

An Introduction to Sequential Monte Carlo for Filtering and Smoothing

Olivier Cappé

LTCI, TELECOM ParisTech & CNRS
<http://perso.telecom-paristech.fr/~cappe/>

ACKNOWLEDGMENT: Eric Moulines (TELECOM ParisTech) & Tobias
Rydén (Lund)

- 1 Bayesian Dynamic Models
 - Hidden Markov Models and State-Space Models
 - Extensions
- 2 The Filtering and Smoothing Recursions
- 3 Sequential Importance Sampling
- 4 Sequential Importance Sampling with Resampling
- 5 The Auxiliary Particle Filter
- 6 Smoothing
- 7 Mixture Kalman Filter

Hidden Markov Model (HMM)

The Hidden State Process $\{X_k\}_{k \geq 0}$ is a Markov chain with initial probability density function (pdf) $t_0(x)$ and transition density function $t(x, x')$ such that*

$$p(x_{0:k}) = t_0(x_0) \prod_{l=0}^{k-1} t(x_l, x_{l+1}) .$$

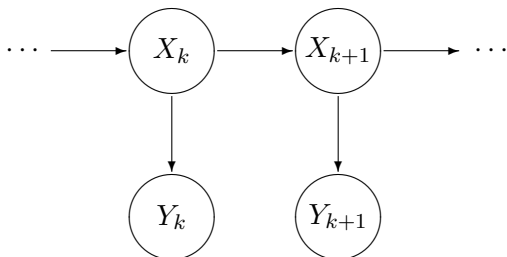
The Observed Process $\{Y_k\}_{k \geq 0}$ is such that the conditional joint density of $y_{0:k}$ given $x_{0:k}$ has the conditional independence (product) form

$$p(y_{0:k} | x_{0:k}) = \prod_{l=0}^k \ell(x_l, y_l) .$$

* $x_{0:k}$ denotes the collection x_0, \dots, x_k .

Graphical Representation of the Dependence Structure

The HMM can be represented pictorially by a **Bayesian network** which depicts the conditional independence relations:



State-Space Form

Here the model is described in a functional form:

$$\begin{aligned}X_{k+1} &= a(X_k, U_k) , \\ Y_k &= b(X_k, V_k) ,\end{aligned}$$

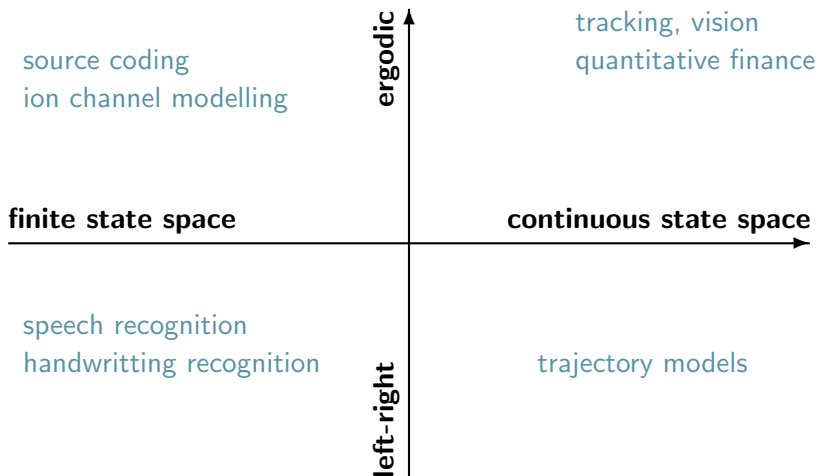
where $\{U_k\}_{k \geq 0}$ and $\{V_k\}_{k \geq 0}$ are mutually independent i.i.d. sequences of random variables (also independent of X_0).

Remark

The term *state-space model* often refers to the case where a and b are linear functions of their arguments (and $\{U_k\}$, $\{V_k\}$, X_0 are jointly Gaussian).

Likewise, the term *HMM* is sometimes used (**not in this talk!**) more restrictively for the case where X is a finite set.

HMM Examples



Beyond HMMs

For sequential Monte Carlo methods, the key point is the **structure of the conditional** $p(x_{0:k}|y_{0:k})$: the methods described in this talk directly apply in cases where the conditional may be factored as

$$p(x_{0:k}|y_{0:k}) = p(x_0|y_0) \prod_{l=0}^{k-1} p(x_{l+1}|x_l, y_{0:l+1})$$

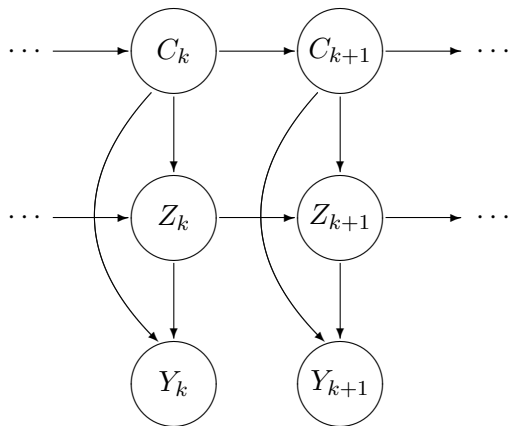
Example: Switching Autoregressive Model

If the observation equation is replaced by

$$Y_k = b(X_k)Y_{k-1} + V_k$$

then $p(x_{0:k}|y_{0:k}) = p(x_0|y_0) \prod_{l=0}^{k-1} p(x_{l+1}|x_l, y_{l+1}, y_l)$

Hierarchical HMMs



Bayesian network for a hierarchical HMM: The state sequence $\{X_k\}$ is composed of two chains $\{C_k\}$ and $\{Z_k\}$ that evolve in parallel and the observation Y_k depends on both component of the chain.

In hierarchical HMMs, sequential Monte Carlo is applicable to the *marginalized posterior* $p(c_{0:k}|Y_{0:k})$ (despite the fact that it doesn't factorizes as previously requested due to the marginalization of $Z_{0:k}$) when

$$p(z_k|c_{0:k}, y_{0:k}) \text{ is computable}$$

Conditionally Gaussian Linear State-Space Model

- Dynamic equation

$$Z_{k+1} = A(C_{k+1})Z_k + R(C_{k+1})U_k$$

- Observation equation

$$Y_k = B(C_k)Z_k + S(C_k)V_k$$

where $\{C_k\}$ is a finite-valued Markov Chain.

- 1 Bayesian Dynamic Models
- 2 The Filtering and Smoothing Recursions
 - Basic Recursions
 - Computational Filtering and Smoothing Approaches
- 3 Sequential Importance Sampling
- 4 Sequential Importance Sampling with Resampling
- 5 The Auxiliary Particle Filter
- 6 Smoothing
- 7 Mixture Kalman Filter

Tasks of interest for HMMs

State Inference How to make probabilistic statements on the state sequence given the model **and the observations?**

Filtering $\pi_{k|k}(x_k) = p(x_k|Y_{0:k})$

Prediction $\pi_{k+1|k}(x_{k+1}) = p(x_{k+1}|Y_{0:k})$

Smoothing $\pi_{0:k|k}(x_{0:k}) = p(x_{0:k}|Y_{0:k})$
(*fixed-interval*: $\pi_{l|k}$ for $l = 0, \dots, k$;
fixed-lag: $\pi_{k|k+\Delta}$ for $k = 0, \dots$)

Parameter Inference How to tune the model parameters based on the observations?

Recursive Structure of the Joint Smoothing Density

By Bayes' rule

$$\begin{aligned}\pi_{0:k+1|k+1}(x_{0:k+1}) &= (\mathbf{L}_{k+1}(Y_{0:k+1}))^{-1} t_0(x_0) \prod_{l=0}^k t(x_l, x_{l+1}) \prod_{l=0}^{k+1} \ell(x_l, Y_l) \\ &= \left(\frac{\mathbf{L}_{k+1}(Y_{0:k+1})}{\mathbf{L}_k(Y_{0:k})} \right)^{-1} \pi_{0:k|k}(x_{0:k}) t(x_k, x_{k+1}) \ell(x_{k+1}, Y_{k+1}),\end{aligned}$$

where the normalization constants \mathbf{L}_k , i.e., the **likelihood** of the observations, is usually not computable.

The Joint Smoothing Recursion

$$\pi_{0:k+1|k+1}(x_{0:k+1}) = \left(\frac{L_{k+1}}{L_k} \right)^{-1} \pi_{0:k|k}(x_{0:k}) t(x_k, x_{k+1}) \ell(x_{k+1}, Y_{k+1})$$

The **marginal recursion** may be decomposed in two steps:

Prediction

$$\pi_{k+1|k}(x_{k+1}) = \int \pi_{k|k}(x_k) t(x_k, x_{k+1}) dx_k$$

Filtering

$$\pi_{k+1|k+1}(x_{k+1}) = \left(\frac{L_{k+1}}{L_k} \right)^{-1} \pi_{k+1|k}(x_{k+1}) \ell(x_{k+1}, Y_{k+1})$$

Exact Implementation of the Filtering and Smoothing Recursions

When X is finite (Baum *et al.*, 1970) The associated computational cost is $|X|^2$ per time index (for the filtering part).

In linear Gaussian state-space models (Kalman & Bucy, 1961) The filtering and prediction recursion is implemented by the *Kalman filter* (L_{k+1}/L_k is interpreted as the likelihood of the $(k+1)$ -th innovation).

Such *finite dimensional filters* exist only in very specific models (see, e.g., Runggaldier & Spizzichino, 2001)

Approximate Implementations of the Filtering and Smoothing Recursions

- EKF (Extended Kalman Filter) Linearization-based approach (for non-linear Gaussian state space models)
- UKF (Unscented Kalman Filter, Julier & Uhlmann, 1997) Point-based approach
- Variational Methods (e.g., Valpola & Karhunen, 2002) Based on parametric density approximation arguments.
- Exact Suboptimal Filters In particular, Kalman filter viewed as minimum mean square error **linear** filtering.

Sequential Monte Carlo (SMC)

- **Sequential Monte Carlo** (sometimes called *particle filtering*) is a method which uses pseudo-random simulations to produce successive populations of weighted “particles” $X_k^{1:n}$ and associated weights $W_k^{1:n}$ such that

$$\sum_{i=1}^n W_k^i f(X_k^i) \approx \int f(x) \pi_{k|k}(x) dx ,$$

for all functions f of interest.

- The SMC process is sequential in the sense that given $X_k^{1:n}$, $W_k^{1:n}$ and the observations $Y_{0:k+1}$, $X_{k+1}^{1:n}$ and $W_{k+1}^{1:n}$ are conditionally independent of previous populations of particles.
- SMC is based on **importance sampling and resampling**.

- 1 Bayesian Dynamic Models
- 2 The Filtering and Smoothing Recursions
- 3 Sequential Importance Sampling
 - Self-Normalized Importance Sampling
 - Sequential Importance Sampling (SIS)
 - Weight Degeneracy
- 4 Sequential Importance Sampling with Resampling
- 5 The Auxiliary Particle Filter
- 6 Smoothing
- 7 Mixture Kalman Filter

Self-Normalized Importance Sampling, or IS (Hammersley & Handscomb, 1964)

IS is a weighted form of Monte Carlo approximation, in which expectations under a target pdf π

$$\pi(f) = \mathbb{E}_{\pi}[f(X)]$$

are estimated as

$$\hat{\pi}_q^n(f) = \sum_{i=1}^n \frac{\omega^i}{\underbrace{\sum_{j=1}^n \omega^j}_{W^i}} f(X^i),$$

where

- $X^i \sim \text{iid } q$,
- $\omega^i = \frac{\pi}{q}(X^i)$.

This form of IS (sometimes also called Bayesian IS) does not necessitate that π be properly normalized.

Performance of IS

Assuming that $E_{\pi}[\frac{\pi}{q}(X)(1 + f^2(X))] < \infty$, $\hat{\pi}_q^n(f)$ is consistent and asymptotically normal, with **asymptotic variance** given by

$$v_q(f) = E_{\pi} \left[\frac{\pi}{q}(X) (f(X) - \pi(f))^2 \right].$$

The asymptotic variance can be estimated from the IS sample by

$$\hat{v}_q^n(f) = n \sum_{i=1}^n (W^i)^2 \{f(X^i) - \hat{\pi}_q^n(f)\}^2,$$

where $W^i = \omega^i / \sum_{j=1}^n \omega^j$ are the **normalized weights**.

