A tutorial on expectation-maximization for nonlinear and/or non-Gaussian state-space models

Joshua T. Vogelstein[1], and Liam Paninski[2]
[1] Department of Neuroscience, Johns Hopkins School of Medicine
[2] Department of Statistics and Center for Theoretical Neuroscience, Columbia University

May 22, 2008

Contents

These Supplementary Materials provide a brief overview of the basic sequential Monte Carlo expectation maximization (SMC-EM) approach that we use in the main text. First, we describe the problem in terms of a *state-space model*. Once in this formalism, we show how an *Expectation Maximization* (EM) algorithm can infer the hidden states, along with the model parameters. Because the standard EM algorithm requires the evaluation of integrals that become intractable for this model, we use an approximation technique called *particle filtering* which sequentially generates Monte Carlo samples (hence, this approach is often referred to as *sequential Monte Carlo* (SMC)), discretizing the state-space, and approximating the problematic integrals by tractable sums. Derivations of each of the equations are provided in the Appendices.

1 State-Space Modeling

A state is anything that is time varying. The time varying states may be divided into those that are *hidden* (denoted by $\boldsymbol{H}_t$) wand those that are *observed* (denoted by $\boldsymbol{O}_t$). If the model also adheres to the following two conditions, then it can be considered a state-space model:

$$P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_{0:T}|\boldsymbol{H}_{0:T}) = \prod_{t=1}^{T} P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_t|\boldsymbol{H}_t) \tag{1}$$

$$P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_{0:T}) = P_{\boldsymbol{\theta}}(\boldsymbol{H}_0) \prod_{t=1}^{T} P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}). \tag{2}$$

where $\boldsymbol{X}_{0:T} = \{\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T\}$. Eq. 1 defines the *observation distribution* $P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_t|\boldsymbol{H}_t)$ by asserting that the probability of obtaining the observation at the current time, $\boldsymbol{O}_t$, is only a function of the hidden states at that time, $\boldsymbol{H}_t$. This distribution is governed entirely by the *observation parameters*, $\boldsymbol{\theta}_o$. Similarly, Eq. 2 defines the *transition distribution* $P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1})$, by asserting that the probability of the hidden state at the current time, $\boldsymbol{H}_t$ is only a function of the previous value of the hidden state, $\boldsymbol{H}_{t-1}$. This distribution is governed entirely by the *transition parameters*, $\boldsymbol{\theta}_{Tr}$. Fig. 1 (**A**) graphically depicts these two assumptions.

Taken together, these two assumptions imply that the *complete likelihood*, $P_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{H})$, i.e., the joint likelihood of the observation and hidden states for *all* time steps, may be simplified:

$$P_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{H}) = P_{\boldsymbol{\theta}_o}(\boldsymbol{O}|\boldsymbol{H}) P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}) = \prod_{t=0}^{T} P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_t|\boldsymbol{H}_t) P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}), \tag{3}$$

where we use the notation $\boldsymbol{X}$ (without a subscript) to indicate a sequence for all time, i.e., $\boldsymbol{X} = \boldsymbol{X}_{0:T}$. Eq. 3 asserts that the complete likelihood is characterized entirely by the observation and transition distributions (we ignore the initial conditions because they contribute relatively little to this likelihood).[1]

2 Expectation Maximization for State-Space Models

An Expectation Maximization (EM) algorithm generally iterates two key operations. First, EM algorithms compute the sufficient statistics for performing an optimal inference, given any setting of the model parameters (this is called the *Expectation* (or E) step). Second, EM algorithms provide the maximum likelihood estimates of the parameters, given the above inference (this is called the *Maximization* (or M) step). More precisely, in the E step one explicitly writes down the expected value of the complete log likelihood, $E_{P_{\boldsymbol{\theta}'}(\boldsymbol{H}|\boldsymbol{O})} \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{H})$, in terms of the model states, given the current parameter estimates. The M step then computes a weighted maximum likelihood estimate of the parameters, given this expectation. The E and M steps are iterated until the parameters converge (1). Note that it is the likelihood of the expected value of the complete log likelihood that is guaranteed to converge, not the likelihood of $P_{\boldsymbol{\theta}}(\boldsymbol{O}_t)$. The E and M step can be formally written as:

[1] For the model in the main text, the observed state is the fluorescence observation: $\boldsymbol{O}_t = F_t$, the hidden states are the calcium concentrations, spikes, and spike history terms: $\boldsymbol{H}_t = \{[\mathrm{Ca}^{2+}]_t, n_t, \boldsymbol{h}_t\}$. Note that the external stimulus $\boldsymbol{x}_t$ is *not* a state. Rather $\boldsymbol{x}_t$ operates on the neuron by modulating the probability of spiking. The observation distribution parameters are $\boldsymbol{\theta}_o = \{\alpha, \beta, \sigma_F, n, k_d\}$. The transition distribution parameters are $\boldsymbol{\theta}_{Tr} = \{\boldsymbol{k}, \omega, \tau_h, \sigma_h, \tau, A, \sigma\}$.

