# Lectures on state space models and particle filters

Hans R. Künsch
Seminar für Statistik, ETHZ

Luminy, September 2001.

**Contents:**



Thanks to Markus Hürzeler, Barry Schouten and Roland Brun



## 1. State Space and Hidden Markov Models

A **general state space model** consists of an unobserved state sequence $(X_t)$ and an observation sequence $(Y_t)$ with the following properties:

**State evolution:** $X_0, X_1, X_2, \ldots$ is a Markov chain with $X_0 \sim k_0(x)d\mu(x)$ and

$$X_t \mid X_{t-1} = x_{t-1} \;\sim\; k_t(x_{t-1}, x)d\mu(x)$$

**Generation of observations:** Conditionally on $(X_t)$, the $Y_t$'s are independent and $Y_t$ depends on $X_t$ only with

$$Y_t \mid X_t = x_t \;\sim\; q_t(x_t, y)d\nu(y).$$

If $X_t$ is discrete, it is usually called a **Hidden Markov Model**.



The dependence structure is represented by the following graph

$$\begin{array}{ccccccccc} \ldots & \rightarrow & X_{t-1} & \rightarrow & X_t & \rightarrow & X_{t+1} & \rightarrow & \ldots \\ & & \downarrow & & \downarrow & & \downarrow & & \\ \ldots & & Y_{t-1} & & Y_t & & Y_{t+1} & & \ldots \end{array}$$

From this, various conditional independence properties can be deduced. For instance

$$X_{t+1} \perp\!\!\!\perp (X_{1:t-1}, Y_{1:t}) \mid (X_t, Y_{t+1:T}).$$

and

$$X_t \perp\!\!\!\perp (X_{t+2:T}, Y_{t+1:T}) \mid (X_{t+1}, Y_{1:t}).$$

Extensions with additional dependencies can be handled either by enlarging the state space, or directly by modifying the formulae.



**History:**
Origin in the 1960's in engineering (Kalman-Bucy, Baum-Welch).
Recognized in time series analysis in 1980's (Akaike, Hannan, Harvey).
Large interest in 1990's (Molecular biology, Monte Carlo methods).

**General references:**
H.-R. K., Chapter 3 in Complex Stochastic Systems, Barndorff-Nielsen et al., eds., Chapman and Hall (2001).
Doucet, de Freitas and Gordon (eds.), Sequential Monte Carlo Methods in Practice, Springer (2001).

**Problems:**

States can be interpreted as latent variables or as time varying parameter. We prefer the former and distinguish between

1. Inference about the states based on a stretch of observed values $y_{s:t} = (y_u, s \le u \le t)$ for a given model (i.e. $k_t$ and $q_t$ known). Denote by $f_{s|t}$ the conditional density of $X_s$ given $y_{1:t}$ and call it **prediction**, **filtering** and **smoothing** density depending on whether $s < t$, $s = t$ or $s > t$. One can write down formulae for $f_{s|t}$ and also for all joint conditional distributions of interest, but exact evaluation is rarely possible.

2. Inference about unknown parameters in $k_t$, $q_t$. For this, one has to first solve the filtering and smoothing problem.



## Examples

The scope of general state space models becomes clear from the representation

$$X_t = F_t(X_{t-1}, U_t), \quad Y_t = H_t(X_t, V_t)$$

where $(U_t)$ and $(V_t)$ are two independent white noises.

**ARMA models:** A Gaussian ARMA($p, q$) process can be represented as a linear state space model:

$$X_t = F X_{t-1} + U_t, \quad Y_t = H X_t + V_t.$$

Heavy tailed distributions for $U_t$ or $V_t$ lead to different outlier models.

**Structural time series:** Superposition of unobservable trend $T_t$, seasonal component $S_t$ and noise $V_t$. Trend is modeled as random walk or integrated random walk, seasonal effect similarly. Heavy tails allow to model for instance level shifts.



**Stochastic volatility:** $(Y_t)$ are the log returns and $(X_t)$ are exogenous stochastic volatilities. Simplest model:

$$X_t = m + \phi X_{t-1} + U_t, \quad Y_t = \exp(X_t)\ V_t$$

where $(U_t)$ and $(V_t)$ are independent Gaussian white noises. By considering $\log|Y_t|$ we obtain a linear, non-Gaussian state space model, but this does not help much.

**Partially observed diffusion models:**

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t, \quad Y_k = h(X_{k\triangle}) + V_k$$

where $W$ is a Brownian motion and $(V_k)$ an independent white noise. Used for instance for interest rates, or to introduce a stochastic component in a deterministic model given by a differential equation.



Example: Oxygen balance model in a batch reactor (Kops et al., Brun). $X_t$ = biomass, $S_t$ = substrate, $O_t$ = oxygen.