Elements of proof

Consistency If $\bar{\pi} = \pi/c$, where $c = \int \pi(x)dx$, is a pdf,

$$\mathbb{E}_q[\omega^i f(X^i)] = \mathbb{E}_q \left[\frac{\pi}{q}(X) f(X) \right] = c \mathbb{E}_{\bar{\pi}}[f(X)].$$

CLT

$$\sqrt{n}(\hat{\pi}_q^n(f) - \bar{\pi}(f)) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \omega^i (f(X^i) - \bar{\pi}(f))}{\frac{1}{n} \sum_{j=1}^n \omega^j}$$

and $\text{Var}[\omega^i (f(X^i) - \pi(f))] = c^2 \mathbb{E}_{\bar{\pi}}[\frac{\bar{\pi}}{q}(X)(f(X) - \pi(f))^2]$.

Empirical estimate

$$\hat{v}_q^n(f) = \sum_{i=1}^n W^i \frac{\frac{\bar{\pi}}{q}(X^i)}{\frac{1}{n} \sum_{j=1}^n \frac{\bar{\pi}}{q}(X^j)} \{f(X^i) - \hat{\pi}_q^n(f)\}^2$$

Empirical diagnostic tools for IS

Effective sample size

$$\text{ESS}_q^n = \left[\sum_{i=1}^n (W^i)^2 \right]^{-1}$$

- 1 $1 \leq \text{ESS}_q^n \leq n$
- 2 n / ESS_q^n is an estimate of $\text{E}_\pi \left[\frac{\pi(X)}{q(X)} \right]$ which is the maximal IS asymptotic variance for functions f such that $|f(x) - \pi(f)| \leq 1$ (note: these functions have maximal Monte Carlo variance of 1 under π).
- 3 $n / \text{ESS}_q^n - 1$ is an estimator of the χ^2 divergence

$$\int \frac{(\pi(x) - q(x))^2}{q(x)} dx$$

($n / \text{ESS}_q^n - 1$ is also the square of the empirical coefficient of variation associated with the normalized weights).

Another Empirical diagnostic tools for IS

(Shannon) Entropy of the Normalized Weights

$$\text{ENT}_q^n = - \sum_{i=1}^N W^i \log(W^i)$$

- 1 $1 \leq \exp(\text{ENT}_q^n) \leq n$
- 2 $\exp(\text{ENT}_q^n)/n$ is an estimate of $\exp[-K(\pi||q)]$, where $K(\pi||q)$ is the Kullback-Leibler divergence between π and q .

Optimal IS Instrumental Density

When considering large classes of functions f , IS generally appears to perform worse than Monte Carlo under π (e.g., for functions such that $|f(x) - \pi(f)| \leq 1$, the maximal Monte Carlo asymptotic variance is 1 whereas, the maximal IS asymptotic variance is $E_{\pi}[\frac{\pi}{q}(X)]$ which is strictly larger than 1 by Jensen's inequality, unless $q = \pi$).

However, for a fixed function f , the optimal IS density is **not** π but

$$\frac{|f(x) - \pi(f)|\pi(x)}{\int |f(x') - \pi(f)|\pi(x')dx'}$$

as Cauchy-Schwarz inequality implies that

$$E_{\pi} \left[\frac{\pi}{q}(X) (f(X) - \pi(f))^2 \right] \underbrace{E_{\pi} \left[\frac{q}{\pi}(X) \right]}_{=1} \geq E_{\pi}^2 [|f(X) - \pi(f)|] .$$

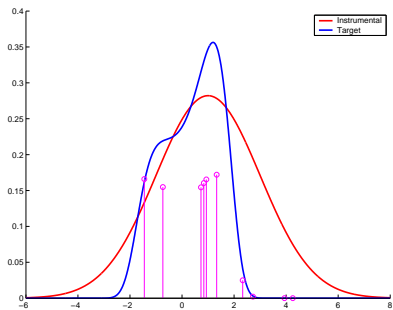
Example (Bayesian Posterior)

Consider a simple model in which

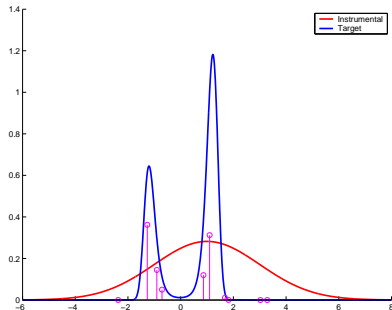
$$X_0 \sim N(1, 2^2) ,$$

$$Y_0|X_0 \sim N(X_0^2, \sigma^2) .$$

And apply IS to compute expectations under the posterior for $Y_0 = 2$, using the prior as instrumental pdf.



$\sigma = 2$	
ESS_q^n / n	0.66
$v_q(x \mapsto x)$	1.69
$v_\pi(x \mapsto x)$	1.25



$\sigma = 0.5$	
ESS_q^n / n	0.25
$v_q(x \mapsto x)$	5.35
$v_\pi(x \mapsto x)$	1.25

Back to the Filtering and Smoothing Problem

How to estimate expectations under the posterior

$\pi_{0:k|k}(x_{0:k}) = p(x_{0:k}|Y_{0:k})$ in the model

$$p(x_{0:k}) = t_0(x_0) \prod_{l=0}^{k-1} t(x_l, x_{l+1}) ,$$

$$p(y_{0:k}|x_{0:k}) = \prod_{l=0}^k \ell(x_l, y_l) ,$$

using a sequential algorithm ?

Sequential Smoothing through IS, or SIS (Handschin & Mayne, 1969-1970)

- Propose n independent particle trajectories $\{X_{0:k+1}^i\}^{1 \leq i \leq n}$ under a Markovian scheme such that

$$p(x_{0:k+1}) = \rho_{0:k+1}(x_{0:k+1}) = q_0(x_0) \prod_{l=1}^k q_l(x_l, x_{l+1}).$$

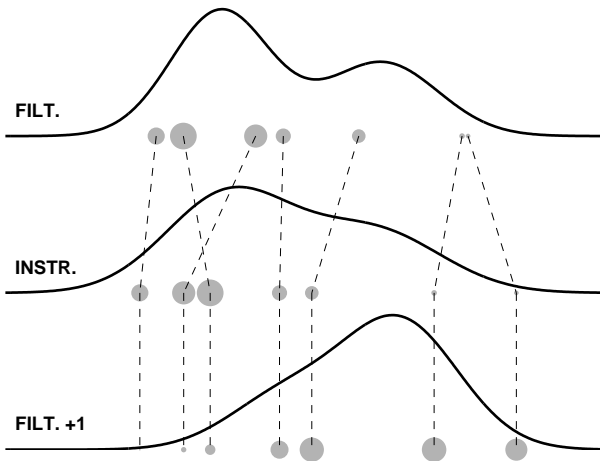
- Compute importance weights sequentially:

$$\omega_{k+1}^i = \frac{\pi_{0:k+1|k+1}(X_{0:k+1}^i)}{\rho_{0:k+1}(X_{0:k+1}^i)} = \omega_k^i \times \frac{t(X_k^i, X_{k+1}^i) \ell(X_{k+1}^i, Y_{k+1})}{q_k(X_k^i, X_{k+1}^i)}.$$

Then,

$$\sum_{i=1}^n \frac{\omega_{k+1}^i}{\sum_{j=1}^n \omega_{k+1}^j} f(X_{0:k+1}^i)$$

is an estimate of $\mathbb{E}[f(X_{0:k+1}) | Y_{0:k+1}]$.



One step of the SIS algorithm with just seven particles.

Choice of the Instrumental Kernel

The so-called “optimal” choice of q_k , consists in setting

$$q_k(x, x') = q_k^{\text{opt}}(x, x') = \frac{t(x, x')\ell(x', Y_{k+1})}{\int t(x, x'')\ell(x'', Y_{k+1})dx''}$$

for which the normalized weights are predictable, i.e. W_{k+1}^i depend on X_k^i but not X_{k+1}^i (hence, for this instrumental kernel the conditional variance cancels).

This is however usually not feasible and common choices include

- 1 the prior $q_k = t$ (and then $\omega_k^i \propto \ell(X_k^i, Y_k)$)
- 2 approximations (sometimes heuristic) to q_k^{opt} (moment matching, use of EKF or UKF, ...),
- 3 tuning parameters of q_k so as to maximize the *effective sample size* or *entropy* criterions

Weight Degeneracy

Empirically, the SIS approach always fail when the time-horizon k is more than a few tens; the IS weights $\omega_k^{1:n}$ usually become very unbalanced with a few weights dominating all the other

To understand why it is the case, consider the (silly) model where

$$\begin{cases} t(x, x') = t(x') = t_0(x'), & \text{(Independent states)} \\ \ell(x, y) = \ell(y), & \text{(Non-informative observations)} \end{cases}$$

and the instrumental kernel is such that $q_l(x, x') = q_0(x') = q(x')$

Weight Degeneracy (Contd.)

For a function of interest f that only depends on the last coordinate x_k of the trajectory $x_{0:k}$, the asymptotic variance of the SIS approximation to $\pi_{k|k}(f) = \mathbb{E}_{\pi_{k|k}}[f(X)]$ is given by

$$\begin{aligned}
 v_k(f) &= \\
 &\int \cdots \int \left(\prod_{l=0}^k \frac{t}{q}(x_l) \right)^2 (f(x_k) - \pi_{k|k}(f))^2 \prod_{l=0}^k q(x_l) dx_0 \dots dx_k \\
 &= \underbrace{\left(\int \frac{t}{q}(x) t(x) dx \right)^k}_{>1} \int \frac{t}{q}(x) (f(x) - \pi_{k|k}(f))^2 t(x) dx .
 \end{aligned}$$

In practise, this situation can usually be detected by monitoring the *effective sample size* or *entropy* criteria, which become abnormally small.

Application to the Stochastic Volatility Model

This is a non-linearly observed state-space model used to represent log-returns in quantitative finance:

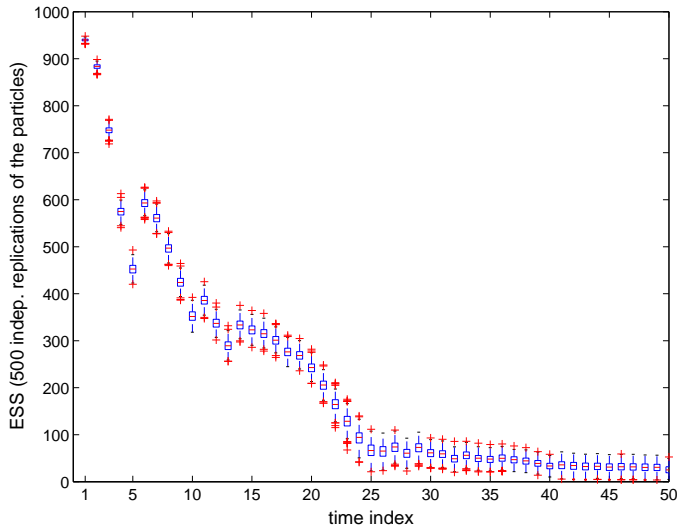
$$\begin{aligned} X_{k+1} &= \phi X_k + \sigma U_k & |\phi| < 1, \\ Y_k &= \beta \exp(X_k/2) V_k, \end{aligned}$$

where

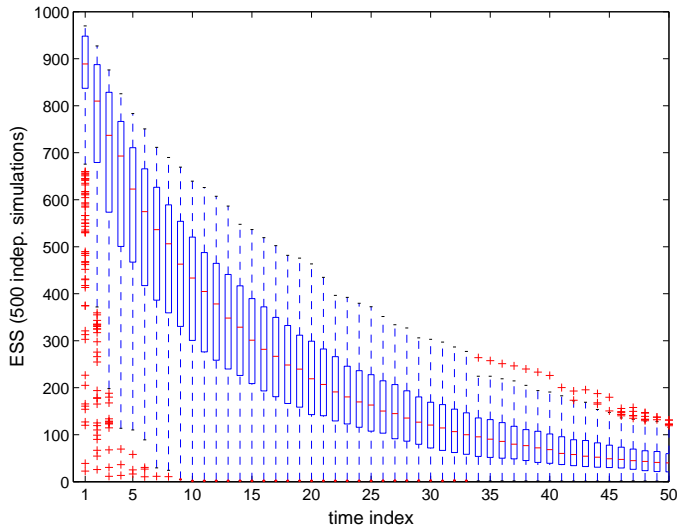
- $\{U_k\}_{k \geq 0}$ and $\{V_k\}_{k \geq 0}$ are independent standard Gaussian white noise processes.
- $X_0 \sim \mathcal{N}(0, \sigma^2 / (1 - \phi^2))$.

