Journal of Machine Learning Research x (2006) x-x Submitted 06/06; Published xx/xx

MCMC for On-line Filtering: The Particle Path Filter

Jesper Ferkinghoff-Borg BORG@ALF.NBI.DK
Niels Bohr Institute
University of Copenhagen
Blegdamsvej 17
2100 Copenhagen Ø, Denmark

Tue Lehn-Schiøler TLS@IMM.DTU.DK
Ole Winther OWI@IMM.DTU.DK
Intelligent Signal Processing
Informatics and Mathematical Modelling
Technical University of Denmark, B321
2800 Lyngby, Denmark

Editor: xxx

Abstract

We propose a novel Monte Carlo (MC) method for on-line filtering of dynamical state-space models called the particle path filter (PPF). The main new feature of the method is the use of a proposal distribution that exploits two key feature of Markovian systems: The decomposability of the posterior probability of the latent variables and the exponential decaying time correlations of the variables. With this proposal distribution, the whole *path of variables affecting the present* is sampled. This should be contrasted with two extremes: Traditional Markov chain MC (MCMC) for filtering draws samples from the latent variables across the whole time-series and particle filters (PFs) only drawing samples at the current time step. In both cases knowledge about the correlations is ignored leading to slow convergence of the Markov chain. We test and compare the PPF with state-of-the-art PFs for two generic 1d dynamical systems with two attractive fix points emphasizing the importance of using correlation time information. For filtering of systems with very short correlation times PFs outperform PPF in terms of the required particles to reach a given accuracy. For systems with long correlations PPF outperforms PFs with orders of magnitude.

Keywords: State-space models, Markov Chain Monte Carlo, particle filters, path sampling, mean first passage-time

1. Introduction

A dynamical system with an observed state variable $\boldsymbol{z}$ and a hidden state variable $\boldsymbol{x}$ can be formulated as

$$\boldsymbol{x}_k = \boldsymbol{f}(\boldsymbol{x}_{k-1}) + \boldsymbol{v}_{k-1} \tag{1a}$$
$$\boldsymbol{z}_k = \boldsymbol{g}(\boldsymbol{x}_k) + \boldsymbol{w}_k \tag{1b}$$

where $\boldsymbol{v}$ and $\boldsymbol{w}$ are the process noise and the observation noise. The state transition density is fully specified by $\boldsymbol{f}$ and the process noise distribution $p_{\boldsymbol{v}}$ and the observation likelihood is fully specified by $\boldsymbol{g}$ and the observation noise distribution $p_{\boldsymbol{w}}$:

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) = p_v(\boldsymbol{x}_k - \boldsymbol{f}(\boldsymbol{x}_{k-1})) \quad (2)$$

$$p(\boldsymbol{z}_k|\boldsymbol{x}_k) = p_w(\boldsymbol{z}_k - \boldsymbol{g}(\boldsymbol{x}_k)) \ . \quad (3)$$

In filtering, the problem is to find the distribution of the process variable at time k ($\boldsymbol{x}_k$) given all observation up to time k ($\boldsymbol{z}_{1:k}$). This marginal distribution is denoted by $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k}) = \int d\boldsymbol{x}_{1:k-1}\, p(\boldsymbol{x}_{1:k}|\boldsymbol{z}_{1:k})$ where the posterior is

$$p(\boldsymbol{x}_{1:k}|\boldsymbol{z}_{1:k}) = \frac{1}{p(\boldsymbol{z}_{1:k})} \prod_{j=1}^{k} \left[p(\boldsymbol{x}_j|\boldsymbol{x}_{j-1})p(\boldsymbol{z}_j|\boldsymbol{x}_j)\right] \ . \quad (4)$$

It is well-known that Kalman filters (Kalman, 1960) are optimal for linear state-space models with Gaussian noise. However, these models are often found to be too restrictive for realistic data analysis.

The various generalizations and alternatives to the Kalman filters fall into three categories 1) deterministic methods: Extended Kalman filters, sigma-point filters (Julier and Uhlmann, 1997), mean field methods (Ghahramani and Jordan, 1995, Jordan et al., 1999, Heskes and Zoeter, 2002), mixture of Gaussians (Gaussian-sum filters) (Alspach and Sorensen, 1972) and, pseudo-Bayes (Bar-Shalom and Li, 1993), 2) sequential (on-line) Monte Carlo methods (SMC) also known as particle filters (PF) (Gordon et al., 1993) including various extension (Pitt and Shephard, 1999, Kotecha and Djuric, 2001, Lehn-Schiøler et al., 2004, Merwe and Wan, 2003) and 3) off-line Markov Chain Monte Carlo methods discussed in more detail below.

Originally the use of Monte Carlo techniques for State-Space Models was introduced by Carlin et al. (1992) and further investigated by Gordon et al. (1993), Shephard (1994). Tanizaki and Mariano (2000) provides a thorough review of the MCMC sampling in non-Gaussian State-Space Models. The MCMC-method has the advantage of directly providing smoothing estimates for the state space process, i.e. $p(\boldsymbol{x}_{k'}|\boldsymbol{z}_{1:k})$ with $k' < k$, but in its traditional form the method suffers from poor convergence properties given the amount of computation typically available in an on-line filtering application. This problem have been held as an argument in favor for the particle filtering-methods (Pitt and Shephard, 1999).

In the particle filter (PF), the marginal density is represented by a weighted sum of δ-distributions so-called "particles". If the particles representing the probability distribution at a given iteration step is left unaltered at subsequent iterations, the effective sample size (i.e. the number of particles with non-negligible weights) will invariably decrease over time, leading to a successively poorer approximation of the true marginal density. The standard method to reduce the decay of the effective sample size is either to improve the proposal distribution implied in the particle updates or to perform a resampling of the particles whenever the effective number fall below a given threshold. The former approach will always be system specific, whereas the latter approach introduces other deficiencies of the sampling. In particular, it reduces the diversity of particle paths and consequently makes any smoothed estimate less reliable.

The purpose of this paper is two-fold; **firstly**, we explain why PF-methods in general will fail for systems with long correlation times. Processes with long correlation times are characteristic of systems with competing meta-stable phases. Such systems are ubiquitous in almost all scientific areas ranging from reaction processes in chemical kinetics, homogenous nucleation and phase transitions in statistical physics, electrical circuit theory and theory of diffusion in solids (Hänggi and Talkner, 1990, Risken, 1996). Meta-stable phases can be found in non-linear state-space models, when the process function has competing fixed points, i.e. when the posterior is multi-modal. **Secondly**, we wish to promote the *particle path filter (PPF)* as a novel MCMC method for online filtering. The method is based on a straight-forward modification of the proposal distribution of off-line MCMC methods: Variables Δt before the present time k: $\boldsymbol{x}_{k-\Delta t}$ are updated using a suitable proposal distribution, but the probability of choosing that variable decays exponentially $\propto \exp -\Delta t/\tau_q$. The name thus derives from the fact that it is the path of the state vector that is sampled. The free parameter τ_q should be chosen to optimize the sampling properties. The correlation time (or "memory") of the dynamical system τ serves as an upper bound for τ_q. With an appropriate choice of τ_q the method has the added advantage of providing running smoothing estimates.

The paper is organized as follows. We shortly introduce the fundamentals of Markov Chain Monte Carlo in Section 2. Particle filters (PFs) and the particle path filter (PPF) are discussed in Sections 3 and 4. In Sections 5 and 6 we present, analyze and give results for two different bimodal models where the correlation time can be controlled in a simple manner. For the first model, which has been studied extensively in the literature (Arulampalam et al., 2002, Carlin et al., 1992, Gordon et al., 1993, Kitagawa, 1996), the observation model is reflection symmetric but contains an explicit time dependent term in the hidden transition probabilities to drive the process across the two modes. This model has very short correlation time making it ideal for PFs. In the second "Mexican hat" model, the transition probabilities are time independent but the observation model distinguishes between the two modes, providing a weak evidence as to which of the two modes the state belongs to. Outlook and conclusion are given in Section 7.

2. Markov Chain Monte Carlo

In the Markov Chain Monte Carlo (MCMC) method, a state space, $\phi \in \Phi$ is sampled according to a given probability distribution, $\phi \sim p(\phi)$, by generating a Markov chain of states, $\{\phi^{(i)}\}_i$, through a fixed matrix of transition probabilities. In state-space tracking a MCMC 'state' is associated with the entire history of the hidden space $p(\boldsymbol{x}_{1:k}|z_{1:k})$, eq. (4).

