A PARTICLE FILTER FOR MODEL BASED AUDIO SOURCE SEPARATION

C. Andrieu, S.J. Godsill

Signal Processing Lab., University of Cambridge
Department of Engineering, Trumpington Street
CB2 1PZ Cambridge, UK
Email: {ca226,sjg}@eng.cam.ac.uk

ABSTRACT

In this paper we present an original modelling of the source separation problem that takes into account all the non-stationarities of the underlying processes. The estimation of the sources then reduces to that of a filtering/fixed-lag smoothing algorithm, for which we propose an efficient numerical solution, relying on particle filter techniques.

1 Introduction

The problem of separating convolutively mixed sources is of interest to many applications such as speech enhancement, crosstalk removal in multichannel communications and multipath channel identification. In this paper we consider the problem of separating audio signals modelled as autoregressive processes. The desired sources are clearly non-stationary as the vocal tract is continually changing, sometimes slowly, sometimes rapidly and the coupling filters may be time varying as the propagation conditions in the medium or more simply the sources/sensors geometry change over time. Here we explicitly model the non-stationarities of both the sources and the propagation medium, allowing us to take into account different stationarity time scales: speech will typically be stationary over periods of 20-40 milliseconds, whereas the transmission channel is expected to be stationary over longer periods. Taking into account these non-stationarities clearly removes some of the problems associated with framed based algorithms that assume piecewise stationarity and result in a delayless algorithm.

The modelling we adopt for the system facilitates a state space representation, and the problem of estimating the sources from observed mixtures then reduces to that of a filtering/fixed-lag smoothing estimation problem. The filtering and smoothing problems, that is the recursive estimation of the sources conditional upon the currently available data, require the evaluation of integrals that do not admit closed-form analytic solutions and approximate methods must be employed. Classical methods to obtain approximations to the desired distributions include analytical approximations, such as the extended Kalman filter [1] the Gaussian sum filter [3], and deterministic numerical integration techniques (see $e.g.$. [5]). The extended Kalman filter and Gaussian sum filter are computationally cheap, but fail in difficult circumstances.

Instead of hand-crafting such algorithms, we propose here the use of an adaptive stochastic grid approximation, which introduces a so-called system of particles. These particles evolve randomly in time in correlation with one another, and either give birth to offspring particles or die according to their ability to represent the different zones of interest of the state space dictated by the observation process and the dynamics of the underlying system. This class of Monte Carlo methods was introduced in the late 60's, but were overlooked mostly because of their computational complexity, see [9] for a review. The advantage of these methods is that they get around the problems faced by classical methods: the adaptativity of the stochastic grid results in an accurate representation of the density of interest at a rate independent of the dimension of the problem [6].

Particle techniques are here applied to obtain filtered and fixed-lag smoothed estimates of the non-stationary sources modelled as time varying autoregressive processes, observed in additive white Gaussian noise. The simulation-based algorithms developed here are not just a straightforward application of the basic methods, but are designed to make efficient use of the analytical structure of the model. At each iteration the algorithm has a computational complexity that is linear in the number of particles, and can easily be implemented on parallel computers, thus facilitating near real-time processing. It is also shown how an efficient fixed-lag smoothing algorithm may be obtained by combining the filtering algorithm with Markov chain Monte Carlo (MCMC) methods (see [12] for an introduction to MCMC methods). Note that the power and flexibility of these techniques can allow for far more complex models than those considered here.

The remainder of the paper is organized as follows. The model specification and estimation objectives are stated in Section 2. In Section 3 sequential simulation-based methods are developed to solve the filtering and fixed-lag smoothing problem and determine the model adequacy. In this section we first introduce Monte Carlo basics, secondly show how to take advantage of the structure of the model and reduce dramatically the dimensionality of the problem, thirdly explain why a selection scheme for the particles is required and why diversity must be introduced, here using MCMC steps. Section 4 presents and discusses simulation results.

2 Model of the data

The problem addressed is the problem of source separation, the n sources being modelled as autoregressive processes, from which we have at each time t, m observations which are convolutive mixtures of the n sources.

2.1 Model for the sources

Source i can be modelled for $t = 1, \ldots$ as:

$$s_{i,t} = \mathbf{a}_{i,t}^{\mathrm{T}} \mathbf{s}_{i,t-1:t-p_i} + \sigma_{v,i,t} v_{i,t}^{s} \tag{1}$$

C. Andrieu is sponsored by AT&T Lab., Cambridge UK.

and p_i is the order of the i^{th} AR model. We assume that $\mathbf{s}_{i,-p_i+1:0} \sim \mathcal{N}(\mathbf{m}_0^s, \mathbf{P}_0^s)$ for $i = 1,\ldots,n$. $\left(v_{i,t}^s\right)_{t=1,\ldots,}$ is a zero mean normalized i.i.d. Gaussian sequence, *i.e.* for $i = 1,\ldots,n$ and $t = 1,\ldots$

$$v_{i,t}^s \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \tag{2}$$

The $\left(\sigma_{v,i,t}^2\right)_{i=1,\ldots,n}$ are the variances of the dynamic noise for each source at time t.