We consider trajectories of length 50 simulated under the model with $\phi = 0.98$, $\sigma = 0.17$ and $\beta = 0.64$ and apply SIS with $q_k = q$ and $n = 1,000$ particles.

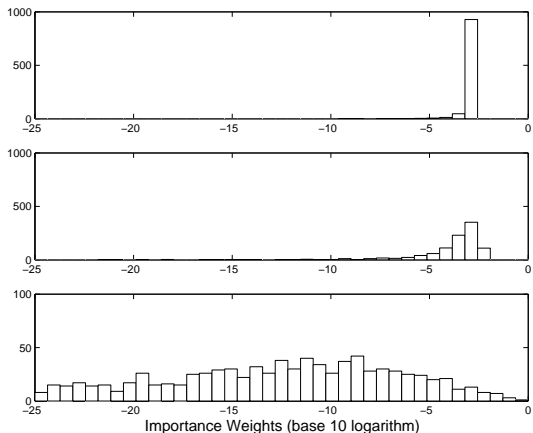
Typical Evolution of ESS for a Single Trajectory



Evolution of ESS When Averaging Over Trajectories

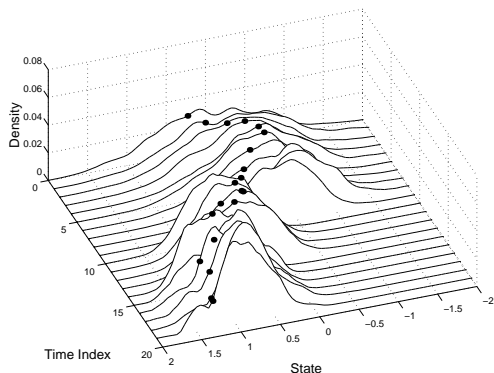


Typical Evolution of the Weight Distribution



Histograms of the base 10 logarithm of the normalized importance weights after (from top to bottom) 1, 10 and 100 iterations for the stochastic volatility model (improved instrumental kernel based on moment matching).

Filtering Results



Waterfall representation of the sequence of estimated filtering pdfs (weighted kernel smooth) together with the actual state for 1,000 particles (improved instrumental kernel based on moment matching).

- 1 Bayesian Dynamic Models
- 2 The Filtering and Smoothing Recursions
- 3 Sequential Importance Sampling
- 4 Sequential Importance Sampling with Resampling
 - Sampling Importance Resampling
 - Sequential Importance Sampling with Resampling (SISR)
 - Case Study
- 5 The Auxiliary Particle Filter
- 6 Smoothing
- 7 Mixture Kalman Filter

In IS, it is indeed possible to **reset the weights to a constant value** at the price of a, usually moderate, increase in variance.

Sampling Importance Resampling (Rubin, 1987)

Replace $\{X^{1:n}, W^{1:n}\}$ by $\{\tilde{X}^{1:\tilde{N}}, \tilde{W}^{1:\tilde{N}}\}$ such that the discrepancy between the resampled weights $\{\tilde{W}^{1:\tilde{N}}\}$ is reduced and $\sum_{i=1}^{\tilde{N}} \tilde{W}_k^i \delta_{\tilde{X}^i}$ is a good approximation to $\sum_{i=1}^n W^i \delta_{X^i}$.

In general the resampling is random and subject to the constraints

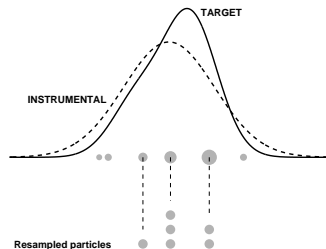
$$\begin{cases} \tilde{N} = n, \\ \tilde{W}^i = 1/\tilde{N}, \\ \mathbb{E} \left[\sum_{i=1}^{\tilde{N}} \mathbb{1}\{\tilde{X}^i = X^j\} \mid X^{1:n}, W^{1:n} \right] = \tilde{N} W^j \quad (1 \leq j \leq n). \end{cases}$$

The last condition is often referred to as *unbiasedness* or *proper weighting*.

Multinomial Resampling

- 1 Draw, conditionally independently given $\{X^{1:n}, W^{1:n}\}$, n discrete random variables (J^1, \dots, J^n) taking their values in the set $\{1, \dots, n\}$ with probabilities (W^1, \dots, W^n) .
- 2 Set, for $i = 1, \dots, n$, $\tilde{X}^i = X^{J^i}$ and $\tilde{W}^i = 1/n$.

Let $C^i = \sum_{j=1}^n \mathbb{1}\{\tilde{X}^j = X^i\}$ ($i = 1, \dots, n$) denote the number of times each particle is duplicated in the resampling process. The counts (C^1, \dots, C^n) follow a multinomial distribution with parameters n , (W^1, \dots, W^n) , conditionally to $\{X^{1:n}, W^{1:n}\}$.



Some Results on SIR

- 1 $\tilde{X}^i \xrightarrow{\mathcal{D}} \pi$ as $n \rightarrow \infty$ (some extensions of this result)
- 2 $\frac{1}{n} \sum_{i=1}^n f(\tilde{X}^i)$ is an asymptotically normal estimator of $\pi(f)$ (assuming $\mathbb{E}_\pi[\frac{\pi}{q}(X)(1 + f^2(X)) + f^2(X)] < \infty$) with asymptotic variance given by

$$\tilde{v}_q(f) = \underbrace{\mathbb{E}_\pi \left[\frac{\pi}{q}(X) (f(X) - \pi(f))^2 \right]}_{v_q(f)} + \underbrace{\mathbb{E}_\pi \left[(f(X) - \pi(f))^2 \right]}_{\text{Var}_\pi[f(X)]}$$

If n is sufficiently large, the cost of resampling is very moderate in situation that are challenging for IS, i.e., when $v_q(f) \gg \text{Var}_\pi[f(X)]$.

Elements of Proof

- For a bounded function f ,

$$\mathbb{E} \left[f(\tilde{X}^i) \right] = \mathbb{E} \left[\mathbb{E} \left(f(\tilde{X}^i) \mid X^{1:n}, W^{1:n} \right) \right] = \mathbb{E} \left[\sum_{j=1}^n W^j f(X^j) \right]$$

and $\sum_{j=1}^n W^j f(X^j) \xrightarrow{\text{a.s.}} \pi(f)$ ($|\sum_{j=1}^n W^j f(X^j)| \leq \|f\|_\infty$).

- For the CLT, an ingredient of the proof is that

$$\begin{aligned} n \operatorname{Var} \left[\frac{1}{n} \sum_{i=1}^n f(\tilde{X}^i) \right] &= n \operatorname{Var} \left[\underbrace{\sum_{i=1}^n W^i f(X^i)}_{\rightarrow v_q(f)} \right] \\ &+ \mathbb{E} \left[\underbrace{\sum_{i=1}^n W^i f^2(X^i) - \left(\sum_{i=1}^n W^i f(X^i) \right)^2}_{\xrightarrow{\text{a.s.}} \operatorname{Var}_\pi[f(X)]} \right] \end{aligned}$$

There Exists Resampling Schemes with Reduced Variance

Residual Resampling

For $i = 1, \dots, n$ set

$$C^i = \lfloor nW^i \rfloor + \bar{C}^i,$$

where $\bar{C}^1, \dots, \bar{C}^n$ are distributed, conditionally to $W^{1:n}$, according to the multinomial distribution $\text{Mult}(n - R, \bar{W}^1, \dots, \bar{W}^n)$ with $R = \sum_{i=1}^n \lfloor nW^i \rfloor$ and

$$\bar{W}^i = \frac{nW^i - \lfloor nW^i \rfloor}{n - R}, \quad i = 1, \dots, n.$$

This scheme is obviously unbiased and induces an additional variance which is always lower than $\text{Var}_\pi[f]$ (the same is true for the bootstrap filter based on residual resampling, Douc & Moulines, 2005).

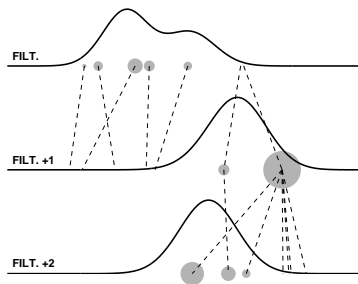
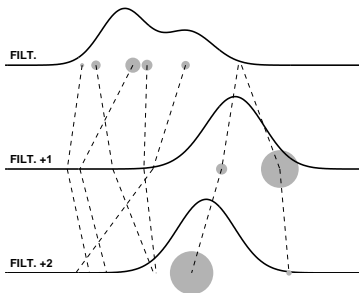
The Simplest Functional Algorithm (Gordon et al., 1993)

Regular resampling is added to avoid **weight degeneracy** and to guarantee the long-term ($k \rightarrow \infty$) stability of the particle filter.

The Bootstrap filter

- 1 Given $\tilde{X}_k^{1:n}$, propose new positions X_{k+1}^i independently **under the prior dynamic** $t(\tilde{X}_k^i, \cdot)$, for $i = 1, \dots, n$;
- 2 Compute the **weights** $\omega_{k+1}^i = \ell(X_{k+1}^i, Y_{k+1})$, for $i = 1, \dots, n$ and normalize them ($W_{k+1}^i = \omega_{k+1}^i / \sum_{j=1}^n \omega_{k+1}^j$);
- 3 **Resample** to obtain $\tilde{X}_{k+1}^{1:n}$, e.g., by drawing independent indices J_{k+1}^i such that $P(J_{k+1}^i = j | W_{k+1}^{1:n}) = W_{k+1}^j$ and setting $\tilde{X}_{k+1}^i = X_{k+1}^{J_{k+1}^i}$ (Multinomial Resampling).

To avoid unnecessary resampling, **3** can be used only when the ESS statistics of the weights $W_{k+1}^{1:n}$ falls below a threshold.



SIS (left) and SISR (right).

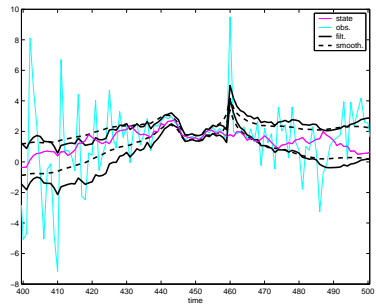
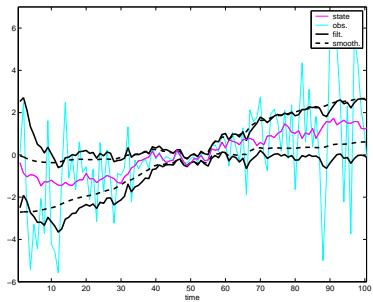
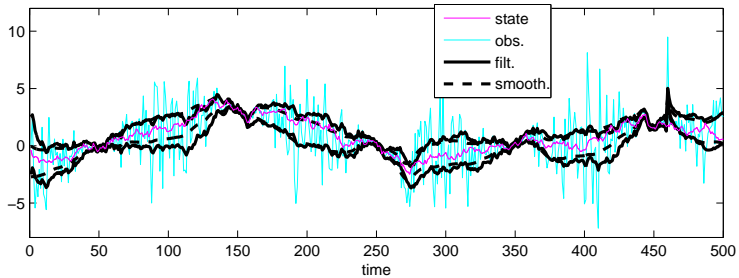
AR(1) Model Observed in Pulsated Noise

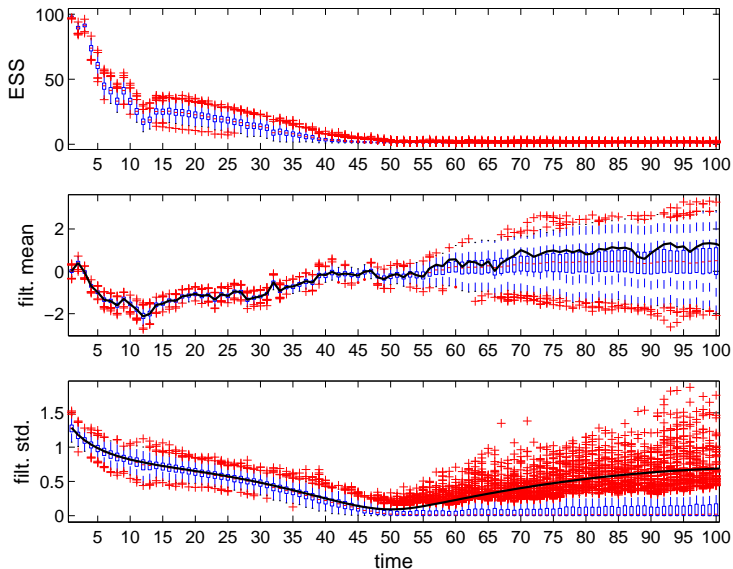
We consider a simple Gaussian linear state-space model to allow for comparison with analytical computations:

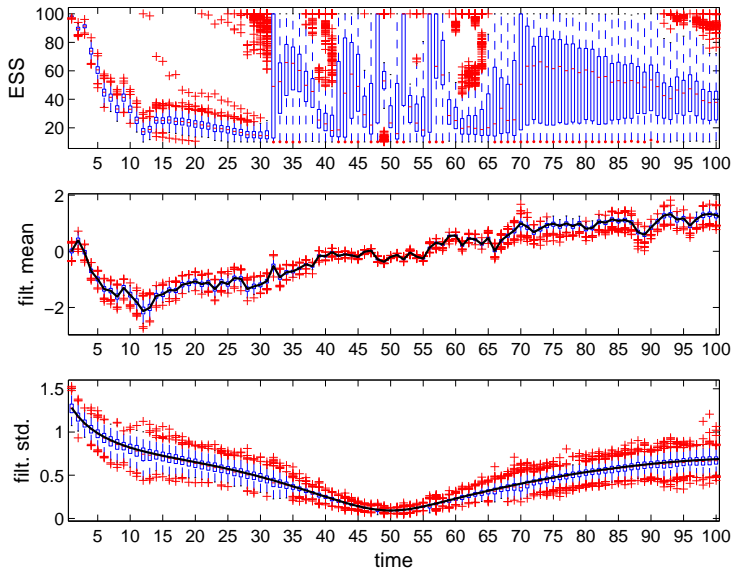
$$\begin{aligned}X_{k+1} &= \phi X_k + \sigma U_k , \\ Y_k &= X_k + \eta_k V_k ,\end{aligned}$$

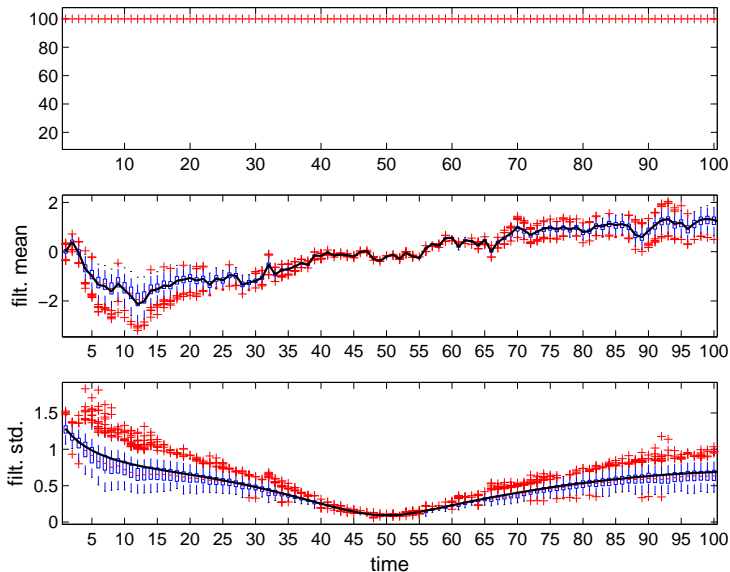
where $\phi = 0.99$, $\sigma = 0.2$, and η_k is varied (*periodically*) between 0.1 and 3.

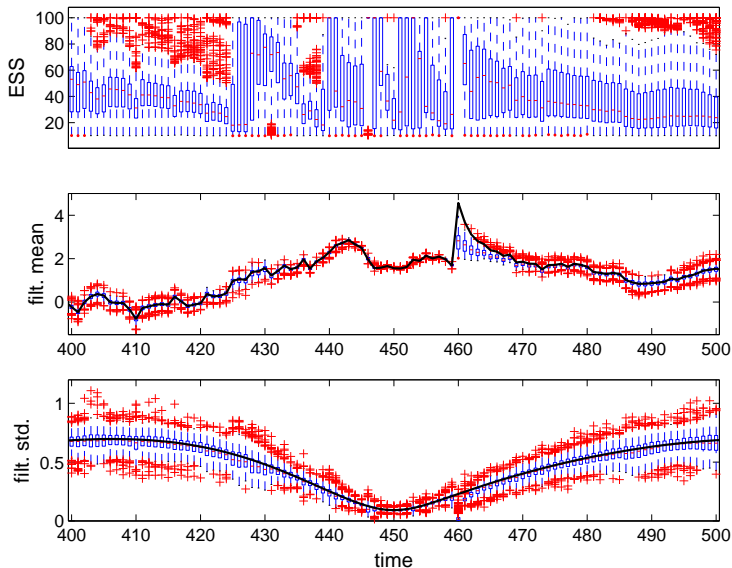
The observation sequence is simulated from the model but we also consider the influence of an out-of-model **outlying observation** at time 460.



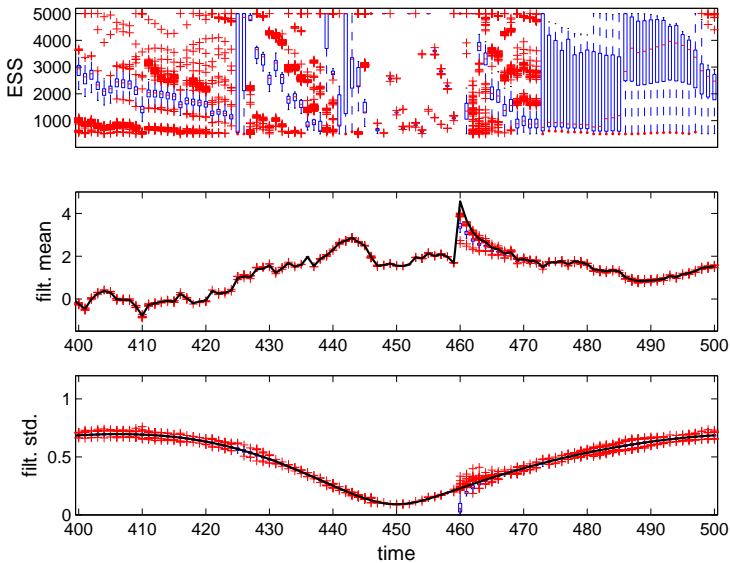
Sequential Importance Sampling (prior kernel, $n = 100$)

Bootstrap Filter (prior kernel, $n = 100$, resamp. ESS < 10)

Bootstrap Filter (prior kernel, $n = 100$, resamp. always)

Bootstrap Filter (prior kernel, $n = 100$, resamp. ESS < 10)

Bootstrap Filter ($n = 5,000$, resamp. ESS < 500)



Conclusions

- With **resampling**, SISR achieves **long-term stability**.
- The increase in variance due to resampling is very moderate, especially when resampling is applied only when needed.
- The method is still sensitive to outliers, model misspecification, etc., which may necessitate the use of more elaborate instrumental kernels.

- 1 Bayesian Dynamic Models
- 2 The Filtering and Smoothing Recursions
- 3 Sequential Importance Sampling
- 4 Sequential Importance Sampling with Resampling
- 5 The Auxiliary Particle Filter
 - A New Interpretation of SISR
 - The Auxiliary Trick
- 6 Smoothing
- 7 Mixture Kalman Filter

Alternatives to SISR

- The resampling step in the SISR algorithm can be seen as a method to sample approximately under the distribution obtained when plugging the current particle approximation into the filtering update.
- This alternative way of thinking about resampling suggests several sequential Monte Carlo variants.

The Filtering Recursion Revisited

Recall that (with more general notations)

$$\pi_{k+1|k+1}(f) = \frac{\int \int f(x') \pi_{k|k}(dx) t(x, dx') \ell(x', Y_{k+1})}{\int \int \pi_{k|k}(dx) t(x, dx') \ell(x', Y_{k+1})}.$$

Now, consider what happens when considering the **empirical filtering distribution**

$$\hat{\pi}_{k|k}^n = \sum_{i=1}^n W_k^i \delta_{X_k^i}$$

as an approximation to $\pi_{k|k}$ and plugging it into the previous relation.

Filtering Target

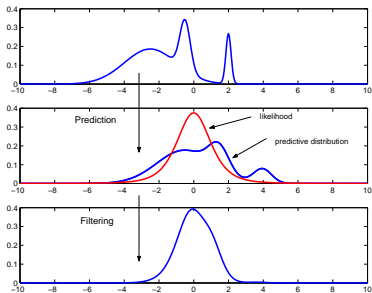
One obtains a mixture target pdf

$$\begin{aligned}\pi_{k+1|k+1}^n(x) &= \frac{\sum_{i=1}^n W_k^i t(X_k^i, x) \ell(x, Y_{k+1})}{\sum_{i=1}^n \int W_k^i t(X_k^i, x) \ell(x, Y_{k+1}) dx} \\ &= \sum_{i=1}^n \frac{W_k^i \gamma_k(X_k^i)}{\sum_{j=1}^n W_k^j \gamma_k(X_k^j)} q_k^{\text{opt}}(X_k^i, x),\end{aligned}$$

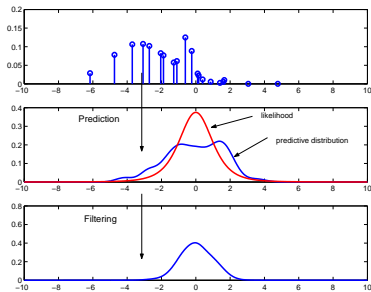
where

$$\begin{aligned}\gamma_k(x) &= \int t(x, dx') \ell(x', Y_{k+1}), \\ q_k^{\text{opt}}(x, x') &= \frac{t(x, x') \ell(x', Y_{k+1})}{\gamma_k(x)}.\end{aligned}$$

Filtering Step



Approximation



The previous reinterpretation of SISR shows that resampling and trajectory update can be integrated into a single **global IS step that targets** $\pi_{k+1|k+1}^n$.

But,

- how to chose the global instrumental kernel $q_k^{\text{glob}}((X_k^{1:n}, W_k^{1:n}), x)$?
- as $\pi_{k+1|k+1}^n$ is an n -mixture density, evaluation of the IS weights may necessitate of the order of n^2 computations.

Remark The global instrumental kernel that corresponds to the bootstrap filter (with resampling) is

$$q_k^{\text{glob}}((X_k^{1:n}, W_k^{1:n}), x) = \sum_{i=1}^n W_k^i t(X_k^i, x) .$$

The Auxiliary Trick

Assume that the global instrumental kernel is a mixture of the form

$$q_k^{\text{glob}}((X_k^{1:n}, W_k^{1:n}), x) = \sum_{i=1}^n \tau_k^i q_k^i(X_k^i, x).$$

To avoid the n^2 update, one can use **data augmentation**, introducing the mixture component as an **auxiliary variable**:

$$q_k^{\text{aux}}((X_k^{1:n}, W_k^{1:n}), (i, x)) = \tau_k^i q_k^i(X_k^i, x)$$

defines a pdf on $\{1, \dots, n\} \times \mathsf{X}$, whose marginal is $q_k((X_k^{1:n}, W_k^{1:n}), x)$.

Likewise,

$$\pi_{k+1|k+1}^{n, \text{aux}}(i, x) = \frac{1}{c} W_k^i t(X_k^i, x) \ell(x, Y_{k+1}),$$

where $c = \sum_{i=1}^n \int W_k^i t(X_k^i, x) \ell(x, Y_{k+1}) dx$, is the auxiliary target with marginal $\pi_{k+1|k+1}^n(x)$.

Auxiliary Sampling Algorithm (Pitt & Shephard, 1999)

Auxiliary Particle Filter

Repeat n times independently, for $i = 1, \dots, n$,

- 1 Sample an index J^i in $\{1, \dots, n\}$ with probabilities $(\tau_k^1, \dots, \tau_k^n)$.
- 2 Sample a position X_{k+1}^i from $q_k(X_k^{J^i}, x)$.
- 3 Compute the (unnormalized) auxiliary IS weight

$$\omega_{k+1}^i = \frac{W_k^{J^i} t(X_k^{J^i}, X_{k+1}^i) \ell(X_{k+1}^i, Y_{k+1})}{\tau_k^{J^i} q_k(X_k^{J^i}, X_{k+1}^i)} .$$

The Auxiliary View Provides New Degrees of Freedom

In the original auxiliary particle filter, Pitt & Shephard used $q_k^i = t$ and proposed an heuristic rule for setting the weights τ_k^i based on X_k^i and Y_{k+1} .