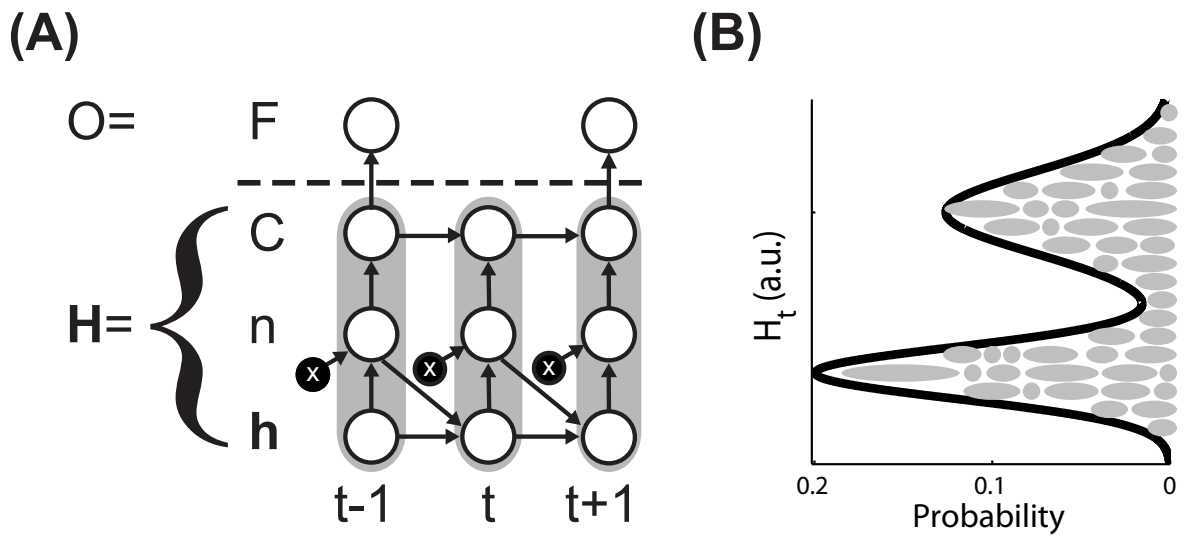

Figure 1: Sequential Monte Carlo Assumptions. **(A)** Directed Acyclic Graphical representation of a state-space model. The horizontal dotted line divides the graph between the observed states (above the line) and hidden states (below the line). The graph depicts the conditional dependencies of the model by drawing directed edges (*arrows*) between the nodes (*open circles*). The time step is indicated on the bottom. For the model in the main text, the observation state is the intermittent fluorescence observations, F_t, and the hidden states (*gray ellipses*) are the time-varying intracellular calcium concentration $[\mathrm{Ca}^{2+}]_t$ spikes n_t, and spike history terms $\boldsymbol{h}_t$. Collectively, the hidden states comprise a Markov process. Note that the external input $\boldsymbol{x}_t$ operates on the spiking probability. **(B)** A set of particles (*gray ellipses*) comprise a discrete approximation to the continuous mixture of Gaussians distribution (*black line*). The size of each ellipse corresponds to its weight, $w_t^{(i)}$, and the position corresponds to its value, $\boldsymbol{H}_t^{(i)}$.

$$\textbf{E step: Compute } Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = E_{P_{\boldsymbol{\theta}'}(\boldsymbol{H}|\boldsymbol{O})} \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{H}) = \int P_{\boldsymbol{\theta}'}(\boldsymbol{H}|\boldsymbol{O}) \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{H}) d\boldsymbol{H}$$

$$\textbf{M step: Compute } \widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

where $\boldsymbol{\theta}'$ is the previous EM iteration's parameter estimate, and $\widehat{\boldsymbol{\theta}}$ is the new estimate. The E step may be expanded using Eq. 3 (See (2) or Appendix A for derivation):

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{t=1}^{T} \iint P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t, \boldsymbol{H}_{t-1}|\boldsymbol{O}) \ln P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) d\boldsymbol{H}_t d\boldsymbol{H}_{t-1} + \sum_{t=0}^{T} \int P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t|\boldsymbol{O}) \ln P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_t|\boldsymbol{H}_t) d\boldsymbol{H}_t. \quad (4)$$

Note that these integrals need not be evaluated, as we approximate them with sums in the next section. Because the transition and observation distributions are given by the model, completing the E step requires computing both (i) the *pairwise joint conditional distributions* (or pairwise joint conditionals), $P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t, \boldsymbol{H}_{t-1}|\boldsymbol{O})$, and (ii) the *marginal conditional distributions* (or marginal conditionals), $P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t|\boldsymbol{O})$. These distributions can be efficiently computed using a forward-backward approach, originally developed for HMMs (Baum-Welch Algorithm (3)) and linear-Gaussian state-space models (4). The forward-backward approach proceeds by adopting a forward recursion to compute the distribution of the hidden state at time step t, given all *previous* observations, $P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}_{0:t})$, which is referred to as the *forward distribution* (or forward filter). Upon arriving at the final time step, one iteratively recurses *backward* to compute the pairwise joint and marginal conditionals for each time step t, which are conditioned on *all* the observations. The forward recursion uses the following update equation (see (2) or Appendix B for derivation):

$$P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}_{0:t}) = \frac{1}{Z} P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_t|\boldsymbol{H}_t) \int P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) P_{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1}|\boldsymbol{O}_{0:t-1}) d\boldsymbol{H}_{t-1}, \tag{5}$$

where Z is a normalization constant required to ensure that the forward distribution integrates to unity. The backward recursion uses the following update equation (see (1) or Appendix C for derivation):

$$P_{\boldsymbol{\theta}}(\boldsymbol{H}_t, \boldsymbol{H}_{t-1}|\boldsymbol{O}) = P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}) \frac{P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) P_{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1}|\boldsymbol{O}_{0:t-1})}{\int P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) P_{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1}|\boldsymbol{O}_{0:t-1}) d\boldsymbol{H}_{t-1}} \tag{6a}$$

$$P_{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1}|\boldsymbol{O}) = \int P_{\boldsymbol{\theta}}(\boldsymbol{H}_t, \boldsymbol{H}_{t-1}|\boldsymbol{O}) d\boldsymbol{H}_t, \tag{6b}$$

yielding the pairwise joint conditionals (Eq. 6a) and marginal conditionals (Eq. 6b). Note that this recursion requires first computing the forward distribution for all t, from which the name "forward-backward" was derived.