Given the chain ϕ a new chain ϕ' can be selected. The transition probabilities, $T(\phi \to \phi')$, are chosen so the condition of *detailed balance* is satisfied

$$p(\phi)T(\phi \to \phi') = p(\phi')T(\phi' \to \phi). \tag{5}$$

Let $p^{(i)}(\phi|\phi^{(0)})$ denote the probability distribution of ϕ for the i'the element of the Markov chain, when it is initialized in state $\phi^{(0)}$. According to the Perron-Frobenius theorem, $p^{(i)}$ will converge to the 'true' distribution $p(\phi)$ independent of the choice of $\phi^{(0)}$;

$$p(\phi) = \lim_{i\to\infty} p^{(i)}(\phi|\phi^{(0)}),$$

provided that T is ergodic and aperiodic (see for example Ferkinghoff-Borg, 2002). In practice, some finite Markov chain of length $\tilde{N}$ is generated where the first $\tilde{n} < \tilde{N}$ states are discarded from the calculation of the relevant state observables, to account for the initial relaxation of the chain.

The transition probabilities are in a computational sense constructed as a product of a proposal probability distribution $q(\phi'|\phi)$, and an acceptance rate $a(\phi'|\phi)$, i.e. $T(\phi \rightarrow \phi') = q(\phi'|\phi)a(\phi'|\phi)$. At the $(i+1)$-step in the MCMC algorithm a trial state ϕ', is drawn according to the distribution $q(\phi'|\phi^{(i)})$ and accepted as the new state $\phi^{(i+1)} = \phi'$ with the probability $a(\phi'|\phi^{(i)})$. Otherwise, one sets $\phi^{(i+1)} = \phi^{(i)}$.

There is a considerable freedom in the choice of a. The standard Metropolis-Hastings algorithm (Hastings, 1970) is to use

$$a(\phi'|\phi) = \min\left\{\frac{p(\phi')q(\phi|\phi')}{p(\phi)q(\phi'|\phi)}, 1\right\}, \tag{6}$$

This prescription automatically satisfies the condition of detailed balance, as verified by direct inspection of eq. (5).

The main deficiency of the MCMC-method in the traditional form outlined above, is its susceptibility to slow relaxation (long correlation times) of the Markov chain. Slow relaxation reduces the effective number of samples and may lead to results which are erroneously sensitive to the particular initialization of the chain.

3. Particle Filters

In the traditional particle filter approach to state-space tracking, information about the system is represented by the marginal density, $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})$ of the current state, $\boldsymbol{x}_k$, only. It is further assumed that the marginals $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})$ and $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$ can be estimated by discrete distributions, a weighted sum of δ-functions

$$p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k}) \approx \sum_{i=1}^{N} w^i \delta(\boldsymbol{x}_k - \boldsymbol{x}_k^i) \ .$$

At each time step these δ-functions (particles) represents the entire knowledge about the system. The idea is to propagate this knowledge through time by moving the particles and updating the weights w^i. The new particles are found by sampling the proposal distribution and the weight of a particle is found by evaluating how likely the particle is given the observation. In the simplest case the proposal density used at time k takes the form

$$q_{PF} = p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) = p_v(\boldsymbol{x}_k - f(\boldsymbol{x}_{k-1})) \ , \tag{7}$$

where the distribution of $\boldsymbol{x}_{k-1}$ is the weighted empirical distribution of the PF sample at $k-1$. The name sequential Monte Carlo filtering arises from the fact that at each time step a Monte Carlo sample is drawn from the distribution moving the filter to next time step.

4. MCMC Techniques and the Particle Path Filter

In applying the MCMC technique to the tracking problem, a state in the Markov chain, ϕ, is identified with the full history of states in the original state-space, $\phi = \boldsymbol{x}_{1:k}$. The Markov

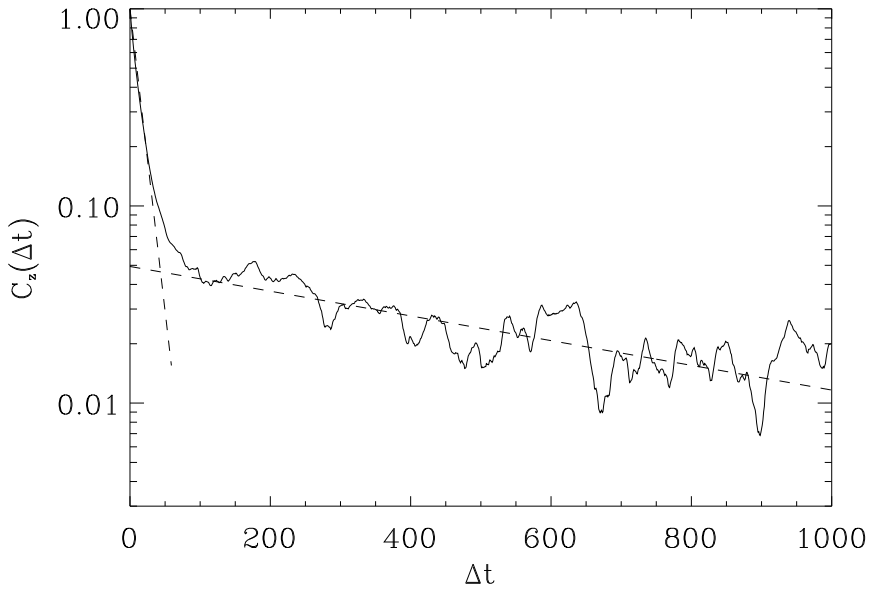

Figure 1: The correlation function $\hat{C}_z(\Delta t)$ as function of Δt (solid). The dashed lines are exponential fits to $C_z(\Delta t)$ for small and large times respectively. The initial fast decay is related to the local fluctuations ($\tau_{loc} \approx 14$) around each stable fixed point, $x = \pm x_f$, whereas the subsequent slow decay is related to the the typical time to make a transition between the fixed points ($\tau \approx 700$).

property of the state transition density and the observation likelihood, eq. (1), implies that the joint posterior density $p(\phi) = p(\boldsymbol{x}_{1:k}|\boldsymbol{z}_{1:k})$ is given by eq. (4). Notice, that the normalization constant $p(\boldsymbol{z}_{1:k})$, cancels out in the Metropolis definition of the acceptance rates, eq. (6).

One obvious advantage of sampling the joint posterior density $p(\boldsymbol{x}_{1:k}|\boldsymbol{z}_{1:k})$ rather than the marginalized posterior density $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})$ is the gain of statistical information. However, when the purpose is on-line filtering, one should design the proposal distribution so that it matches the dynamical properties of the state-space system. An important property of a dynamical system is the (auto)-correlation function of the characteristic observables, for example $\boldsymbol{z}$, $C_z(\Delta t)$. For a finite sequence of T observations, it can be estimated from

$$\hat{C}_z(\Delta t) = c_0^{-1}\left[\frac{1}{T-\Delta t}\sum_{i=1}^{T-\Delta t} z_i z_{i+\Delta t} - \frac{1}{(T-\Delta t)^2}\left(\sum_{i=1}^{T-\Delta t} z_i\right)\left(\sum_{i=\Delta t}^{T} z_i\right)\right], \qquad (8)$$

where $c_0 = \frac{1}{T}\sum_{i=1}^{T} z_i^2 - \left(\frac{1}{T}\sum_{i=1}^{T} z_i\right)^2$ is a normalization constant. Figure 1 gives an example for the correlation function of the bimodal system studied in Section 5. Most stochastic processes encountered in physics and chemistry display correlations decaying

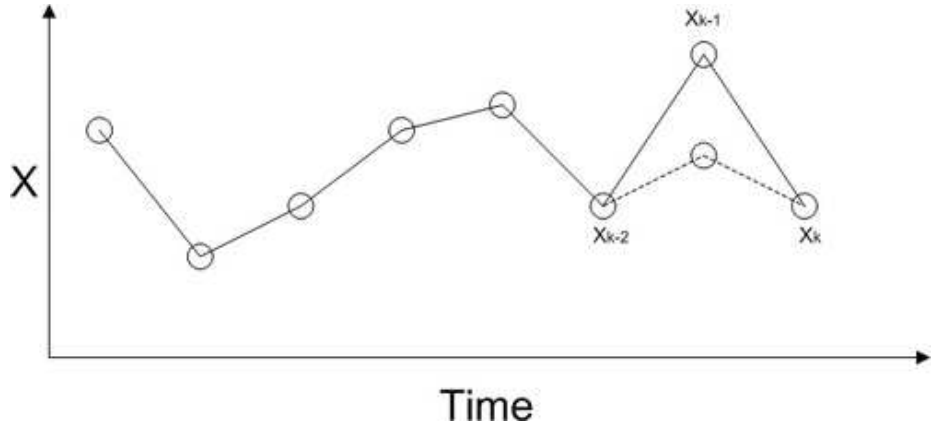

Figure 2: Choosing a new path involves selecting a point according to eq. (10), in this case $\boldsymbol{x}_{k-1}$. Once the point is selected a move is proposed according to eq. (11). The new sequence can be accepted or rejected according to eq. (12)

exponentially in time (van Kampen, 1981), with some characteristic decay constant τ, so $C_z(\Delta t) \sim \exp(-\Delta t/\tau_z)$ for a given component of $\boldsymbol{z}$. Important exceptions are provided by critical phenomena, i.e. processes occurring in the vicinity of second order phase transitions, where correlations typically decay algebraically in time (van Kampen, 1981, Mezard et al., 1987).

