PARTICLE FILTERS FOR PARTIALLY OBSERVED MARKOV CHAINS

Natacha Caylus, Arnaud Guyader, François LeGland

IRISA / Université de Rennes 1, Université de Rennes 2 and INRIA
Campus de Beaulieu
35042 RENNES Cédex, France

ABSTRACT

We consider particle filters in a model where the hidden states and the observations form jointly a Markov chain, which means that the hidden states alone do not necessarily form a Markov chain. This model includes as a special case non–linear state–space models with correlated Gaussian noise. Our contribution is to study propagation of errors, stability properties of the filter, and uniform error estimates, using the framework of LeGland and Oudjane [5].

1. EXTENSIONS OF HIDDEN MARKOV MODELS

In the classical HMM situation, the hidden state sequence $\{X_k\,,\,k \geq 0\}$ is a Markov chain taking values in the space E. It is not observed, but instead an observation sequence $\{Y_k\,,\,k \geq 0\}$ taking values in the space F is available, with the property that given the hidden states $\{X_k\,,\,k \geq 0\}$, the observations $\{Y_k\,,\,k \geq 0\}$ are mutually independent, and the conditional probability distribution of Y_k depends only on the hidden state X_k at the same time instant. In addition, when $x \in E$ varies, all the conditional probability distributions $\mathbb{P}[Y_k \in dy \mid X_k = x]$ are assumed absolutely continuous w.r.t. a nonnegative measure $\lambda_k^F(dy)$ on F which does not depend on x. The situation is completely described by the initial distribution and local characteristics

$$\mathbb{P}[X_0 \in dx] = \mu_0(dx) \;,$$

$$\mathbb{P}[X_k \in dx' \mid X_{k-1} = x] = Q_k(x, dx') \;,$$

$$\mathbb{P}[Y_k \in dy \mid X_k = x] = g_k(x,y)\, \lambda_k^F(dy) \;.$$

1.1. Conditionally Markovian observations

Alternatively, the following more general assumption could be made : given the hidden states $\{X_k\,,\,k \geq 0\}$, the observations $\{Y_k\,,\,k \geq 0\}$ form a Markov chain, and the conditional probability distribution of Y_k given Y_{k-1} depends only on the hidden state X_k at the same time instant. The situation is completely described by the joint initial distribution and local characteristics

$$\mathbb{P}[X_0 \in dx, Y_0 \in dy] = \mu_0(dx)\, g_0(x,y)\, \lambda_0^F(dy) \;,$$

$$\mathbb{P}[X_k \in dx' \mid X_{k-1} = x] = Q_k(x, dx') \;,$$

$$\mathbb{P}[Y_k \in dy' \mid Y_{k-1} = y, X_k = x'] = g_k(x', y, y')\, \lambda_k^F(y, dy') \;.$$

Particle filters for these models, which include switching autoregressive models, have already been investigated in Cappé [1] and in Del Moral and Jacod [2].

1.2. Jointly Markovian hidden states and observations

Even more generally, the following assumption could be made that hidden states $\{X_k\,,\,k \geq 0\}$ and observations $\{Y_k\,,\,k \geq 0\}$ form jointly a Markov chain, and that the transition kernel can be factorized as

$$\begin{aligned} &\mathbb{P}[X_k \in dx', Y_k \in dy' \mid X_{k-1} = x, Y_{k-1} = y] \\ &\qquad = R_k(x, y, y', dx')\; \lambda_k^F(y, dy') \;, \end{aligned} \tag{1}$$

where $R_k(x, y, y', dx')$ is a nonnegative measure on E for any $x \in E$ and any $y, y' \in F$, and where $\lambda_k^F(y, dy')$ is a nonnegative measure on F for any $y \in F$. Particle filters for these models have already been investigated in Crişan and Doucet [3], where even the joint Markov property is removed, and in Desbouvries and Pieczynski [4]. Notice that when $x \in E$ and $y \in F$ vary, all the conditional probability distributions $\mathbb{P}[Y_k \in dy' \mid X_{k-1} = x, Y_{k-1} = y]$ are absolutely continuous w.r.t. a nonnegative measure $\lambda_k^F(y, dy')$ on F which does not depend on x. Indeed, integrating (1) w.r.t. $x' \in E$ yields

$$\begin{aligned} &\mathbb{P}[Y_k \in dy' \mid X_{k-1} = x, Y_{k-1} = y] \\ &\qquad = R_k(x, y, y', E)\; \lambda_k^F(y, dy') \;. \end{aligned} \tag{2}$$