The auxiliary IS weight may also be rewritten in normalized form as

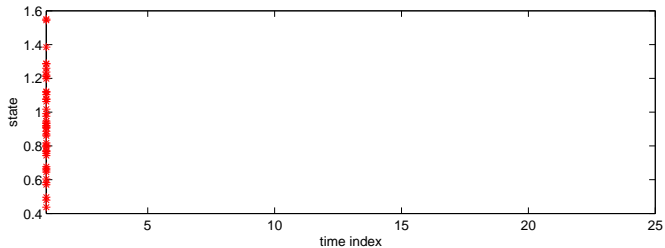
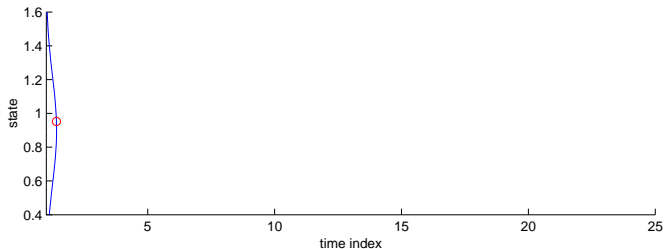
$$\omega_{k+1}^i = \frac{W_k^{J^i} \gamma_k(X_k^{J^i}) q_k^{\text{opt}}(X_k^{J^i}, X_{k+1}^i)}{\tau_k^{J^i} t(X_k^{J^i}, X_{k+1}^i)}.$$

showing that the optimal choice of τ_k^i , in the sense of minimizing the **conditional** (to $X_k^{1:n}, W_k^{1:n}$) variance of $\sum_{i=1}^n W_{k+1}^i f(X_{k+1}^i)$ for a function f , is an instance of the Neyman allocation problem. And thus the optimal choice is

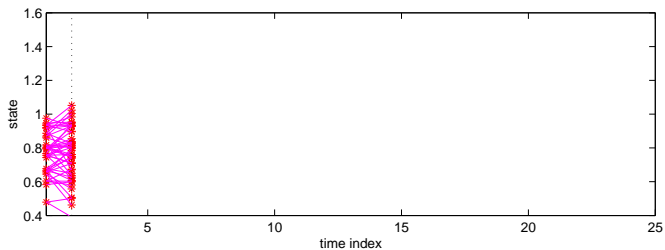
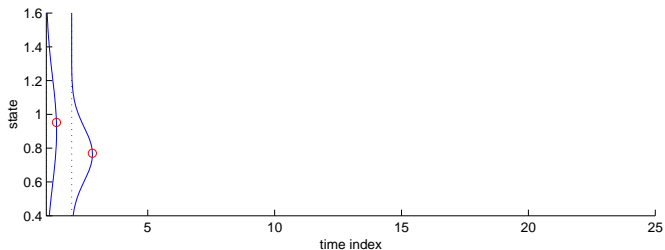
$$\tau_k^i \propto W_k^i \gamma_k(X_k^i) \sqrt{v^i(f)}$$

where $v^i(f)$ is the asymptotic variance of IS for f , target $q_k^{\text{opt}}(X_k^i, x)$, and, instrumental pdf $t(X_k^i, x)$ (Olsson *et al.*, 2007).

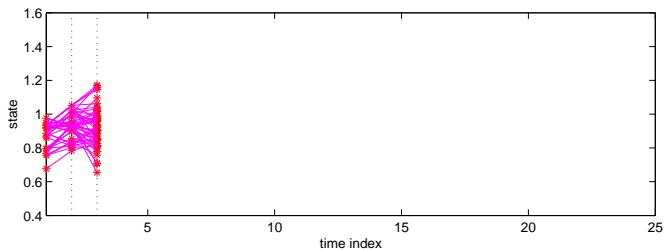
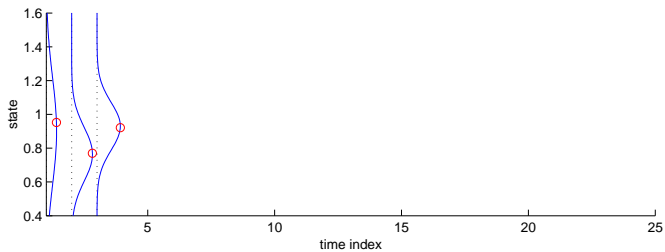
- 1 Bayesian Dynamic Models
- 2 The Filtering and Smoothing Recursions
- 3 Sequential Importance Sampling
- 4 Sequential Importance Sampling with Resampling
- 5 The Auxiliary Particle Filter
- 6 Smoothing
 - Using the Ancestry Tree
 - Backward Reweighting
 - Smoothing for Sum Functionals
- 7 Mixture Kalman Filter



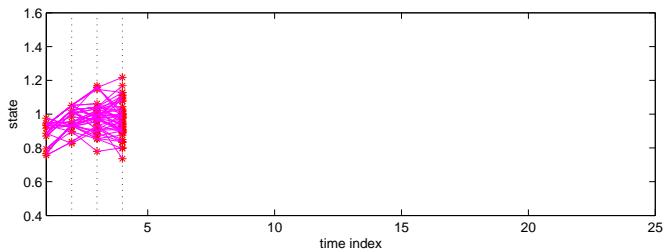
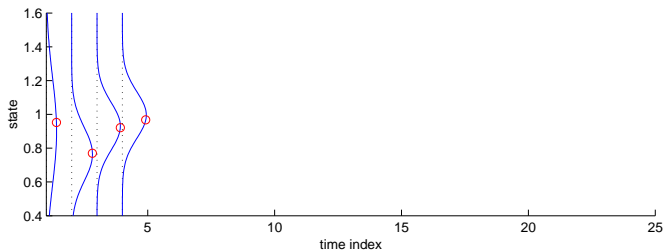
Predictive densities and evolution of the particle ancestry tree



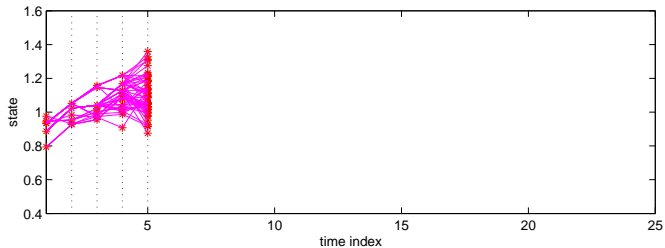
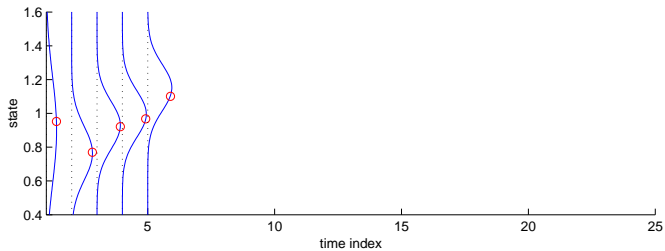
Predictive densities and evolution of the particle ancestry tree



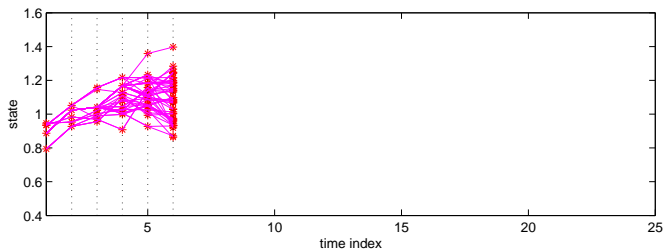
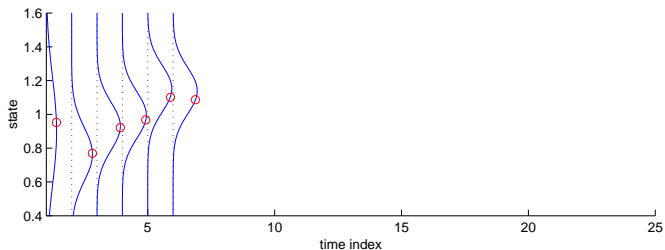
Predictive densities and evolution of the particle ancestry tree



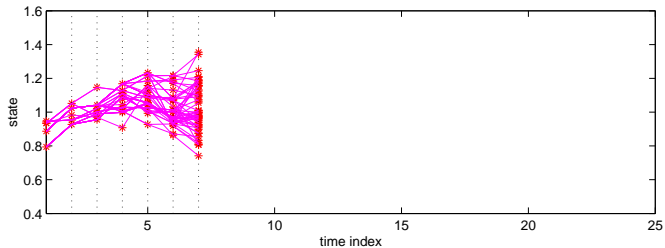
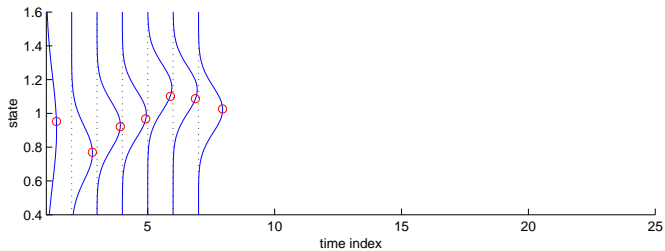
Predictive densities and evolution of the particle ancestry tree



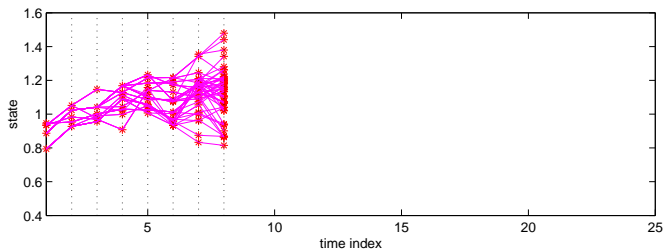
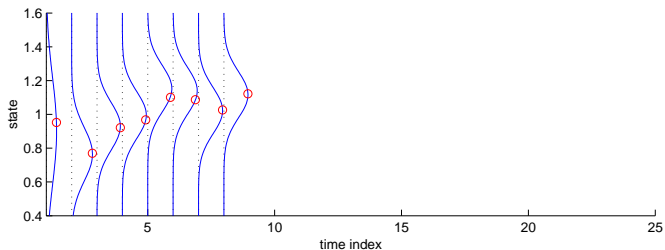
Predictive densities and evolution of the particle ancestry tree



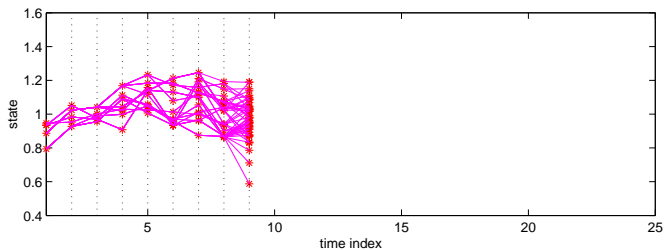
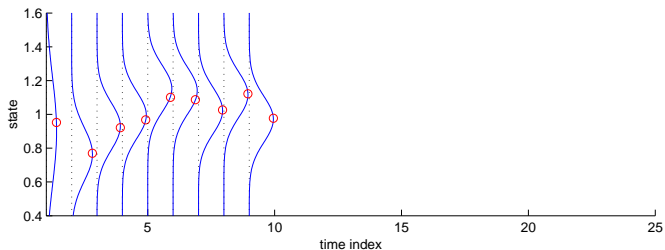
Predictive densities and evolution of the particle ancestry tree



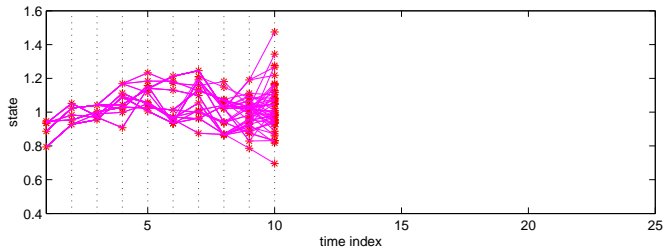
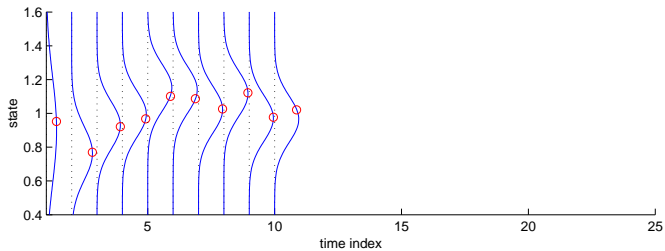
Predictive densities and evolution of the particle ancestry tree



