Sequential inference and state number determination for discrete state-space models through particle filtering

Nicolas CHOPIN
Laboratoire de Statistique, CREST, INSEE,
Timbre J120, 75675 Paris cedex 14, France
chopin@ensae.fr

Abstract We investigate the application of particle filtering methodology to discrete state-space models, that is models which involve a conditioning on an unobserved discrete process. We first focus on the case where the states are Markov (hidden Markov models). In that setting, following Doucet et al. (2000), we show how to marginalize out the states z_t of the considered posterior distribution, in order to reduce the volatility of the particle weights. The corresponding algorithm can be seen as a Monte Carlo generalization of the HMM filter. We then propose a sequential state number determination procedure, in order to detect the number of distinct states that have actually appeared at time t. Within this approach, we show how to make an improper prior modeling possible. Finally we consider to which extent these results apply to hidden semi-Markov models, and to change-point models.
Key-words: change-point models, HMM filter, Hidden Markov Models, Hidden Semi-Markov Models, Metropolis-Hastings, MCMC.

Résumé Nous nous intéressons à l'application de la méthodologie des filtres particulaires aux modèles à espace-état discret (modèles conditionnés à un processus discret non observé). Lorsque le processus discret (z_t) est Markovien, nous montrons, à la suite de Doucet et al. (2000), comment intégrer sur ces z_t la loi a posteriori considérée. L'algorithme correspondant est une généralisation Monte Carlo du filtre HMM. Une procédure de détermination du nombre d'états est proposée, qui permet d'évaluer le nombre de régimes ayant réellement apparu dans la série étudiée. Nous prouvons que cette approche est compatible avec une modélisation a priori impropre. Une généralisation de ces résultats aux modèles semi-Markoviens, et aux modèles à points de changement est considérée.
Mots-clefs: filtre HMM, MCMC, Metropolis-Hastings, modèles à chaîne de Markov cachée, modèles à points de changement.

1 Introduction

Discrete state-space models have proven to be a valuable tool in a variety of areas, ranging from econometrics to genetics, from finance to speech processing. They are mostly designed to conveniently capture homogeneous subsequences in time series featuring global heterogeneity. The core hypothesis for this class of models is the existence of an underlying, unobserved discrete process (z_t), which gives the state of the system (regime) at time t. Each time z_t change its value, a shift in the behaviour of the observed process occurs. In some applications, the hidden process has a clear interpretation, but in others, it is just an artifact for modeling heterogeneity.

More formally, we adopt the following mixture representation. Given the state z_t, the observed process at time t, $t \geq 1$, verifies

$$y_t | \{z_t = k, y_1, ..., y_{t-1}\} \sim f_{\xi_k}(y_t | y_{1:t-1}), \tag{1}$$

where $\xi_1, ..., \xi_K$ are parameters corresponding to a given parametric family $\{f_\xi(.|.), \xi \in \Xi\}$ of conditional densities (with respect to an appropriate measure), K is the cardinal of the state-space, and $y_{1:t-1}$ stands for the sequence of observations $y_1, ..., y_{t-1}$ (for $\emptyset$ if $t = 1$). In the same manner, $z_{1:t}$ denotes the sequence $z_1, ..., z_t$. Most obviously, (1) also includes conditionally independent models. In that case, $f_\xi(y_t | y_{1:t-1})$ reduces to $f_\xi(y_t)$.

Equation (1) is commonly denoted the *observation equation*. The model specification is completed by the *system equation*, which states the inner dynamics of the hidden process (z_t). For instance, if the z_t's are Markov, we have

$$P(z_{t+1} = l | z_t = k) = p_{kl}, \tag{2}$$

where $(p_{kl})_{1 \leq k,l \leq K}$ is a given transition probability matrix (hidden Markov models), but other dynamics are possible (semi-Markov models, change point models) as we will see further.

Discrete state-space models, being dynamic models, lend themselves naturally to a Bayesian sequential analysis. In this setting, a new inference is drawn at each time t, which takes account of the available observations $y_1, ..., y_t$. More specifically, a sequential analysis may comprise at time t estimations of the past states $z_1, ..., z_{t-1}$ (*smoothing*), the current state z_t (*filtering*), the next state z_{t+1} (*forecast*), and the fixed estimates (forming a vector denoted θ).

The algorithms presented in this paper are devoted to a sequential Bayesian analysis of the discrete state-space models: the posterior distributions of

the form $\pi(z_{1:t}, \theta | y_{1:t})$, or some marginals of these distributions, are sequentially approximated by Monte Carlo sampling schemes. Note however that these algorithms are perfectly suited for a "direct" analysis of data, that is in applications where the whole sample $y_1, ..., y_T$ is available at once, and the only distribution of interest is $\pi(z_{1:T}, \theta | y_{1:T})$. In that case, only the final output of the algorithm is considered. In such a setting, the methods proposed here seem to outperform classical inference algorithms (namely the MCMC techniques in a Bayesian framework) in many cases, in terms of execution time.

Another interesting problem is to determine the cardinal of the discrete state-space, which gives the number of distinct regimes. In a sequential setting, this problem is rather intricate, since at time t, it may be that only some of the regimes appeared, while the others will appear later. Thus, it seems more sensible to evaluate the number which appeared *for the time being*. We will call such a procedure "sequential state number determination". As we will see, a strong appeal of this approach is that it overcomes in some cases the well-known incompatibility of discrete state-space models with improper priors, and therefore allows for a less informative prior modeling.

This paper is structured as follows. Section 2 recalls the basics of the particle filter methodology. Section 3 describes the Monte Carlo HMM filter, a particle filter method specially dedicated to the sequential analysis of the hidden Markov models. Section 4 presents the sequential state number determination procedure, and shows how to handle an improper prior within such a procedure. Section 5 provides some illustrative applications. Section 6 investigates how these results may be extended to semi-Markov and change-point models.

2 Particle filters

We give now a brief overview of particle filter methodology. For a more extensive survey, see for instance Doucet et al. (2001). A particle filter is an algorithm able to provide iterative Monte Carlo approximations for a given sequence of distributions of interest (π_t). In particular, a particle filter makes possible a sequential analysis of dynamic models. In such a setting, the distributions of interest are the posterior distributions of the form $\pi_t = \pi(\theta, z_{1:t} | y_{1:t})$ (following the notations of the previous section). Other applications exist (such as the estimation of static models, see Chopin, 2000), but will not be treated here.

More formally, a particle filter algorithm produces a sequence of weighted

Monte Carlo realizations (the particles), denoted here $(\theta^{(j)}, z_{1:t}^{(j)})$, with weights w_j, which, at every time t, *targets* the distribution π_t, in the sense that

$$\lim_{H \to +\infty} \frac{\sum_{j=1}^{H} w_j h(\theta^{(j)}, z_{1:t}^{(j)})}{\sum_{j=1}^{H} w_j} = E_{\pi_t}[h(\theta, z_{1:t})] \text{ almost surely} \qquad (3)$$

holds for any function h such that the expectation in the limit term is defined.

The simplest particle filter scheme is the sequential importance sampling algorithm (SIS). It consist in iterating two operations:

(E) extend the space: simulate $z_{t+1}^{(j)} \sim q(.|\theta^{(j)}, z_{1:t}^{(j)})$ and add this component to the particles

$$(\theta^{(j)}, z_{1:t}^{(j)}) \to (\theta^{(j)}, z_{1:t+1}^{(j)}).$$

(R) reweight: compute the *incremental weights*

$$u_{t+1}(\theta^{(j)}, z_{1:t+1}^{(j)}) = \frac{\pi_{t+1}(\theta^{(j)}, z_{1:t+1}^{(j)})}{\pi_t(\theta^{(j)}, z_{1:t}^{(j)}) q(z_{t+1}^{(j)}|\theta^{(j)}, z_{1:t}^{(j)})},$$

and update the weights

$$w_j \to w_j \times u_{t+1}(\theta^{(j)}, z_{1:t+1}^{(j)}).$$

Particles are usually initially simulated from π_0 (in this case, weights are initialized to 1). Note that $q(z_{t+1}|\theta^{(j)}, z_{1:t}^{(j)})$ is merely a instrumental distribution, in that it can be set to virtually any imaginable distribution. An "ideal" choice is the true conditional density $P(z_{t+1}|y_{1:t+1}, \theta)$, since it simplifies the computation of the incremental weights, and minimizes the conditional variance of the weights (in order to reduce the system degeneracy, see further), but this distribution is not always tractable. For more details on the SIS filter and derived methods, see Liu and Chen (1998).