This work was partially supported by CNRS, under the *MathSTIC* project *Chaînes de Markov Cachées et Filtrage Particulaire*, and under the *AS–STIC* project *Méthodes Particulaires* (AS 67)

Not only is the decomposition (2) necessary, but it is also a sufficient condition for the decomposition (1) to hold. Indeed, if the decomposition (2) holds, then the decomposition (1) holds with

$$R_k(x,y,y',dx') = \widehat{g}_k(x,y,y')\ \widehat{Q}_k(x,y,y',dx')\ , \quad (3)$$

where by definition

$$\widehat{Q}_k(x,y,y',dx') = \frac{R_k(x,y,y',dx')}{R_k(x,y,y',E)}$$

$$= \mathbb{P}[X_k \in dx' \mid X_{k-1}=x, Y_{k-1}=y, Y_k=y']\ ,$$

and

$$\widehat{g}_k(x,y,y') = R_k(x,y,y',E)\ ,$$

for any $x \in E$ and any $y,y' \in F$. In full generality, for any $x \in E$ and any $y,y' \in F$, the nonnegative measure $R_k(x,y,y',dx')$ can be factorized as

$$R_k(x,y,y',dx') = W_k(x,y,y',x')\ P_k(x,y,y',dx')\ , \quad (4)$$

into the product of a nonnegative importance weight function $W_k(x,y,y',x')$, and an importance probability distribution $P_k(x,y,y',dx')$. The decomposition (4) is clearly not unique. As much as possible, a clear distinction should be made between results and estimates

- which depend only on the nonnegative kernel R_k,
- which depend on the *specific* importance decomposition (W_k,P_k) of the nonnegative kernel R_k.

In practice, the importance decomposition should be such that, for any $x \in E$ and any $y,y' \in F$, it is *easy*

- to *evaluate* the weight function $W_k(x,y,y',x')$,
- to *simulate* a r.v. X according to the probability distribution $P_k(x,y,y',dx')$,

Another meaningful criterion for the choice of the importance decomposition is the optimization of error estimates for associated particle schemes, see Remark 4.3 below.

2. OPTIMAL BAYESIAN FILTER AND FEYNMAN–KAC FORMULAS

For any test function f defined on E^{n+1}

$$\mathbb{E}[f(X_{0:n}) \mid Y_{0:n}]$$

$$= \frac{\displaystyle\int_E \cdots \int_E f(x_{0:n})\ R_0(dx_0)\ \prod_{k=1}^n R_k(x_{k-1},dx_k)}{\displaystyle\int_E \cdots \int_E R_0(dx_0)\ \prod_{k=1}^n R_k(x_{k-1},dx_k)}\ ,$$

where by definition and with an abuse of notation

$$\begin{aligned} R_0(dx) &= R_0(Y_0,dx)\ , \\ R_k(x,dx') &= R_k(x,Y_{k-1},Y_k,dx')\ . \end{aligned} \quad (5)$$

Given the observations, the objective of filtering is to estimate the hidden states, and to this effect the probability distribution

$$\mu_k(dx) = \mathbb{P}[X_k \in dx \mid Y_0,\cdots,Y_k]\ ,$$

is introduced. The evolution of the sequence $\{\mu_k\ ,\ k \geq 0\}$ taking values in the space $\mathcal{P}(E)$ of probability distributions on E, is very easily derived using the following Feynman–Kac formula. Let

$$\langle \gamma_n, f\rangle = \int_E \cdots \int_E f(x_n)\ \ R_0(dx_0)\ \prod_{k=1}^n R_k(x_{k-1},dx_k)$$

for any test function f defined on E. Clearly

$$\langle \mu_n, f\rangle = \frac{\langle \gamma_n, f\rangle}{\langle \gamma_n, 1\rangle} \qquad \text{and} \qquad \langle \gamma_n, f\rangle = \langle \gamma_{n-1}, R_n\, f\rangle\ ,$$