We assume that the evolving autoregressive model follows a linear Gaussian state-space representation,

$$\begin{aligned} \mathbf{a}_{i,t+1} &= \mathbf{A}_i^{\mathbf{a}}\mathbf{a}_{i,t} + \mathbf{B}_i^{\mathbf{a}}\mathbf{v}_{i,t+1}^{\mathbf{a}} \\ \phi_{v,i,t+1} &= \mathbf{A}_i^{\phi_v}\phi_{v,i,t} + \mathbf{B}_i^{\phi_v}\mathbf{v}_{i,t+1}^{\phi_v} \end{aligned} \tag{3}$$

with $\phi_{v,i,t} \triangleq \log\left(\sigma_{v,i,t}^2\right)$,

$$\mathbf{v}_{i,t}^{\mathbf{a}} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}_{p_i}, \mathbf{I}_{p_i}) \text{ and } \mathbf{v}_{i,t}^{\phi_v} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \tag{4}$$

and $\mathbf{a}_{i,0} \sim \mathcal{N}(\mathbf{m}_0^{\mathbf{a}}, \mathbf{P}_0^{\mathbf{a}})$. A future version of the model will include switching between different sets of matrices $\mathbf{A}_i^{\mathbf{a}}, \mathbf{B}_i^{\mathbf{a}}$ to take into account silences, for example. Typically $\mathbf{A}_i^{\mathbf{a}} = \mathbf{I}_{p_i}$ and $\mathbf{B}_i^{\mathbf{a}} \propto \mathbf{I}_{p_i}$.

2.2 Model of the mixing system and noise

The mixing model is assumed to be a multidimensional time varying FIR filter. More precisely we assume that the sources are mixed in the following manner, and corrupted by an additive Gaussian i.i.d. noise sequence: at the j^{th} sensor, and for $j = 1,\ldots,m$

$$y_{j,t} = \sum_{i=1}^{n} \mathbf{h}_{i,j,t}\mathbf{s}_{i,t:t-l_{i,j}+1} + \sigma_{w,j,t}w_{j,t} \tag{5}$$

where $l_{i,j}$ is the length of the filter from source i to sensor j. The series $(w_{j,t})_{t=1,\ldots}$ is a zero mean normalized i.i.d. Gaussian sequence, *i.e.* for $j = 1,\ldots,n$ and $t = 1,\ldots$

$$w_{j,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \tag{6}$$

The quantities $\left(\sigma_{w,j,t}^2\right)_{j=1,\ldots,m}$ are the variances of the observation noise for each sensor at time t. The $w_{j,t}$ are assumed independent of the excitations of the AR models. We impose the constraint $[\mathbf{h}_{i,j,t}]_{1,1} = 1$ and $[\mathbf{h}_{i,j,t}]_{k,l} = 0$ for $j = i \mod m$ and $i = 1,\ldots,n$. In the case $m = n$ this constraint corresponds to $[\mathbf{h}_{i,j,t}]_{1,1} = 1$ and $[\mathbf{h}_{i,j,t}]_{k,l} = 0$ for $j = i$. As for the model of the sources, we assume that the observation system also follows a state-space representation: writing $\phi_{w,j,t} \triangleq \log\left(\sigma_{w,j,t}^2\right)$

$$\begin{aligned} \mathbf{h}_{i,j,t+1} &= \mathbf{A}_{i,j}^{\mathbf{h}}\mathbf{h}_{i,j,t} + \mathbf{B}_{i,j}^{\mathbf{h}}\mathbf{v}_{i,j,t+1}^{\mathbf{h}} \\ \phi_{w,j,t+1} &= \mathbf{A}_j^{\phi_w}\phi_{w,j,t} + \mathbf{B}_j^{\phi_w}\mathbf{v}_{j,t+1}^{\phi_w} \end{aligned} \tag{7}$$

with

$$\mathbf{v}_{i,j,t}^{\mathbf{h}} \overset{i.i.d.}{\sim} \mathcal{N}\left(\mathbf{0}_{l_{i,j}\times 1}, \mathbf{I}_{l_{i,j}}\right), \; \mathbf{v}_{j,t}^{\phi_w} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \tag{8}$$

and $\mathbf{h}_{i,j,0} \sim \mathcal{N}\left(\mathbf{m}_0^{\mathbf{h}}, \mathbf{P}_0^{\mathbf{h}}\right)$.

2.3 State-space representations

2.3.1 State-space representation of the sources

For each of the source i the signal can be rewritten in the following form:

$$\mathbf{s}_{i,t:t-\lambda_i+1} = \mathbf{A}_{i,t}^{\mathbf{s}}\mathbf{s}_{i,t-1:t-\lambda_i} + \mathbf{B}_{i,t}^{\mathbf{s}}v_{i,t}^{\mathbf{s}} \tag{9}$$

where $\lambda_i \triangleq \max\left\{p_i, \max_j \{l_{i,j}\}\right\}$ and $\mathbf{A}_{i,t}^{\mathbf{s}}$ is a $\lambda_i \times \lambda_i$ matrix defined as:

$$\mathbf{A}_{i,t}^{\mathbf{s}} \triangleq \begin{bmatrix} \mathbf{a}_{i,t}^{\mathrm{T}} & \mathbf{0}_{1\times(\lambda_i-p_i)} \\ \mathbf{I}_{\lambda_i-1} & \mathbf{0}_{(\lambda_i-1)\times 1} \end{bmatrix}, \; \mathbf{B}_{i,t}^{\mathbf{s}} \triangleq \left(\sigma_{v,i,t}, \mathbf{0}_{1\times(\lambda_i-1)}\right)^{\mathrm{T}} \tag{10}$$