Deterministic model:

$$\begin{aligned}
\frac{\mathrm{d}X_t}{\mathrm{d}t} &= \mu \frac{S_t}{K_S + S_t} X_t - b X_t \\
\frac{\mathrm{d}S_t}{\mathrm{d}t} &= -\frac{1}{Y}\mu \frac{S_t}{K_S + S_t} X_t \\
\frac{\mathrm{d}O_t}{\mathrm{d}t} &= K_{ex}(O^{sat} - O_t) - \frac{1-Y}{Y}\mu \frac{S_t}{K_S + S_t} X_t \\
&\quad -(1 - f_I) b X_t
\end{aligned}$$

The only variable which can be observed easily is $O_t$.

The model has 7 parameters ($\mu$, $K_S$, $b$, $Y$, $K_{ex}$, $O^{sat}$ and $f_I$) and 3 initial conditions. Assuming only measurement error, one can fit the model by least squares, but residuals typically show systematic behavior.

Model with randomly varying parameters:

Writing $\mu_t = \bar{\mu}$ + "white noise", $b_t = \bar{b}$ + "white noise" and interpreting "white noise" as the "derivative" of Brownian motion leads to a diffusion model for $\mathbf{Z}_t = (X_t, S_t, O_t)^T$ with drift $f(Z_t)$ as in the deterministic case and

$$\sigma(t, \mathbf{Z}_t) = \begin{bmatrix} \sigma_\mu \frac{S_t}{K_S + S_t} X_t & -\sigma_b X_t \\ -\frac{1}{Y} \sigma_\mu \frac{S_t}{K_S + S_t} X_t & 0 \\ -\frac{1-Y}{Y} \sigma_\mu \frac{S_t}{K_S + S_t} X_t & -(1 - f_I)\sigma_b X_t \end{bmatrix}$$

The highly irregular fluctuations somehow contradict biological intuition. Another possibility is to model $\mu_t$ and $b_t$ as Ornstein-Uhlenbeck processes.



**Biology:**
Ion channels are proteins located in the cell membrane that can be open or closed. The simplest model has two different means depending on whether the channel is open or closed plus an additive noise. A more realistic model allows for several internally different states of the channel, colored noise with state dependent variance and the effect of filters (de Gunst et al., 2001):

$$Y_t = \sum_{k=-r}^{r} \gamma_k \mu(X_{t-k}) + C_t + \sigma(X_t)\delta_t$$

Here $C_t$ is a Gaussian AR process, $\delta_t$ is white noise and $X_t$ is a continuous time Markov chain with the following structure:

$$C_1 \underset{q_2}{\overset{4q_1}{\rightleftharpoons}} C_2 \underset{2q_2}{\overset{3q_1}{\rightleftharpoons}} C_3 \underset{3q_2}{\overset{2q_1}{\rightleftharpoons}} C_4 \underset{4q_2}{\overset{q_1}{\rightleftharpoons}} O \underset{q_4}{\overset{q_3}{\rightleftharpoons}} C_5$$



DNA and protein sequences: One of the simplest examples are CG-islands. A homogeneous Markov chain for the sequence of letters from the DNA alphabet $\{A, C, G, T\}$ is not adequate since the frequencies of letters and transition changes. In $CG$ islands the pairs $CG$ occur more frequently. Two different transition regimes $P_0$ and $P_1$ can be obtained by letting the state take values in $\{A, C, G, T\} \times \{0, 1\}$ with transition matrix

$$\begin{pmatrix} (1-\varepsilon)P_0 & \varepsilon Q \\ \eta Q & (1-\eta)P_1 \end{pmatrix}.$$

$\varepsilon$ and $\eta$ are the probabilities for switching between the two regimes.



**Engineering:**

Many examples with tracking and control applications can be found in Doucet et al. (2001).

Usually, state = position and velocity of a moving object (satellite, robot etc.). Observations = partial, noisy information about the state, e.g. position only or angle only. State transitions usually depend also on a control variable $u_t$ which is a function of $y_{1:t-1}$.

One of the earliest uses of hidden Markov models is in speech analysis. In isolated word recognition, one has a hidden Markov model for each word, the states being the different phonems or different stages of the phonems. Transition probabilities satisfy $k(i, j) = 0$ for $j < i$. Observations = features extracted from the acoustic signal over short overlapping intervals. One then computes the likelihood $p(y_{1:t})$ for each word and applies Bayes rule.

**Geophysics:**

Observations = daily rainfall at a network of stations; states = weather types. Given the state, the different stations are independent.

Data assimilation: States = atmospheric variables on a grid used for numerical integration of partial differential equations. State transitions are determinstic, filtering is used to combine forecasts with actual observations to obtain initial conditions for next integration period.