The finite correlation time $\tau = \max_z \tau_z$ for 'non-critical' processes implies that the marginal distribution is not dependent upon all the observations but only the most recent: $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k}) \approx p(\boldsymbol{x}_k|\boldsymbol{z}_{k-\tau':k})$, where τ' is a few times τ. This has some important implications for how we should design sampling schemes: Sampling from only the marginal of the current state, as in PF, will lead to slow relaxation because we cannot correct for weak evidence that has build up over time, i.e. correlations beyond one time step are underestimated. On the other hand performing off-line MCMC is wasteful because data beyond the time horizon (determined by the correlation time of $\boldsymbol{z}$) cannot affect the current state.

A simple way to extend the proposal distribution to take time correlation structure into account is to decompose it into a time (T) and space (X) mixture:

$$q_k(\boldsymbol{x}'_{1:k}|\boldsymbol{x}_{1:k}, \boldsymbol{z}_{1:k}) = \sum_{t=1}^{k} q_T(t|k) q_X^{(t)}(\boldsymbol{x}'_{1:k}|\boldsymbol{x}_{1:k}, \boldsymbol{z}_{1:k}) \ . \tag{9}$$

In effect, sampling from this mixture is a two step process: first a time index is selected, $1 \leq t \leq k$, independent of the current state $\boldsymbol{x}_{1:k}$, according to the probability distribution $q_T(t|k)$. Then, a trial path is drawn according to the spatial proposal distribution, $q_X^{(t)}(\boldsymbol{x}'_{1:k}|\boldsymbol{x}_{1:k}, \boldsymbol{z}_{1:k})$. Figure 2 gives a schematic view of the sampling. We will specify the spatial distribution below.

Since the Markov process is expected to generate states with exponentially decaying time-correlations, a natural form for $q_T(t|k)$ is the exponential distribution, $q_T(t|k) \sim$

$\exp((t-k)/\tau_q)$. Here, τ_q equals the average size of the back-propagating step in the path-space sampling following an observation at time k. In order to model the short time scale correlations (see Figure 1) and equilibrate the chain according to the new information available, an extra emphasis should be put on the sampling of the latest state, $\boldsymbol{x}_k$. Therefore the following definition of q_T are proposed

$$q_T(t|k) = \begin{cases} 0 & t > k \\ q_{now}\delta_{t,k} + (1-q_{now})\frac{1}{N_k}\exp((t-k)/\tau_q) & 0 < t \leq k \end{cases} \tag{10}$$

Here, q_{now} is the probability of attempting a change to the latest state $\boldsymbol{x}_k$ only, and N_k is a normalization constant, $N_k = \sum_{t=1}^{k}\exp(t-k)/\tau_q = \frac{1-\exp(-k/\tau_q)}{1-\exp(-1/\tau_q)}$. The algorithm is quite insensitive to the choice of q_{now} as long as it is non-negligible (we use $q_{now} = 0.1$). The same can not be said for τ_q. An upper bound for the necessary τ_q is the correlation time of the process because data beyond a few times τ will have no effect on the present state. In our experience using $\tau_q \geq \gamma\tau$ with $\gamma \sim 1/5$ will give the performance of the traditional MCMC (but at much lower computational cost), see Section 6. When $\tau_q < \gamma\tau$ the performance of PPF approaches that of the PF methods.

The most direct approach to the spatial proposal distribution $q_X^{(t)}(\boldsymbol{x}'_{1:k}|\boldsymbol{x}_{1:k}, \boldsymbol{z}_{1:k})$, is simply to fix all variables except $\boldsymbol{x}_t$ and adopt the proposal distribution applied in a given PF-method to $\boldsymbol{x}_t$:

$$q_X^{(t)}(\boldsymbol{x}'_{1:k}|\boldsymbol{x}_{1:k}, \boldsymbol{z}_{1:k}) = \delta(\boldsymbol{x}'_{1:t-1} - \boldsymbol{x}_{1:t-1})q_{PF}(\boldsymbol{x}'_t|\boldsymbol{x}_{t-1}, \boldsymbol{z}_t)\delta(\boldsymbol{x}'_{t+1:k} - \boldsymbol{x}_{t+1:k}) \ , \tag{11}$$

where q_{PF} is given by eq. (7). With the above choices of q_T and q_X the acceptance probability in the MCMC method, eq. (6), takes the particular simple form for $1 \leq t < k$

$$a(\boldsymbol{x}'_t|\boldsymbol{x}_{1:k}, \boldsymbol{z}_{1:k}) = \min\left\{\frac{p(\boldsymbol{z}_t|\boldsymbol{x}'_t)p(\boldsymbol{x}'_t|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_{t+1}|\boldsymbol{x}'_t)q_{PF}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{z}_t)}{p(\boldsymbol{z}_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)q_{PF}(\boldsymbol{x}'_t|\boldsymbol{x}_{t-1}, \boldsymbol{z}_t)}, 1\right\} \ . \tag{12}$$

For $t = k$, the ratio $\frac{p(\boldsymbol{x}_{t+1}|\boldsymbol{x}'_t)}{p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)}$ should be omitted in the above expression. In PPF we thus exploit the Markov property such that an update is only slightly more expensive than one particle update in PF. On top of that we will use a second type of "global move" specifically designed for dynamical systems with known symmetries, see appendix A. This corresponds to using a proposal that takes the Likelihood term $p(\boldsymbol{z}_t|\boldsymbol{x}_t)$ into account without explicitly including the Likelihood in the proposal distribution (Arulampalam et al., 2002).

In essence the on-line version of MCMC selects a single sample from the sequence, propose a change of that sample and accept it according to eq. (12). Samples near the current time are selected with higher probability because it is expected that new observations will be more likely to influence them. Finally, the type of moves used can be expanded using knowledge about symmetries of the dynamical system.

5. Two Bimodal Models

In order to compare the performance of various particle filtering methods with the PPF-method two different bimodal models are examined. The first model, which we will refer

to as the periodically driven (PD) model, has been analysed before in many publications (Arulampalam et al., 2002, Carlin et al., 1992, Gordon et al., 1993, Kitagawa, 1996):

$$\begin{aligned} x_k &= f_{PD}(x_{k-1}, k) + v_k && \text{(13a)} \\ f_{PD}(x,k) &= f_{x,PD}(x) + f_{k,PD}(k) = \frac{x}{2} + \frac{25x}{1+x^2} + 8\cos(1.2k) \\ z_k &= g_{PD}(x_k) + w_k && \text{(13b)} \\ g_{PD}(x) &= \frac{x^2}{20} \end{aligned}$$

The map $f_{x,PD}(x)$ has two attractive fixed points at $x = \pm 7$ and a repulsive fixed point at $x = 0$, implying that the state-space is divided into two basins, $B_- = \{x|x < 0\}$ and $B_+ = \{x|x > 0\}$. The noise terms, v_k and w_k are zero mean Gaussian random variables with variances $\sigma_v^2 = 10$ and $\sigma_w^2 = 1$, respectively (Arulampalam et al. (2002)). The reflection asymmetry of the posterior distribution is provided by the explicit driving term, $f_{k,PD}$, which forces the system to periodically switch between the two basins. With the above choice of parameters the correlation time is essentially set by the driving frequency, $\tau \simeq \frac{\pi}{1.2}$, i.e. a very short correlation time.

In the second model, which we refer to as the Mexican Hat (MH) model, the reflection asymmetry of the posterior distribution is given by the asymmetry of the observation function:

$$\begin{aligned} x_k &= f_{MH}(x_{k-1}) + v_k && \text{(14a)} \\ f_{MH}(x) &= x - \frac{2h}{x_f}\left(\left(\frac{x}{x_f}\right)^3 - \left(\frac{x}{x_f}\right)\right) \\ z_k &= g_{MH}(x_k) + w_k && \text{(14b)} \\ g_{MH}(x) &= x^2 + \epsilon x \end{aligned}$$

The map $f_{MH}(x)$ has attractive fixed points at $\pm x_f$ and a repulsive fixed point at $x = 0$. Consequently, the process will spend most of the time fluctuating around x_f or $-x_f$. The parameter h determines the probability of crossing from one basin to the other. In our experiments $x_f = 10$, $\epsilon = 1$, the noise contributions v_k and w_k are normal zero mean with variance 1. The value of h is varied between 2.5 and 4.5.