Having the pairwise joint and marginal conditionals from the backward recursion completes the E step, as one can now explicitly write out $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ for this particular model using Eq. 4. Importantly, these conditionals perform a double duty. First, they are the sufficient statistics for performing the optimal inference. To see this, note that one could estimate H_t by simply computing the conditional mean,

$$E(H_t) = \int H_t P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}) dH_t. \tag{7}$$

Second, these pairwise joint and marginal distributions provide the sufficient statistics for computing the maximum likelihood estimators for the model parameters. For state-space-models, the maximization breaks down into two separate maximizations, one for the transition distribution parameters $\boldsymbol{\theta}_{Tr}$, and one for the observation distribution parameters $\boldsymbol{\theta}_o$, which follows directly from the expansion in Eq. 4. Therefore, maximizing with respect to the transition distribution parameters requires only the pairwise joint conditionals:

$$\widehat{\boldsymbol{\theta}}_{Tr} = \underset{\theta_{Tr}}{\operatorname{argmax}} \sum_{t=1}^{T} \iint P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t, \boldsymbol{H}_{t-1}|\boldsymbol{O}) \ln P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) d\boldsymbol{H}_t d\boldsymbol{H}_{t-1}, \tag{8}$$

and maximizing with respect to the observation distribution parameters requires only the marginal conditionals:

$$\widehat{\boldsymbol{\theta}}_o = \underset{\theta_o}{\operatorname{argmax}} \sum_{t=1}^{T} \int P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t|\boldsymbol{O}) \ln P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_t|\boldsymbol{H}_t) d\boldsymbol{H}_t. \tag{9}$$

3 SMC Forward Recursion

While the EM algorithm for state-space models provides the distributions of interest, the integral in Eq. 5 is often difficult to compute.[2] Instead, we will use an approximate (SMC) method to perform the forward recursion (5, 6). This forces minor modifications to the backward recursion and the M-step.

The SMC idea is quite simple. Instead of integrating over all possible hidden states $\boldsymbol{H}_t$ at each time step, one integrates over some finite set of *particles*, $\{\boldsymbol{H}_t^{(1)}, \ldots, \boldsymbol{H}_t^{(N)}\}$, intelligently chosen to approximate the entire distribution. At each time step, each particle has an associated weight, $w_t^{(i)}$, which together comprise the forward distribution approximation:

$$P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}_{0:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta\big(\boldsymbol{H}_t - \boldsymbol{H}_t^{(i)}\big), \tag{10}$$

[2]Technically, we could evaluate the integral in Eq. 5 for the state space model in the main text, since the computations involve a conceptually simple mixture of Gaussians. However, since the number of distributions in the mixture doubles with each time step, evaluating the integral after many time steps becomes computationally intractable.

where $\delta(X)$ is the Dirac delta function, taking value one when $X = 0$ and zero otherwise (for a proof that as $N \to \infty$, this approximation becomes exact, and other convergence results, see (5)). The pair $\left(\boldsymbol{H}_t^{(i)}, w_t^{(i)}\right)$ indicates that at time step t, the probability of the hidden state taking value $\boldsymbol{H}_t^{(i)}$ is $w_t^{(i)}$. This set of particles and weights then acts as a discrete approximation of the forward distribution, as depicted in Fig. 1 (**B**). Substituting $\sum_{j=1}^{N} w_{t-1}^{(j)} \delta(\boldsymbol{H}_{t-1}^{(j)} - \boldsymbol{H}_{t-1})$ for $P_{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1}|\boldsymbol{O}_{0:t-1})$ in Eq. 5 yields a particle analog to the forward update equation:

$$w_t^{*(i)} = \frac{1}{Z} P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_t|\boldsymbol{H}_t^{(i)}) \sum_{j=1}^{N} P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(j)}) w_{t-1}^{(j)}. \tag{11}$$