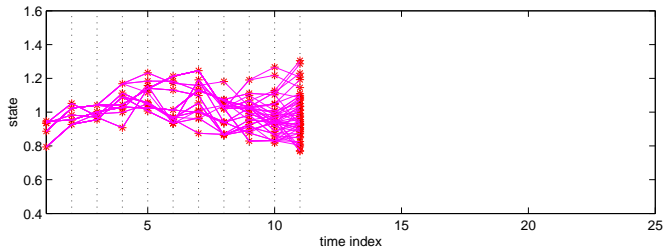
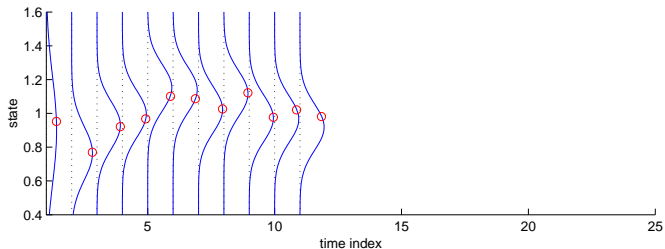
Predictive densities and evolution of the particle ancestry tree



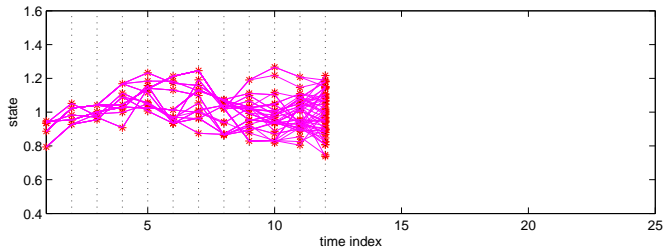
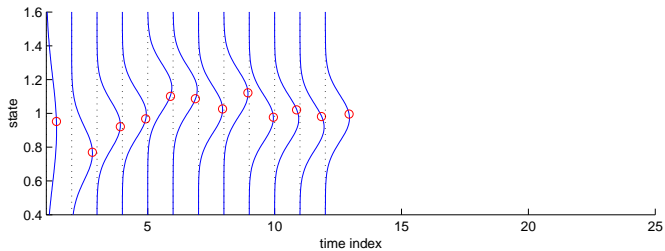
Predictive densities and evolution of the particle ancestry tree



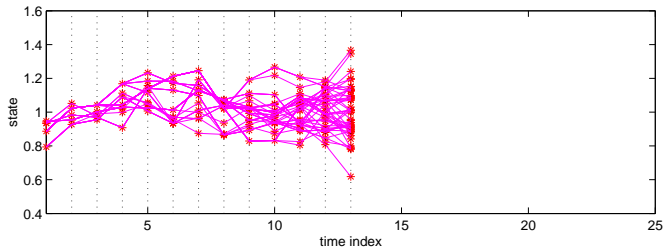
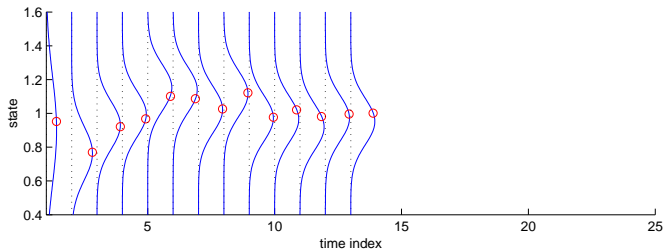
Predictive densities and evolution of the particle ancestry tree



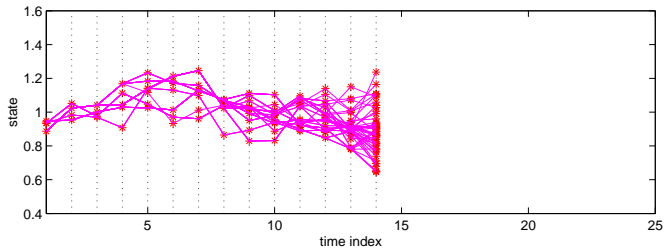
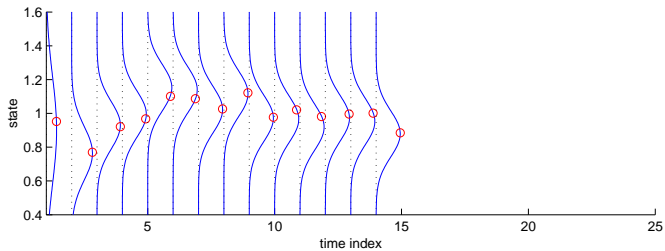
Predictive densities and evolution of the particle ancestry tree



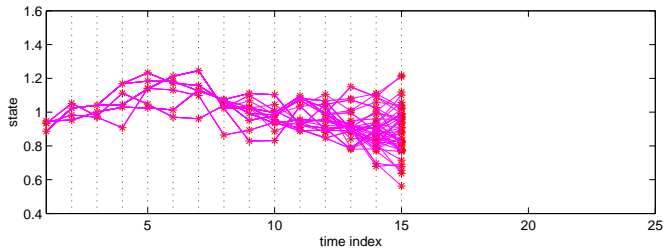
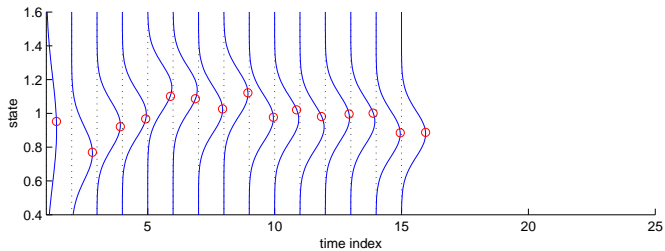
Predictive densities and evolution of the particle ancestry tree



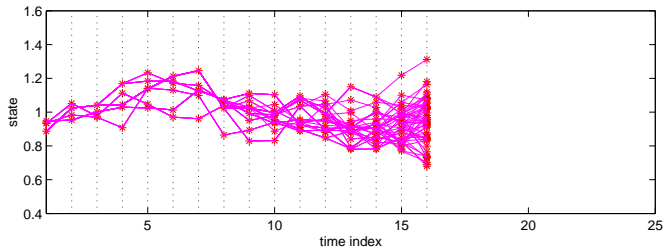
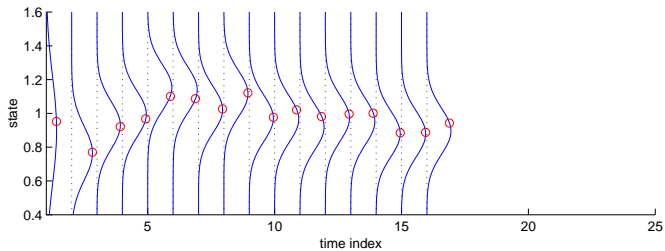
Predictive densities and evolution of the particle ancestry tree



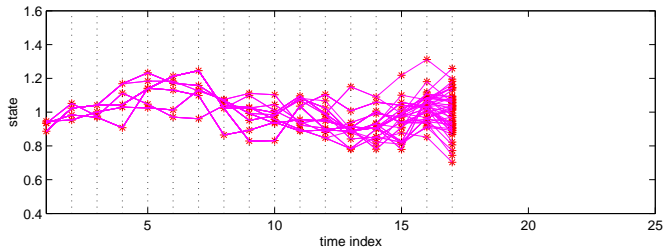
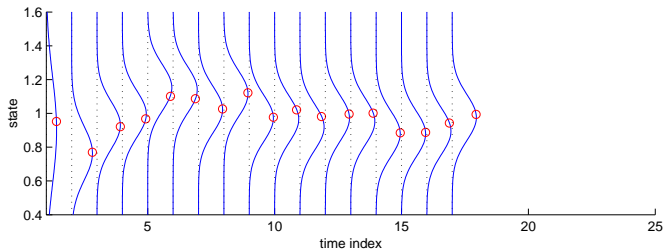
Predictive densities and evolution of the particle ancestry tree



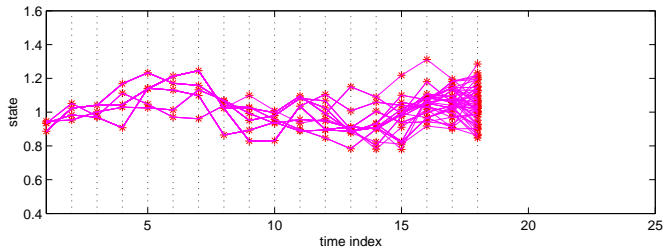
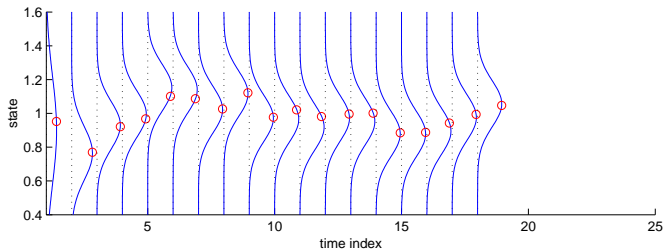
Predictive densities and evolution of the particle ancestry tree



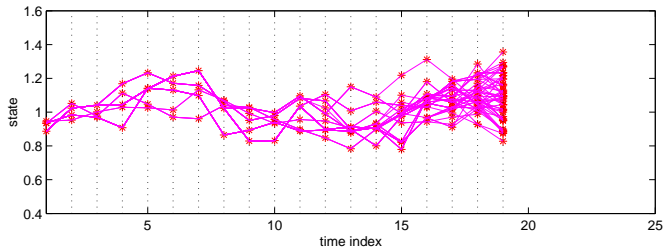
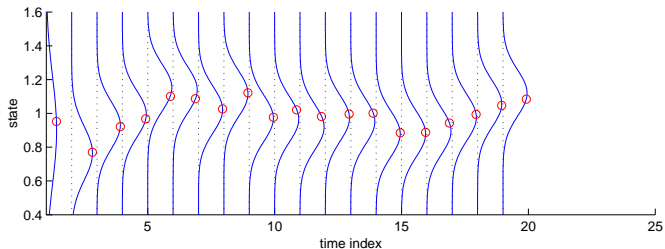
Predictive densities and evolution of the particle ancestry tree



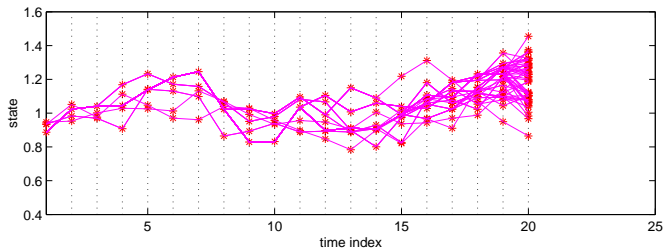
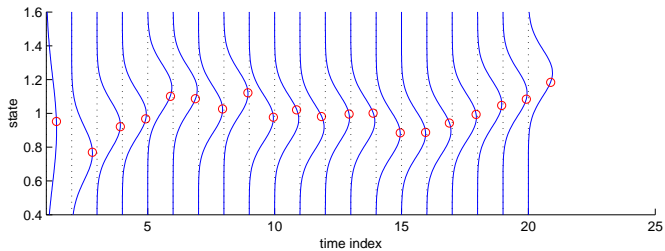
Predictive densities and evolution of the particle ancestry tree



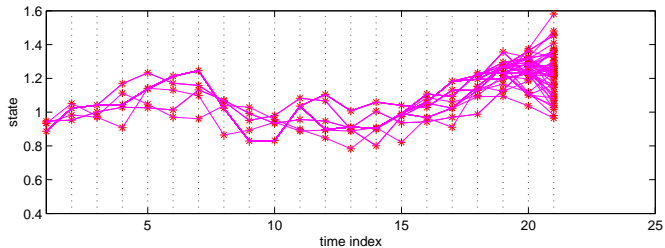
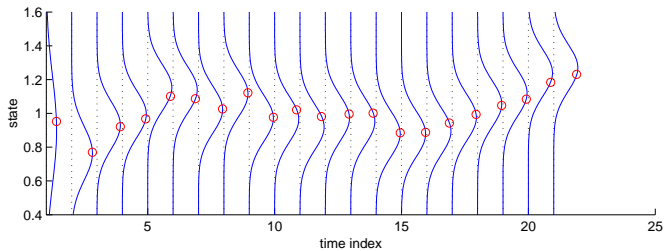
Predictive densities and evolution of the particle ancestry tree