Then one can rewrite the dynamic system equations:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}_{t+1}^{\mathbf{x}}\mathbf{x}_t + \mathbf{B}_{t+1}^{\mathbf{x}}\mathbf{v}_{t+1}^{\mathbf{x}} \\ \mathbf{y}_t &= \mathbf{C}_t^{\mathbf{x}}\mathbf{x}_t + \mathbf{D}_t^{\mathbf{x}}\mathbf{w}_t^{\mathbf{x}} \end{aligned} \tag{11}$$

with

$$\begin{aligned} &\mathbf{A}_t^{\mathbf{x}} \triangleq diag\left(\mathbf{A}_{1,t},\cdots,\mathbf{A}_{m,t}\right), \; \mathbf{B}_t^{\mathbf{x}} \triangleq diag\left(\mathbf{B}_{1,t},\cdots,\mathbf{B}_{m,t}\right) \\ &\mathbf{v}_t^{\mathbf{x}} = \left(v_{1,t}^{\mathbf{s}} \; \cdots \; v_{m,t}^{\mathbf{s}}\right)^{\mathrm{T}}, \; \mathbf{x}_t \triangleq \left(\mathbf{s}_{1,t:t-\lambda_1+1}^{\mathrm{T}} \; \cdots \; \mathbf{s}_{m,t:t-\lambda_m+1}^{\mathrm{T}}\right)^{\mathrm{T}}, \\ &[\mathbf{C}_t^{\mathbf{x}}] \text{ contains the mixing system} \\ &\text{and } \mathbf{w}_t^{\mathbf{x}} = \left(w_{1,t} \; \cdots \; w_{n,t}\right)^{\mathrm{T}}, \; \mathbf{D}_t^{\mathbf{x}} \triangleq diag\left(\sigma_{w,1} \; \cdots \; \sigma_{w,n}\right). \end{aligned} \tag{12}$$

2.3.2 State-space representation of the parameters

Defining the "stacked" parameter vectors

$$\begin{aligned} &[\mathbf{a}_t]_{\sum_{k=1}^{i-1} p_k+j,1} \triangleq [\mathbf{a}_{i,t}]_{j,1} \;\; i = 1,\ldots,m; \; j = 1,\ldots,p_i \\ &[\mathbf{h}_t]_{\sum_{u=1}^{i-1}\sum_{v=1}^{j_i-1} l_{uv}+k,1} \triangleq [\mathbf{h}_{i,j,t}]_{k,1} \quad \begin{array}{l} i = 1,\ldots,m; \\ j_i = 1,\ldots,n; \\ k = 1,\ldots,l_{ij} \end{array} \end{aligned} \tag{13}$$

and $l_{\mathbf{a}} \triangleq \sum_{i=1}^{m} p_i$, $l_{\mathbf{h}} \triangleq \sum_{j=1}^{n}\sum_{i=1}^{m} l_{ij} = \sum_{j=1}^{n} l_{\mathbf{h},j}$ it is possible to consider the following state space representations

$$\begin{aligned} \mathbf{h}_{t+1} &= \mathbf{A}_{t+1}^{\mathbf{h}}\mathbf{h}_t + \mathbf{B}_{t+1}^{\mathbf{h}}\mathbf{v}_{t+1}^{\mathbf{h}} \\ \mathbf{y}_t &= \mathbf{C}_t^{\mathbf{h}}\mathbf{h}_t + \mathbf{D}_t^{\mathbf{h}}\mathbf{w}_t^{\mathbf{h}} \end{aligned} \tag{14}$$

where $\mathbf{A}_t^{\mathbf{h}}, \mathbf{B}_t^{\mathbf{h}} \in \mathbb{R}^{l_{\mathbf{h}}\times l_{\mathbf{h}}}$,[1] $\mathbf{C}^{\mathbf{h}} \in \mathbb{R}^{n\times ml_{\mathbf{h}}}$, $\mathbf{D}_t^{\mathbf{h}} \in \mathbb{R}^{n\times n}$

$$\mathbf{C}_t^{\mathbf{h}} \triangleq diag\left\{\mathbf{x}_{1,t:t-l_{1,i}+1} \; \cdots \; \mathbf{x}_{m,t:t-l_{m,i}+1}\right\}_{i=1,\ldots,n} \tag{15}$$

and

$$\begin{aligned} \mathbf{a}_{t+1} &= \mathbf{A}_{t+1}^{\mathbf{a}}\mathbf{a}_t + \mathbf{B}_{t+1}^{\mathbf{a}}\mathbf{v}_{t+1}^{\mathbf{a}} \\ \mathbf{x}_t &= \mathbf{C}_t^{\mathbf{a}}\mathbf{a}_t + \mathbf{D}_t^{\mathbf{a}}\mathbf{w}_t^{\mathbf{a}} \end{aligned} \tag{16}$$

with $\mathbf{A}_t^{\mathbf{a}}, \mathbf{B}_t^{\mathbf{a}} \in \mathbb{R}^{l_{\mathbf{a}}\times l_{\mathbf{a}}}$, $\mathbf{C}_t^{\mathbf{a}} \in \mathbb{R}^{m\times l_{\mathbf{a}}}$ where

$$\mathbf{C}_t^{\mathbf{a}} = diag\left\{\mathbf{x}_{i,t-1:t-p_i}\right\}_{i=1,\ldots,n} \tag{17}$$

This is of practical interest as it allows us to integrate out the mixing filters and autoregressive filters, as conditional upon the variances and states $\mathbf{x}_t$ the systems (14) and (16) are linear Gaussian state space models. Together with (11) they define a bilinear Gaussian process.

[1] In fact one can be more general and this can be replaced by $\mathbb{R}^{l_{\mathbf{h}}\times u}$ for $\mathbf{B}_t^{\mathbf{h}}$.