Because the sum in Eq. 11 requires computing the transition distribution for each pair of particles, one typically approximates Eq. 11 with

$$\bar{w}_t^{(i)} = \frac{1}{Z} P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_t|\boldsymbol{H}_t^{(i)}) P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(i)}) w_{t-1}^{(i)}, \tag{12}$$

which is accurate when the transition distribution $P_{\boldsymbol{\theta}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(i)})$ is highly concentrated at $\boldsymbol{H}_t^{(i)} = \boldsymbol{H}_{t-1}^{(i)}$ (e.g., this is a good approximation when either the time step or transition noise is small). To compute Eq. 12, one must sample $\boldsymbol{H}_t^{(i)}$ from some distribution, which we will call the *sampling distribution*, $q(\boldsymbol{H}_t)$, though it is also known as the importance or proposal distribution. An importance sampling argument informs us that upon approximating a distribution by sampling, one must normalize the likelihood by the probability of having sampled that value (see (5) or Appendix D for justification). Therefore, one updates the *importance weights* (or weights) using

$$\widetilde{w}_t^{(i)} = \frac{P_{\boldsymbol{\theta}_o}(\boldsymbol{O}_t|\boldsymbol{H}_t^{(i)}) P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(i)}) w_{t-1}^{(i)}}{q(\boldsymbol{H}_t^{(i)})} \tag{13a}$$

$$w_t^{(i)} = \frac{\widetilde{w}_t^{(i)}}{\sum_j \widetilde{w}_t^{(j)}}. \tag{13b}$$

If possible, one samples from the so-called optimal sampling distribution, $q\big(\boldsymbol{H}_t^{(i)}\big) = P_{\boldsymbol{\theta}}\big(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}^{(i)}, \boldsymbol{O}_t\big)$.

One final note on the use of SMC algorithms relates to the issue of *resampling*. Because particles are sampled, some may have weights close to zero. When this happens, one can sample particles with replacement according to their weights, a process called resampling. This tends to drop the unlikely particles and replicate the very likely ones. Although many resampling strategies are available (7), we use *stratified resampling* because of its efficiency and simplicity (see Appendix E for details). Upon resampling, all the weights are set to $1/N$. Thus, to complete the SMC approximation to the forward recursion, first initialize a set of N particles to take some reasonable starting value, and assign each an equal weight. Then, at each time step, (i) update the position (in hidden space) of each particle by sampling from the sampling distribution; (ii) use Eq. 13 to compute the weight of each particle; and (iii) resample. One iterates these three steps (together called Sequential Importance Sampling with Resampling (5)) until arriving at $t = T$, at which time the forward recursion is complete.

4 SMC Backward Recursion

Having completed the SMC forward recursion, the SMC approximation to the backward recursion proceeds by substituting these weights into Eq. 6 to get the particle analog for the backward recursion:

$$J_{t,t-1}^{(i,j)} = P_{\boldsymbol{\theta}}(\boldsymbol{H}_t^{(i)}, \boldsymbol{H}_{t-1}^{(j)}|\boldsymbol{O}) = P_{\boldsymbol{\theta}}\big(\boldsymbol{H}_t^{(i)}|\boldsymbol{O}\big) \frac{P_{\boldsymbol{\theta}_{Tr}}\big(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(j)}\big) w_{t-1}^{(j)}}{\sum_j P_{\boldsymbol{\theta}_{Tr}}\big(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(j)}\big) w_{t-1}^{(j)}} \tag{14a}$$

$$M_{t-1}^{(j)} = P_{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1}^{(j)}|\boldsymbol{O}) = \sum_{i=1}^{N} J_{t,t-1}^{(i,j)}, \tag{14b}$$

where $J_{t,t-1}^{(i,j)}$ is the pairwise joint likelihood of particle i taking value $\boldsymbol{H}_t^{(i)}$ at time t and particle j taking value $\boldsymbol{H}_{t-1}^{(j)}$ at time $t-1$, conditioned on *all* the observations. Similarly, $M_t^{(i)}$ is the marginal likelihood of particle i taking value $\boldsymbol{H}_t^{(i)}$ at time t, conditioned on *all* the observations. One therefore completes the SMC approximation to the backward recursion by initializing $M_T^{(j)} = w_T^{(j)}$ for all j, and then recursing *backward* using Eq. 14 to compute the pairwise joint and marginal conditional likelihoods until $t = 0$. At this point, the particle approximation of the E step is complete, and one may proceed to the M step. Because each forward recursion takes $O(TN)$ time, and each backward recursion takes $O(TN^2)$ time (due to the pairwise transitions, $P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(j)})$), each E step takes $O(TN^2)$ time. Note that without the approximation in Eq. 12, the forward recursion would take $O(TN^2)$ steps, though this could potentially be reduced (6).

5 SMC M Step

Having the particle approximation to marginal and joint conditional distributions, they may be plugged into Eq. 8 and Eq. 9 to find the maximum likelihood estimates of the transition distribution and observation distribution parameters

$$\widehat{\boldsymbol{\theta}}_{Tr} = \underset{\boldsymbol{\theta}_{Tr}}{\operatorname{argmax}} \sum_{t=1}^{N} \sum_{i,j=1}^{N} J_{t,t-1}^{(i,j)} \ln P_{\boldsymbol{\theta}_{Tr}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(j)}) \tag{15}$$

$$\widehat{\boldsymbol{\theta}}_{o} = \underset{\boldsymbol{\theta}_{o}}{\operatorname{argmax}} \sum_{t=1}^{N} \sum_{i,j=1}^{N} M_t^{(i)} \ln P_{\boldsymbol{\theta}_{o}}(\boldsymbol{O}_t|\boldsymbol{H}_t^{(i)}), \tag{16}$$

completing one SMC-EM iteration. Any SMC-EM algorithm therefore proceeds in a similar fashion as an EM algorithm for state-space models, but one must replace the forward, backward, and M steps with their corresponding SMC approximations. Upon convergence, the inferences follows as in Eq. 7, but using the particle approximation:

$$E(H_t) = \sum_i H_t^{(i)} P_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{O}). \tag{17}$$

Table 1 provides pseudocode for using a SMC-EM algorithm to perform these inferences.