# 2. Filtering and smoothing recursions

By Bayes formula:

$$p(x_{0:t} \mid y_{1:t}) = \frac{p(y_t \mid x_{0:t}, y_{1:t-1}) p(x_{0:t} \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})}$$

and by the definition of conditional densities

$$p(x_{0:t} \mid y_{1:t-1}) = p(x_t \mid x_{0:t-1}, y_{1:t-1}) p(x_{0:t-1} \mid y_{1:t-1})$$

Thus by conditional independence built into the model:

$$\begin{aligned} & p(x_{0:t} \mid y_{1:t}) \\ & \quad = \frac{q_t(x_t, y_t) k_t(x_{t-1}, x_t)}{p(y_t \mid y_{1:t-1})} p(x_{0:t-1} \mid y_{1:t-1}). \end{aligned}$$

This is a recursion with increasing dimensions.



For recursions with fixed dimensions, we can combine a recursion for the filter density $f_{t|t}$, that is the density of $X_t$ given $y_{1:t}$ with an expression for the backward transitions
$p(x_s \mid x_{s+1:t}, y_{1:t})$.

The same arguments as before give the two-step filter recursion:

**Propagation** From filter to prediction density

$$f_{t|t-1}(x_t \mid y_{1:t-1}) = \int f_{t-1|t-1}(x \mid y_{1:t-1}) k_t(x, x_t) d\mu(x$$

**Update** From prediction to filter density

$$\begin{aligned} f_{t|t}(x_t \mid y_{1:t}) &= \frac{f_{t|t-1}(x_t \mid y_{1:t-1}) q_t(x_t, y_t)}{p(y_t \mid y_{1:t-1})} \\ &\propto f_{t|t-1}(x_t \mid y_{1:t-1}) q_t(x_t, y_t) \end{aligned}$$



**Smoothing**

Conditional on $y_{1:t}$, $(X_0, X_1, \ldots, X_t)$ is an inhomogeneous Markov chain with backward transition densities

$$\begin{aligned} p(x_s \mid x_{s+1}, y_{1:t}) &= p(x_s \mid x_{s+1}, y_{1:s}) \\ &\propto k_{s+1}(x_s, x_{s+1}) f_{s|s}(x_s \mid y_{1:s}). \end{aligned}$$

Thus the conditional distribution of the states given the observations is:

$$p(x_{0:t} \mid y_{1:t}) = f_{t|t}(x_t \mid y_{1:t}) \prod_{s=0}^{t-1} p(x_s \mid x_{s+1}, y_{1:s}).$$

In order to compute the marginal densities $f_{s|t}$ of $X_s$ given $y_{1:t}$, we combine the forward filter recursion with the backward recursion

$$\begin{aligned} & f_{s|t}(x_s \mid y_{1:t}) \\ & \quad = \int p(x_s \mid x_{s+1}, y_{1:t}) f_{s+1|t}(x_{s+1} \mid y_{1:t}) d\mu(x_{s+1}). \end{aligned}$$

For **simulation** according to $p(x_{0:t} \mid y_{1:t})$, use the forward recursion for the filter densities and the backward recursion for the transition densities and the actual simulation. This depends of course on the ability to actually compute the recursions.

Alternatively one can use the Gibbs sampler (or another MCMC algorithm) where one needs the full conditionals of $p(x_{0:t} \mid y_{1:t})$. These are easily obtained

$$\begin{aligned} & p(x_s \mid x_{0:s-1}, x_{s+1:t}, y_{1:t}) \\ & \quad = \; p(x_s \mid x_{s-1}, x_{s+1}, y_s) \\ & \quad \propto \; k_s(x_{s-1}, x_s) k_{s+1}(x_s, x_{s+1}) q_s(x_s, y_s). \end{aligned}$$



## Prediction of observations and likelihood

Because the denominator in the update step of the filter recursion is just a normalization:

$$p(y_t \mid y_{1:t-1}) = \int f_{t|t-1}(x \mid y_{1:t-1}) q_t(x, y_t) d\mu(x).$$

To obtain the likelihood of observed data, use

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t \mid y_{1:t-1}).$$

A different representation of the likelihood is

$$p(y_{1:T}) = \int k_0(x_0) \prod_{t=1}^{T} k_t(x_{t-1}, x_t) q_t(x_t, y_t) \prod_{t=0}^{T} d\mu(x_t).$$



## Viterbi algorithm

This is an dynamic programming method to compute

$$\arg\max_{x_0^T} p(x_0^T \mid y_1^T),$$

the most likely state sequence. Like smoothing, it involves a forward followed by a backward recursion.



## Recursive filtering in operator notation

A compact notation for the filter recursion is

$$f_{t|t}(. \mid y_{1:t}) = B(K_t^* f_{t-1|t-1}(. \mid y_{1:t-1}), q_t(., y_t)).$$

Here

$$K_t^* f(x) = \int f(x') k_t(x', x) d\mu(x')$$

is the Markov transition operator that assigns to a density $f$ the density one time step later, and

$$B(f, q)(x) = \frac{f(x) q(x)}{\int f(x) q(x) d\mu(x)}$$

is the Bayes operator that assigns to a prior $f$ and a likelihood $q$ the posterior.