It is well known that the SIS filter suffers from a progressive *degeneracy*: each iteration adds to the variability of the estimates. In fact, since particles always keep the same values, less and less particles are in a region of high π_t probability, when π_t is moving away from π_0, while more and more get an insignificant weight, and therefore contribute very weakly to the current inference.

The SIS filter can be accelerated by occasional *resampling* of the particles. Resampling consists in replacing each particle $(\theta^{(j)}, z_{1:t}^{(j)})$ by n_j of its replicates (n_j may be equal to zero). The new weights are set to 1. The n_j must be determined in such a way that condition (3) is still fulfilled. The most famous selection scheme is the multinomial resampling (Gordon et al., 1993), but more recent schemes (residual resampling, Liu and Chen, 1998; stratified resampling, Carpenter et al., 1999) seem today preferable, for they reduce the Monte Carlo variability of the estimates.

Resampling the particles from time to time discards particles of insignificant weight, and therefore saves further execution time. However, it does not prevent the particle system from degenerating, since no new particle value is created. To counter this, Gilks and Berzuini (2001) proposed to *move* the resampled particles, by applying a transition kernel with stationary distribution π_t. This kernel is usually chosen in the standard toolkit of the MCMC (Monte Carlo Markov Chains) methodology (see Robert and Casella, 1999, for a thorough presentation). Moving the particles is clearly an appealing idea, since it replaces n_j identical replicates of a single particle by the same number of "fresh" particles, without modifying the current target π_t of the system.

To conclude, particle refer algorithms now refer to iterative algorithms alternating the three following stages:

---

1. Extend and reweight:
   Apply steps (E) and (R) $(t' - t)$ times, from date t to date $t' > t$

2. Resample (according to a given selection scheme):

$$(\theta^{(j)}, z_{1:t'}^{(j)}, w_j) \rightarrow (\widetilde{\theta}^{(j)}, \widetilde{z}_{1:t'}^{(j)}, 1),$$

   for $j = 1, ..., H$, where the $(\widetilde{\theta}^{(j)}, \widetilde{z}_{1:t'}^{(j)})$'s are the resampled particles.

3. Move:

$$(\widetilde{\theta}^{(j)}, \widetilde{z}_{1:t'}^{(j)}) \rightarrow K_{t'}(\widetilde{\theta}^{(j)}, \widetilde{z}_{1:t'}^{(j)}),$$

   for $j = 1, ..., H$,where $K_{t'}$ is a transition kernel of invariant measure $\pi_{t'}$

Important parameters for these algorithms are the choice of proposal distribution for the space extension $q(.|\theta, z_{1:t})$, and of the MCMC kernel for the

move step, which both strongly affect the efficiency of the algorithm. Another tuning to consider is the resample-move schedule. Usually, particles are resampled-moved when some empirical degeneracy criterion is fulfilled, for instance when the empirical variance of the weights reaches a given threshold (Liu and Chen, 1998). For more elaborate criteria, see for instance Carpenter et al. (1999, §5) or Chopin (2000, §4.2).

3 The Monte Carlo HMM filter

We now restrict ourselves to hidden Markov models. Thus, we consider a model specified by equations (1) and (2). The vector θ of the fixed parameters comprises the mixture parameters $\xi_1, ..., \xi_K$ and the components of the transition matrix $P = (p_{kl})_{1 \leq k,l \leq K}$.

3.1 HMM filter

Suppose θ is known, so that we merely need to infer the states (past, present or future) from the current observations. Remarkably, the corresponding distributions can be derived *exactly*, through iterative formulae (Hamilton, 1989). More precisely, since a state $z_{t'}$ is discrete, it follows, conditionally on a given set of observations $y_{1:t}$, a multinomial distribution $\mathcal{S}_t^{t'}(\theta)$. Denote $S_t^{t'}(\theta)$ the vector of the corresponding probabilities $P(z_{t'} = k | y_{1:t}, \theta)$, for $k = 1, ..., K$.

At time t, the forecast probabilities $S_t^{t+1}(\theta)$ can be derived from the filtering probabilities $S_t^{t+1}(\theta)$ with

$$S_t^{t+1}(\theta) = P' S_t^t(\theta), \tag{4}$$

where P is the transition matrix defined by the corresponding components of θ. Then, the filter probabilities at time $t+1$ can be obtained from the forecast $S_t^t(\theta)$

$$S_{t+1}^{t+1}(\theta) \propto O_{t+1}(\theta) \otimes S_t^{t+1}(\theta), \tag{5}$$

where $O_{t+1}(\theta)$ is the vector containing the observation densities $f_{\xi_k}(y_{t+1} | y_{1:t}, \theta)$, for $k = 1, ..., K$, and $\otimes$ denotes the element-by-element product of two vectors. The later formula gives $S_{t+1}^{t+1}(\theta)$ up to a multiplicative constant which can retrieved by normalization (the probabilities sum to one). The right term $O_{t+1}(\theta) \otimes S_t^{t+1}(\theta)$ is a vector containing the densities $P(z_{t+1} = k, y_{t+1} | y_{1:t}, \theta)$, for $k = 1, \ldots, K$, and equation (5) is a Bayes formula in disguise.

Finally, the state z_t can be smoothed at time $t+1$ (Kitagawa, 1987) with

$$S_{t+1}^{t}(\theta) \propto S_{t}^{t}(\theta) \otimes [PO_{t+1}(\theta)], \tag{6}$$

and more generally, for $k \geq 0$,

$$S_{t+1}^{t-k}(\theta) \propto S_{t-k}^{t-k}(\theta) \otimes \left\{ P\left[S_{t+1}^{t-k+1}(\theta) \oslash S_{t-k}^{t-k+1}(\theta) \right] \right\}, \tag{7}$$

where $S_{t+1}^{t-k+1}(\theta) \oslash S_{t-k}^{t-k+1}(\theta)$ denotes the element-by element division of $S_{t+1}^{t-k+1}(\theta)$ by $S_{t-k}^{t-k+1}(\theta)$. The algorithm which sequentially forecasts and filters through formulae (4) and (5) is usually referred to the HMM filter. If smoothing steps are added, we will rather speak of the Kitagawa-HMM filter. The matrix formulation of equations (4) to (7) is adapted from Ryden (2000).

3.2 Particle filtering for the HMMs

Suppose now θ is unknown, and assign some prior distribution π. In comparison with the "naïve" particle filter proposed in §2, we present here an *integrated* filter, which marginalizes out the states z_t of the sequence of the target distributions. The key idea is to follow the evolution of numerous HMM filters running in parallel, each of which is initiated with a distinct value $\theta^{(j)}$ for the fixed parameter θ. These $\theta^{(j)}$, along with the corresponding HMM filters, are the particles of our system. We call this integrated particle filter algorithm the Monte Carlo HMM filter (MCHF). Note this marginalizing technique, along with the Rao-Blackwell argument presented in §3.3, can be seen as a particular case of the more general particle filter Rao-Blackwellisation scheme developed in Doucet et al. (2000).

More formally, we design a particle filter algorithm which tracks the sequence of posterior distributions $\pi(\theta|y_{1:t})$. Thus, the considered distributions and the corresponding particles are now of *constant* dimension, and the (E) step (Extend the space) presented in §2 is no longer necessary. The first stage of a particle filter algorithm reduces in this setting to the (R) step (reweight), which consists in multiplying the weights by some incremental weights $u_{t+1}(\theta^{(j)})$,

$$\begin{aligned} u_{t+1}(\theta^{(j)}) &\propto \frac{\pi(\theta^{(j)}|y_{1:t+1})}{\pi(\theta^{(j)}|y_{1:t})} \\ &\propto P(y_{t+1}|y_{1:t}, \theta^{(j)}) \propto \sum_{k=1}^{K} P(z_{t+1} = k, y_{t+1}|y_{1:t}, \theta^{(j)}), \end{aligned}$$

and this sum is derived in the following way. Let $(\theta^{(j)}, w_j)_{j=1,\dots,H}$, a set of particles currently targeting $\pi(\theta|y_{1:t})$, and suppose we have at our disposal the corresponding filter probabilities $S_t^t(\theta^{(j)})$. Compute the forecast probabilities $S_t^{t+1}(\theta^{(j)})$, through (4), and, when the next observation y_{t+1} is available (at time $t+1$) compute the new filter probabilities $S_{t+1}^{t+1}(\theta^{(j)})$, with (5). Then $u_{t+1}(\theta^{(j)})$ is a direct by-product of the later computation, since it is the multiplicative constant induced by the normalization of $S_{t+1}^{t+1}(\theta)$ in (5) (see previous subsection). The parallel HMM filters and the weights of the particles are updated simultaneously.