2.4 Objectives

Our aim is, given the number of sources m, p_i and $l_{i,j}$ to estimate sequentially the sources $(\mathbf{x}_t)_{t=1,\ldots}$ and their parameters $\boldsymbol{\theta}_t \triangleq \{\mathbf{a}_t, \mathbf{h}_t, \boldsymbol{\sigma}^2_{t,1:m,v}, \boldsymbol{\sigma}^2_{t,1:n,w}\}$ from the observations $y_{j,t}$. More precisely, in the framework of Bayesian estimation, one is interested in the recursive, in time, estimation of posterior distributions of the type $p\left(d\boldsymbol{\theta}_t, d\mathbf{x}_t \middle| \mathbf{y}_{1:t+L}\right)$: when $L = 0$ this corresponds to the filtering distribution and when $L > 0$ this corresponds to the fixed-lag smoothing distribution. This is a very complex problem that does not admit any analytical solution, and one has to resort to numerical methods. In the next section we develop such a numerical method based on Monte Carlo simulation. Subsequently we will use the following notation: $\boldsymbol{\alpha}_t \triangleq \{\mathbf{a}_t, \mathbf{h}_t\}$, $\boldsymbol{\beta}_t \triangleq \{\boldsymbol{\sigma}^2_{t,1:m,v}, \boldsymbol{\sigma}^2_{t,1:n,w}\}$ and $\boldsymbol{\gamma}_t \triangleq \{\mathbf{x}_t, \boldsymbol{\sigma}^2_{t,1:m,v}, \boldsymbol{\sigma}^2_{t,1:n,w}\}$.

3 A Simulation-Based Optimal Filter/Fixed-lag Smoother

This section develops a simulation-based optimal filter/fixed-lag smoother to obtain filtered/fixed-lag smoothed estimates of the unobserved sources and their parameters of the type

$$I_L(f_t) \triangleq \int f_t(\boldsymbol{\theta}_t, \mathbf{x}_t)\, p\left(d\boldsymbol{\theta}_t, d\mathbf{x}_t \middle| \mathbf{y}_{1:t+L}\right) \tag{18}$$

The standard Bayesian importance sampling method is first described, and then we show how it is possible to take advantage of the analytical structure of the model by integrating out the parameters $\mathbf{a}_t$ and $\mathbf{h}_t$ which can be high dimensional, using Kalman filter related algorithms. This leads to an elegant and efficient algorithm for which the only tracked parameters are the sources and the noise variances. Then a sequential version of Bayesian importance sampling for optimal filtering is presented, and it is shown why it is necessary to introduce selection as well as diversity in the process. Finally, a Monte Carlo filter/fixed-lag smoother for our problem is described.

3.1 Monte Carlo Simulation for Optimal Estimation

For any f_t it will subsequently be assumed that $|I_L(f_t)| < +\infty$. Suppose that it is possible to sample N *i.i.d.* samples, called particles, $\left(\mathbf{x}^{(i)}_{0:t+L}, \boldsymbol{\theta}^{(i)}_{0:t+L}\right)$ according to $p\left(\mathbf{x}_{0:t+L}, \boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$. Then an empirical estimate of this distribution is given by

$$\begin{aligned}\widehat{p}_N&\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)\\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \delta_{\mathbf{x}^{(i)}_{0:t+L}, \boldsymbol{\theta}^{(i)}_{0:t+L}}\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L}\right)\end{aligned} \tag{19}$$

so that a Monte Carlo approximation of the marginal distribution $p\left(d\mathbf{x}_t, d\boldsymbol{\theta}_t \middle| \mathbf{y}_{1:t+L}\right)$ follows as

$$\widehat{p}_N\left(d\mathbf{x}_t, d\boldsymbol{\theta}_t \middle| \mathbf{y}_{1:t+L}\right) = \tfrac{1}{N}\textstyle\sum_{i=1}^N \delta_{\mathbf{x}^{(i)}_t, \boldsymbol{\theta}^{(i)}_t}\left(d\mathbf{x}_t, d\boldsymbol{\theta}_t\right) \tag{20}$$

Using this distribution, an estimate of $I_L(f_t)$ for any f_t may be obtained as

$$\begin{aligned}\widehat{I}_{L,N}(f_t) &= \textstyle\int f_t(\mathbf{x}_t, \boldsymbol{\theta}_t)\, \widehat{p}_N\left(d\mathbf{x}_t, d\boldsymbol{\theta}_t \middle| \mathbf{y}_{1:t+L}\right)\\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N f_t\left(\mathbf{x}^{(i)}_t, \boldsymbol{\theta}^{(i)}_t\right).\end{aligned} \tag{21}$$

This estimate is unbiased and from the strong law of large numbers, $\widehat{I}_{L,N}(f_t) \underset{N\to+\infty}{\overset{a.s.}{\to}} I_L(f_t)$. Under additional assumptions the estimates satisfy a central limit theorem. The advantage of the Monte Carlo method is clear. It is easy to estimate $I_L(f_t)$ for any f_t, and the rate of convergence of this estimate does not depend on t or the dimension of the state space, but only on the number of particles N and the characteristics of the function f_t. Unfortunately, it is not possible to sample directly from the distribution $p\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$ at any t, and alternative strategies need to be investigated.