$K_t^*$ and $B(., q)$ map the space of densities into itself. They can be extended to operate on the space of probability distributions.

Usually, the observations $y_{1:t}$ are fixed and will be dropped from the notation.

### Implementation of recursions

If $X_t$ is discrete with $M$ possible values, integrals are sums and the recursions need $O(TM^2)$ operations.

In a linear Gaussian state space model, all $f_{t|s}$ are Gaussian, and their means and variances are computed with the Kalman filter and smoother.

In practically all other cases, the recursions are difficult to compute. Analytical approximations like the extended Kalman filter are not satisfactory, and numerical integration is problematic in high dimensions.

Much current interest focuses on Monte Carlo methods. Standard Markov chain Monte Carlo can be used, but they lack a recursive implementation. Convergence can be slow unless one can perform multiple updates by conditioning on some components of the state vector.



## 3. Recursive Monte Carlo filter

Key observation: The Markov transition $K_t^* f$ is difficult to compute, but easy to sample from if we can sample from $f$ and $k_t(x, .)$.

Generate recursively a sequence of samples ("particles") $(x_{j,t}; j = 1, \dots N, t = 0, 1, \dots)$ with approximate distribution $f_{t|t}$:

If $(x_{j,t-1})$ is available, we can replace

$$K_t^* f = \int f_{t-1|t-1}(x \mid y_{1:t-1}) k_t(x, x_t) d\mu(x)$$

by

$$\frac{1}{N} \sum_{j=1}^{N} k_t(x_{j,t-1}, x).$$

Therefore we sample $(x_{j,t})$ from

$$f_{t|t}^N(x) \propto q_t(x, y_t) \frac{1}{N} \sum_{j=1}^{N} k_t(x_{j,t-1}, x).$$



**Simple algorithms** to sample from a density

$$f^N(x) \propto q(x) \sum_{j=1}^{N} k_j(x).$$

$\sum k_j(x)/N$ is the prior (equal to the approximate predicion in the filtering context), $q(x)$ is the likelihood.

**Rejection sampling** with proposal $\rho$ and acceptance probability

$$\frac{q(x) \sum_{j=1}^N k_j(x)}{\rho(x)} \left( \sup_x \frac{q(x) \sum_{j=1}^N k_j(x)}{\rho(x)} \right)^{-1}.$$

**Sampling importance resampling.** Generate $(z_k; 1 \le k \le R)$ according to some proposal $\rho$ and select from these a sample of size $N$ with inclusion probabilities

$$\pi(z_k) \propto \frac{q(z_k) \sum_{j=1}^N k_j(z_k)}{\rho(z_k)}.$$



The standard proposal for both methods is the prior, making the evaluation of acceptance/inclusion probabilities simple.

Without the resampling step, the weights concentrate quickly on very few values.

The resampling step produces ties, but these separate at the next proposal step if the state transition and the proposal are absolutely continuous. $R > N$ should reduce the number of ties.

Rejection sampling produces no ties if the state transitions are absolutely continuous.

Both methods have difficulties if the likelihood is less spread out than the prior or if prior and likelihood are in conflict.

**Monte Carlo smoothing:**

There is a similar recursive simulation method that generates samples $(x_{j,s|t})$ from the conditional distribution of $X_{0:t}$ given $Y_{1:t} = y_{1:t}$:

$$\begin{aligned} x_{j,t|t} &= x_{j,t} \\ x_{j,s|t} &\sim const.k_{s+1}(x, x_{j,s+1|t}) q_s(x, y_s) \cdot \\ & \frac{1}{N} \sum_{i=1}^{N} k_s(x_{i,s-1}, x) d\mu(x). \end{aligned}$$

Again we can use rejection sampling.

If $t$ increases by one, we need in principle to do a full backward iteration $s = t, t-1, \dots 0$. Usually, stopping at $s = t - m$ with a fixed $m$ is sufficient.



Alternatively we can try to use directly the recursion for $p(x_{0:t} \mid y_{1:t})$.

Sampling importance resampling implementation:

- For $1 \le j \le N$, set $z_{j,0:t} = x_{j,0:t|t}$ and propose $z_{j,t+1} \sim k_t(x_{j,t|t}, x) d\mu(x)$.

- Compute inclusion probabilities

$$\pi_j \propto q_{t+1}(z_{j,t+1}).$$

- Select from the $(z_{j,0:t+1})$ a sample of size $N$ with the above inclusion probabilities.



Rejection sampling implementation:

- Choose $I$ from $\{1, \dots N\}$, set $z_{j,0:t} = x_{I,0:t|t}$ and propose $z_{j,t+1} \sim k_t(x_{I,t|t}, x) d\mu(x)$.