The Monte Carlo HMM filter will consist in iterating the reweighting scheme described above. It can also include occasional resample-move steps, as explained in §2. In that case, kernels of invariant measure of the form $\pi(\theta|y_{1:t})$ must be designed (see §3.4).

The Monte Carlo HMM filter provides sequential estimations for the fixed parameter θ, but also for the filter and forecast distributions. For θ, any expectation of the form $E_{\pi(\theta|y_{1:t})}[h(\theta)]$ can be consistently estimated by the corresponding Monte Carlo weighted average $\sum_{j=1}^H w_j h(\theta^{(j)}) / \sum_{j=1}^H w_j$. The filter distribution $\pi(z_t|y_{1:t})$ and the forecast distribution $\pi(z_{t+1}|y_{1:t})$ are multinomial distributions, and the vectors of the corresponding probabilities, respectively denoted S_t^t and S_t^{t+1}, are consistently estimated by :

$$\widehat{S}_t^t = \frac{\sum_{j=1}^H w_j S_t^t(\theta^{(j)})}{\sum_{j=1}^H w_j}, \qquad \widehat{S}_t^{t+1} = \frac{\sum_{j=1}^H w_j S_t^{t+1}(\theta^{(j)})}{\sum_{j=1}^H w_j}. \tag{8}$$

Note that it is also possible to estimate the smoothing distributions $\pi(z_{t-k}|y_{1:t})$ (for $k > 0$), by applying the smooth step of the Kitagawa-HMM filter, presented in the previous section, and then marginalizing out θ, as in (8). In that case, we will rather speak of the Monte Carlo Kitagawa-HMM filter (MCKHF). This second algorithm is more memory demanding (the intermediary $S_t^t(\theta^{(j)})$'s must be stored) and is more intensive (at each iteration t, step (7) is applied t times). In applications where the execution time of one iteration must be kept constant, it is more reasonable to restrict the smoothing to the m later states $z_{t-m:t-1}$ (fixed-lag smoothing).

3.3 Advantages of the integrated filters

The integrated filters (MCHF and MCKHF) must be preferred to the equivalent "standard" particle filters (as described in §2) for two reasons. First, they strongly reduce the Monte Carlo sampling space dimension. This implies considerable memory savings, especially for the MCHF: for standard filters, particle dimension increases linearly in time, while for integrated filters,

it keeps a constant value. In fact, to be more precise, it is possible to devise a constant dimension standard filter in case smoothing is not necessary: in that case, $\pi(z_t, \theta|y_{1:t})$ is the target at time t, and the reweighting from t to $t+1$ consist this time in drawing a $z_{t+1}^{(j)}$ for each particle, then computing the corresponding incremental weight, and finally discard the component $z_t^{(j)}$ for each particle. However, as we will explain more broadly in next subsection, a move strategy is hardly implementable with such a filter. Alternatively, for the MCKHF, the memory requirements keep increasing linearly, since the vectors $S_t^t(\theta^{(j)})$ must be stored along with the particles $\theta^{(j)}$, for further smoothing operations. However, if only a constant-horizon smoothing is applied, important memory savings are still achievable.

The second reason for preferring the integrated filters is that they allow for a smaller variability of the particle weights, and therefore are likely to degenerate more slowly. This can be seen by a simple Rao-Blackwell argument. Suppose we have, at time 0, particles $\theta^{(j)}$ targeting $\pi(\theta)$ (the prior distribution). The first iteration of the algorithm induces a reweighting with incremental weights $u_I(\theta^{(j)}) = \pi(\theta^{(j)}|y_1)/\pi(\theta^{(j)})$, for an integrated filter, and $u_S(\theta^{(j)}, z_1^{(j)}) = \pi(\theta^{(j)}, z_1^{(j)}|y_1)/q(z_1^{(j)}|\theta^{(j)})\pi(\theta^{(j)})$ for a standard filter, where $z_1^{(j)}$ was simulated from some given proposal conditional distribution $q(.|\theta^{(j)})$. We have

$$E[u_S(\theta^{(j)}, z_1^{(j)})|\theta^{(j)}] = \int \frac{\pi(\theta^{(j)}, z_1^{(j)}|y_1)}{q(z_1^{(j)}|\theta^{(j)})\pi(\theta^{(j)})} q(z_1^{(j)}|\theta^{(j)}) dz_1^{(j)} = u_I(\theta^{(j)})$$

and therefore $V[u_I(\theta^{(j)})] = V\{E[u_S(\theta^{(j)}, z_1^{(j)})|\theta^{(j)}]\} \leq V[u_S(\theta^{(j)}, z_1^{(j)})]$ (Rao-Blackwell inequality).

Obviously, the inequality $V[u_I(\theta^{(j)})] \leq V[u_S(\theta^{(j)}, z_{1:t}^{(j)})]$ stands more generally at any time t. Since the incremental weights at each stage shows a lesser volatility in an integrated filter, the weights themselves tend to vary less.

This second reason can be stated more intuitively. In fact, both kinds of filters perform very similar computations of probabilities of multinomial distributions. The difference is that an integrated filter carries forward these probabilities and use them in later computations, while a standard filter only keeps one simulated realization from this multinomial distribution and discards these probabilities. Obviously, more precise results are expected from a method able to profit from a richer information, such as for instance when taking into account the exact knowledge of a given distribution rather than a single realization of it.

3.4 Move strategies

For the move step, various advantages of moving the particles through an independent Hastings-Metropolis kernel, rather than any other MCMC strategy, have been discussed in Chopin (2000). Among others, it allows for an easier and more efficient control of the current level of degeneracy of the particle system. The proposal distribution for Hastings-Metropolis kernels may be set according to the considered problem, but a convenient all-purpose proposal, such as for instance a joint distribution with Gaussian parts for the ξ_i's and Dirichlet parts for the lines of the transition matrix, properly fitted through the moments of the particle system itself (in the same manner than in Chopin, 2000) should do in many settings. Besides, such an instrumental distribution makes the algorithm a "black-box", in that the internal machinery of the algorithm is not model-dependent, and therefore the adaptation cost to a new model reduces to modify some external routines dedicated to evaluating the observation likelihood $f_\xi(.|.)$ or the prior density.

Notice that performing a Hastings-Metropolis move of particles targeting $\pi(\theta|y_{1:t})$ requires to be able to compute this density for any θ (up to a multiplicative constant), in order to derive the acceptance probability. This is done by computing iteratively the ratio $\pi(\theta|y_{1:t'+1})/\pi(\theta|y_{1:t'})$, whose derivation was already described in §3.2. Besides, as we indicated in previous subsection, a move strategy would be hardly implementable for a constant dimension standard filter, that is targeting at each time t $\pi(\theta, z_t|y_{1:t})$: a Hasting-Metropolis move would require to compute the density $\pi(\theta, z_t|y_{1:t})$, which cannot be done anyhow without marginalizing the previous states as in an integrated filter, whereas a Gibbs move would need in most cases to simulate the previous states $z_{1:t-1}$ (hence the memory saving would be lost).

4 Sequential state number determination

4.1 Prior modeling for the hidden Markov models

Recall the parameter θ comprises for the hidden Markov models the mixture components $\xi_1, ..., \xi_K$ and the transition probabilities $(p_{kl})_{1\leq k,l\leq K}$ of the hidden Markov process. The kth line $(p_{k1}, ..., p_{kK})$ of the transition matrix will be denoted $p_{k|}$ from now on. We suppose the prior distribution on θ factorizes in the following way:

$$\pi(\theta) \propto \prod_{k=1}^{K} \pi(\xi_k) \prod_{k=1}^{K} \pi(p_{k|}) D(\xi_1, ..., \xi_K), \tag{9}$$

where $D(\xi_1, ..., \xi_K)$ is a *discriminating factor* between the K components, that is a continuous function which verifies $D(\xi_1, ..., \xi_K) \in [0, 1]$ and $D(\xi_1, ..., \xi_K) = 0$ as soon as two components take the same value. As discussed more broadly in §4.4, such a factor allows for penalizing weak identifiability regions in the parameter space. However, introducing $D(\xi_1, ..., \xi_K)$ in the prior does not modify its major features, notably its (im)propriety, since it is a bounded quantity.