The model is instructive because the stationary probability distribution, $W_0(x)$, and the correlation time, τ_x, of the process can be varied in a controlled manner by changing h. Here, $W_0(x)$ represents the probability density that $x_k = x$ at an arbitrary point, $k \gg \tau_x$, in time. Approximate expressions for W_0 and τ_x can be obtained by mapping eq. (14) to a Fokker-Planck (FP) equation, see appendix B. The FP-equation depends on two functions, $D_1(x)$ and $D_2(x)$, which represent respectively the drift and the diffusion of the process. As described in the appendix, $D_1(x) \hat{=} f(x) - x$ and $D_2 \hat{=} \sigma_v^2/2$, where $\sigma_v^2 = 1$ is the variance of the random variable $v = v_k$. The FP-equation is an accurate description of the process provided that the characteristic length scale, l_D for the variation of D_1 is much larger than the local length scale, $l(x) \simeq \sqrt{\sigma_v^2 + D_1(x)^2}$, associated with the change of x in eq. (14), i.e. $l_D \gg l(x)$ for all x. Here, $l_D = x_f$ and $l(x) \simeq 1$, so this condition is satisfied. Consequently,

h	τ	τ_x	$(\tau_x)_{th}$
2.5	480 ± 30	585 ± 20	541
3.0	700 ± 60	910 ± 35	860
3.5	1300 ± 100	1400 ± 80	1201
4.0	1600 ± 150	1900 ± 100	1707
4.5	2200 ± 200	2550 ± 130	2473

Table 1: Correlation times obtained from state space process, eq. (14) for various values of the barrier heights h. $(\tau_x)_{th}$ is the theoretical value τ is estimated from correlations in the observation sequence $z_{1:T}$ and τ_x is estimated from correlations in the hidden sequence. For all values of h there is good correspondence between theory and numerics.

according to eq. (33) the stationary probability distribution is given by

$$W_0(x) = \frac{1}{\mathcal{N}} \exp\left(-\frac{2U(x)}{\sigma_v^2}\right) , \tag{15}$$

where $\mathcal{N} = \int_{-\infty}^{\infty} e^{-\frac{1}{2}U(x)} dx$ is a normalization constant and

$$U(x) \hat{=} -\int^x D_1(x')dx' = 2h\left[\frac{1}{4}\left(\frac{x}{x_f}\right)^4 - \frac{1}{2}\left(\frac{x}{x_f}\right)^2\right] \tag{16}$$

represents the driving potential of the process. From eqs. (15) and (16) calculations show that for a given process noise, σ_v^2, the probability to be at the unstable fixed point, $x = 0$, relative to the stable ones, $x = \pm x_f$, is solely determined by h, i.e. $\frac{W_0(0)}{W_0(x_f)} = \exp(-h/\sigma_v^2)$. This implies that the observation noise will be low compared to the process noise most of the time, since an uncertainty, δz, in the observation variable z is related to an uncertainty $\delta x = \frac{\delta z}{2x+\epsilon}$ in the state variable x. For $|x| \simeq x_f$ one obtains $\delta x \simeq \frac{\sigma_w}{2x_f} \ll \sigma_v$.

According to eq. (34), the correlation time, τ_x, of the state space process is approximately given by the theoretical expression (th)

$$(\tau_x)_{th} = \frac{2}{\sigma_v^2}\left[\int_0^\infty dx \exp\left(2U(x)/\sigma_v^2\right) \int_x^\infty dy \exp\left(-2U(y)/\sigma_v^2\right)\right] . \tag{17}$$

The correlation time equals half the average time spend in each basin and it sets the maximum relevant value for the time scale, τ_q, in the proposal distribution, eq. (10). In Table 1 the estimated correlation times, τ and τ_x, obtained from an exponential fit to the correlation functions, $C_z(\Delta t)$ and $C_x(\Delta t)$ respectively, is listed for different values of h. The predicted value, $(\tau_x)_{th}$, obtained from a numerical integration of eq. (17) (numerically infinity is $\simeq 3x_f$) is given in the last row. The Table shows that the two correlation times, τ and τ_x are of the same order and reasonably well estimated by the theoretical expression, eq. (17). An example of a typical correlation profile in the present model is shown in Figure 1 for $h = 3.0$.

6. Simulation Results

To quantify the results of the PPF method compared to the traditional PF-methods two error measures are studied. The traditional root-mean-square error is given by

$$RMSE = \sqrt{\frac{1}{T}\sum_{1}^{T}(x_k - \langle x_k \rangle)^2}$$

where T is the total number of steps and $\langle x_k \rangle$ is the posterior average of the state variable at time k estimated through a given algorithm. In addition to this the Basin Error (BE) defined as

$$BE = \frac{1}{2}(1 - \frac{1}{T}\sum_{1}^{T}(\text{sign}(x_k)\text{sign}(\langle x_k \rangle)))$$

is used. It quantifies the fraction of times the algorithm predicts a wrong sign for the state variable x. A value of $BE = 0.5$ means that the performance of the algorithm in resolving the basin-state of the system is the same as by guessing at random.

6.1 Periodically Driven Model

In Figure 3, we show in solid line the filtered RMS-error of the PPF-method as function of the number of trial states, $\tilde{N}$, generated at each time, k. The first half of the trial states are discared in the calculation of the posterior average $\langle x_k \rangle$. The RMSE values are calculated in the same manner as in (Arulampalam et al., 2002), as the average over 100 MC-runs each of length $T = 100$. The parameter, τ_q, in the time proposal function, q_T is set to $\tau_q = 3$. However, due to the small correlation times of the PD-model the PPF-method gives similar results for all $\tau_q < \tau'$ with $\tau' \sim 10$ (data not shown). For the spatial proposal function, q_X, the standard particle filter proposal distribution has been adopted, in conjunction with the global move, $x \to -x$, chosen with probability $q_\pm = 0.15$, see appendix A.

From Figure 3 we observe that the limiting performance is obtained around $\tilde{N} \simeq 2000$. The RMS-error of the standard particle filter algorithm with $N = 50$ particles and resampling at each step is $RMSE = 5.54$ (Arulampalam et al., 2002), which for the PPF-method is obtained around $\tilde{N} \simeq 400$ trial steps. In terms of computational time to reach a certain accuracy of the filtered estimates the PPF-method is ~ 4 or 8 times slower, depending on whether the resampling step in the particle filter algorithm is included in the comparison or not. The PFs thus more or less reach the limiting performance of filtering with only $N = 50$ particles. So in a model like this with short correlation time and no large barrier between fixed points the PF is very effective. Figure 4 (left) illustrates a typical behaviour of the PF-method applied to the PD-model. As shown, the marginal probability distribution is centered around the true state value most of the time.

It should be emphasized that since PPF-method in principle samples from the total joint posterior density, $p(\boldsymbol{x}_{1:k}|\boldsymbol{z}_{1:k})$, rather than the marginalized posterior density alone, $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})$, it directly facilities the calculation of smoothed estimates; something that is difficult to achieve with Particle Filters. The RMS-error of the smoothed estimates is shown with dashed line in Figure 3. The limiting value of the smoothed RMSE corresponds to a reduction of the basin error from $BE = 0.2 \pm 0.004$ (filtered estimates) to $BE = 0.024 \pm 0.002$.

This means that if we can wait only $\tau \simeq 3$ time steps before making predictions we can gain an order of magnitude in precision.

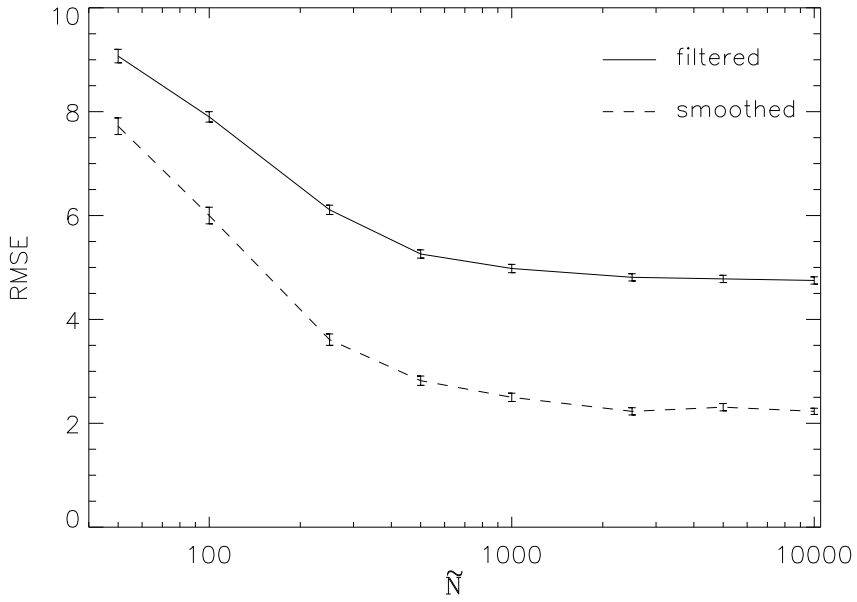

Figure 3: RMS-error as function of the number of trial states, $\tilde{N}$, in the periodically driven model using the PPF-algorithm. The solid line is the error of the filtered estimates and dashed line is the error of the smoothed estimates.

