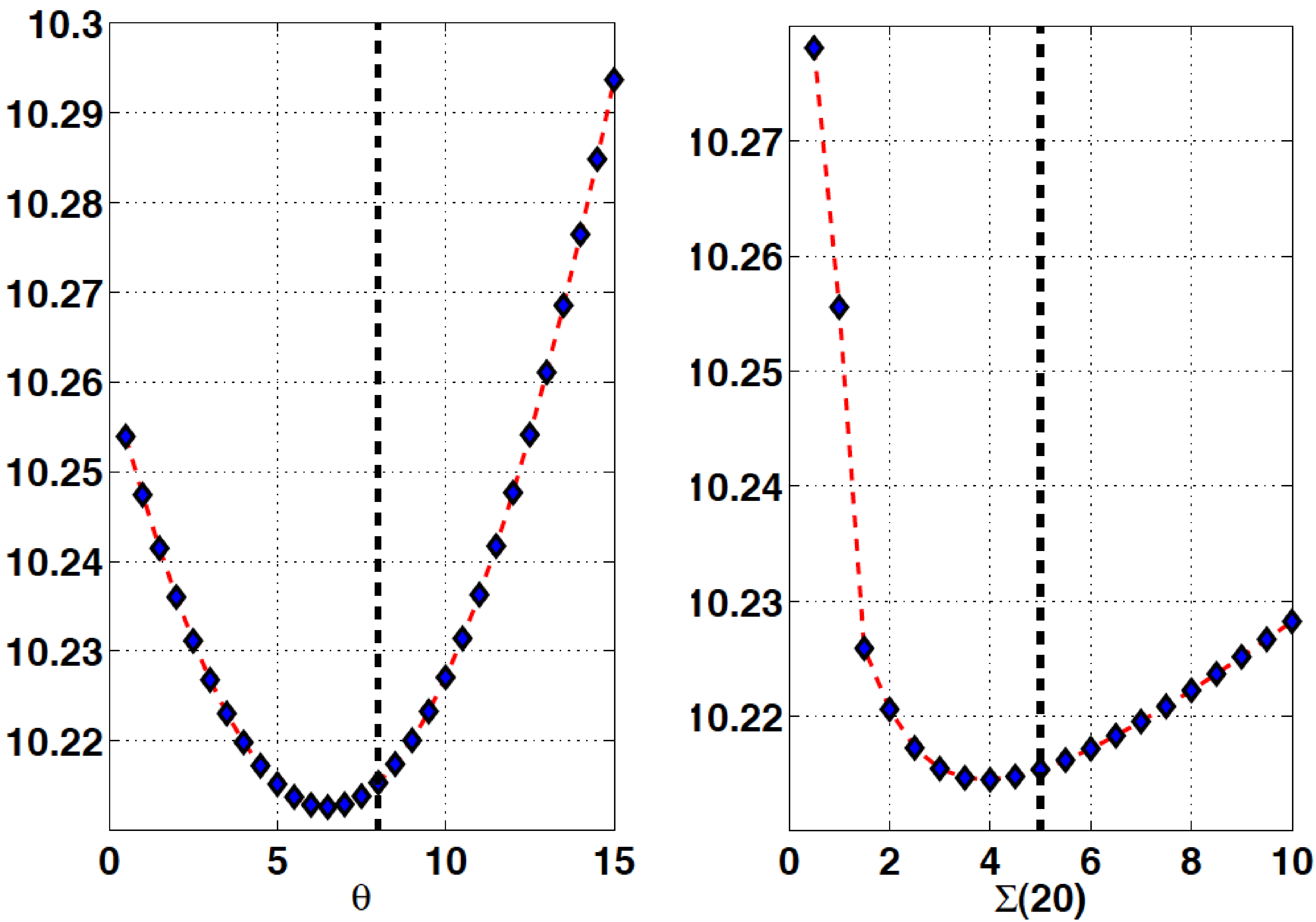
$x = (x^1, \ldots, x^{40})$ with drift

$$f_i(x_t) = \left(x_t^{i+1} - x_t^{i-2}\right) x_t^{i-1} - x_t^i + \theta$$

$\Sigma = 5$ and $N = 90$ observations.