Two problems arises with such a prior distribution. First, a mixture model is invariant by permutation of its components, and therefore is not fully identified with (9). Usually a ordering constraint on the ξ_k's ($\xi_1 < ... < \xi_K$) or a given coordinate of the ξ_k's (if they are multi-dimensional) is added to force the distinction between components. Unfortunately, such a constraint often hinders inference (Celeux et al., 2000). Secondly, a mixture model is in most cases incompatible with a fully non-informative approach. In particular, an improper distribution for $\pi(\xi)$ in (9) will commonly lead to an improper posterior distribution.

When no prior information is available on the parameters, an "objective" inference can still be achieved by specifying a partially proper prior, in the same spirit than Diebolt and Robert (1994) and Wasserman (2000). Given a prior $\pi(\xi)$ for the components, which is supposed to be *improper of order* k, that is k is the smallest integer such that, for any $y_{1:t}$, $1 \leq t_1 \leq ... \leq t_k = t$,

$$\int \pi(\xi) \prod_{i=1}^{k} f_\xi(y_{t_i}|y_{1:t_i-1})\, d\xi < +\infty$$

the whole posterior distribution is made proper by conditioning on the fact that at least k observations of the whole sample are assigned to each component (see the appendix for a proof of this claim). Wasserman (2000) showed that this approach was equivalent to specifying a data-dependent prior.

The two raised problems are more acute in a sequential context. At early stages, when only a small amount of observations is at our disposal, it is likely that some of the K components have not appeared yet. An ordering constraint has no clear meaning in that setting, since only the already detected components can be ordered, and we have no clue where the later components should be inserted in this partial ordering. Thus, the posterior distribution will show some intricate mixture structure which does not have a clear interpretation. But above all, a data-dependent prior has not sense here, since all data are not available at once.

Clearly, a new prior specification is required in our setting. It is in fact more sensible to sort the components by *order of appearance*, and to

provide, at time t, a joint estimation of m, the number of components that have appeared for the time being, and the corresponding $\xi_1, ..., \xi_m$. We will call this procedure *sequential state number determination*. We present this procedure in the following section, and show it makes less informative prior specification possible. Such a procedure can even be made deliver an inference on K itself, the total number of components, as we will see in §4.4.

4.2 Sequential reparameterization

We now propose a sequential reparameterization of the hidden Markov models. The observation equation is unchanged

$$y_t|\{z_t = k, y_1, ..., y_{t-1}\} \sim f_{\xi_k}(y_t|y_{1:t-1}),$$

but the system equation now features an additional hidden process (M_t), where M_t stands for the number of components appeared at time t

$$\begin{aligned} z_1 &= 1, \\ P(z_{t+1} = l|z_t = k, M_t = m) &= \begin{cases} p_{kl} & \text{if } k, l \leq m \leq K, \\ \sum_{l'=m+1}^{M} p_{kl'} & \text{if } l = m+1 \leq K, \\ 0 & \text{otherwise,} \end{cases} \\ M_1 &= 1, \\ M_{t+1} &= \max(M_t, z_{t+1}). \end{aligned}$$

When at time t, with $z_t = k$ and $M_t = m$ $(k \leq m)$, the next regime can be either an already visited regime l $(l \leq m)$ with probability p_{kl}, or a new regime, which will be labelled $m+1$. Since the remaining regimes are not distinguishable at time t, the probability of appearance is indeed $\sum_{l'=m+1}^{M} p_{kl'}$. If a new regime appears, we have $M_{t+1} = z_{t+1} = m+1$, if not, $M_{t+1} = m$, hence in general $M_{t+1} = \max(M_t, z_{t+1})$.

Since the hidden process (z_t, M_t) is clearly Markov and discrete (lying in a space of cardinal $K(K+1)/2$), this reparameterized model can still be analyzed as a hidden Markov model. But this time the complete posterior $\pi(\theta|y_{1:t})$ is not a distribution of direct interest. Rather, following the lines of the sequential state number determination procedure presented in the previous section, and denoting $\theta_{1:m}$ the partial vector $(\xi_1, ..., \xi_m, p_{1|}, ..., p_{m|})$, we merely need to evaluate the following marginals, for $1 \leq m \leq K$

$$\pi(M_t = m|y_{1:t}), \qquad \pi(\theta_{1:m}|M_t = m, y_{1:t}), \tag{10}$$

in order to recover the components that have appeared for the time being.

4.3 Towards a non-informative prior modeling

If we specify a proper prior distribution for the ξ_k's, the sequential state number determination procedure reduces to apply the Monte Carlo HMM filter on the reparameterized model. In particular, the probabilities $\pi(M_t = m|y_{1:t})$ can be estimated when filtering the state (M_t, z_t) (see §4.5 for implementation details). However, our aim here is to show more generally that this new writing of the model allows for less informative prior specification. In fact, we show now that an improper distribution $\pi(\xi)$ for the components, while making the complete posterior distribution improper, can still allow for a satisfying definition of the true quantities of interest in (10). Let

$$\pi(\theta_{1:m}) \propto \prod_{k=1}^{m} \pi(\xi_k) \prod_{k=1}^{m} \pi(p_{k|}) D_m(\xi_1, ..., \xi_m), \tag{11}$$

$$\pi(\theta_{1:m}|M_t = m, y_{1:t}) \propto \pi(\theta_{1:m}) P(M_t = m|\theta_{1:m}) P(y_{1:t}|M_t = m, \theta_{1:m}), \tag{12}$$

$$\pi(M_t = m|y_{1:t}) \propto \int \pi(\theta_{1:m}) P(M_t = m|\theta_{1:m}) P(y_{1:t}|M_t = m, \theta_{1:m})\, d\theta_{1:m}, \tag{13}$$

where $D_m(\xi_1, ..., \xi_m)$ is a partial discriminating factor between the m first components (see §4.4), $\pi(\theta_{1:m})$ is somehow the marginal prior distribution of $\theta_{1:m}$ (while rigorously speaking the marginals of an improper distribution are not defined), $P(M_t = m|\theta_{1:m})$ in fact only depends on $p_{1|}, ..., p_{m|}$, and is given by the system equation, and $P(y_{1:t}|M_t = m, \theta_{1:m})$ is the likelihood of the observations, given that the m first components are visited at least once, and the $(K - m)$ remaining components are not.

Lemma 1 *Provided $\pi(\xi)$ is improper of order 1, the distributions $\pi(M_t|y_{1:t})$ and $\pi(\theta_{1:m}|M_t = m, y_{1:t})$, as defined through equations (11) to (13), are proper.*

A proof is given in the appendix.

Alternatively, a non-informative inference can be managed through a smart hierarchical modeling. Consider for instance this appealing Gaussian hierarchical prior, in case the ξ_k's are univariate:

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}, \qquad \xi_1, ..., \xi_K|\{\mu, \sigma^2\} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2).$$

This hierarchical structure generally does not allow for a proper inference in a mixture context, but through the simple reformulation

$$\pi(\xi_1) \propto 1, \quad \pi(\sigma^2|\xi_1) \propto \frac{1}{\sigma^2}, \quad \mu|\{\xi_1, \sigma^2\} \sim \mathcal{N}(\xi_1, \sigma^2),$$

$$\xi_2, ..., \xi_K|\{\mu, \sigma^2, \xi_1\} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2),$$

we get, as shown by the following lemma, consistent definitions for the quantities produced by our sequential order choice procedure. Note that μ was kept in the second formulation only for an illustrative purpose, but we now discard it, by marginalizing μ out in the last equation:

$$\pi(\xi_1) \propto 1, \quad \pi(\sigma^2|\xi_1) \propto \frac{1}{\sigma^2}, \quad \xi_2, ..., \xi_K|\{\xi_1, \sigma^2\} \overset{iid}{\sim} \mathcal{N}(\xi_1, 2\sigma^2) \tag{14}$$

More generally, if we want to assign to the components a hierarchical improper prior of the form

$$\pi(\xi_1)\pi(\lambda|\xi_1) \prod_{k=2}^{K} \pi(\xi_k|\xi_1, \lambda),$$

where λ is the hyper-parameter, the sequential state number determination procedure be adapted in the following way. Let $\pi(\theta_1) = \pi(\xi_1)\pi(p_{1|})$ and for $m > 1$,

$$\pi(\theta_{1:m}, \lambda) = \pi(\xi_1)\pi(\lambda|\xi_1) \prod_{k=2}^{m} \pi(\xi_k|\xi_1, \lambda) \prod_{k=1}^{m} \pi(p_{k|}) D_m(\xi_1, ...\xi_m),$$

then replace in equations (12) and (13) the term $\theta_{1:m}$ by $(\theta_{1:m}, \lambda)$ when $m > 1$.