6.2 Mexican Hat Model

The success of the PF-method compared to the PPF-method in the previous example is most likely due to the small correlation time of the state process. In order to test this hypothesis we will in the following focus on the Mexican hat (MH) model, where the correlation times are long and the observation model only provides weak evidence as to which of the two basins the state belongs to.

For each value of h in the MH-model, 10 independent realizations of the state process, eq. (14), is generated starting from $x_0 = 0$. The process in each realization is iterated $T = 15000$ times to ensure a non-vanishing number of transitions between the basins for all h, cf. eq. (17). For each realization, a corresponding observation path $z_{1:T}$ is generated. All algorithms discussed below are tested on this fixed set of state and observation realizations.[1]

In Table 2 the RMSE and BE of the various sequential filtering algorithms for $h = 3.0$ are listed. The ReBEL toolbox[2] by van der Merwe and Wan was used to perform

1. Programs and this benchmark data set are available from the authors.
2. `http://choosh.ece.ogi.edu/rebel/`

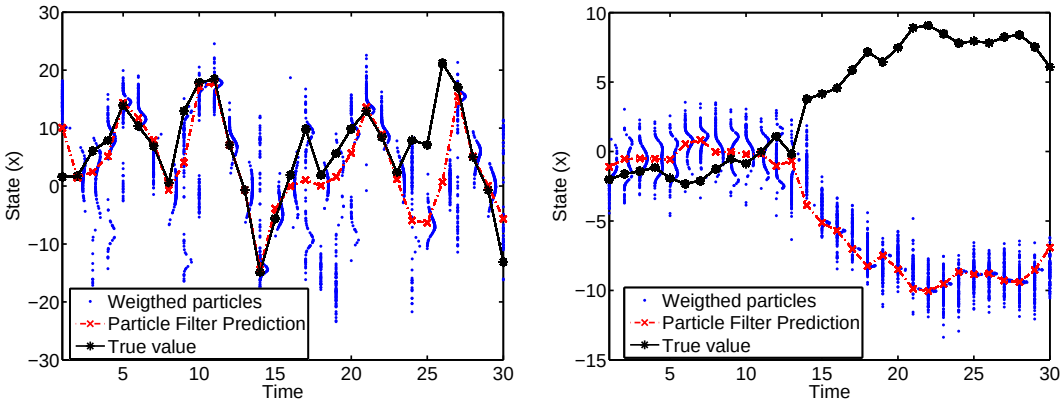

Figure 4: True and estimated time course using a standard particle filter. The dots are particles, and their position relative to the time index illustrate the particle weight. In problems with small correlation length (like the periodical driven system of eq. (13), left plot) the particle filter performs well but as the correlation length increases (as in problems of the Mexican hat type eq. (14), right plot) the particle filter fails.

Method	basin error	STD	RMS error	STD
particle filter (SPF)	0.50	0.05	13.3	0.7
Sigma Point particle filter	0.47	0.04	13.0	0.3
Gaussian Sum particle filter	0.59	0.05	14.7	0.7
SRCDKF †	0.55	0.04	14.9	0.7
particle path filter (PPF)	**0.51**	0.04	**13.3**	0.47
particle filter global move (SPF*)	0.44	0.05	12.3	0.7
particle path filter global move (PPF*)	**0.14**	0.002	**6.41**	0.05

Table 2: Errors obtained with different filtering methods. The ReBEL toolbox was used to perform the experiments. 1000 particles were used in the PF methods. † Three times the output was NaN.

the experiments. The entries give the estimated average error and the uncertainty of the estimate (STD) based on the 10 realizations and using $N = 1000$ particles.

The accuracy of the present PPF-method is also given in Table 2 (third last row), where the standard particle filter (SPF) proposal function has been adopted in the definition of the spatial proposal distribution, eq. (11). The time-scale, τ_q for the proposal distribution, eq. (10), is set to $\tau_q = 250$ which is approximately one-third of the observed correlation times τ for $h = 3$, cf. Table 17. However, without the global move the performance is insensitive to this choice. The number of trial states, $\tilde{N}$, generated at each time, k, is chosen equivalent to the number of particles in the PF-methods, i.e. $\tilde{N} = 1000$. As before, we discard the first half of these in the calculation of the posterior average $\langle x_k \rangle$.

For $N = 1000$ number of particles (or number of trial states) none of the methods performs significantly better in estimating the basin than by guessing at random. This leaves to the conclusion that the accuracy of the various PF-algorithms are more or less identical for the model at hand, and in the following focus will be on just one of these; the SPF-method. Figure 4 (right) shows a typical case where the SPF-method fails to predict the correct basin of the state variable for the MH-model. The total weight of the particles belonging to the correct basin 'accidentially' decays to zero in a few time steps after the system passes the transition region between the two basins. The PF approximation to the marginal probability distribution fails to recreate its bimodal shape at subsequent iteration steps.

As discussed in the previous section, one obvious remedy is to complement the proposal distribution with a move which explicitly carries out the transitions between the two basins. The second last row of Table 2 gives the accuracy of the SPF method when this operation is added to the sampling, chosen with the probability $q_{\pm} = 0.05$. The abbreviation SPF* is used for this modified algorithm. Only a marginal improvement of the algorithm is observed, which nevertheless indicates that the failure of the method is related to the small transition probabilities between the basins. However, as shown in the last row of Table 2 the error reduction is dramatic when the same move is added to the PPF-method, subsequently abbreviated as PPF*.

To appreciate the order of the improvement provided by the PPF*-method we show in Table 3 how the accuracy of the SPF* scales with the number of particles for various choices of h. Two interesting observations can be made. First, a very large number of particles, N_{lim}, are in general needed to reach the limiting accuracy. Secondly, N_{lim} increases with increasing h, corresponding to longer correlation times or smaller transition probabilities, cf. eq. (17). In fact, the limiting accuracy has not yet been reached at $N = 10^6$ for $h > 3$.

In Table 4 we show the performance of the PPF*-method for various choices of h using $\tilde{N} = 1000$ trial states and $\tau_q = 250$. For all h the method gives significantly better results than the SPF*-method with the equivalent number of particles, $N = 1000$. In fact, for $h > 2.5$ the results of the PPF*-method compares favorably with the SPF*-method even in the case where $N = 10^6$ number of particles are used. This corresponds to at least a three-order of magnitude improvement in terms of the computational time required to reach a certain accuracy. For $h = 2.5$ the limiting accuracy obtained by SPF*-method at $N = 10^5$ is reached with the PPF*-method using $\tilde{N} \approx 10^4$ trial states.

In Figure 5 the dependency of the basin error on how far back in time samples ares changed (the choice of τ_q) is shown in solid line. In the limit $\tau_q \to 1$ the accuracy is com-

	100	1000	10000	100000	1000000
2.5	0.53 ± 0.05	0.55 ± 0.03	0.30 ± 0.03	0.17 ± 0.02	0.17 ± 0.02
3.0	0.45 ± 0.05	0.44 ± 0.05	0.30 ± 0.04	0.18 ± 0.02	0.13 ± 0.02
3.5	0.60 ± 0.03	0.58 ± 0.06	0.35 ± 0.04	0.17 ± 0.02	0.12 ± 0.02
4.0	0.54 ± 0.06	0.44 ± 0.05	0.32 ± 0.05	0.14 ± 0.05	0.09 ± 0.02
4.5	0.50 ± 0.07	0.58 ± 0.07	0.30 ± 0.07	0.08 ± 0.03	0.09 ± 0.03

Table 3: Experiments with particle filter using global move. The Basin Error for varying barrier heights (h) and number of particles. A very large number of particles are needed to reach the limiting accuracy. Also note that the algorithm performs worse for small values of h, corresponding to larger transition probabilities between the basins.