Table 1: Pseudocode for SMC-EM

1. Initialize $\boldsymbol{\theta}'$ using some good guess of the initial parameters.
2. Call the parameters from the previous EM iteration $\boldsymbol{\theta}'$.
   - **Expectation Step:** The expectation step simplifies to a forward recursion and a backward recursion.
     - **Forward:** Initialize particle, meaning choose a value for each particle at time $t = 0$, and assign each a weight of $1/N$. Then, for $i \in \{1, \ldots, N\}$ and $t = 1, \ldots, T$:
       (a) update particles by sampling from the sampling distribution $\boldsymbol{H}_t^{(i)} \sim q(\boldsymbol{H}_t)$,
       (b) update weights using Eq. 13,
       (c) if necessary, stratified resample and set $w_t^{(i)} = 1/N$ for all i.
     - **Backward:** For $t = T, \ldots, 1$
       (a) Compute the pairwise joint conditional likelihoods using Eq. 14a,
       (b) Compute the marginal conditional likelihoods using Eq. 14b.
   - **Maximization Step:**
     - Find $\widehat{\boldsymbol{\theta}}_{Tr}$, the maximum likelihood estimates of the transition distribution parameters, using Eq. 15, and let $\boldsymbol{\theta}'_{Tr} \rightarrow \widehat{\boldsymbol{\theta}}_{Tr}$ for the next iteration.
     - Find $\widehat{\boldsymbol{\theta}}_o$, the maximum likelihood estimates of the observation distribution parameters using Eq. 16, and let $\boldsymbol{\theta}'_o \rightarrow \widehat{\boldsymbol{\theta}}_o$ for the next iteration.
3. Repeat the EM steps until convergence. Then, perform the desired inferences by plugging the final parameter estimates into equations such as 17.

A Expectation Step for State-Space Models

Our goal here is to evaluate the following expectation:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \boldsymbol{E}_{P_{\boldsymbol{\theta}'}(\boldsymbol{H}|\boldsymbol{O})} \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{H}) = \int \cdots \int \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{H}) P_{\boldsymbol{\theta}'}(\boldsymbol{H}|\boldsymbol{O}) d\boldsymbol{H}_0 \cdots d\boldsymbol{H}_T \tag{A.1}$$

We do so invoking two mathematical tricks. First, by making use of the conditional independencies inherent in our model, we can write the whole joint probability as a product of conditional probabilities:

$$P_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{H}) = P_{\boldsymbol{\theta}}(\boldsymbol{O}|\boldsymbol{H}) P_{\boldsymbol{\theta}}(\boldsymbol{H}) = P_{\boldsymbol{\theta}}(\boldsymbol{H}_0) \prod_{t=1}^{T} P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) \prod_{t=0}^{T} P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t), \tag{A.2}$$

or equivalently, the log as a set of sums:

$$\ln P_{\boldsymbol{\theta}}(\boldsymbol{O}, \boldsymbol{H}) = \ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_0) + \sum_{t=1}^{T} \ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) + \sum_{t=0}^{T} \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t). \tag{A.3}$$

Substituting (A.3) into (A.1) yields:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \int \cdots \int \Big(\ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_0) + \sum_{t=1}^{T} \ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) + \sum_{t=0}^{T} \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t) \Big) P_{\boldsymbol{\theta}'}(\boldsymbol{H}|\boldsymbol{O}) d\boldsymbol{H}_0 \cdots d\boldsymbol{H}_T. \tag{A.4}$$

Second, we use the rules for maginalizing densitites to simplify the products in Eq. A.4:

$$\int \cdots \int P_{\boldsymbol{\theta}'}(\boldsymbol{H}|\boldsymbol{O}) \times \ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_0) d\boldsymbol{H}_0 \cdots d\boldsymbol{H}_T = \int P_{\boldsymbol{\theta}'}(\boldsymbol{H}_0|\boldsymbol{O}) \times \ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_0) d\boldsymbol{H}_0 \tag{A.5}$$

$$\int \cdots \int \sum_{t=1}^{T} P_{\boldsymbol{\theta}'}(\boldsymbol{H}|\boldsymbol{O}) \times \ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) d\boldsymbol{H}_0 \cdots d\boldsymbol{H}_T = \sum_{t=1}^{T} \iint P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t, \boldsymbol{H}_{t-1}|\boldsymbol{O}) \times \ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) d\boldsymbol{H}_t d\boldsymbol{H}_{t-1} \tag{A.6}$$

$$\int \cdots \int \sum_{t=0}^{T} P_{\boldsymbol{\theta}'}(\boldsymbol{H}|\boldsymbol{O}) \times \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t) d\boldsymbol{H}_0 d\boldsymbol{H}_T = \sum_{t=0}^{T} \int P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t|\boldsymbol{O}) \times \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t) d\boldsymbol{H}_t, \tag{A.7}$$

Therefore, Eq. A.4 becomes:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \int P_{\boldsymbol{\theta}'}(\boldsymbol{H}_0|\boldsymbol{O}) \times \ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_0) + d\boldsymbol{H}_0 \sum_{t=1}^{T} \iint P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t, \boldsymbol{H}_{t-1}|\boldsymbol{O}) \times \ln P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) d\boldsymbol{H}_t d\boldsymbol{H}_{t-1} + \sum_{t=0}^{T} \int P_{\boldsymbol{\theta}'}(\boldsymbol{H}_t|\boldsymbol{O}) \times \ln P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t) d\boldsymbol{H}_t. \tag{A.8}$$