- Accept $z_{j,0:t+1}$ with probability

$$q_{t+1}(z_{j,t+1}) / \sup q_{t+1}(x).$$

  If accepted, increase $j$ by 1.

- Repeat until $j = N$.

Note that for $s \le t$ $\{x_{j,s|t+1}\}$ is a subset of $\{x_{j,s|t}\}$ which can be problematic.



**Modifications of simple rejection sampling**

Reducing randomness: Instead of selecting the index in the mixture $\sum k_j(x)/N$ at random, choose each $j$ once to generate a proposal and make individual decisions about acceptance. Repeat the procedure until the total of accepted values exceeds $N$ for the first time.

Better proposals: In general difficult to implement. Made easier by proposing both an index $J$ and a variable $X$ according to $\tau_j \rho_j(x)$ and accepting $X$ with probability

$$\frac{k_j(x) q(x)}{\tau_j \rho_j(x)} \left( \sup_{j,x} \frac{k_j(x) q(x)}{\tau_j \rho_j(x)} \right)^{-1}.$$

(Pitt and Shephard, 1999).

Maximizing the acceptance rate gives

$$\tau_j \propto \sup_x \frac{k_j(x) q(x)}{\rho_j(x)}.$$

and $\rho_j(x)$ should be a good proposal for $k_j(x) q(x)$.

**Modification of sampling importance resampling**

Reducing randomness: Select each component in the mixture once for the proposal. Do the resampling step not at random, but in a more systematic way: Circular sampling or tree-based sampling.

Better proposals: Again propose both an index $J$ and a value $Z$ with distribution $\tau_j \rho_j(z)$. In particular, one can choose $\rho_j(x) = k_j(x)$ and unequal $\tau_j$'s, choosing those indices more frequently which tend to give higher values for $q(z_k)$.



# 4. Parameter estimation

Assume $k_t(x_{t-1}, x_t)$ depends on a parameter $\tau$ and $q_t(x_t, y_t)$ depends on a parameter $\eta$. Together $\theta = (\tau, \eta)$.

Bayesian methods include $\theta$ in the state vector $X_t$. However, since $\theta$ does not change over time, the particle filter may suffer from too many ties (sample depletion). One can add some noise, but it is not clear how large the noise should be and how we should transform the parameters before adding noise.

We discuss here approximations of the likelihood function and/or the MLE. Remember the two formulae for the likelihood given earlier when discussing filter and smoothing recursions.



By the definition of the likelihood

$$L(\theta \mid y_{1:T}) = \prod_{t=1}^{T} \mathbf{E}_\theta[q_{t,\eta}(X_t, y_t) \mid y_{1:t-1}].$$

Hence log likelihood is approximately equal to

$$const. + \sum_{t=1}^{T} \log \left( \sum_{j=1}^{N} q_{t,\eta}(x_{j,t|t-1}, y_t) \right)$$

where $x_{j,t|t-1} \sim k_{t,\tau}(x_{j,t-1}, x) d\mu(x)$.

However, this approximation is only pointwise for one value of $\theta$. If one wants the likelihood at a different value of $\theta$, a new particle filter has to be generated. One should maybe indicate in the notation that the samples $(x_{j,t})$ depend also on $\theta$.



In order to obtain an approximation of log likelihood in a whole neighborhood of $\theta_0$ based on a particle filter with parameter $\theta_0$, one can try to use importance sampling.

One can write down a recursion for the importance weights

$$\frac{f_{t|t-1,\theta}(x_t \mid y_{1:t})}{f_{t|t-1,\theta_0}(x_t \mid y_{1:t})}$$

and approximate all expectations with averages from the filter samples.

However, the final recursions need $O(TN^2)$ operations for evaluation. This severely limits the usefulness of this approach.

Approximations based on smoother samples are more promising. The key is the following formula:

$$\log L(\theta) - \log L(\theta_0) = \log \left( \frac{p_\theta(y_{1:T})}{p_{\theta_0}(y_{1:T})} \right)$$
$$= \log \left( \mathbf{E}_{\theta_0} \left[ \frac{p_\theta(X_{0:T}, y_{1:T})}{p_{\theta_0}(X_{0:T}, y_{1:T})} \mid y_{1:T} \right] \right).$$

Hence this is approximately equal to

$$\log \left( \frac{1}{N} \sum_{j=1}^{N} \frac{k_{0,\tau}(x_{j,0|T})}{k_{0,\tau_0}(x_{j,0|T})} \cdot \prod_{t=1}^{T} \frac{k_{t,\tau}(x_{j,t-1|T}, x_{j,t|T}) q_{t,\eta}(x_{j,t|T}, y_t)}{k_{t,\tau_0}(x_{j,t-1|T}, x_{j,t|T}) q_{t,\eta_0}(x_{j,t|T}, y_t)} \right).$$

Need to update $\theta_0$ as soon as ratios above differ too much from one.