Lemma 2 *The sequential state number determination for an hierarchical improper prior, as defined above, provides proper definitions for the distributions $\pi(M_t|y_{1:t})$ and $\pi(\theta_{1:m}, \lambda|M_t = m, y_{1:t})$, provided that, for any $y_{1:t}$, $t \geq m$, $1 < t_2 < ... < t_m \leq t$, $m = 2, ..., K$,*

$$\int \pi(\xi_1) f_{\xi_1}(y_1)\, d\xi_1 < +\infty,$$

$$\int \pi(\xi_1)\pi(\lambda|\xi_1) \prod_{k=2}^{m} \pi(\xi_k|\xi_1, \lambda) D_m(\xi_1, ...\xi_m) \prod_{k=1}^{m} f_{\xi_k}(y_{t_k}|y_{1:t_k-1})\, d\lambda d\xi_1 ... d\xi_m < +\infty.$$

In particular, for the Gaussian hierarchical prior of (14), where $\lambda = \sigma^2$, the previous conditions are equivalent to, respectively,

$$\int f_{\xi_1}(y_{t_1}|y_{1:t_1-1})\,d\xi_1 < +\infty,$$

$$\int \left[\sum_{k=2}^{m}(\xi_k - \xi_1)^2\right]^{-\frac{m-1}{2}} D_m(\xi_1,...,\xi_K)\prod_{k=1}^{m} f_{\xi_k}(y_{t_k}|y_{1:t_k-1})\,d\xi_1...d\xi_m < +\infty,$$

for any $y_{1:t}$, $1 < t_1 < ... < t_m \leq t$, $m = 2,...,K$.

For a proof, see the appendix.

4.4 From sequential state number determination choice to global state number determination

The design of the sequential state number procedure we presented above is in fact incomplete. Since a state number determination is performed at each stage, a balance between the fit of the data and the parsimony of the model must be devised. If not, our sequential state number determination procedure may tend to saturate quicker that necessary the number of components M_t to K, by creating vaguely distinct components. This would eventually endanger the following estimation stages.

From a decisional point of view, it is the decision maker task to specify to which extent two components can be distinguished, for instance by introducing in the prior distribution a discriminating factor of the like

$$D(\xi_1,...,\xi_K) = \prod_{1\leq i<j\leq K}[1 - \exp(-C\Delta(\xi_i,\xi_j)^\nu)], \tag{15}$$

where $\Delta(\xi_i,\xi_j)$ is some distance between components ξ_i and ξ_j, and C and ν are tuning parameters.

Accordingly, the partial discriminating factor $D_m(\xi_1,...,\xi_m)$ needed in the non-informative state number determination procedure of §4.2 may be defined as the marginal discriminating factor of the m first components, by marginalizing out the supplementary components against some proper distribution $q(\xi)$:

$$D_m(\xi_1,...,\xi_m) = \int D(\xi_1,...,\xi_K)q(\xi_{m+1})...q(\xi_K)\,d\xi_{m+1}...d\xi_K. \tag{16}$$

This may seem an arbitrary definition at first, for it depends of the choice of q. This dependence will be negligible however, provided the support of q is

large enough, and besides all it allows for a much simpler implementation, as we will see in §4.5.

Note Lemma 2 in the previous section gives another justification for introducing a discriminating factor between the components, since the related integrals whose finiteness ensure the consistency of a hierarchical improper prior modeling diverge as $\xi_k \to \xi_1$, $k > 1$, without a (correctly set) partial discriminating factor $D_m(\xi_1, ..., \xi_m)$ in the integrand.

A sensible choice for $\Delta(\xi_i, \xi_j)$ is the (expected) Kullback-Leibler divergence between the observation densities $f_{\xi_i}(y_t|y_{1:t-1})$ and $f_{\xi_j}(y_t|y_{1:t-1})$, provided this quantity is tractable. By Kullback-Leibler divergence between two densities g and h, we mean the sum of the Kullback informations $I_K(g|h) + I_K(h|g)$, where $I_K(g|h) = E_g \log[g(X)/h(X)]$. The quantity $\Delta(\xi_i, \xi_j)$ somehow measures to which extent two distinct components ξ_i and ξ_j predict a distinct behaviour for the observations. See §5 for a derivation of $\Delta(\xi_i, \xi_j)$ in a practical case.

When studying a finite sample of size T, a "global" state number determination can be performed by simply estimating M_T at the last stage of the algorithm. Note however that such a procedure will provide biased results if the p_{kl}'s are assigned "standard" prior distributions, such as Dirichlet distributions. With such a prior, the event that the observed sequence is produced from a sub-model of order $K' < K$, that is the probabilities p_{kl}, for $K' < l \leq K$, are null, is assigned a null probability, and accordingly M_T is assumed to converge to K as T goes toward infinity with probability one. In this connection, a mixture Dirichlet prior for the p_{kl}'s of the form

$$K' \sim \mathcal{U}[1, K], \text{ and, conditionally on K'=k',}$$

$$(p_{k1}, ...p_{kk'}) \sim \mathcal{D}(\alpha_{k1}^{(k')}, ..., \alpha_{kk'}^{(k')}),\ p_{k(k'+1)} = ... = p_{kK} = 0 \text{ almost surely}$$

is more appropriate for managing a global state number determination.

When dealing with vectors of transition probabilities, most authors prescribe a symmetric prior such like a Dirichlet $\mathcal{D}(1, ..., 1)$. We feel however that the probabilities p_{kk} of *staying* in a given state k should be distinguished from the probabilities p_{kl}, $l \neq k$ of *leaving* this state k. In fact, it is common (prior) knowledge that the diagonal terms of the transition matrix are close to 1 in most interesting settings, or, to put it in another way, that components with a small staying probability p_{kk}, and therefore a very short staying time, would be hardly identifiable and interpretable in practice. In this connection, we advice to set $\alpha_{kl}^{(k')} = \alpha_{\rightarrow}^{(k')}$, for $k \neq l$, and $\alpha_{kk}^{(k')} = \alpha_{\circlearrowleft}^{(k')}$, with $\alpha_{\circlearrowleft}^{(k')} \gg \alpha_{\rightarrow}^{(k')}$.

4.5 Implementation issues

In the simplest setting, when the prior distribution is proper, the Monte Carlo HMM filter can be applied straightforwardly to the reparameterized version of a hidden Markov model. In particular, by summing over k the filtering densities $P(z_t = k, M_t = m|\theta^{(j)}, y_{1:t})$, $1 \leq k \leq m \leq K$, and then marginalizing out θ by computing the corresponding weighted Monte Carlo average $\sum w_j P(M_t = m|\theta^{(j)}, y_{1:t}) / \sum w_j$, we get a consistent estimate for the quantity $P(M_t = m|y_{1:t})$, $m = 1, ..., K$. Moreover, since, for any m, $1 \leq m \leq K$,

$$\frac{\pi(\theta|M_t = m, y_{1:t})}{\pi(\theta|y_{1:t})} \propto P(M_t = m|\theta, y_{1:t}),$$

we can derive from the current particle system $(\theta^{(j)}, w_j)$, targeting $\pi(\theta|y_{1:t})$, a supplementary partial particle system $(\theta^{(j)}_{1:m}, w^{(m)}_j)$, for each m, targeting the conditional posterior distribution $\pi(\theta_{1:m}|M_t = m, y_{1:t})$, through the simple reweighting scheme $w^{(m)}_j = w_j P(M_t = m|\theta^{(j)}, y_{1:t})$.

For handling an improper prior $\pi(\xi)$ for the components, in the lines of the non-informative inference procedure presented in §4.2, a distinctive strategy must be derived. First consider an improper prior $\pi(\xi)$ which fulfills conditions of Lemma 1. Replace $\pi(\xi)$ by a proper, instrumental prior $q(\xi)$, and apply the Monte Carlo HMM Filter. Denote $q(\theta|y_{1:t})$ the target at time t, that is the posterior distribution corresponding to prior q, then through the following reweighting scheme

$$u^{(m)}_t(\theta) \propto \frac{\pi(\theta_{1:m}|M_t = m, y_{1:t})}{q(\theta|y_{1:t})} \propto \frac{\pi(\xi_1)...\pi(\xi_m)}{q(\xi_1)...q(\xi_m)} P(M_t = m|\theta, y_{1:t})$$

we get, as above, consistent inference from the partial posteriors $\pi(\theta_{1:m}|M_t = m, y_{1:t})$. Note that we assume the instrumental prior q is also the proper distribution from which are derived in (16) the marginal discriminating factors $D_m(\xi_1, ..., \xi_m)$ (see §4.4). Besides, the $\pi(M_t = m|y_{1:t})$'s are also easily evaluated through the following lemma.