B Forward Recursion Derivation

First, we simplify $P(\boldsymbol{H}_t|\boldsymbol{O}_{0:t})$ by applying Bayes' Rule, followed by the laws governing marginal probabilities, and Bayes' Rule again:

$$P(\boldsymbol{H}_t|\boldsymbol{O}_{0:t}) = \frac{P(\boldsymbol{H}_t, \boldsymbol{O}_{0:t})}{P(\boldsymbol{O}_{0:t})} \tag{B.1a}$$

$$= \frac{1}{P(\boldsymbol{O}_{0:t})} \int d\boldsymbol{H}_{0:k-1} P(\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:t}) \tag{B.1b}$$

$$= \frac{1}{P(\boldsymbol{O}_{0:t})} \int d\boldsymbol{H}_{0:k-1} P(\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:k-1}) P(\boldsymbol{O}_t|\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:k-1}). \tag{B.1c}$$

Then, we note that because of the model assumptions, we can simplify $P(\boldsymbol{O}_t|\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:k-1})$ using the following:

Lemma B.1.

$$P(\boldsymbol{O}_t|\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:k-1}) = P(\boldsymbol{O}_t|\boldsymbol{H}_t) \tag{B.2}$$

Proof.

$$P(\boldsymbol{O}_t|\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:k-1}) = \frac{P(\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:t})}{P(\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:k-1})} \tag{B.3a}$$

$$= \frac{P(\boldsymbol{O}_{0:t}|\boldsymbol{H}_{0:t}) P(\boldsymbol{H}_{0:t})}{\int d\boldsymbol{O}_t P(\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:k-1}, \boldsymbol{O}_t)} \tag{B.3b}$$

$$= \frac{P(\boldsymbol{O}_{0:t}|\boldsymbol{H}_{0:t}) P(\boldsymbol{H}_{0:t})}{\int d\boldsymbol{O}_t P(\boldsymbol{O}_{0:t}|\boldsymbol{H}_{0:t}) P(\boldsymbol{H}_{0:t})} \tag{B.3c}$$

$$= \frac{P(\boldsymbol{H}_{0:t}) \prod_{s=1}^{t} P(\boldsymbol{O}_s|\boldsymbol{H}_s)}{P(\boldsymbol{H}_{0:t}) \int d\boldsymbol{O}_t \prod_{s=1}^{t} P(\boldsymbol{O}_s|\boldsymbol{H}_s)} \tag{B.3d}$$

$$= \frac{P(\boldsymbol{H}_{0:t}) \prod_{s=1}^{t} P(\boldsymbol{O}_s|\boldsymbol{H}_s)}{P(\boldsymbol{H}_{0:t}) \prod_{s=1}^{t-1} P(\boldsymbol{O}_s|\boldsymbol{H}_s)} \tag{B.3e}$$

$$= P(\boldsymbol{O}_t|\boldsymbol{H}_t) \tag{B.3f}$$

□

Therefore,

$$P(\boldsymbol{H}_t|\boldsymbol{O}_{0:t}) = \frac{1}{P(\boldsymbol{O}_{0:t})} \int d\boldsymbol{H}_{0:k-1} P(\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:k-1}) P(\boldsymbol{O}_t|\boldsymbol{H}_t) \tag{B.4a}$$

$$= \frac{1}{P(\boldsymbol{O}_{0:t})} P(\boldsymbol{O}_t|\boldsymbol{H}_t) \int d\boldsymbol{H}_{0:k-1} P(\boldsymbol{H}_{0:t}, \boldsymbol{O}_{0:k-1}) \tag{B.4b}$$

$$= \frac{1}{P(\boldsymbol{O}_{0:t})} P(\boldsymbol{O}_t|\boldsymbol{H}_t) P(\boldsymbol{H}_t, \boldsymbol{O}_{0:k-1}) \tag{B.4c}$$

$$= \frac{1}{P(\boldsymbol{O}_{0:t})} P(\boldsymbol{O}_t|\boldsymbol{H}_t) P(\boldsymbol{H}_t|\boldsymbol{O}_{0:k-1}) P(\boldsymbol{O}_{0:k-1}) \tag{B.4d}$$

$$= \frac{1}{P(\boldsymbol{O}_t|\boldsymbol{O}_{0:k-1})} P(\boldsymbol{O}_t|\boldsymbol{H}_t) P(\boldsymbol{H}_t|\boldsymbol{O}_{0:k-1}), \tag{B.4e}$$

where all the equalities follow from Bayes's rule or the definition of conditional and marginal densities. We now need to simplify the so-called "one-step predictor", $P(\boldsymbol{H}_t|\boldsymbol{O}_{0:k-1})$, termed so because it predicts the likelihood of the

hidden state at time step t given observations up to, but not including, time step t:

$$P(\boldsymbol{H}_t|\boldsymbol{O}_{0:k-1}) = \int d\boldsymbol{H}_{0:k-1} P(\boldsymbol{H}_{0:t}|\boldsymbol{O}_{0:k-1}) \tag{B.5a}$$

$$= \int d\boldsymbol{H}_{0:k-1} P(\boldsymbol{H}_t|\boldsymbol{H}_{0:k-1}, \boldsymbol{O}_{0:k-1}) P(\boldsymbol{H}_{0:k-1}|\boldsymbol{O}_{0:k-1}) \tag{B.5b}$$

$$= \int d\boldsymbol{H}_{t-1} P(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) P(\boldsymbol{H}_{t-1}|\boldsymbol{O}_{0:k-1}), \tag{B.5c}$$

where the first two equalities follow from basic probability theory and the third equality follows from the following:

Lemma B.2.

$$P(\boldsymbol{H}_t|\boldsymbol{H}_{0:k-1}, \boldsymbol{O}_{0:k-1}) = P(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}). \tag{B.6}$$

which follows from a similar proof as Lemma . This result together with Eq. B.4e, provides the following recursive relationship, often referred to as the Chapman-Kolmogorov equations:

$$P(\boldsymbol{H}_t|\boldsymbol{O}_{0:t}) = \frac{1}{P(\boldsymbol{O}_t|\boldsymbol{O}_{0:k-1})} P(\boldsymbol{O}_t|\boldsymbol{H}_t) P(\boldsymbol{H}_t|\boldsymbol{O}_{0:k-1}) \tag{B.7}$$

$$P(\boldsymbol{H}_t|\boldsymbol{O}_{0:k-1}) = \int d\boldsymbol{H}_{t-1} P(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) P(\boldsymbol{H}_{t-1}|\boldsymbol{O}_{0:k-1}). \tag{B.8}$$

Note that for the rest of this text, we let $P(\boldsymbol{O}_t|\boldsymbol{O}_{0:k-1})$ act as a normalizing constant as it is not a function of $\boldsymbol{H}_t$, and therefore, denote it by Z.

C Backward Recursion Derivation

To estimate the joint posterior hidden probabilities, we apply Bayes' rule several times to simplify:

$$P(\boldsymbol{H}_t, \boldsymbol{H}_{t+1}|\boldsymbol{O}) = P(\boldsymbol{H}_{t+1}|\boldsymbol{O}) P(\boldsymbol{H}_t|\boldsymbol{H}_{t+1}, \boldsymbol{O}) \tag{C.1a}$$

$$= P(\boldsymbol{H}_{t+1}|\boldsymbol{O}) P(\boldsymbol{H}_t|\boldsymbol{H}_{t+1}, \boldsymbol{O}_{0:t}) \tag{C.1b}$$

$$= P(\boldsymbol{H}_{t+1}|\boldsymbol{O}) \frac{P(\boldsymbol{H}_t, \boldsymbol{H}_{t+1}, \boldsymbol{O}_{0:t})}{P(\boldsymbol{H}_{t+1}, \boldsymbol{O}_{0:t})} \tag{C.1c}$$

We can then simplify the numerator by applying Bayes rule and our model assumptions a couple times:

$$P(\boldsymbol{H}_t, \boldsymbol{H}_{t+1}, \boldsymbol{O}_{0:t}) = P(\boldsymbol{H}_t, \boldsymbol{H}_{t+1}|\boldsymbol{O}_{0:t}) P(\boldsymbol{O}_{0:t}) \tag{C.2a}$$

$$= P(\boldsymbol{H}_{t+1}|\boldsymbol{H}_t) P(\boldsymbol{H}_t|\boldsymbol{O}_{0:t}) P(\boldsymbol{O}_{0:t}) \tag{C.2b}$$

$$= P(\boldsymbol{H}_{t+1}|\boldsymbol{H}_t) P(\boldsymbol{H}_t, \boldsymbol{O}_{0:t}) \tag{C.2c}$$

Substituting the result from Eq. into Eq. , yields:

$$P(\boldsymbol{H}_t, \boldsymbol{H}_{t+1}|\boldsymbol{O}) = P(\boldsymbol{H}_{t+1}|\boldsymbol{O}) \frac{P(\boldsymbol{H}_{t+1}|\boldsymbol{H}_t) P(\boldsymbol{H}_t, \boldsymbol{O}_{0:t})}{P(\boldsymbol{H}_{t+1}, \boldsymbol{O}_{0:t})} \tag{C.3a}$$

$$P(\boldsymbol{H}_t, \boldsymbol{H}_{t+1}|\boldsymbol{O}) = P(\boldsymbol{H}_{t+1}|\boldsymbol{O}) \frac{P(\boldsymbol{H}_{t+1}|\boldsymbol{H}_t) P(\boldsymbol{H}_t, \boldsymbol{O}_{0:t})}{\int P(\boldsymbol{H}_{t+1}|\boldsymbol{H}_t) P(\boldsymbol{H}_t|\boldsymbol{O}_{0:t}) d\boldsymbol{H}_t} \tag{C.3b}$$

D Importance Sampling

The idea for importance sampling comes from the following intuition. We can define the expected value of some discrete-valued function, $\phi(x)$, with respect to the distribution $p(x)$, as:

$$E_p(\phi(x)) = \sum_x \phi(x)p(x). \tag{D.1}$$

Therefore, if we sample uniformly from $p(x)$ to generate particles, $x^{(i)}$, we can approximate this expected value as:

$$E_p(\phi(x)) \approx \sum_i \phi(x^{(i)})p(x^{(i)}) = \frac{1}{N}\sum_i \phi(x^{(i)}), \tag{D.2}$$

where $p(x^{(i)}) = 1/N$ for all i by definition, as we have sampled from it. This "uniform sampling" strategy can be very inefficient for certain kinds of distributions. Instead of uniformly sampling from $p(x)$, we could sample from some other distribution, $q(x)$. Note that the expectation can be redefined accordingly by multipling and dividing by $q(x)$:

$$E_p(\phi(x)) = \sum_x \phi(x)\frac{p(x)}{q(x)}q(x) = E_q\left(\phi(x)\frac{p(x)}{q(x)}\right). \tag{D.3}$$