One solution to estimate $p\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$ and $I_L(f_t)$ is the well-known Bayesian importance sampling method [9]. This method assumes the existence of an arbitrary importance distribution $\pi\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$ which is easily simulated from, and whose support contains that of $p\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$. Using this distribution $I_L(f_t)$ may be expressed as

$$I_L(f_t) = \frac{\int \pi\left(d\mathbf{x}_{1:t+L}, d\boldsymbol{\theta}_{1:t+L} \middle| \mathbf{y}_{1:t+L}\right) f_t(\mathbf{x}_t, \boldsymbol{\theta}_t)\, w\left(\mathbf{x}_{0:t+L}, \boldsymbol{\theta}_{0:t+L}\right)}{\int \pi\left(d\mathbf{x}_{1:t+L}, d\boldsymbol{\theta}_{1:t+L} \middle| \mathbf{y}_{1:t+L}\right) w\left(\mathbf{x}_{0:t+L}, \boldsymbol{\theta}_{0:t+L}\right)}, \tag{22}$$

where the importance weight $w(\mathbf{x}_{0:t+L}, \boldsymbol{\theta}_{0:t+L})$ is given by

$$w(\mathbf{x}_{0:t+L}, \boldsymbol{\theta}_{0:t+L}) \propto \frac{p\left(\mathbf{x}_{0:t+L}, \boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)}{\pi\left(\mathbf{x}_{0:t+L}, \boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)}. \tag{23}$$

The importance weight can normally only be evaluated up to a constant of proportionality, since, following from Bayes' rule,

$$\begin{aligned}p&\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)\\ &= \frac{p\left(\mathbf{y}_{1:t+L} \middle| \mathbf{x}_{0:t+L}, \boldsymbol{\theta}_{0:t+L}\right) p\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L}\right)}{p\left(\mathbf{y}_{1:t+L}\right)},\end{aligned} \tag{24}$$

where the normalizing constant $p(\mathbf{y}_{1:t+L})$ can typically not be expressed in closed-form.

If N i.i.d. samples $\left(\mathbf{x}^{(i)}_{0:t+L}, \boldsymbol{\theta}^{(i)}_{0:t+L}\right)$ can be simulated according to a distribution $\pi\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$, a Monte Carlo estimate of $I_L(f_t)$ in (22) may be obtained as

$$\begin{aligned}\widehat{I}^1_{L,N}(f_t) &= \frac{\sum_{i=1}^N f_t\left(\mathbf{x}^{(i)}_t, \boldsymbol{\theta}^{(i)}_t\right) w\left(\mathbf{x}^{(i)}_{0:t+L}, \boldsymbol{\theta}^{(i)}_{0:t+L}\right)}{\sum_{i=1}^N w\left(\mathbf{x}^{(i)}_{0:t+L}, \boldsymbol{\theta}^{(i)}_{0:t+L}\right)}\\ &= \textstyle\sum_{i=1}^N \overline{w}^{(i)}_{0:t+L} f_t\left(\mathbf{x}^{(i)}_t, \boldsymbol{\theta}^{(i)}_t\right),\end{aligned} \tag{25}$$

where the normalized importance weights are given by

$$\overline{w}^{(i)}_{0:t+L} = \frac{w\left(\mathbf{x}^{(i)}_{0:t+L}, \boldsymbol{\theta}^{(i)}_{0:t+L}\right)}{\sum_{j=1}^N w\left(\mathbf{x}^{(j)}_{0:t+L}, \boldsymbol{\theta}^{(j)}_{0:t+L}\right)}. \tag{26}$$

This method is equivalent to a point mass approximation of the target distribution of the form

$$\begin{aligned}\widehat{p}_N&\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)\\ &= \textstyle\sum_{i=1}^N \overline{w}^{(i)}_{0:t+L} \delta_{\mathbf{x}^{(i)}_{0:t+L}, \boldsymbol{\theta}^{(i)}_{0:t+L}}\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L}\right),\end{aligned} \tag{27}$$

The perfect simulation case, when $\pi\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right) = p\left(d\mathbf{x}_{0:t+L}, d\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$, corresponds to $\overline{w}^{(i)}_{0:t+L} = N^{-1}$, $i = 1, \ldots, N$. In practice, the importance distribution will be chosen to be as close as possible to the target distribution in a given sense. For finite N, $\widehat{I}^1_{L,N}(f_t)$ is biased, since it involves a ratio of estimates, but asymptotically, according to the strong law of large numbers, $\widehat{I}^1_{L,N}(f_t) \underset{N\to+\infty}{\overset{a.s.}{\to}} I_L(f_t)$. Under additional assumptions a central limit theorem also holds [9].

3.2 Analytical integrations

It is possible to reduce the estimation of $p\left(d\mathbf{x}_t, d\boldsymbol{\theta}_t \middle| \mathbf{y}_{1:t+L}\right)$ and $I_L(f_t)$ to one of sampling from $p\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$, where we recall that $\boldsymbol{\gamma}_t \triangleq \{\mathbf{x}_t, \boldsymbol{\sigma}^2_{t,1:m,v}, \boldsymbol{\sigma}^2_{t,1:n,w}\}$. Indeed,

$$\begin{aligned}p\left(d\boldsymbol{\alpha}_t, d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right) &= p\left(d\boldsymbol{\alpha}_t \middle| \boldsymbol{\gamma}_{0:t+L}, \mathbf{y}_{1:t+L}\right)\\ &\times p\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)\end{aligned} \tag{28}$$

where $p\left(d\boldsymbol{\alpha}_t \middle| \boldsymbol{\gamma}_{0:t+L}, \mathbf{y}_{1:t+L}\right)$ is a Gaussian distribution whose parameters may be computed using Kalman filter type techniques. Thus, given an approximation of $p\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$, an approximation of $p\left(d\mathbf{x}_t, d\boldsymbol{\theta}_t \middle| \mathbf{y}_{1:t+L}\right)$ may straightforwardly be obtained. Defining the marginal importance distribution and associated importance weight as

$$\begin{array}{l}\pi\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right) = \int \pi\left(d\boldsymbol{\alpha}_{0:t+L} d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right) \\ \quad w\left(\boldsymbol{\gamma}_{0:t+L}\right) \propto \frac{p\left(\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)}{\pi\left(\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)},\end{array} \tag{29}$$