Lemma 3 *A consistent estimate of $\pi(M_t = m|y_{1:t})$ (up to a multiplicative constant, which does not depend on m) is the weighted average*

$$\frac{\sum_{j=1}^{H} w^{(m)}_j P(M_t = m|\theta^{(j)}, y_{1:t})}{\sum_{j=1}^{H} w^{(m)}_j},$$

where the particle system $(\theta^{(j)}, w_j)$ is assumed to target $q(\theta|y_{1:t})$ and the

$w_j^{(m)}$*'s verify*

$$w_j^{(m)} \propto w_j \frac{\pi(\xi_1)...\pi(\xi_m)}{q(\xi_1)...q(\xi_m)}$$

See the appendix for a proof.

The instrumental distribution q should be set so that the incremental weights defined above does not vary too much over the support of $q(\theta|y_{1:t})$. A minimal requirement seems to check the existence of moment of order 2 of $u_t^{(m)}(\theta)$ over $q(\theta|y_{1:t})$. In that case, we will speak of a *valid* reweighting operation. The following lemma gives a sufficient condition for ensuring the validity of the reweighting operation from q to π.

Lemma 4 *Assume $\pi(\xi)$ is improper of order 1, and*

$$\int \frac{\pi(\xi)^2}{q(\xi)} f_\xi(y_t|y_{1:t-1})\, d\xi < +\infty$$

holds for any sequence $y_{1:t}$, then the reweighting operation defined above is valid.

For a proof, see the appendix.

Alternatively, for a hierarchical improper prior which fulfills conditions of Lemma 2, start again the algorithm with a proper instrument $q(\xi)$. At time t, extend the space in λ, by drawing for each particle $\theta^{(j)}$ a $\lambda^{(j)}$ from some conditional instrumental distribution $q_m(\lambda|\xi_1, ..., \xi_m)$, and reweight through

$$\frac{\pi(\theta_1|M_t = 1, y_{1:t})}{q(\theta|y_{1:t})} \propto \frac{\pi(\xi_1)}{q(\xi_1)} P(M_t = 1|\theta, y_{1:t}),$$

$$\frac{\pi(\theta_{1:m}, \lambda|M_t = m, y_{1:t})}{q(\theta|y_{1:t}) q_m(\lambda|\xi_1, ..., \xi_m)} \propto \frac{\pi(\xi_1)\pi(\lambda|\xi_1) \prod_{k=2}^m \pi(\xi_k|\xi_1, \lambda)}{q(\xi_1)...q(\xi_m) q_m(\lambda|\xi_1, ..., \xi_m)} P(M_t = m|\theta, y_{1:t}),$$

in order to infer from the partial posteriors $\pi(\theta_1|M_t = 1, y_{1:t})$ and $\pi(\theta_{1:m}, \lambda|M_t = m, y_{1:t})$, $m > 1$. Similarly, consistent estimates of the $\pi(M_t = m|y_{1:t})$ are derived through the same lines as in Lemma 3, where this time the weights $w_j^{(m)}$ are set to

$$w_j^{(1)} \propto w_j \frac{\pi(\xi_1)}{q(\xi_1)}, \qquad w_j^{(m)} \propto w_j \frac{\pi(\xi_1)\pi(\lambda|\xi_1) \prod_{k=2}^m \pi(\xi_k|\xi_1, \lambda)}{q(\xi_1)...q(\xi_m) q(\lambda|\xi_1, ..., \xi_m)}.$$

Finally, a sufficient condition for the validity of this reweighting operation is given by the following lemma.

Lemma 5 *Suppose conditions of Lemma 2 are fulfilled, and*

$$\int \frac{\pi(\xi_1)^2}{q(\xi_1)} f_{\xi_1}(y_1)\, d\xi_1 < +\infty,$$

$$\int \frac{\left[\pi(\xi_1)\pi(\lambda|\xi_1)\prod_{k=2}^{m}\pi(\xi_k|\xi_1,\lambda)\right]^2}{q(\xi_1)...q(\xi_m)q_m(\lambda|\xi_1,...,\xi_m)} D_m(\xi_1,...,\xi_K) \prod_{k=1}^{m} f_{\xi_k}(y_{t_k}|y_{1:t_k-1})\, d\lambda d\xi_1...d\xi_m < +\infty.$$

hold for any $y_{1:t}$, $t \geq m$, $1 < t_2 < ... < t_m \leq t$, $m = 2,...,K$, *then the reweighting strategy defined above for a hierarchical improper prior is valid. In particular, for the Gaussian hierarchical prior of (14),* $\lambda = \sigma^2$, *if we set* $q_m(\sigma^2|\xi_1,...\xi_m)$ *so that*

$$\frac{1}{\sigma^2}|\{\xi_1,...\xi_m\} \sim \Gamma\left((m-1)/2, \sum_{k=2}^{m}(\xi_k-\xi_1)^2/2\right),$$

the previous conditions are equivalent to, respectively,

$$\int \frac{1}{q(\xi_1)} f_{\xi_1}(y_1)\, d\xi_1 < +\infty,$$

$$\int \left[\sum_{k=2}^{m}(\xi_k-\xi_1)^2\right]^{-(m+1)} \frac{D_m(\xi_1,...\xi_m)}{q(\xi_1)...q(\xi_m)} \prod_{k=1}^{m} f_{\xi_k}(y_{t_k}|y_{1:t_k-1})\, d\lambda d\xi_1...d\xi_m < +\infty.$$

For a proof, see the appendix. Note that the proposed $q_m(\sigma^2|\xi_1,...,\xi_m)$ in the gaussian case is the "ideal" instrumental distribution, in the sense given in §2, since it is the exact conditional posterior distribution $\pi(s^2|\theta_{1:m}, y_{1:t})$.

Finally, the Monte Carlo HMM filter may be quite intensive if the number of components K is important, since vector of probabilities of size $K(K+1)/2$ must be manipulated. Execution time savings can be obtained by occasional *conditioning* of the particles: draw for each particle an integer $M_t^{(j)}$ from the conditional distribution $P(M_t|\theta^{(j)}, y_{1:t})$, attach this additional component to $\theta^{(j)}$, and carry on the computations for this particle *conditionally* on the value taken by $M_t^{(j)}$. In that case, the filtering probabilities corresponding to the states such that $M_t < m$, where m is the value taken by $M_t^{(j)}$, are discarded, and the size of the vector of probabilities is reduced. This conditioning increases the conditional variance of the weights, in the same manner than a standard particle filter produces more volatile weights than an integrated filter (see §3), and for this reason it should be performed only from time to time.

5 Illustrative results

We consider a mean-switching auto-regressive model of order 1, with observation equation,

$$y_1 = \frac{\xi_1}{1-\rho} + \frac{s}{(1-\rho^2)^{1/2}}\varepsilon_1, \quad \varepsilon_1 \sim \mathcal{N}(0,1) \tag{17}$$

and, for $t \geq 2$, conditionally on $z_t = k$, $k = 1, ..., K$

$$y_t = \xi_k + \rho y_{t-1} + s\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1). \tag{18}$$

A very similar model was introduced by Billio et al. (1999). The first observation follows in fact the invariant distribution corresponding to the auto-regressive model without switching, and z_1 is set to 1, as required in our sequential state number determination procedure. Apart from the means $\xi_1, ..., \xi_K$ and the transitions probabilities, we will also suppose ρ and s to be unknown, and include them in the vector parameter θ, but obviously the Monte Carlo HMM filter is easily adapted to the case where additional parameters are considered. For this model, the derivation of $\Delta(\xi_i, \xi_j)$ as the Kullback-Leibler divergence between $f(y_t|y_{1:t-1}, \xi_i)$ and $f(y_t|y_{1:t-1}, \xi_j)$ (see §4.4), is straightforward:

$$\Delta(\xi_i, \xi_j) = (\xi_i - \xi_j)^2/s^2. \tag{19}$$

Therefore, we consider an improper prior distribution of the form

$$\pi(\theta) \propto \prod_{k=1}^{K} \pi(p_{k|}) s^{2(\alpha_0+1)} e^{-\beta_0/s^2} D(\xi_1, ..., \xi_K),$$

where the close form of $D(\xi_1, ..., \xi_K)$ was given by (15), so that, without this discriminating factor, each part would marginally follow $\pi(\xi_k) \propto 1$, $\pi(\rho) \propto \mathcal{U}]-1,1[$ and $1/s^2 \sim \Gamma(\alpha_0, \beta_0)$. For the transition probabilities we take the Dirichlet mixture prior presented in §4.4 , with $\alpha_{\rightarrow}^{(k')} = 1$ and $\alpha_{\circlearrowleft}^{(k')} = 3(k'-1)$, for $k' = 1, ..., K$.