hence

$$\langle \mu_n, f\rangle = \frac{\langle \gamma_n, f\rangle}{\langle \gamma_n, 1\rangle} = \frac{\langle \gamma_{n-1}, R_n\, f\rangle}{\langle \gamma_{n-1}, R_n\, 1\rangle} = \frac{\langle \mu_{n-1}, R_n\, f\rangle}{\langle \mu_{n-1}, R_n\, 1\rangle}\ ,$$

and the transition from μ_{n-1} to μ_n is described by the following diagram

$$\mu_{n-1} \longrightarrow \mu_n = \frac{\mu_{n-1}\, R_n}{(\mu_{n-1}\, R_n)(E)} = \bar{R}_n(\mu_{n-1})\ .$$

Remark 2.1. Proceeding as in LeGland and Oudjane [5, Remark 2.1], it can be shown that the normalizing constant $(\mu_{n-1}\, R_n)(E)$ is a.s. positive, hence the probability distribution $\bar{R}_n(\mu_{n-1})$ is well–defined. Moreover, the likelihood of the model is given by

$$R_0(E)\ \prod_{k=1}^n (\mu_{k-1}\, R_k)(E)\ , \quad (6)$$

with the usual abuse of notation (5).

3. PARTICLE APPROXIMATION

By definition, and for a given importance decomposition (4)

$$\mu\, R_k(dx') = \int_E W_k(x,x')\ \mu(dx)\, P_k(x,dx')\ ,$$

with the usual abuse of notation. On the product space $E \times E$, let $\pi\ :\ (x,x') \longmapsto x'$ denote the projection on the (second) space E. For any probability distribution μ on the space E, the probability distribution $\mu \otimes P_k$ is defined on the product space $E \times E$ by

$$(\mu \otimes P_k)(dx, dx') = \mu(dx)\ P_k(x, dx') \ .$$

It follows that

$$\mu\, R_k(dx') \ = \ \int_E W_k(x, x')\ (\mu \otimes P_k)(dx, dx')$$

$$= \ (W_k\ (\mu \otimes P_k)) \circ \pi^{-1}(dx')\ ,$$

i.e. the nonnegative measure $\mu\, R_k$ on the space E is the marginal of the nonnegative measure $W_k\ (\mu \otimes P_k)$ on the product space $E \times E$, with importance weight function W_k and importance probability distribution $\mu \otimes P_k$. It follows also that

$$\bar{R}_k(\mu)(dx') = (W_k \cdot (\mu \otimes P_k)) \circ \pi^{-1}(dx')\ ,$$

where $\cdot$ denotes the projective product. The weighted particle approximation of the probability distribution $W_k \cdot (\mu \otimes P_k)$ is defined by

$$W_k \cdot (\mu \otimes P_k) \ \approx \ W_k \cdot S^N(\mu \otimes P_k)$$

$$= \ \sum_{i=1}^N \frac{W_k(\widehat{\xi}_{k-1}^i, \xi_k^i)}{\displaystyle\sum_{j=1}^N W_k(\widehat{\xi}_{k-1}^j, \xi_k^j)}\ \delta_{(\widehat{\xi}_{k-1}^i, \xi_k^i)}\ ,$$

where $\{\, (\widehat{\xi}_{k-1}^i, \xi_k^i)\, ,\ i = 1, \cdots, N\}$ is an N–sample with probability distribution $\mu \otimes P_k$, which can be achieved in the following manner : independently for any $i = 1, \cdots, N$

$$\widehat{\xi}_{k-1}^i \sim \mu(dx) \qquad \text{and} \qquad \xi_k^i \sim P_k(\widehat{\xi}_{k-1}^i, dx')\ ,$$

and the corresponding particle approximation for the marginal probability distribution $\bar{R}_k(\mu) = (W_k \cdot (\mu \otimes P_k)) \circ \pi^{-1}$ is defined by

$$\bar{R}_k(\mu) \ \approx \ (W_k \cdot S^N(\mu \otimes P_k)) \circ \pi^{-1}$$

$$= \ \sum_{i=1}^N \frac{W_k(\widehat{\xi}_{k-1}^i, \xi_k^i)}{\displaystyle\sum_{j=1}^N W_k(\widehat{\xi}_{k-1}^j, \xi_k^j)}\ \delta_{\xi_k^i}\ .$$