and assuming that a set of samples $\boldsymbol{\gamma}_{0:t+L}^{(i)}$ distributed according to $\pi\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$ is available, an alternative Bayesian importance sampling estimate of $I_L\left(f_t\right)$ follows as

$$\begin{array}{l}\widehat{I}_N^2\left(f_t\right) = \frac{\sum_{1=1}^N p\left(\boldsymbol{\alpha}_t^{(i)} \middle| \boldsymbol{\gamma}_{0:t+L}^{(i)}, \mathbf{y}_{1:t+L}\right) f_t\left(\mathbf{x}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right) w\left(\boldsymbol{\gamma}_{0:t+L}^{(i)}\right)}{\sum_{i=1}^N w\left(\boldsymbol{\gamma}_{0:t+L}^{(i)}\right)} \\ \quad = \sum_{i=1}^N \widetilde{w}_{0:t+L}^{(i)} p\left(\boldsymbol{\alpha}_t^{(i)} \middle| \boldsymbol{\gamma}_{0:t+L}^{(i)}, \mathbf{y}_{1:t+L}\right) f_t\left(\mathbf{x}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right),\end{array} \tag{30}$$

provided that $p\left(\boldsymbol{\alpha}_t \middle| \boldsymbol{\gamma}_{0:t+L}, \mathbf{y}_{1:t+L}\right) f_t\left(\mathbf{x}_t, \boldsymbol{\theta}_t\right)$ can be evaluated in a closed-form expression. In (30) the marginal normalized importance weights are given by

$$\widetilde{w}_{0:t+L}^{(i)} = \frac{w\left(\boldsymbol{\gamma}_{0:t+L}^{(i)}\right)}{\sum_{j=1}^N w\left(\boldsymbol{\gamma}_{0:t+L}^{(j)}\right)}, \quad i = 1, \ldots, N. \tag{31}$$

Intuitively, to reach a given precision, $\widehat{I}_{L,N}^2\left(f_t\right)$ will need a reduced number of samples over $\widehat{I}_{L,N}^1\left(f_t\right)$, since it only requires samples from the marginal distribution $\pi\left(d\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right)$. It can be proved that the variance of the estimates is subsequently reduced [8]. In our case this is important as at each time instant the number of parameters is (when assuming that all mixing filters and AR processes have the same length),

- $m^2L - mL$ parameters for the mixing filters, where L can be large.
- m or 1 parameter(s) for the observation noise.
- $nl + n$ parameters for the autoregressive processes.
- n parameters for the sources.

It is not clear which integration will allow for the best variance reduction [8], but at least in terms of search in the parameters space the integration of the mixing filters and autoregressive filters seems preferable.

Given these results, the subsequent discussion will focus on Bayesian importance sampling methods to obtain approximations of $p\left(d\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right)$ and $I_L\left(f_t\right)$ using an importance distribution of the form $\pi\left(d\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right)$. The methods described up to now are batch methods. The next section illustrates how a sequential method may be obtained.

3.3 Sequential Bayesian Importance Sampling

The importance distribution at discrete time t may be factorized as

$$\begin{array}{l}\pi\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right) \\ \quad = \pi\left(d\boldsymbol{\gamma}_0 \middle| \mathbf{y}_{1:t+L}\right) \prod_{k=1}^{t+L} \pi\left(d\boldsymbol{\gamma}_k \middle| \boldsymbol{\gamma}_{0:k-1}, \mathbf{y}_{1:t+L}\right).\end{array} \tag{32}$$

The aim is to obtain at any time t an estimate of the distribution $p\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$ and to be able to propagate this estimate in time without modifying subsequently the past simulated trajectories $\boldsymbol{\gamma}_{0:t+L}^{(i)}$. This means that $\pi\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$ should admit $\pi\left(d\boldsymbol{\gamma}_{0:t-1+L} \middle| \mathbf{y}_{1:t-1+L}\right)$ as marginal distribution. This is possible if the importance distribution is restricted to be of the general form

$$\begin{array}{l}\pi\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right) \\ \quad = \pi\left(d\boldsymbol{\gamma}_0\right) \prod_{k=1}^{t+L} \pi\left(d\boldsymbol{\gamma}_k \middle| \boldsymbol{\gamma}_{0:k-1}, \mathbf{y}_{1:k}\right),\end{array} \tag{33}$$

Such an importance distribution allows recursive evaluation of the importance weights, *i.e.* $w\left(\boldsymbol{\gamma}_{0:t+L}\right) = w\left(\boldsymbol{\gamma}_{0:t-1+L}\right) w_{t+L}$, and in our particular case

$$\begin{array}{l}\frac{p\left(\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)}{\pi\left(\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)} \propto \frac{p\left(\boldsymbol{\gamma}_{0:t+L-1} \middle| \mathbf{y}_{1:t+L-1}\right)}{\pi\left(\boldsymbol{\gamma}_{0:t+L-1} \middle| \mathbf{y}_{1:t+L-1}\right)} \\ \times \frac{p\left(\mathbf{y}_{t+L} \middle| \boldsymbol{\gamma}_{0:t+L}\right) p\left(\mathbf{x}_{t+L} \middle| \mathbf{x}_{t+L-1}, \boldsymbol{\beta}_{t+L}\right) p\left(\boldsymbol{\beta}_{t+L} \middle| \boldsymbol{\beta}_{t+L-1}\right)}{\pi\left(\boldsymbol{\gamma}_{t+L} \middle| \boldsymbol{\gamma}_{0:t+L-1}, \mathbf{y}_{1:t+L}\right)}\end{array} \tag{34}$$

The quantity $p\left(d\mathbf{x}_{t+L} \middle| \mathbf{x}_{t+L}, \boldsymbol{\beta}_{t+L}\right)$ can be computed up to a normalizing constant using a one step ahead Kalman filter for the system given by Eq. (16) and $p\left(\mathbf{y}_{t+L} \middle| \mathbf{x}_{0:t+L}, \boldsymbol{\beta}_{t+L}\right)$ can be computed using a one step ahead Kalman filter of the system given by Eq. (14).