Since (17) holds, replacing the improper $\pi(\xi_1) \propto 1$ by the proper conditional distribution $\pi(\xi_1|s^2, \rho) \sim \mathcal{N}\left(y_1(1-\rho), s^2(1-\rho)^2/(1-\rho^2)\right)$, transforms the prior distribution above into the posterior distribution at stage 1. Thus, with this replacement, and starting the algorithm at stage $t = 2$, only components ξ_2 to ξ_K are affected an improper distribution. While Lemmas 1 and 4 do not apply directly here, since the observation likelihood $f_{\xi,s,\rho}(.|.)$

this time also depends on extra parameters s and ρ, it is very easy to see that the corresponding sufficient conditions simply generalize by incorporating s and ρ in the integration variables of the related integrals. In this manner, through easy derivations, we get that the outputs of our sequential state number determination procedure are consistently defined, and that proper distributions for the components of the form, for $k = 2, ..., K$:

$$q(\xi_k) \propto \mathcal{C}(\mu_0, \sigma_0^2)$$

allow for a valid reweighting operation, as soon as $\alpha_0 > 1$.

The studied sample is a simulated sequence of $T = 200$ points, featuring 4 distinct components, and drawn from the model with parameters

$$s^{(0)} = 0.2, \quad \rho^{(0)} = 0.3, \quad \begin{pmatrix} \xi_1^{(0)} \\ \xi_2^{(0)} \\ \xi_3^{(0)} \\ \xi_4^{(0)} \end{pmatrix} = \begin{pmatrix} 0.7 \\ 1.4 \\ 2.1 \\ 2.8 \end{pmatrix}, \quad P^{(0)} = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.1 & 0.85 & 0.05 & 0 \\ 0 & 0.05 & 0.9 & 0.05 \\ 0 & 0 & 0.1 & 0.9 \end{pmatrix}.$$

The hyper-parameters of the prior were set to: $K = 5$, $\alpha_0 = 2$, $\beta_0 = 1/8$, $\mu_0 = 0$, and $\sigma_0 = 1$, the tuning parameters of the discriminating factor to $C = (1/2)^\nu$ and $\nu = 4$.

Figures 1 gives the plot of the sequence $y_{1:200}$ over time, along with the true values of the states $z_{1:200}$. Figure 2 gives the evolution over time of the filtered expectations of M_t, that is $E_\pi(M_t|y_{1:t}) = \sum_{m=1}^K m\pi(M_t = m|y_{1:t})$, as estimated by the MCHF algorithm, against the true values of this process. Figure 3 provides the weighted histogram of the particle system targeting $\pi(\theta_{1:4}|M_t = 4, y_{1:t})$, whereas the estimated value for $\pi(M_{200} = 4|y_{1:200})$ is 0.999. The MCHF algorithm was run with $H = 10000$ particles. As one can see, results are more than satisfactory.

6 Extension to other discrete state-space models

Suppose now the hidden process (z_t) is semi-Markov, that is it features successive regimes $s_1, ..., s_i, ...$, whose durations are random variables $\tau_1, ...\tau_i, ...$

$$z_t = s_u, \text{ for } t \text{ s.t. } \sum_{i=1}^{u-1} \tau_i < t \leq \sum_{i=1}^{u} \tau_i.$$

The s_i's form a Markov chain of order K (with transitions probabilities $(p_{kl})_{1 \leq k,l \leq K}$) which never stay in the same state twice in a row ($p_{kk} = 0$,

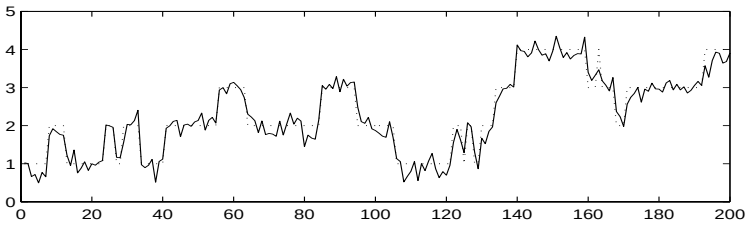

Figure 1: Plot of the simulated sequence $y_{1:200}$ (solid line), and the corresponding states $z_{1:200}$ (dotted line)

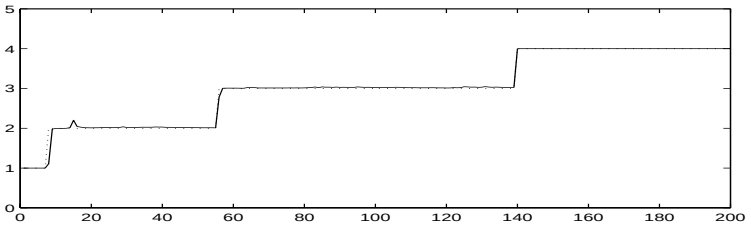

Figure 2: Plot of the estimates of $E_\pi(M_t|y_{1:t})$ (solid line), against the true values of $M_{1:200}$ (dotted line)

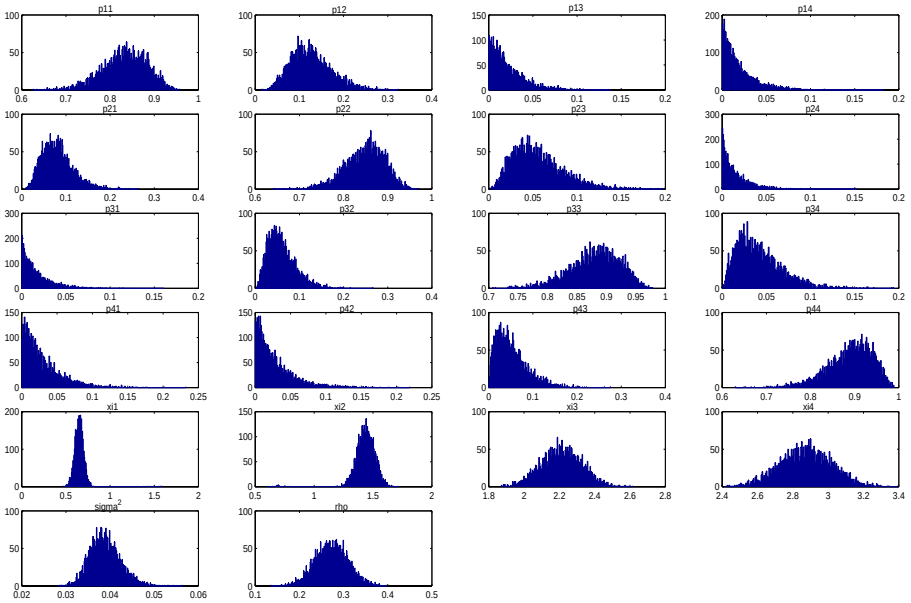

Figure 3: Weighted histogram of the particle system targeting $\pi(\theta_{1:4}|M_t = 4, y_{1:t})$, $t = 200$

$k = 1, ..., K$). In general, the duration in a given regime is assumed to be Poisson-distributed:

$$\tau_i | \{s_i = k\} \sim \mathcal{P}(\lambda_k),$$

and the global parameter comprises this time the ξ_k's (featured by the observation equation), the λ_k's and the p_{kl}'s. It seems impossible to generalize the state marginalization implemented in the Monte Carlo HMM filter to the semi-Markov models, and therefore we have to fall back on the standard particle filters (§2) to estimate this class of models. With such a filter, it is possible to simulate the next state z_{t+1} (when extending the space in the reweighting step) from the true conditional distribution $P(z_{t+1}|\theta, z_{1:t}, y_{1:t+1})$, since it merely consists in drawing a multinomial realization, with probabilities, for $l = 1, ..., K$,

$$P(z_{t+1} = l|\theta, z_{1:t}, y_{1:t+1}) \propto f_{\xi_l}(y_{t+1}|y_{1:t}) P(z_{t+1} = l|z_{1:t}, \theta),$$

where the proportionality constant is again recovered by normalization. The quantities $P(z_{t+1} = l|z_{1:t}, \theta)$ are derived as follows. Suppose $z_{1:t}$ is such that the later regime $s_i = k$ has appeared at time $t' + 1$, then

$$P(z_{t+1} = l|z_{1:t'-1}, z_{t'} \neq k, z_{t'+1} = ... = z_t = k, \theta) = \begin{cases} P(\tau_i > t - t'|s_i = k, \theta) & \text{if } k = l, \\ p_{kl} P(\tau_i = t - t'|s_i = k, \theta) & \text{otherwise.} \end{cases}$$

With such a conditional distribution for extending the space, the incremental weights verify

$$\begin{aligned} u_{t+1}(\theta, z_{1:t+1}) &\propto \frac{\pi(z_{1:t+1}|y_{1:t+1}, \theta)}{\pi(z_{1:t}, \theta|y_{1:t}) P(z_{t+1}|z_{1:t}, y_{1:t}, \theta)} \\ &\propto \frac{P(\theta, z_{1:t+1}|y_{1:t+1}, \theta)}{P(\theta, z_{1:t+1}|y_{1:t}, \theta)} \\ &\propto P(y_{t+1}|z_{1:t+1}, y_{1:t}, \theta) \end{aligned}$$

and are given by the observation equation.