	Basin error	STD	RMS error	STD
2.5	0.24	0.005	8.51	0.38
3.0	0.140	0.002	6.41	0.05
3.5	0.090	0.001	5.18	0.04
4.0	0.056	0.002	4.14	0.08
4.5	0.079	0.002	4.95	0.07

Table 4: Experiments with PPF-method using global moves and $\tau_q = 250$ for different barrier heights (h). Compared to the particle filter in Table 3 the errors are very small given that only $\tilde{N} = 1000$ particles were used.

parable to the results of the SPF*-method. As τ_q is increased the error drops significantly until a limiting value is reached around $\tau_q \approx 150 - 200$. This value is lower than one might expect from the observed correlation times, cf. Table 1. However, the correlation time only sets the maximum relevant time scale for the proposal distribution. In the present case, the error saturation around $\tau_q \simeq 150 - 200$ simply reflects the typical number of observations needed to accumulate evidence as to which of the two basins the state belongs to.

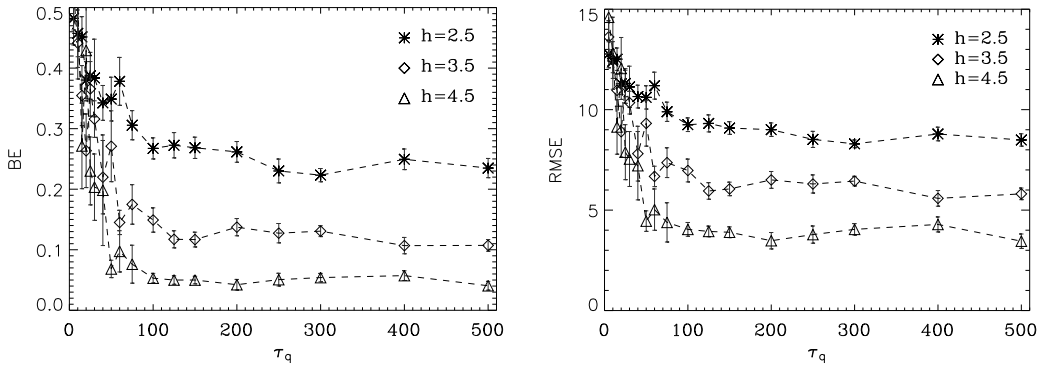

Figure 5: The filtering error as function of the time scale, τ_q, in the PPF proposal distribution. The errors are calculated for different barrier heights (h). Left plot shows the Basin Error (BE), the right plot shows the root-mean-square error (RMSE). In both error measures a sharp decrease of the error is observed as τ_q is increased. The error saturates around $\tau_q \simeq 150 - 200$.

Again we can get smoothing estimates. In Figure 6 we show the BE- and the RMSE-error of the smoothed estimates after $k = T = 15000$ as function of τ_q for different choices of h. As expected, the error of the smoothed estimates is considerably reduced compared to the error of the filtered estimates for all $\tau_q \gg 1$.

7. Conclusion and Outlook

We have demonstrated that it is possible to formulate a Markov chain Monte Carlo (MCMC) algorithm, the particle path filter (PPF), that explicitly uses the time-correlation structure of the dynamical system we are filtering. The main problem with MCMC for online filtering is the slow relaxation of the Markov chain and thus a prohibitive amount of computation needed in order to give a proper sampling. The key point made in this article is that we can avoid this by only considering the states in the past that are actually relevant for the present state. A correlation analysis gives the information we need to define the "temporal" component of the proposal distribution, i.e. which state to change using the "spatial" proposal. After this temporal selection process we can use the same spatial proposal distribution as in the particle filter (PF) method. The temporal proposal distribution we used was a simple mixture of choosing the present and an exponential for past states. One can imagine more refined distributions such as sums of exponentials that reflects the different

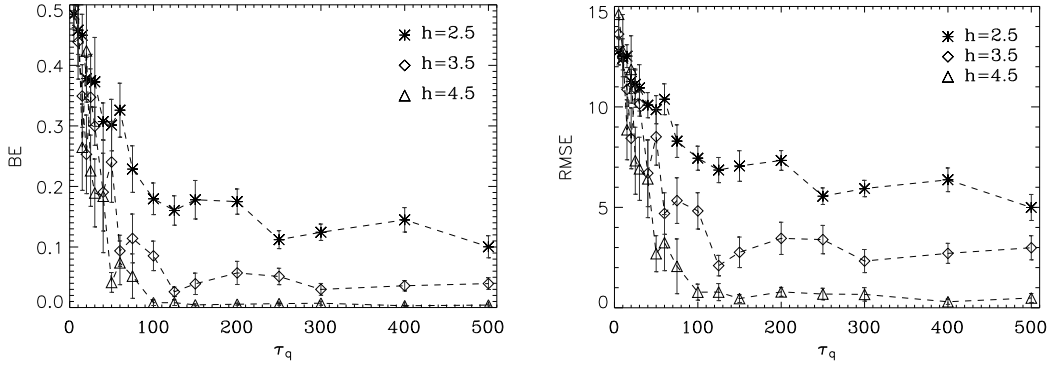

Figure 6: The smoothing error as function of the time scale, τ_q, in the PPF proposal distribution. The errors (Basin error left and RMS-error right) are calculated for different barrier heights (h). As expected, for all $\tau_q \gg 1$ the errors are significantly reduced compared to the filtered estimates (Figure 5).

time-scales of the dynamical system, for example short time adaptation within a basin and inter-basin dynamics.

It has been shown that there are no hinderance in using MCMC in online applications and the experiments indicate that with the same computational complexity MCMC methods can produce much superior results. The reason for the success of the MCMC methods is the ability to accumulate evidence over several time steps, thus utilizing the small differences in posterior probabilities. Whether a particle filter approach is sufficient depends crucially on the temporal correlations present in the dynamical system. Performing a correlation analysis will thus provide valuable information: If the correlation time is short, say $1-10$ time steps – like the periodic driven model considered in this paper – PFs outperform the PPF in terms of computation needed in order to achieve a given error level. On the other hand, if the correlation time is long, 100+ time steps, the PFs will typically fail as illustrated by the Mexican hat model.

Besides handling long correlation times there are more added benefits to using a MCMC method: 1) We get smoothing (back in time) estimates for free since we are in principle sampling the whole chain and 2) we can use standard ways of improving the performance of MCMC methods such as parallel tempering and bridging (Iba, 2001) which can also give us marginal likelihood estimates.

Acknowledgments

J F-B would like to acknowledge Carlsberg Fonden for financial support. The work was funded (in part) by the Danish Technical Research Council project No. 26-04-0092 Intelligent Sound (www.intelligentsound.org) and the PASCAL Network of Excellence (www.pascal-network.org).

Appendix A. Global Moves and Extended Ensembles

Knowledge about the global symmetries of the system at hand can be incorporated into the sampling procedure. For example, for the periodically driven (PD) model discussed in Section 5, the state process, $f_{PD}(x)$, and the observation model, $g_{PD}(x)$, are reflection symmetric. Thus, for a given time-step Δt back from the present time, t, the spatial proposal function is naturally augmented with a proposal to change the part, $\boldsymbol{x}_{(t-\Delta t):(t)}$, of the sequence according to

$$\boldsymbol{x}_{(t-\Delta t):t} \to \left(-x_{(t-\Delta t)}; -x_{(t-\Delta t+1)}; \ldots; -x_t\right). \tag{18}$$

The Mexican hat model displays a reflection symmetry, $f_{MH}(-x) = -f_{MH}(x)$, for the state process, f_{MH}, and a shifted reflection symmetry $g_{MH}(-x-\epsilon) = g_{MH}(x)$ for the observation model, g_{MH}. In principle both transformations could be incorporated in the sampling procedure. As discussed in Section 5, the observation noise is low compared to the process noise. Therefore, for the MH-model we should expect the latter transformation, $x \to -x-\epsilon$, to be more effective in reducing the relaxation time of the sampling procedure. Consequently, we use the global move

$$\boldsymbol{x}_{(t-\Delta t):t} \to \left(-x_{(t-\Delta t)} - \epsilon; -x_{(t-\Delta t+1)} - \epsilon; \ldots; -x_t - \epsilon\right). \tag{19}$$

In both models the global move is chosen with some low probability $q_{\pm}$ and is to be accepted with an acceptance rate similar to eq. (12). Note, that for the particle filter methods this move can only be applied to the latest state, x_t. Exploiting symmetries is a computational cheap version of the Likelihood particle filer (Arulampalam et al., 2002) which uses $p(\boldsymbol{x}_t|\boldsymbol{z}_t, \boldsymbol{x}_{t-1}) \propto p(\boldsymbol{z}_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ as proposal.