3.3.1 Choice of the Importance Distribution

There is an unlimited number of choices for the importance distribution $\pi\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$, the only restriction being that its support includes that of $p\left(d\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$. Two possibilities are considered next. A possible strategy is to choose at time $t + L$ the importance distribution that minimizes the variance of the importance weights given $\boldsymbol{\gamma}_{0:t-1}$ and $\mathbf{y}_{1:t}$. The importance distribution that satisfies this condition is [9] $p\left(d\boldsymbol{\gamma}_{t+L} \middle| \boldsymbol{\gamma}_{0:t-1+L}, \mathbf{y}_{1:t+L}\right)$, with the associated incremental importance weight given by

$$\begin{array}{l}w_{t+L} \propto p\left(\mathbf{y}_{t+L} \middle| \boldsymbol{\gamma}_{0:t-1+L}, \mathbf{y}_{1:t+L-1}\right) \\ \quad = \int p\left(\mathbf{y}_{t+L} \middle| \boldsymbol{\gamma}_{0:t+L}, \mathbf{y}_{1:t+L-1}\right) p\left(d\boldsymbol{\gamma}_{t+L} \middle| \boldsymbol{\gamma}_{t-1+L}\right).\end{array} \tag{35}$$

Direct sampling from the optimal importance distribution is difficult, and evaluating the importance weight is analytically intractable. The aim is thus in general to mimic the optimal distribution by means of tractable approximations, typically local linearization $p\left(d\boldsymbol{\gamma}_t \middle| \boldsymbol{\gamma}_{0:t-1}, \mathbf{y}_{1:t}\right)$. Instead, here we describe a mixed suboptimal method. We propose to sample the particles at time t according to two importance distributions π_1 and π_2 with proportions α and $1-\alpha$ such that the importance weights $w\left(\boldsymbol{\gamma}_{0:t+L}^{(i)}\right)$ have now the form (note that it would be possible to draw N_1 and N_2 randomly according to a Bernoulli distribution with parameter α, but this would increase the estimator variance)

$$\left\{\begin{array}{c}\alpha \frac{p\left(\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right)}{\pi_1\left(\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right)} / \mathbb{E}_{\pi_1}\left[\frac{p\left(\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)}{\pi_1\left(\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)}\right] \\ (1-\alpha) \frac{p\left(\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right)}{\pi_2\left(\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right)} / \mathbb{E}_{\pi_2}\left[\frac{p\left(\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)}{\pi_2\left(\boldsymbol{\gamma}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)}\right]\end{array}\right. \tag{36}$$

which in practice is estimated as $\overline{w}_{0:t+L}^{(i)}$

$$\left\{\begin{array}{c}\alpha \frac{p\left(\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right) / \pi_1\left(\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right)}{\sum_{j=1}^{N\alpha} p\left(\boldsymbol{\gamma}_{0:t+L}^{(j)} \middle| \mathbf{y}_{1:t+L}\right) / \pi_1\left(\boldsymbol{\gamma}_{0:t+L}^{(j)} \middle| \mathbf{y}_{1:t+L}\right)} \\ (1-\alpha) \frac{p\left(\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right) / \pi_2\left(\boldsymbol{\gamma}_{0:t+L}^{(i)} \middle| \mathbf{y}_{1:t+L}\right)}{\sum_{j=\alpha N+1}^{N} p\left(\boldsymbol{\gamma}_{0:t+L}^{(j)} \middle| \mathbf{y}_{1:t+L}\right) / \pi_2\left(\boldsymbol{\gamma}_{0:t+L}^{(j)} \middle| \mathbf{y}_{1:t+L}\right)}\end{array}\right. \tag{37}$$

The importance distribution $\pi_1\left(d\mathbf{x}_{t+L}\middle|\,\mathbf{x}_{1:t+L-1},\boldsymbol{\beta}_{1:t+L},\mathbf{y}_{1:t+L}\right)$ will be taken to be a normal distribution centered around zero with variance $\sigma_{\mathbf{x}}^2$, and $\pi_2\left(d\mathbf{x}_{t+L}\middle|\,\mathbf{x}_{1:t+L-1},\boldsymbol{\beta}_{1:t+L},\mathbf{y}_{1:t+L}\right)$ is taken to be $p\left(d\mathbf{x}_{t+L}^{(i)}\middle|\,\mathbf{m}_{t+L|t+L}^{\mathbf{a},\mathbf{h}\ (i)},\mathbf{P}_{t+L|t+L}^{\mathbf{a},\mathbf{h}\ (i)},\boldsymbol{\beta}_{t+L}^{(i)},\mathbf{y}_{1:t+L}\right)$ which is a Gaussian distribution obtained from a one step ahead Kalman filter for the state space model described in (11) with $\mathbf{m}_{t+L|t+L}^{\mathbf{a}\ (i)}$ and $\mathbf{m}_{t+L|t+L}^{\mathbf{h}\ (i)}$ as values for $\mathbf{a}^{(i)}$ and $\mathbf{h}^{(i)}$ and initial variances $\mathbf{P}_{t+L|t+L}^{\mathbf{a}}$ and $\mathbf{P}_{t+L|t+L}^{\mathbf{h}}$. The variances are sampled from their prior distributions, and expression (34) is used to compute $\overline{w}_{0:t+L}^{(i)}$. Note that other importance distributions are possible, but this approach yields good results and seems to preserve diversity of the samples.