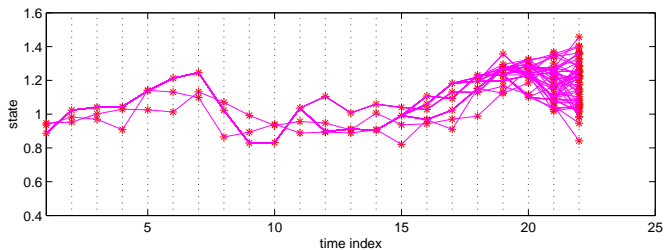
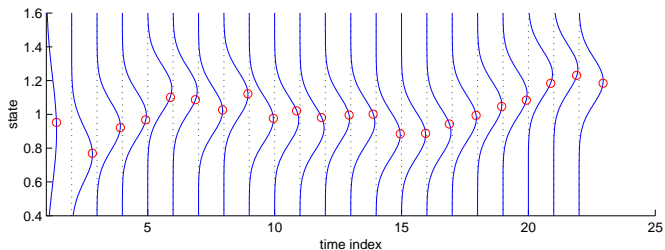
Predictive densities and evolution of the particle ancestry tree



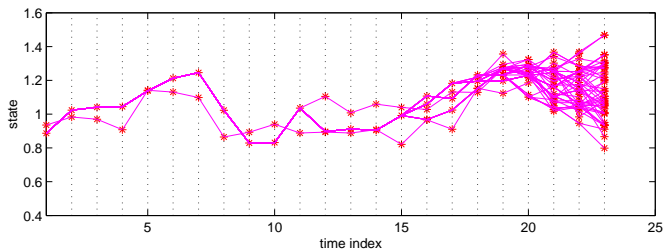
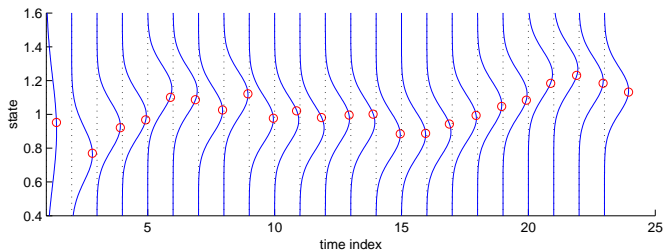
Predictive densities and evolution of the particle ancestry tree



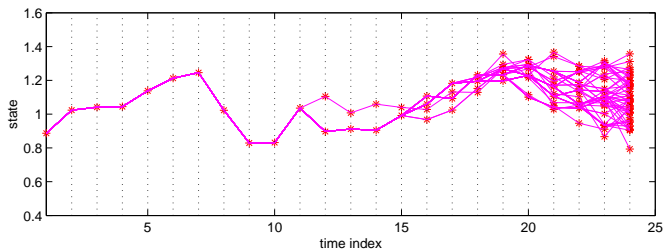
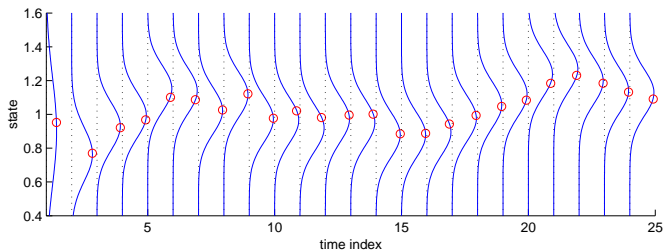
Predictive densities and evolution of the particle ancestry tree



Predictive densities and evolution of the particle ancestry tree



Predictive densities and evolution of the particle ancestry tree



Predictive densities and evolution of the particle ancestry tree

Backward Smoothing Recursion

There are several options for computing the marginal smoothing pdfs $\pi_{l|k}$ such that

$$\pi_{l|k}(x) = p(x_l | Y_{0:k})$$

for $l = 0, \dots, k$.

The **backward smoothing recursion** is based on the observation that the conditional time-reversed state sequence has a non-homogeneous Markovian structure (for conditionally Gaussian linear state-space models, this is known as RTS, or Rauch-Tung-Striebel, 1965, smoothing).

Backward Smoothing Recursion

Define the backward smoothing kernels:

$$b_l(x_{l+1}, x_l) = \frac{\pi_{l|l}(x_l)t(x_l, x_{l+1})}{\int \pi_{l|l}(x)t(x, x_{l+1})dx} ,$$

for $l = k - 1, k - 2, \dots, 0$.

Then

$$\pi_{l|k}(x_l) = \int \pi_{l+1|k}(x_{l+1})b_l(x_{l+1}, x_l)dx_{l+1} .$$

As

$$p(x_l, x_{l+1}|Y_{0:k}) = \underbrace{p(x_l|x_{l+1}, Y_{0:k})}_{b_l(x_{l+1}, x_l)} \underbrace{p(x_{l+1}|Y_{0:k})}_{\pi_{l+1|k}(x_{l+1})}$$

Particle Reweighting Scheme

The natural approximation to the backward smoothing recursion consists in computing **smoothing weights** $W_{l|k}^i$ backwards according to the recursion

- $W_{k|k}^i = W_k^i$, for $i = 1, \dots, n$.
- For $l = k - 1, k - 2, \dots, 0$,

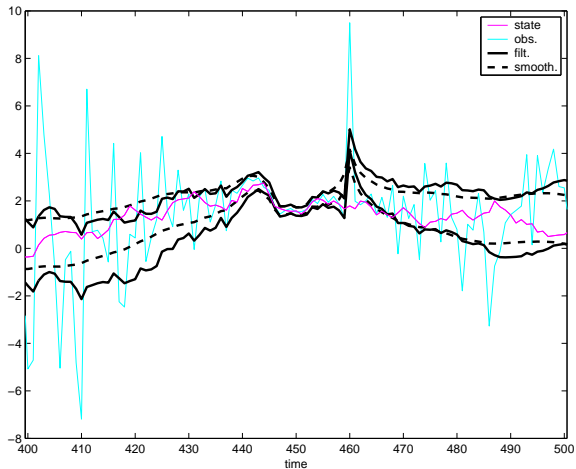
$$W_{l|k}^i = \sum_{j=1}^n W_{l+1|k}^j \frac{W_l^i t(X_l^i, X_{l+1}^j)}{\sum_{i'=1}^n W_l^{i'} t(X_l^{i'}, X_{l+1}^j)},$$

for $i = 1, \dots, n$.

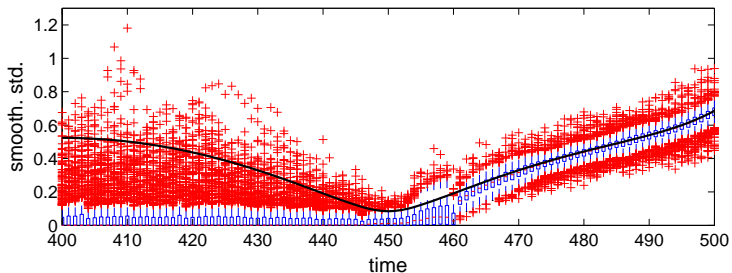
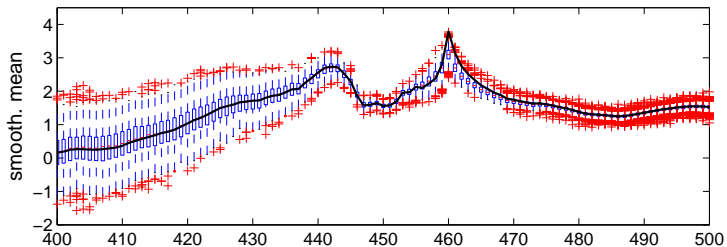
The approximation to $E[f(X_l)|Y_{0:k}]$ is given by $\sum_{i=1}^n W_{l|k}^i f(X_l^i)$.

Back to Our Case Study

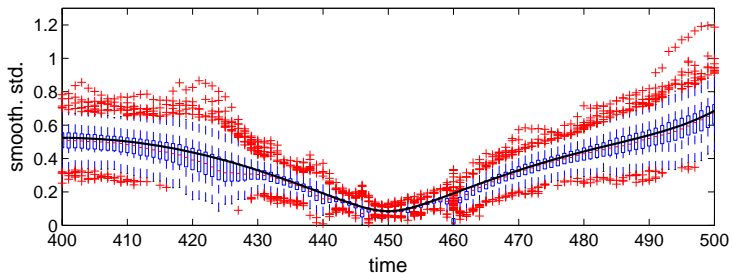
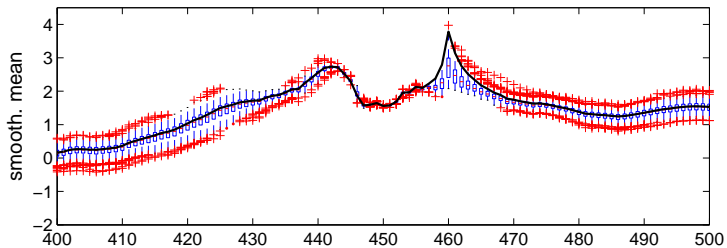
AR(1) Model Observed in Pulsated Noise



Using the Ancestry Tree ($n = 100$)



With Backward Reweighting ($n = 100$)



Backward Particle Reweighting

Appears to be efficient and stable in the long term (although this hasn't been proved yet).

Yet,

- it is not sequential (in particular, one needs to store all particle positions and weights);
- it has a potential numerical complexity proportional to the number n of particles *squared* (but not further likelihood evaluation is needed).

Sum Functionals and Parameter Estimation

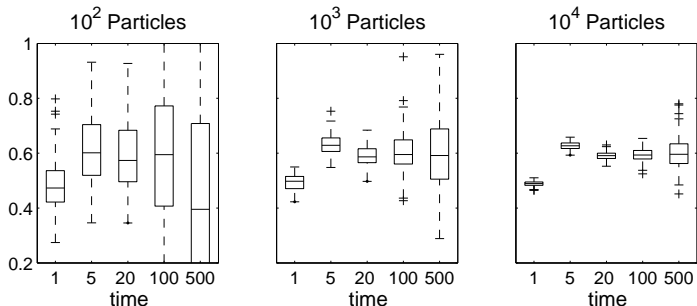
For parameter estimation, one requires smoothing of particular sum functionals of the hidden states.

In the example of the **stochastic volatility model**, one needs to evaluate $E[s_i(X_{0:n})|Y_{0:n}]$, for $0 \leq i \leq 4$ with

$$s_0(x_{0:n}) = x_0^2, \quad s_1(x_{0:n}) = \sum_{k=0}^{n-1} x_k^2, \quad s_2(x_{0:n}) = \sum_{k=1}^n x_k^2, \\ s_3(x_{0:n}) = \sum_{k=1}^n x_k x_{k-1}, \quad s_4(x_{0:n}) = \sum_{k=0}^n Y_k^2 \exp(-x_k).$$

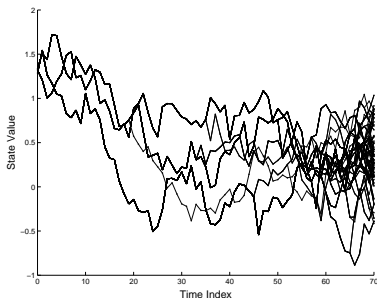
We consider the simple case of $s_0(X_0)$ using $\sum_{i=1}^N W_n^i \{X_{0:n}^i(0)\}^2$ as the sequential Monte Carlo estimate of $E(s_0(X_0)|Y_{0:n})$ (for sum functionals, the same applies for each term in the sum).

Smoothing for $s_0(X_0)$



Box and whisker plots of particle estimates of $\int x^2 \pi_{0|k}(dx)$ for $k = 1, 5, 20, 30$ and 500, and particle population sizes $n = 10^2$, 10^3 and 10^4 .

Smoothing for $s_0(X_0)$, Contd.



Particle trajectories at time $n = 70$ for the stochastic volatility model with $n = 100$ particles and systematic resampling.

Using $k < n$ is beneficial! Properly setting k corresponds to a bias-variance tradeoff: $k \uparrow$ bias decreases, $k \downarrow$ variance decreases.