Let $\{\mu_n^N\, ,\ n \geq 0\}$ denote the particle filter approximation, associated with the importance decomposition (4), to the optimal filter $\{\mu_n\, ,\ n \geq 0\}$. The transition from μ_{n-1}^N to μ_n^N is described by the following diagram

$$\mu_{n-1}^N \longrightarrow \mu_n^N = (W_n \cdot S^N(\mu_{n-1}^N \otimes P_n)) \circ \pi^{-1}\ .$$

In practice, the particle approximation

$$\mu_k^N = \sum_{i=1}^N w_k^i\ \delta_{\xi_k^i}\ ,$$

is completely described by the set $\{\xi_k^i, w_k^i\, ,\ i = 1, \cdots, N\}$ of particles locations and weights, and the transition from $\{\xi_{k-1}^i, w_{k-1}^i\, ,\ i = 1, \cdots, N\}$ to $\{\xi_k^i, w_k^i\, ,\ i = 1, \cdots, N\}$ consists of the following steps

1. Independently for any $i = 1, \cdots, N$, generate

$$\widehat{\xi}_{k-1}^i \sim \mu_{k-1}^N(dx) \qquad \text{and} \qquad \xi_k^i \sim P_k(\widehat{\xi}_{k-1}^i, dx')\ .$$

2. For any $i = 1, \cdots, N$, compute the weight

$$w_k^i = W_k(\widehat{\xi}_{k-1}^i, \xi_k^i)\ /\ [\sum_{j=1}^N W_k(\widehat{\xi}_{k-1}^j, \xi_k^j)\,]\ ,$$

and set

$$\mu_k^N = \sum_{i=1}^N w_k^i\ \delta_{\xi_k^i}\ .$$

4. ERROR ESTIMATES

From now on, mathematical expectation $\mathbb{E}[\,\cdot\,]$ is taken only w.r.t. the additional randomness coming from the simulated r.v.'s, but not w.r.t. the observations. The following bias estimate does not depend on the importance decomposition (4).

Lemma 4.1. *For any (possibly random) probability distributions μ, μ' on E, it holds*

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}|\, \langle \mu \otimes P_k - \mu' \otimes P_k, W_k\ (\phi \circ \pi) \rangle\, |$$

$$\leq \ \sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}|\, \langle \mu - \mu', \phi \rangle\, |\ \sup_{x \in E} R_k(x, E)\ .$$

In contrast, the following variance estimate depends explicitly on the specific importance decomposition (4).

Lemma 4.2. *For any (possibly random) probability distribution μ on E, it holds*

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}|\, \langle S^N(\mu \otimes P_k) - \mu \otimes P_k, W_k\ (\phi \circ \pi) \rangle\, |$$

$$\leq \ \frac{1}{\sqrt{N}}\ [\, \sup_{x, x' \in E} W_k(x, x')\ (\mu\, R_k)(E)\,]^{1/2}\ .$$

Remark 4.3. In statistical applications, it is important to accurately estimate the likelihood (6) of the model, i.e. to estimate $(\mu_{k-1}\, R_k)(E) = \langle \mu_{k-1}\, R_k, 1 \rangle$, for the test function $\phi \equiv 1$. It is easy to show that the L^2–error for the particle approximation of (6) is minimum for the particle scheme associated with the decomposition (3) of the nonnegative kernel $R_k(x, dx')$, i.e. for the decomposition

$$R_k(x, dx') = \widehat{g}_k(x)\ \widehat{Q}_k(x, dx') \ .$$

with the usual abuse of notation.

Assumption A The importance weight function is bounded

$$\sup_{x,x' \in E} W_k(x, x') < \infty \ .$$

4.1. Rough estimates on a finite time horizon

If Assumption A holds, then the following notations are introduced

$$\gamma_k = \frac{\sup_{x \in E} R_k(x, E)}{(\mu_{k-1}\, R_k)(E)} \leq \frac{\sup_{x,x' \in E} W_k(x, x')}{(\mu_{k-1}\, R_k)(E)} = \gamma_k(W) \ ,$$

and in view of Remark 2.1, γ_k and $\gamma_k(W)$ are a.s. finite.