A hidden semi-Markov model complies with a sequential reparamaterization, in the same manner than the hidden Markov models. This time, the reparameterization affects the sequence of regimes (s_i), with

$$\begin{aligned} s_1 &= 1, \\ P(s_{i+1} = l|s_i = k, M_i = m) &= \begin{cases} p_{kl} & \text{if } k, l \leq m, \\ \sum_{l'=k+1}^{K} p_{kl'} & \text{if } k < l = m + 1, \\ 0 & \text{otherwise,} \end{cases} \\ M_1 &= 1, \\ M_{i+1} &= \max(M_i, s_{i+1}), \end{aligned}$$

where, in that case, M_i stands for the number of components already appeared after i state shifts.

Therefore, the semi-Markov lend themselves to the sequential state number determination procedure defined in §4. The reparameterized version of the model can still be analyzed as a hidden semi-Markov model (with a state-space cardinal of $K(K+1)/2$), hence a sequential state number determination can be implemented by a simple particle filter for this class of model, such as the one presented before. Moreover, the non-informative prior approach we developed in §4.2 also applies to semi-Markov models, since the proofs of the corresponding results do not suppose any particular structure for the hidden process.

The change-point models can be seen as a degenerated version of the hidden semi-Markov models, with forced transitions from regime i to regime $i+1$, that is $s_i = i$ with probability 1. While somehow artificial, this definition clearly indicates that change-point models also allows for a sequential state number determination procedure, exactly in the same manner than the semi-Markov models.

7 Conclusion

This paper provides new algorithmic and theoretical tools for the inference of discrete state-space models, and most especially of hidden Markov models. For the technical part, the Monte Carlo HMM filter seems to be a quicker and more powerful alternative to the currently proposed MCMC based algorithms. In particular, it is more flexible than methods related to Gibbs sampling techniques, for its internal structure is mostly model-independent. On a theoretical ground, the (informative or non-informative) sequential state number determination procedure seems a promising step towards a more refined analysis of sequential heterogeneous data, and obviously calls to further research on its applicability to other classes of models.

A Proofs

First, note the likelihood $P(y_{1:t}|\theta)$ of a mixture model with observation equation (1) takes the following form:

$$P(y_{1:t}|\theta) = \sum_{z_{1:t}\in[1,K]^t} P(z_{1:t}|P) \prod_{i=1}^{t} f_{\xi_{z_i}}(y_i|y_{1:i-1})$$

where the $P(z_{1:t}|P)$'s are given by the transition equation. Therefore, if an improper prior $\pi(\xi)$ is assigned to the components, that is $\int \pi(\xi)\, d\xi = +\infty$, we have

$$\int \pi(\theta) P(y_{1:t}|\theta)\, d\theta = \sum_{z_{1:t} \in [1,K]^t} \int \pi(\theta) P(z_{1:t}|P) \prod_{i=1}^{t} f_{\xi_{z_i}}(y_i|y_{1:i-1}) d\theta$$

where $\pi(\theta)$ factorises as in (9) and, as it is easy to see, the terms corresponding to the sequences $z_{1:t}$ such that one of the components is not visited gives infinite integrals.

Denote E_t^k the event "$\forall i \in [1, K]$, at least k of the t states takes the value i". Then, for an improper prior of order k, the conditional posterior distribution $\pi(\theta|E_t^k, y_{1:t})$ is proper, since in the corresponding likelihood

$$P(y_{1:t}|E_t^k, \theta) = \sum_{z_{1:t} \in [1,K]^t} P(z_{1:t}|E_t^k, P) \prod_{i=1}^{t} f_{\xi_{z_i}}(y_i|y_{1:i-1})$$

the sequences $z_{1:t}$ that would produce infinite integrals are exactly such that $P(z_{1:t}|E_t^k, P) = 0$. This to clarify the claim of §4.1 that a proper inference can be led from an improper prior of order k by conditioning on the event E_t^k.

In the same manner, conditioning on $M_t = m$ in Lemma 1 is equivalent to conditioning on the event "The m first components are visited at least once, the others are not.", and the partial likelihood $P(y_{1:t}|M_t = m, \theta_{1:m})$, which verify

$$P(y_{1:t}|M_t = m, \theta_{1:m}) = \sum_{z_{1:t} \in [1,m]^t} P(z_{1:t}|M_t = m, P) \prod_{i=1}^{t} f_{\xi_{z_t}}(y_i|y_{1:i-1})$$

obviously only needs to depend on m first components, and is such that the distribution defined in (12) is proper. Proofs of Lemmas 2, 4 and 5 are straightforward through similar likelihood decompositions. Note we are able to withdraw the partial discriminating factors $D_m(\xi_1, ..., \xi_m)$ in the integrals of Lemmas 1 and 4, since it is a bounded quantity, but in Lemmas 2 and 5, the $D_m(\xi_1, ..., \xi_m)$'s remain, since the corresponding integrals cannot be definite without a discriminating factor. The application to the Gaussian hierarchical prior is obtained by marginalizing out σ^2 in the corresponding integrals, through the formula $\int x^{\alpha-1} e^{-\beta x}\, dx = \Gamma(\alpha)/\beta^\alpha$, with $x = 1/\sigma^2$, $\alpha = (m-1)/2$, and $\beta = \sum_{k=2}^{m} (\xi_k - \xi_1)^2/2$.

For Lemma 3, note that

$$\begin{aligned} q(\theta|y_{1:t})P(M_t = m|y_{1:t}, \theta) &\propto q(\theta)P(y_{1:t}, M_t = m|\theta) \\ &\propto q(\theta)P(M_t = m|\theta)P(y_{1:t}|M_t = m, \theta), \end{aligned}$$

where $q(\theta)$ is the global prior for θ, obtained by replacing $\pi(\xi)$ by $q(\xi)$ in equation (9), and since $P(M_t = m|\theta) = P(M_t = m|P)$, and $P(y_{1:t}|M_t = m, \theta) = P(y_{1:t}|M_t = m, \theta_{1:m})$, as explained above, the integral

$$\int \frac{\pi(\xi_1)...\pi(\xi_m)}{q(\xi_1)...q(\xi_m)} q(\theta|y_{1:t})P(M_t = m|y_{1:t}, \theta)\, d\theta$$

is clearly proportional to $\pi(M_t = m|y_{1:t})$, as defined in §4.2. Same remarks apply in the hierarchical case.

References

Billio, M., Monfort, A., and Robert, C. P. (1999). Bayesian estimation of switching ARMA models. *J. Econometrics*, 93(2):229–255.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proc-Radar, Sonar Navigation*, 146(1):2–7.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.*, 95:957–970.

Chopin, N. (2000). A sequential particle filter for static models. *CREST Working Paper*, 2000-45.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *J. Roy. Statist. Soc. B*, 56:363–375.

Doucet, A., de Freias, J. F. G., and Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag: New York, to appear April 2001.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.*, 10(3):197–208.

Gilks, W. R. and Berzuini, C. (2001). Following a moving target - Monte Carlo inference for dynamic Bayesian models. *J. Roy. Stat. Soc. B*, 63:127–146.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-gaussian Bayesian state estimation. *J. Amer. Statist. Assoc.*, 140(2):107–113.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Kitagawa (1987). Non-gaussian state-space modeling of nonstationary time series. *J. Amer. Statist. Assoc.*, 82:1032–1063.

Liu, J. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, 93:1032–1044.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New-York.

Ryden, T. (2000). Statistics for hidden Markov models. *Graduate Lectures at ENSAE, PARIS*.

Wasserman, L. (2000). Asymptotic inference for mixture models using data dependent priors. *J. Roy. Statist. Soc. B*, 62:159–180.