Hence, the general principle for *sequential* smoothing of sum functionals: use fixed-lag smoothing (Olsson *et al.*, 2008).

- 1 Bayesian Dynamic Models
- 2 The Filtering and Smoothing Recursions
- 3 Sequential Importance Sampling
- 4 Sequential Importance Sampling with Resampling
- 5 The Auxiliary Particle Filter
- 6 Smoothing
- 7 Mixture Kalman Filter**
 - The Mixture Kalman Filter (MKF) Algorithm
 - An Application to Change Point Detection

Filtering in Conditionally Gaussian Linear State-Space Models

There are several cases of interest where the “particles” are more complicated than just elements of the state-space X .

Conditionally Gaussian Linear State-Space Model

- Dynamic equation

$$Z_k = A(C_k)Z_{k-1} + R(C_k)U_{k-1}$$

- Observation equation

$$Y_k = B(C_k)Z_k + S(C_k)V_k$$

where $\{C_k\}$ is itself a finite-valued Markov Chain.

Many applications: Non-Gaussian noises modelled as mixtures, switching models, applications in digital communications, etc.

The Exact Filtering Recursions

- Given $C_{0:k}$, $p(z_k | Y_{0:k}, C_{0:k})$ is a Gaussian distribution with mean vector $\hat{Z}_{k|k}(Y_{0:k}, C_{0:k})$ and covariance matrix $\Sigma_{k|k}(Y_{0:k}, C_{0:k})$, which can be determined recursively using the Kalman filter.
- Hence $p(x_k | Y_{0:k})$ is the mixture of $|\mathcal{C}|^{k+1}$ Gaussian densities

$$p(x_k | Y_{0:k}) \propto \sum_{c_{0:k} \in |\mathcal{C}|^{k+1}} L_k(Y_{0:k} | c_{0:k}) p(c_{0:k}) \times \\ N\left(x_k; \hat{Z}_{k|k}(Y_{0:k}, c_{0:k}), \Sigma_{k|k}(Y_{0:k}, c_{0:k})\right) .$$

This is obviously of no use when k is not very small.

Mixture Kalman Filtering

- Using (Z_k, C_k) as state variable, one can use the SMC approaches described so far with $(Z \times C)$ -valued particles.
- One can also use a related algorithm, where each particle represents a trajectory $C_{0:k}$ summarized by the Kalman statistics $\hat{Z}_{k|k}(Y_{0:k}, C_{0:k})$ and $\Sigma_{k|k}(Y_{0:k}, C_{0:k})$.

The resulting algorithm mixes **systematic exploration of the trajectory continuations (in C)**, **Kalman update to compute the likelihood factor** and **resampling**.

MKF: computation of the weights

For $i = 1, \dots, n$ and $j = 1, \dots, r$, compute

$$\hat{Z}_{k+1|k}(C_{0:k}^i, j) = A(j)\hat{Z}_{k|k}(C_{0:k}^i),$$

$$\Sigma_{k+1|k}(C_{0:k}^i, j) = A(j)\Sigma_{k|k}(C_{0:k}^i)A^t(j) + R(j)R^t(j),$$

$$\hat{Y}_{k+1|k}(C_{0:k}^i, j) = B(j)\hat{Z}_{k+1|k}(C_{0:k}^i, j),$$

$$\Gamma_{k+1}(C_{0:k}^i, j) = B(j)\Sigma_{k+1|k}(C_{0:k}^i, j)B^t(j) + S(j)S^t(j),$$

$$\omega_{k+1}^{i,j} = W_k^i \mathbf{N}(Y_{k+1}; \hat{Y}_{k+1|k}(C_{0:k}^i, j), \Gamma_{k+1}(C_{0:k}^i, j)) Q_C(C_k^i, j).$$

MKF: Importance Sampling Step

For $i = 1, \dots, n$, draw J_{k+1}^i with probabilities proportional to $\omega_{k+1}^{i,1}, \dots, \omega_{k+1}^{i,r}$, conditionally independently of the particle history, and set

$$C_{0:k+1}^i = (C_{0:k}^i, J_{k+1}^i),$$

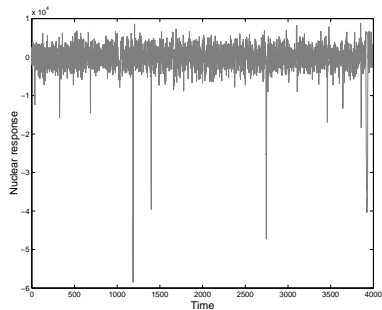
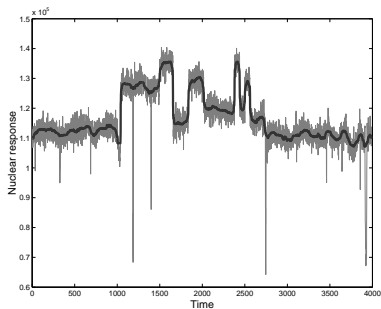
$$W_{k+1}^i = \frac{\sum_{j=1}^r \omega_{k+1}^{i,j}}{\sum_{i=1}^n \sum_{j=1}^r \omega_{k+1}^{i,j}},$$

$$K_{k+1}(C_{0:k+1}^i) = \Sigma_{k+1|k}(C_{0:k}^i, J_{k+1}^i) B^t(J_{k+1}^i) \Gamma_{k+1}^{-1}(C_{0:k+1}^i, J_{k+1}^i),$$

$$\begin{aligned} \hat{Z}_{k+1|k+1}(C_{0:k+1}^i) &= \hat{Z}_{k+1|k}(C_{0:k}^i, J_{k+1}^i) \\ &\quad + K_{k+1}(C_{0:k+1}^i) \{Y_{k+1} - \hat{Y}_{k+1|k}(C_{0:k}^i, J_{k+1}^i)\}, \end{aligned}$$

$$\Sigma_{k+1|k+1}(C_{0:k+1}^i) = \{I - K_{k+1}(C_{0:k+1}^i) B(J_{k+1}^i)\} \Sigma_{k+1|k}(C_{0:k}^i, J_{k+1}^i).$$

Application to a Change Point Detection Task



Left: well-log data waveform with a median smoothing estimate of the state. Right: median smoothing residual.

Change Point Modelling

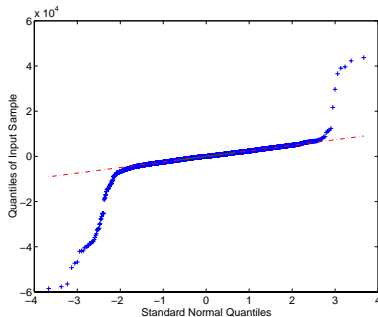
- To model this situation we put $C = \{0, 1\}$, where $C_k = 0$ means that there is no change point at time index k whereas $C_k = 1$ means that a change point has occurred.
- The state space model is

$$\begin{aligned}Z_{k+1} &= A(C_{k+1})Z_k + R(C_{k+1})U_k, \\Y_k &= Z_k + V_k,\end{aligned}$$

where $A(0) = I$, $R(0) = 0$ and $A(1) = 0$ and $R(1) = R$.

- The simplest model consists in taking for $\{C_k\}_{k \geq 0}$ an i.i.d. sequence of Bernoulli random variables with probability of success p . The time between two change points (period of time during which the state variable is constant) is then distributed as a geometric random variable with mean $1/p$.

Outliers Modelling



Quantile-quantile regression of empirical quantiles of the well-log data residuals with respect to quantiles of the standard normal distribution.

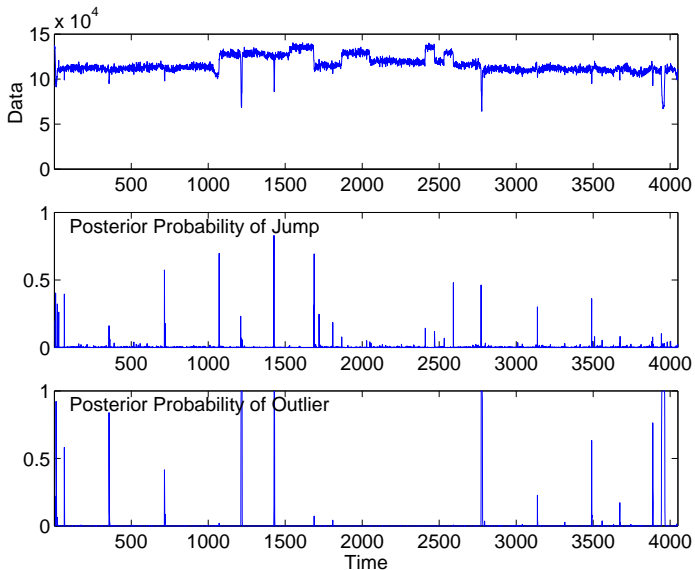
Outliers Modelling, Contd.

- The normal distribution does not fit the measurement noise well in the tails.
- We model the measurement noise as a mixture of two Gaussian distributions:

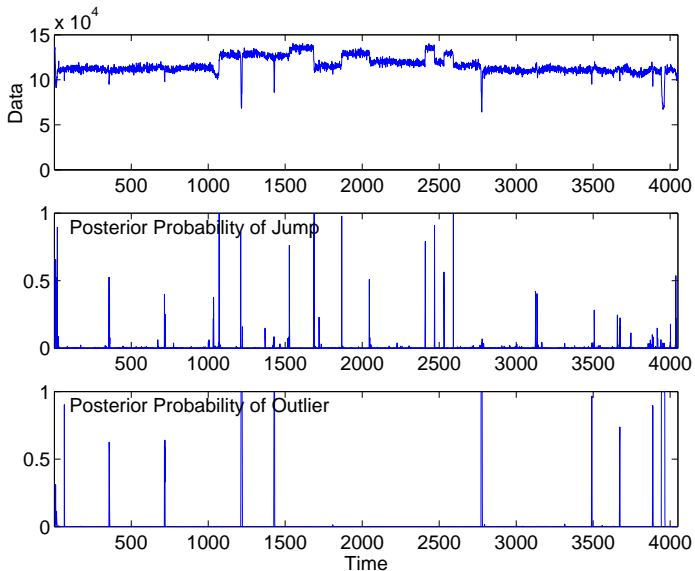
$$\begin{aligned} Z_{k+1} &= A(C_{k+1,1})Z_k + R(C_{k+1,1})U_k, & U_k &\sim \mathbf{N}(0, 1), \\ Y_k &= \mu(C_{k,2}) + B(C_{k,2})Z_k + S(C_{k,2})V_k, & V_k &\sim \mathbf{N}(0, 1), \end{aligned}$$

where $C_{k,1} \in \{0, 1\}$ and $C_{k,2} \in \{0, 1\}$ are indicators of the presence of a change point and of an outlier, respectively.

- $\{C_{k,2}\}$ is modelled as a two-state Markov chain which represents the clustering behavior of outliers.



On-line analysis of the well-log data, using 100 particles with detection delay $d = 0$. Top: data; middle: posterior probability of a jump; bottom: posterior probability of an outlier.



On-line analysis of the well-log data, using 100 particles with detection delay $d = 5$ (same display as above).

There are many more important issues in SMC such as

- **Analysis of Performance** Convergence of more elaborate algorithms, less restrictive conditions for long-term stability, analysis of smoothing algorithms
- **Resampling Variants** Conditional variance reduction schemes, triggered resampling, varying number of particles . . .
- **Choice of the Instrumental Moves** Lookahead moves, combination with deterministic approximation, adaptive moves

Some General References on SMC

- Doucet, A., De Freitas, N. and Gordon, N. (eds.) (2001) *Sequential Monte Carlo Methods in Practice*. Springer.
- Ristic, B., Arulampalam, M. and Gordon, A. (2004) *Beyond Kalman Filters: Particle Filters for Target Tracking*. Artech House.
- Cappé, O., Moulines, E. and Rydén, T. (2005) *Inference in Hidden Markov Models*. Springer.
- Doucet, A., Godsill, S. and Andrieu, C. (2000) On sequential Monte-Carlo sampling methods for Bayesian filtering. *Stat. Comput.*, **10**, 197-208.
- Arulampalam, M., Maskell, S., Gordon, N. and Clapp, T. (2002) A tutorial on particle filters for on line non-linear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.*, **50**, 241–254.
- Cappé, O., Godsill, S. J. and Moulines, E. (2007) An overview of existing methods and recent advances in sequential Monte Carlo, *IEEE Proc.*, **95**, 899–924.