Likelihoods



Other applications of variational approach: Model for transcriptional regulation:

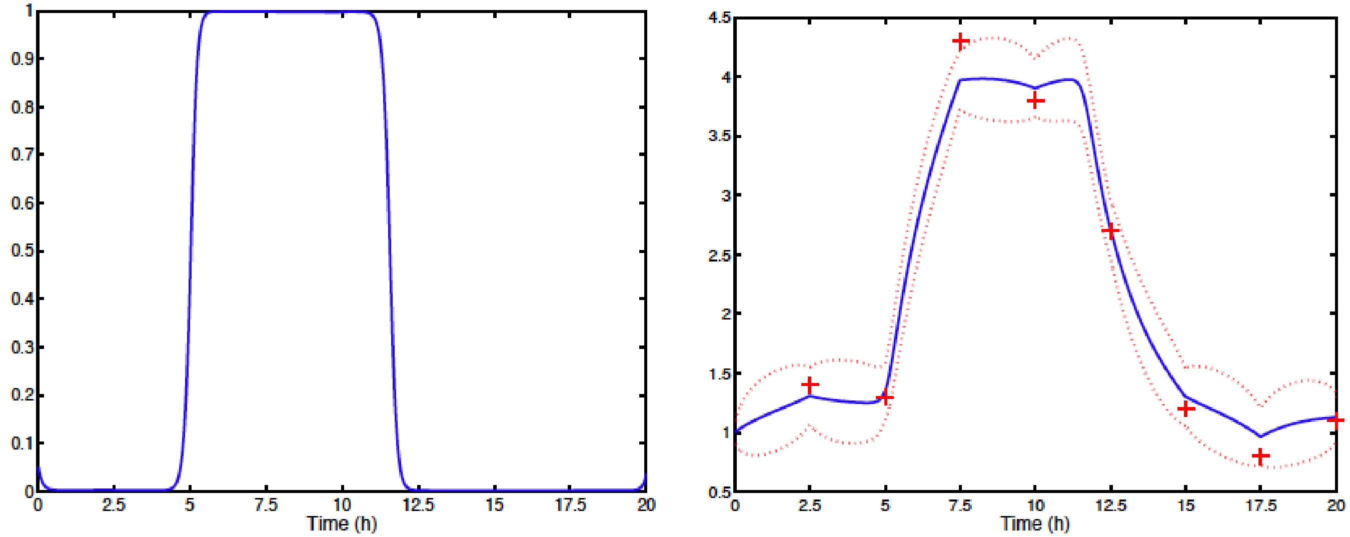
- $x_i(t)$ = mRNA concentration of target gene i. modelled by an Ornstein - Uhlenbeck process

$$dx = (a\mu(t) + c - \lambda x)dt + \sigma dW(t)$$

- $\mu(t)$ = fast switching transcription factor activity (unobserved) modelled by $\mu(t) \sim \mathcal{TP}\left(f_{\pm}\right)$ a random telegraph process.

- Variational approximation

$$q\left(x_{0:T}, \mu_{0:T}\right) = q_x\left(x_{0:T}\right) q_\mu\left(\mu_{0:T}\right).$$



(Opper, Ruttor & Sanguinetti 2010)

Summary

- Posterior probability as the solution of a variational problem involving the relative entropy

- As a byproduct we get a bound on the parameter likelihood

- The relative entropy can be computed analytically for path probabilities of stochastic differential equations

- A Gaussian approximation to path probabilities can be used for smoothing and parameter estimation.

- The Gaussian approximation cannot be applied to state dependent diffusions.

Present & Future work

- Variational path densities as proposal for MCMC

- Perturbative corrections (estimate for error)

- Find good parametric forms for large covariance matrices (projections, low rank representations ?)

- Variational approach to problems with state dependent noise