Theorem 4.4. *If for any $k \geq 1$, Assumption A holds, then*

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}|\ \langle \mu_k^N - \mu_k, \phi \rangle\ |$$

$$\leq\ \frac{1}{\sqrt{N}}\ \sum_{k=0}^{n} 2^{n-k+1}\ \gamma_{n:k+1}\ \sqrt{\gamma_k(W)} \ ,$$

where $\gamma_{n:k+1} = \gamma_n \cdots \gamma_{k+1}$, and with the convention that $\gamma_{n:n+1} = 1$.

4.2. Stability and uniform estimates

Without any assumption on the nonnegative kernel R_k, the error estimate obtained in Theorem 4.4 grows exponentially with the time horizon n. If the nonnegative kernel R_k is *mixing*, then the local errors are forgotten exponentially fast, and it is possible, proceeding as in LeGland and Oudjane [5, Section 4], to obtain error estimates which are uniform w.r.t. the time index n.

Definition 4.5. *The nonnegative kernel R_k is* mixing*, if there exist a constant $0 < \varepsilon_k \leq 1$, and a nonnegative measure λ_k defined on E, possibly depending on (Y_{k-1}, Y_k), such that*

$$\varepsilon_k\ \lambda_k(A) \leq R_k(x, A) \leq \frac{1}{\varepsilon_k}\ \lambda_k(A) \ ,$$

for any $x \in E$ and any Borel subset $A \subset E$, and let

$$\tau_k = (1 - \varepsilon_k^2)\ /\ (1 + \varepsilon_k^2) < 1 \ .$$

If R_k is mixing, then

$$(\mu\, R_k)(E) \geq \varepsilon_k^2\, (\mu_{k-1}\, R_k)(E) \ ,$$

for any probability distribution μ on E, hence a.s.

$$\inf_{\mu \in \mathcal{P}(E)} (\mu\, R_k)(E) > 0 \ ,$$

in view of Remark 2.1. If in addition Assumption A holds, then the following notation is introduced

$$\rho_k(W) = \frac{\sup_{x,x' \in E} W_k(x, x')}{\inf_{\mu \in \mathcal{P}(E)} (\mu\, R_k)(E)} \ ,$$

and $\rho_k(W)$ is a.s. finite.

Theorem 4.6. *If for any $k \geq 1$, Assumption A holds, and the nonnegative kernel R_k is mixing, then*

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}|\ \langle \mu_n^N - \mu_n, \phi \rangle\ |$$

$$\leq\ \frac{1}{\sqrt{N}}\ [\ \delta_n + 2\, \frac{\delta_{n-1}}{\varepsilon_n^2} + \frac{4}{\log 3}\ \sum_{k=1}^{n-2} \tau_{n:k+3}\ \frac{\delta_k}{\varepsilon_{k+2}^2\, \varepsilon_{k+1}^2}\ ] \ ,$$

where $\tau_{n:k+3} = \tau_n \cdots \tau_{k+3}$, and $\delta_k = 2\, \sqrt{\rho_k(W)}$.

5. REFERENCES

[1] O. Cappé, "Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation," *Monte Carlo Methods and Applications*, vol. 7, no. 1–2, pp. 81–92, 2001.

[2] P. Del Moral and J. Jacod, "Interacting particle filtering with discrete observations," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds., Statistics for Engineering and Information Science, chapter 3, pp. 43–75. Springer–Verlag, New York, 2001.

[3] D. Crişan and A. Doucet, "Convergence of sequential Monte Carlo methods," Technical Report CUED/F-INFENG/TR381, Department of Engineering, University of Cambridge, 2000.

[4] F. Desbouvries and W. Pieczynski, "Particle filtering with pairwise Markov processes," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong–Kong 2003*. IEEE–SPS, May 2003, vol. VI, pp. 705–708.

[5] F. Le Gland and N. Oudjane, "Stability and uniform approximation of nonlinear filters using the Hilbert metric, and application to particle filters," *The Annals of Applied Probability*, to appear.