Therefore, sampling from $q(x)$ enables a different approximation of the expectation:

$$E_p(\phi(x)) \approx \frac{1}{Z}\sum_i \phi(x^{(i)})\frac{p(x^{(i)})}{q(x^{(i)})}q(x^{(i)}) \tag{D.4a}$$

$$= \frac{1}{Z}\sum_i \phi(x^{(i)})\frac{p(x^{(i)})}{q(x^{(i)})} \tag{D.4b}$$

$$= \sum_i \phi(x^{(i)})w^{(i)}. \tag{D.4c}$$

Therefore, one can approximate an expectation of $\phi(x)$ by sampling from some distribution, $q(x)$, and using it to compute the weights, $w^{(i)}$. While, this might seem like a trick only useful for approximating expectations, by employing a trick that enables any likelihood to be written as an expectation, one can use this strategy to sample from any distribution. In particular, one can write

$$P(X = x) = E_P(\delta(X, x)) = \int_{y\in\mathcal{Y}} P(X = x)\delta(x, y)dy = \sum_{y\in\mathcal{Y}} P(X = x)\delta(x, y), \tag{D.5}$$

where y is a dummy variable that can take any value in $\mathcal{Y}$. Note that if the space of y does not include some possible values in the space of x, i.e., $\exists\ x$ such that $x \notin \mathcal{Y}$, then the above equalities are actually approximations, and a normalization constant must be introduced to account for this:

$$P(X = x) \approx \frac{1}{Z}\sum_{y\in\mathcal{Y}} P(X = x)\delta(x, y), \tag{D.6}$$

where $Z = \sum_y P(X = y)$ ensures that the distribution $P(X)$ integrates to unity. This is precisely the same trick that enables justification of any sampling procedure. In other words, we could write:

$$P(X = x) \approx \frac{1}{Z}\sum_i P(X = x)\delta(x, x^{(i)}). \tag{D.7}$$

E Stratified Resampling

Stratified resampling is based on based on ideas used in survey sampling(7). One first prepartions the unit interval, into N disjoint sets, $(0,1) = (0,1/N) \cup \cdots ((N-1)/N, 1)$. Then, each $U^{(i)}$ is drawn independently from its associated sub-interval, $U^{(i)} \sim \mathcal{U}\{((i-1)/N, i/N)\}$, where $\mathcal{U}\{a,b\}$ denotes the uniform distribution on the interval (a,b). Next, one generates a cumulative sum of weights, comprising a set now on the same interval. Each sample then corresponds to a particle, the one that sets the lower bound within the interval defined by that particle and the next one. Table 2 provides a detailed algorithm. This approach is more efficient than multinomial resampling in that the conditional variance of stratified resampling is always smaller than that of multinomial resampling.

Table 2: Pseudocode for Stratified Resampling

- Having particle values, $\breve{\boldsymbol{H}}_t^{(i)}$ and associated weights, $\breve{w}_t^{(i)}$, pre-partion the interval $(0,1)$ into N disjoint sets, i.e., $(0,1) = (0,1/N) \cup \cdots \cup (\{N-1\}/N, 1)$
- Draw N samples, one from each subinterval, i.e. $\mathcal{U}^{(i)} \sim \mathcal{U}\left((\{i-1\}/N, i/N)\right)$ where $\mathcal{U}\left((a,b)\right)$ indicates the uniform distribution on the interval (a,b)
- Compute the cumulative sum of weights, i.e., $c^{(i)} = \sum_{t=1}^{i} \breve{w}_t^{(k)}$
- For each sample i, find the interval labeled by j that $\mathcal{U}^{(i)}$ is between, i.e., $\mathrm{find}_j\, c^{(j)} < \mathcal{U}^{(i)} < c^{(j+1)}$
- For each sample i, let the particle values have indexes as chosen by the above, i.e., $\boldsymbol{H}_t^{(i)} = \breve{\boldsymbol{H}}_t^{(j)}$
- Reset all the weights, i.e., $w_t^{(i)} = 1/N$

References

[1] Shumway, R., and D. Stoffer, 2006. Time Series Analysis and Its Applications. Springer, 2nd edition.

[2] Rabiner, L. R., 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* 72:257–286.

[3] Baum, L., T. Petrie, G. Soules, and N. Weiss, 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics* 41:164–171.

[4] Kalman, R., 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 82:35–45.

[5] Smith, A., A. Doucet, N. de Freitas, and N. Gordon, 2001. Sequential Monte Carlo Methods in Practice. Springer.

[6] Klaas, M., N. de Freitas, and A. Doucet, 2005. Toward Practical N2 Monte Carlo: the Marginal Particle Filter. *In* Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05). AUAI Press, Arlington, Virginia, 308–31.

[7] Douc, R., O. Cappe, and E. Moulines, 2005. Comparison of Resampling Schemes for Particle Filtering. *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on* 64–69.