The global move that is augmented here to the local sampling procedure reflects a symmetry property of the system at hand which is obviously not generic. It is possible, though, to circumvent the need of ingenious and system-specific move-schemes altogether and make use of "extended" type of ensembles instead (Iba, 2001, Ferkinghoff-Borg, 2002). This approach, which has proven very successful in various problems in statistical physics, refers to a family of algorithms where the probability distribution of interest, p, is replaced with an "artificial" distribution, $\tilde{p}$, constructed either as an extension or by composition of the original ensemble. The extended distribution, $\tilde{p}$, acts as a 'bridge' from the ensemble where the M arkov chain suffers from slow relaxation to an ensemble where the sampling is free from such problems. An instructive example of an extended ensemble is given by the parallel tempering (PT) algorithm (Iba, 2001), see the online Appendix.

Appendix B. Mapping from discrete to continous processes

In this appendix we describe how to obtain the stationary probability distribution, $W_0(x)$, eq. (15) and the relevant time-scales for a state space process on the form

$$x_{t+1} = f(x_t) + v(x_t), \qquad v(x) \sim N(\mu(x), \sigma^2(x)), \tag{20}$$

where $v(x)$ is a Gaussian distributed stochastic variable with mean $\mu(x)$ and variance $\sigma^2(x)$. Since the process is easier to analyse in the continous time-limit, we extend eq. (20) to any

finite time-step Δt in the following way

$$x_{t+\Delta t} = x_t + D_1(x_t)\Delta t + \sqrt{2D_2(x_t)\Delta t}\gamma_t. \tag{21}$$

Here, γ_t is a Gaussian random variable with zero mean and δ-correlated in the chosen time discretization, $\langle \gamma_t \rangle = 0$ and $\langle \gamma_t \gamma_{t'} \rangle = \delta_{tt'}$. The functions D_1 and D_2, which characterize respectively the drift and the diffusion of the process, are here defined so as to match eq. (20) when $\Delta t = 1$;

$$D_1(x) = f(x) - x, \quad D_2(x) = \sigma^2(x)/2. \tag{22}$$

The evolution of the probability distribution $W(x,t)$ for the stochastic variable x in eq. (21) is given by

$$W(x, t + \Delta t) = \int P_{\Delta t}(x|x')W(x', t)dx', \tag{23}$$

where the transition probabilities are

$$P_{\Delta t}(x|x') = \frac{1}{\sqrt{4\pi D_2(x')\Delta t}} \exp\left(\frac{-(x - x' - D_1(x')\Delta t)^2}{4D_2(x')\Delta t}\right). \tag{24}$$

The advantage of studying the state process, eq. (20), in the continuous time limit is that the integral equation, eq. (23), can be approximated by a differential equation in both t and x. This is accomplished in two steps. First, the integral operator on the right hand side of eq. (23) can be expressed as a differential operator in x by rewriting the transition probability function in terms of its moments, $M_n(x', \Delta t)$,

$$M_n(x'; \Delta t) \hat{=} \int (x - x')^n P_{\Delta t}(x|x')dx. \tag{25}$$

Following Risken (1996) the inversion of this expression is most easily done by noting that the characteristic function

$$C(u, x'; \Delta t) \hat{=} \int e^{iu(x-x')} P_{\Delta t}(x|x')dx \tag{26}$$

is the generating function for the moments $M_n(x'; \Delta t) = (-i)^n \left.\frac{\partial^n C(u,x';\Delta t)}{\partial^n u}\right|_{u=0}$. Consequently, a Taylor expansion of eq. (26) around $u = 0$ gives

$$C(u, x'; \Delta t) = 1 + \sum_{n=1}^{\infty} \frac{(iu)^n}{n!} M_n(x'; \Delta t) \ .$$

Since the transition probabilities is the inverse Fourier transform of the characteristic function one obtains

$$P_{\Delta t}(x|x') = \frac{1}{2\pi} \int e^{-iu(x-x')} \left[1 + \sum_{n=1}^{\infty} \frac{(iu)^n}{n!} M_n(x'; \Delta t)\right] du$$

The integral over u can be rewritten by applying

$$\frac{1}{2\pi} \int (iu)^n e^{-iu(x-x')} du = (-1)^n \frac{\partial^n}{\partial x^n} \frac{1}{2\pi} \int e^{-iu(x-x')} du = (-1)^n \frac{\partial^n}{\partial x^n} \delta(x - x')$$

and since $f(x')\delta(x-x') = f(x)\delta(x-x')$ one finally obtains

$$P_{\Delta t}(x|x') = \left[1 + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n M_n(x;\Delta t)\right] \delta(x-x') \ . \tag{27}$$

Inserting this equation into eq. (23) leads to

$$W(x,t+\Delta t) = W(x,t) + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial x^n}\big(M_n(x;\Delta t)W(x,t)\big) \ . \tag{28}$$

The mapping of eq. (28) to a differential equation in t is facilitated by Taylor expanding eq. (28) to first order in Δt,

$$\frac{\partial W(x,t)}{\partial t} = \sum_{n=1}^{\infty} (-1)^n \frac{\partial^n}{\partial x^n}\big(D_n(x)W(x,t)\big) \ . \tag{29}$$

Here, $D_n(x) \hat{=} \frac{1}{n!}\lim_{\Delta t\to 0}\frac{M_n(x;\Delta t)}{\Delta t}$ are known as the *Kramer-Moyal* expansion coefficients. Note, that the two functions, D_1 and D_2, entering the state space process, eq. (21), are indeed the first and second coefficients in this expansion. Furthermore, due to the particular simple form of the transition probabilities, eq. (24), $D_n = 0$ for all $n > 2$. The equation obtained by truncating Kramer-Moyals expansion to $n = 2$ is generally known as the Fokker-Planck (FP) equation:

$$\frac{\partial W(x,t)}{\partial t} = \mathcal{L}_{FP} W(x,t), \tag{30}$$

where

$$\mathcal{L}_{FP} W(x,t) = -\frac{\partial}{\partial x}\big(D_1(x)W(x,t)\big) + \frac{\partial^2}{\partial x^2}\big(D_2(x)W(x,t)\big). \tag{31}$$

The mapping from eq. (20) to eq. (30) can relatively straight forward be generalized to the multivariate case as well. However, in both cases the accuracy FP-equation to describe probability evolution of the original discrete process, on the form of eq. (20), relies on the approximate constancy of the drift and diffusion function(s) on the length scale(s), $l(x,\Delta t)$ of the process associated with $\Delta t = 1$ and with the spatial domain, x, of interest. In the 1D case, $l(x,\Delta t = 1) \simeq \sqrt{2D_2(x) + D_1(x)^2}$. If the length scale associated with the variation of D_1 and D_2 is denoted l_D, the requirement would be $l(x,\Delta t) \ll l_D$ for all x.

Assuming the FP-equation to be a reasonable approximation all relevant information of the dynamics of eq. (20) is contained in the spectral decomposition of $\mathcal{L}_{FP}$. In particular, since the total probability is conserved under the action of $\mathcal{L}_{FP}$, its largest eigenvalue is $\lambda_0 \leq 0$. Therefore, if a stationary distribution, $W_0(x)$, exists, $\lambda_0 = 0$, then $W(x,t) \to W_0(x)$ at large times. The solution to $\mathcal{L}_{FP}W_0(x) = 0$ yields

$$W_0(x) = \frac{1}{\mathcal{N}} \exp\left(\int^x \frac{D_1(x')}{D_2(x')} dx'\right) \ , \tag{32}$$

where $\mathcal{N}$ is the normalization constant. In order words, W_0 exists provided that $\mathcal{N} < \infty$. Defining $U(x) \hat{=} -\int^x D_1(x')dx'$ one obtains

$$W_0(x) = \frac{1}{\mathcal{N}} \exp\left(-\frac{U(x)}{D}\right) \tag{33}$$

for a constant $D_2 = D$.[3]. From the formal equivalence between eq. (33) and the Boltzmann probability distribution of a thermal ensemble we may view U as a potential energy function and D as the effective temperature.

The second largest eigenvalue, $\lambda_1 < 0$ of $\mathcal{L}_{FP}$ determines the relaxation rate towards the steady state, in the sence that any deviation $\delta W(x,t) = W(x,t) - W_0(x)$ will decay (for large times) according to $\delta W(x,t) \propto \exp(\lambda_1 t)$. Therefore, the second largest eigenvalue defines the correlation time τ_x of the process, $\tau_x = |\lambda_1|^{-1}$. However, in most cases only an approximative expression for λ_1 can be given. For a bistable symmetric potential with minima in $= \pm x_f$ and local maximum in $x = 0$ it can be shown Risken (1996) that for large barrier heights, $\Delta U/D = (U(x_0) - U(x_f))/D \gg 1$ the eigenvalue is approximately given by

$$|\lambda_1| \approx D \left[\int_0^{x_\infty} dx \exp\left(\frac{U(x)}{D}\right) \int_x^{x_\infty} dy \exp\left(-U(y)/D\right) \right]^{-1}, \tag{34}$$

where the point $x_{inf} > x_f$ is chosen so $\exp(-U(x_\infty)/D) \approx 0$. In the bistable symmetric potential the correlation time, τ_x equals half the average time it takes to make a transition from one minimum to the other.