**Stochastic EM-algorithm**

In the E-step we compute

$$\begin{aligned} Q(\theta|\theta_0) &= \mathbf{E}_{\theta_0}[\log p_\theta(X_{0:T}, y_{1:T}) \mid y_{1:T}] \\ &= \mathbf{E}_{\theta_0}[\log k_{0,\tau}(X_0) \mid y_{1:T}] \\ &+ \sum_{t=1}^{T} \mathbf{E}_{\theta_0}[\log k_{t,\tau}(X_{t-1}, X_t) \mid y_{1:T}] \\ &+ \sum_{t=1}^{T} \mathbf{E}_{\theta_0}[\log q_{t,\eta}(X_t, y_t) \mid y_{1:T}]. \end{aligned}$$

We can approximate the expectations by arithmetic means with the smoother samples.

The M-step maximizes $Q(\theta|\theta_0)$ with respect to $\theta$ in order to obtain the next $\theta_0$.



# 5. Convergence Analysis

**Operator notation for the particle filter:**

Denote by $S_N(f)$ the empirical distribution of a sample of size $N$ from $f$. Then the approximate filter density is

$$f_{t|t}^N = B(K_t^* S_N(f_{t-1|t-1}^N), q_t(., y_t)).$$

(Note that $K_t^*$ converts a discrete distribution to one haveing a density). It has to be compared with

$$f_{t|t} = B(K_t^* f_{t-1|t-1}, q_t(., y_t)).$$



**Straightforward approach**

We can write

$$\begin{aligned} & f_{t|t}^N - f_{t|t} \\ &= B(K_t^* S_N(f_{t-1|t-1}^N), q_t) - B(K_t^* f_{t-1|t-1}^N, q_t) \\ &+ B(K_t^* f_{t-1|t-1}^N, q_t) - B(K_t^* f_{t-1|t-1}, q_t). \end{aligned}$$

The first term is the error due to sampling at time $t-1$ (propagated once) and the second term is the propagation of the error at time $t-1$.

We study Lipschitz-continuity of Bayes and Markov operators with respect to the $L_1$-distance between densities:

$$\begin{aligned} & ||f - g||_1 = \int |f(x) - g(x)| d\mu(x) \\ &= 2 \int (f(x) - g(x))^+ d\mu(x) = 2 \sup_B |P_f[B] - P_g[B]|. \end{aligned}$$

Finally, we control the sampling error.

**Continuity of Markov transitions:**

The following result goes back to at least Dobrushin (1956):

$$||K^*f - K^*g||_1 \leq \rho(K^*)\ ||f-g||_1$$

where

$$\rho(K^*) = \frac{1}{2}\sup_{x,x'} ||k(x,\cdot) - k(x',\cdot)||_1 \leq 1.$$

Moreover, for compact state space typically $\rho(K^*) < 1$. i.e. the Markov operator is contracting.



**Continuity of Bayes' formula w.r. to prior:** We have

$$||B(f,q) - B(g,q)||_1 \leq \beta(f,q)||f-g||_1$$

where

$$\beta(f,q) = \frac{\sup_x q(x)}{\int q(x)f(x)d\mu(x)} \in \left(1, \frac{\sup_x q(x)}{\inf_x q(x)}\right].$$

This bound cannot be improved in general, i.e. the Bayes operator is not contracting: To any prior there is a direction where the Bayes operator is expanding.

**Lemma:** If $\mu$ is Lebesgue measure on an interval and $f$ and $q$ are continuous

$$\sup_{g:||f-g||_1\leq\delta} \frac{||B(f,q)-B(g,q)||_1}{||f-g||_1} \stackrel{\delta\to 0}{\longrightarrow} \beta(f,q).$$



**Idea of proof:**

W.l.o.g. assume $\int q(x)f(x)dx = 1$. Let $x_1 = \arg\sup q(x)$ and $x_2$ with $q(x_2) < q(x_1)$. Then construct $g$ by taking away the mass $\delta^2$ from $f$ in $(x_1-\delta, x_1+\delta)$ and adding the same mass in $(x_2-\delta, x_2+\delta)$. Then by direct computation the upper bound is attained as $\delta$ goes to zero.



**Bound on sampling error:**

**Lemma:** If $x \to k(x,.)$ is continuous with respect to the $L_1$-norm, then under i.i.d. sampling from $g$

$$\mathbb{P}[||K^*S_N(g) - K^*g||_1 > \varepsilon] \stackrel{N\to\infty}{\longrightarrow} 0$$

exponentially fast in $N$ for any $\varepsilon > 0$, uniformly for all $g$ with $\int_C g(x)d\mu(x) > 1 - \varepsilon/5$ for some fixed compact set $C$.

This is essentially the same proof as the one for consistency of kernel density estimators in the $L_1$-norm.

**Theorem:** If $x \to k_t(x,.)$ is continuous and for all $t$, all $y_t$ $0 < q_t(x,y_t) \leq C(t,y_t) < \infty$, then for all $t$ and all $y_1^t$

$$||f^N_{t|t} - f_{t|t}||_1 \longrightarrow 0$$

in probability as $N \to \infty$.

However, the required sample size $N$ grows exponentially with $t$, and depends also on $y_1^t$. Assuming a compact state space does not help.

Note that the problem comes from the unexpected behavior of the Bayes formula which does not exclude the possibility that errors propagate. For a better result we need in particular an argument that shows that the (exact) filter density forgets the starting value.



**Better argument:**

Look at error propagation over more than one time step:

$$
\begin{aligned}
& f_{t|t}^N - f_{t|t} \\
& \quad = \; B(K_t^* S_N(f_{t-1|t-1}^N), q_t) \\
& \quad - \; B(K_t^* f_{t-1|t-1}^N, q_t) \\
& \quad + \; \ldots \\
& \quad - \; \ldots \\
& \quad + \; B(K_t^* \cdots B(K_{s+1}^* S_N(f_{s|s}^N), q_{s+1}) \ldots, q_t) \\
& \quad - \; B(K_t^* \cdots B(K_{s+1}^* f_{s|s}^N, q_{s+1}) \ldots, q_t) \\
& \quad + \; B(K_t^* \cdots B(K_{s+1}^* f_{s|s}^N, q_{s+1}) \ldots, q_t) \\
& \quad - \; B(K_t^* \cdots B(K_{s+1}^* f_{s|s}, q_{s+1}) \ldots, q_t).
\end{aligned}
$$

Last difference = Error at time $s$, propagated over $t-s$ steps.

Other differences = Errors due to sampling at time $s \leq u < t$, propagated over $t-u$ steps.



For a bound on the error propagated over $k$ steps that is better than the sum of errors over $k$ single steps, observe

$$
f_{t|t}(x_t) = \int f_{s|t}(x_s) \cdot \prod_{u=s+1}^{t} p(x_u \mid x_{u-1}, y_{u:t}) d\mu(x_{u-1}).
$$

Moreover,

$$
f_{s|t} = B(f_{s|s}, p(y_{s+1:t}|x_s)).
$$

Therefore the transition from $f_{s|s}$ to $f_{t|t}$ is achieved by applying one Bayes operator, followed by $t-s$ Markov operators.

Hence we need a bound for the contraction coefficient of the Markov operator with transition kernel $p(x_u \mid x_{u-1}, y_{u:t}) =$

$$
\frac{k_u(x_{u-1}, x_u) q_u(x_u, y_u) p(y_{u+1:t} \mid x_u)}{p(y_{u:t} \mid x_{u-1})}.
$$



**Lemma:** If

$$
C_K := \sup_{t,x,x',x''} \frac{k_t(x, x'')}{k_t(x', x'')} < \infty,
$$

then $\rho(K_t) \leq 1 - 1/C_K < 1$ and

$$
\begin{aligned}
& ||B(K_t^* \cdots B(K_{s+1}^* f_{s|s}^N, q_{s+1}), \ldots q_t) \\
& \quad - \; B(K_t^* \cdots B(K_{s+1}^* f_{s|s}, q_{s+1}), \ldots q_t)||_1 \\
& \quad \leq \; C_K(1 - 1/C_K^2)^{t-s}.
\end{aligned}
$$

(No conditions on $q$ are needed).

This lemma shows also that under the above condition the filter forgets the initial condition $f_{0|0}$ exponentially fast.

**Proof:** First note that $f(x)/g(x) \leq c$ implies

$$
\begin{aligned}
||f - g||_1 &= 2 \int (f(x) - g(x))^+ d\mu(x) \\
&\leq 2(1 - 1/c) \int f(x) d\mu(x).
\end{aligned}
$$

Hence in order to bound the contraction coefficient for the transition $X_{u-1} \to X_u$ given $y_{u:t}$, we have to bound

$$\frac{p(x_u \mid x_{u-1}, y_{u:t})}{p(x_u \mid x'_{u-1}, y_{u:t})} = \frac{k_u(x_{u-1}, x_u)}{k_u(x'_{u-1}, x_u)} \frac{p(y_{u:t} \mid x_{u-1})}{p(y_{u:t} \mid x'_{u-1})}.$$

The first ratio on the right is at most $C_K$ by assumption, and for the second ratio, we use

$$p(y_{u:t} \mid x_{u-1}) = \int k_u(x_{u-1}, u_s) q_u(x_u, y_u) p(y_{u+1:t} \mid x_u) d\mu(x_u).$$

From this it follows that also the second ratio on the right is bounded by $C_K$. Moreover, the same ratio measures the maximal expansion in the Bayes step going from $f_{s|s}$ to $f_{s|t}$.