Appendix C. Online Appendix: Extended Ensembles

The state space in the parallel tempering (PT) algorithm is composed of R replicas of the original state space, so a PT-state is a family of R states, $\tilde{\phi}_{PT} = \{\phi_r\}_{r=1}^R$, where $\phi_r = [\boldsymbol{x}_{1:k}]_r$ in the present case. The target distribution in the PT-algorithm takes the form

$$p_{PT}(\tilde{\phi}|\boldsymbol{z}_{1:k}) = \prod_r p_r([x_{1:k}]_r \,|\boldsymbol{z}_{1:k}), \tag{35}$$

where the probability distribution associated to the r'th replica is defined in terms of an inverse temperature β_r, i.e.

$$p_r(\boldsymbol{x}_{1:k}|\boldsymbol{z}_{1:k}) = Z_r^{-1} \prod_{j=1}^{k} \left[p(\boldsymbol{x}_j|\boldsymbol{x}_{j-1})p(\boldsymbol{z}_j|\boldsymbol{x}_j)\right]^{\beta_r}. \tag{36}$$

Here, $Z_r = \int d\boldsymbol{x}_{1:k} \prod_{j=1}^k \left[p(\boldsymbol{x}_j|\boldsymbol{x}_{j-1})p(\boldsymbol{z}_j|\boldsymbol{x}_j)\right]^{\beta_r}$, is the normalization constant. The original posterior distribution is recovered for $\beta = 1$. By setting $\beta_1 = 1$ and $\beta_1 > \beta_2 > \cdots > \beta_R \simeq 0$, the distributions, p_r, will become successively flatter and consequently give rise to faster and faster relaxation times, as r increases. The 'bridging' between the different distributions is provided by augmenting the proposal density distribution, q_r, for each replica with a *replica-exchange* move chosen with some prescribed probability q_{ex}. In the traditional form of this move candidates of new states ϕ'_{r1} and ϕ'_{r2} are defined by the exchange of states of the two replica, $\phi'_{r1} = \phi_{r2}$ and $\phi'_{r2} = \phi_{r1}$. If the acceptance probability, a, is on the Metropolis form, $a = \min\{1, \tilde{a}\}$, the rate will —according to eq. (35)— be given by

$$\tilde{a} = \frac{p_{r1}(\phi'_{r1})p_{r2}(\phi'_{r2})}{p_{r1}(\phi_{r1})p_{r2}(\phi_{r2})}. \tag{37}$$

3. In the 1D case, $D_2(x)$ can always be transformed to an arbitrary constant, D, by the transformation $\tilde{x}(x) = \int^x \sqrt{D/D_2(x')}dx'$ In effect, we may assume D_2 to be constant without loss of generality Risken (1996).

Note that in analogy to the acceptance rate, eq. (6), the normalization constants, Z_{r1} and Z_{r2} cancel out in this expression. It can be shown (Iba, 2001) that the number, R, of distributions required to cover the state space with finite replica exchange rates scales with the system size, k, as $R \propto k^{1/2}$. It implies that the traditional replica-exchange move in the PT-algorithm can not directly be applied to the on-line filtering problem. However, the idea underlying the PPF-approach suggests to exchange only the last Δt part of the sequence associated with each of the two replica. In effect, for a given backward step in time, Δt (chosen according to eq. (10)), candidates of the new states ϕ'_{r1} and ϕ'_{r2} are chosen according to $\phi'_{r1} = (\left[\boldsymbol{x}_{1:(k-\Delta t-1)}\right]_{r1}, \left[\boldsymbol{x}_{(k-\Delta t):k}\right]_{r2})$ and $\phi'_{r2} = (\left[\boldsymbol{x}_{1:(k-\Delta t-1)}\right]_{r2}, \left[\boldsymbol{x}_{(k-\Delta t):k}\right]_{r1})$. This ensures that number of required replica can be kept fixed during the on-line filtering and that the computational cost of evaluating the acceptance rate, Eq. (37), is independent of k (for $k \gg \tau_q$).

The PT-approach for the present model does indeed circumvent the need of the system specific global move, Eq. (19), though the efficiency of the replica exchange move in reducing the relaxation times is inferior to Eq. (19). Preliminary runs for $h = 3$ with sampling parameters $(\tilde{N}, \tau_q, q_{ex}) = (5 \cdot 10^3, 700, 0.1)$ and with 10 replicas in the interval $[\beta_{10}, \beta_1] = [0.3; 1]$ gives $BE = 0.31 \pm 0.02$ and $RMSE = 9.62 \pm 0.25$ for the filtered estimates, and $BE = 0.23 \pm 0.03$ and $RMSE = 8.7 \pm 0.5$ for the smoothed estimates. Further improvement of the application of the PT-algorithm to the on-line filtering problem should be possible and will be the subject of a separate work. In the present context is suffices to say that all algorithms based on the extended ensemble approach facilitate fast relaxation times in the ensemble of interest by the propagation of states from high to low temperatures, which in effect is an elegant way of improving the proposal density distribution in the original ensemble. As we shall see in Section 6, the direct extension of the proposal density distribution in the form of eq. (19) is much better exploited by the PPT-method compared to traditional PF-methods. We expect this also to be true in general with the type of move-class extensions provided by the extended ensemble approach.

References

D. Alspach and H. Sorensen. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, August 1972.

S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, February 2002. URL `citeseer.nj.nec.com/article/arulampalam01tutorial.html`.

Y. Bar-Shalom and X.R. Li. *Estimation and Tracking: Principles, Techniques, and Software*. Artech House Norwood, MA., 1993.

B. P. Carlin, N. G. Polson, and D. S. Stoffer. A monte carlo approach to nonnormal and non-linear state-space modelling. *Journal of the American Statistical Association*, 87 (418):493–500, 1992.

J. Ferkinghoff-Borg. *Monte Carlo Methods in Complex Systems*. PhD thesis, Graduate School of Biophysics, Niels Bohr Institute and Risø National Laboratory, Faculty of Science University of Copenhagen, 2002.

Z. Ghahramani and M. I. Jordan. Factorial Hidden Markov Models. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 8, pages 472–478. MIT Press, 1995.

N. Gordon, D. Salmond, and A. F. M. Smith. Novel approach to non-linear and non-gaussian bayesian state estimation. *IEE Proceedings-F*, 140:107–113, 1993.

P. Hänggi and P. Talkner. Reaction-rate theory: fifty years after kramers. *Reviews of Modern Physics*, 62(2), 1990.

W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In A. Darwiche and N. Friedman, editors, *Proceedings UAI-2002*, pages 216–233, 2002.

Y. Iba. Extended ensemble monte carlo. *Int. J. Mod. Phys. C*, 12:623, 2001.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1999.

S. Julier and J Uhlmann. A new extension of the kalman filter to nonlinear systems. *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*, 1997.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

G. Kitagawa. Monte carlo filter and smoother for non-gaussian non-linear state space models. *Journal of Computational and Graphical Statistics*, 5:1–25, 1996.

J.H. Kotecha and P.M. Djuric. Gaussian sum particle filtering for dynamic state space models. *Acoustics, Speech, and Signal Processing, 2001. Proceedings. 2001 IEEE International Conference on*, 6:3465 –3468 vol.6, 2001.

T. Lehn-Schiøler, D. Erdogmus, and J. C. Principe. Parzen particle filters. *Proc. ICASSP*, 5:781–784, May 2004.

R. van der Merwe and Eric Wan. Gaussian mixture sigma-point particle filters for sequential probabilistic inference in dynamic state-space models. *Speech and Signal Processing (ICASSP)*, April 2003.

M. Mezard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyound*. World Scientific, 1987.

M. K. Pitt and N. Shephard. Theory and methods - filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

H. Risken. *The Fokker-Planck Equation*. Springer-Verlag, 1996.

N. Shephard. Partial non-gaussian state space. *Biometrika*, 81(1):115–131, 1994.

H. Tanizaki and R. Mariano. Nonlinear and non-gaussian state-space modeling with monte-carlo simulations. *Journal of Econometrics*, pages 263–290., 2000.

N.G. van Kampen. *Stochastic processes in physics and chemistry*. North Holland, 1981.