Errors due to sampling can be handled as before, leading to

**Theorem** If in addition to the condition on $k_t$ also

$$C_q := \sup_{t,x,x',y} \frac{q_t(x,y)}{q_t(x',y)} < \infty, \tag{1}$$

then to any $\varepsilon > 0$ there are constants $c_1$ and $c_2$ such that for all $t$ and all $N$

$$P[||f^N_{t|t} - f_{t|t}||_1 > \varepsilon] \le c_1 \exp(-c_2 N).$$



It would be nice to relax the condition on $k_t$ since it is even stronger than uniform ergodicity. For instance, if there exists a density $\phi$ and $\varepsilon > 0$ such that

$$\varepsilon\phi(x') \le k_t(x, x')$$

for all $t$, $x$, $x'$, then also $\rho(K_t) \le 1 - \varepsilon$. But with this condition, the above methods seem not to work.

One might try a different metric. Atar and Zeitouni (1997) and Le Gland and Oudjane (2001) use

$$d_H(f,g) = \log\left(\sup_x \frac{f(x)}{g(x)} \sup_x \frac{g(x)}{f(x)}\right),$$

called the Hilbert metric. (It maybe equal to $+\infty$).



Clearly, with respect to $d_H$ any Bayes and any Markov operator are Lipschitz with Lipschitz constant 1. A deeper result due to Birkhoff says:

If $K$ is a Markov kernel with

$$H(K) = \sup_{f,g} d_H(K^* f, K^* g) < \infty,$$

then $K^*$ is in fact Lipschitz with Lipschitz constant $\tanh(H(K)/4) < 1$.

Unfortunately, in order to verify $H(K) < \infty$, one needs essentially the same condition as before.

## Central limit theorems

Del Moral and Guionnet, Ann. Appl. Probab. (1999), have shown the following.

Under suitable conditions, for any fixed $t$, any $y_{1:t}$ and square integrable functions $\phi_s$ (w.r. to $f_{s|s}$):

$$\sqrt{N}\left(\frac{1}{N}\sum_j \phi_s(x_{j,s}) - \int \phi_s(x) f_{s|s}(x) d\mu(x)\right)_{0\le s\le t}$$

is asymptotically centered normal.

The proof is a nice application of contiguity: Denote by $P_N$ the distribution where all $x_{j,s}$ are independent and have density $f_{s|s}$, and by $Q_N$ the actual distribution of the particle filter samples.



Asymptotic normality under $P_N$ is obvious and the theorem is proved by studying the asymptotics of the log likelihood ratio $\log \frac{dQ_N}{dP_N}$ under $P_N$. The key part in contiguity is to show that it converges in distribution to a random variable $V$ with $\mathbf{E}[\exp(V)] = 1$.

In contrast to standard examples in statistics, the limit $V$ is not normal with mean equal to minus a half the variance. Rather the log likelihood ratio is asymptotically equivalent to a sum of degenerate $U$-statistics of order 2. The limit is therefore a linear combination of centered chisquare variables.



Can we use a more direct approach ?

When conditioning on the previous filter sample $(x_{j,t-1})$, we expect a CLT to hold (Lindeberg's condition should take care of the problem that the distribution of particles changes as $N$ changes). There is however a bias because we have to center the variables at the expectation w.r. to $f_{t|t}^N$, not $f_{t|t}$. But the expectation w.r. to $f_{t|t}^N$ is a ratio of the averages of two functions of $(x_{j,t-1})$, so it should be possible to use induction.

In order to make this argument rigorous, set

$$\begin{aligned} S_{N,t}(\phi) &= \frac{1}{N}\sum_{j=}^{N} \phi(x_{j,t}), \\ m_{N,t}(\phi) &= \int \phi(x) f_{t|t}^N(x) d\mu(x), \\ m_t(\phi) &= \int \phi(x) f_{t|t}(x) d\mu(x). \end{aligned}$$



Then

$$\begin{aligned} &\mathbf{E}[\exp(i\lambda\sqrt{N}(S_{N,t}(\phi) - m_t(\phi)))] = \\ &\quad \mathbf{E}[\mathbf{E}[\exp(i\lambda\sqrt{N}(S_{N,t}(\phi) - m_{N,t}(\phi))) \mid (x_{j,t-1})] \cdot \\ &\quad \exp(i\lambda\sqrt{N}(m_{N,t}(\phi) - m_t(\phi)))]. \end{aligned}$$

The inner expectation will converge to

$$\exp(-\sigma_t(\phi)^2\lambda^2/2)$$

for all $(x_{j,t-1})$ such that $||f_{t|t}^N - f_{t|t}||_1$ goes to zero. The outer expectation will converge to a normal characteristic function by the induction assumption.

Still need to check the details !