3.3.2 Degeneracy of the Algorithm

For importance distributions of the form specified by (33) the variance of the importance weights can only increase (stochastically) over time, see [9] and references therein. It is thus impossible to avoid a degeneracy phenomenon. Practically, after a few iterations of the algorithm, all but one of the normalized importance weights are very close to zero, and a large computational effort is devoted to updating trajectories whose contribution to the final estimate is almost zero. For this reason it is of crucial importance to include selection and diversity. This is discussed in more detail in the following section.

3.4 Selection and diversity

The purpose of a selection (or resampling) procedure is to discard particles with low normalized importance weights and multiply those with high normalized importance weights, so as to avoid the degeneracy of the algorithm. A selection procedure associates with each particle, say $\widetilde{\boldsymbol{\gamma}}_{0:t}^{(i)}$, a number of children $N_i \in \mathbb{N}$, such that $\sum_{i=1}^{N} N_i = N$, to obtain N new particles $\boldsymbol{\gamma}_{0:t}^{(i)}$. If $N_i = 0$ then $\widetilde{\boldsymbol{\gamma}}_{0:t}^{(i)}$ is discarded, otherwise it has N_i children at time $t+1$. After the selection step the normalized importance weights for all the particles are reset to N^{-1}, thus discarding all information regarding the past importance weights. Thus, the normalized importance weight prior to selection in the next time step is proportional to (34). These will be denoted as $\widetilde{w}_t^{(i)}$, since they do not depend on any past values of the normalized importance weights. If the selection procedure is performed at each time step, then the approximating distribution before the selection step is given by $\widetilde{p}_N\left(d\boldsymbol{\gamma}_{0:t}\middle|\,\mathbf{y}_{1:t}\right)=\sum_{i=1}^{N}\widetilde{w}_t^{(i)}\delta_{\widetilde{\boldsymbol{\gamma}}_{0:t}^{(i)}}\left(d\boldsymbol{\gamma}_{0:t}\right)$, and the one after the selection step follows as $\widehat{p}_N\left(d\boldsymbol{\gamma}_{0:t}\middle|\,\mathbf{y}_{1:t}\right)=N^{-1}\sum_{i=1}^{N}\delta_{\boldsymbol{\gamma}_{0:t}^{(i)}}\left(d\boldsymbol{\gamma}_{0:t}\right)$. We choose systematic sampling [8] for its good variance properties.

However selection poses another problem. During the resampling stage any particular particle with a high importance weight will be duplicated many times. In particular, when $L>0$, the trajectories are resampled L times from time $t+1$ to $t+L$ so that very few distinct trajectories remain at time $t+L$. This is the classical problem of depletion of samples. As a result the cloud of particles may eventually collapse to a single particle. This degeneracy leads to poor approximations of the distributions of interest. Several suboptimal methods have been proposed to overcome this problem and introduce diversity amongst the particles. Most of these are based on kernel density methods [9, 10], which approximate the probability distribution using a kernel density estimate based on the current set of particles, and sample a new set of distinct particles from it. However, the choice and configuration of a specific kernel are not always straightforward. Moreover, these methods introduce additional Monte Carlo variation. In the next subsection it is shown how MCMC methods may be combined with sequential importance sampling to introduce diversity amongst the samples without increasing the Monte Carlo variation.

An efficient way of limiting sample depletion consists of simply adding a MCMC step to the simulation-based filter/fixed-lag smoother (see Berzuini and Gilks in [9], and [12] for an introduction to MCMC methods). This introduces diversity amongst the samples and thus drastically reduces the problem of depletion of samples. Assume that, at time $t+L$, the particles $\boldsymbol{\gamma}_{0:t+L}^{\prime(i)}$ are marginally distributed according to $p\left(d\boldsymbol{\gamma}_{0:t+L}\middle|\,\mathbf{y}_{1:t+L}\right)$. If a transition kernel $K\left(\boldsymbol{\gamma}_{0:t+L}\middle|\,d\boldsymbol{\gamma}_{0:t+L}^{\prime}\right)$ with invariant distribution $p\left(d\boldsymbol{\gamma}_{0:t+L}\middle|\,\mathbf{y}_{1:t+L}\right)$ is applied to each of the particles, then the new particles $\boldsymbol{\gamma}_{0:t+L}^{(i)}$ are still distributed according to the distribution of interest. Any of the standard MCMC methods, such as the Metropolis-Hastings (MH) algorithm or Gibbs sampler, may be used. However, contrary to classical MCMC methods, the transition kernel does not need to be ergodic. Not only does this method introduce no additional Monte Carlo variation, but it improves the estimates in the sense that it can only reduce the total variation norm [12] of the current distribution of the particles with respect to the target distribution.

3.5 Implementation Issues

3.5.1 Algorithm

Given at time $t+L-1$, $N\in\mathbb{N}^*$ particles $\boldsymbol{\gamma}_{0:t+L-1}^{(i)}$ distributed approximately according to $p\left(d\boldsymbol{\gamma}_{0:t+L-1}\middle|\,\mathbf{y}_{1:t+L-1}\right)$, the Monte Carlo fixed-lag smoother proceeds as follows at time $t+L$.

Monte Carlo filter/fixed-lag smoother

Sequential Importance Sampling Step

- For $i=1,\ldots,N$, $\widetilde{\boldsymbol{\gamma}}_{t+L}^{(i)}\sim\pi\left(d\boldsymbol{\gamma}_{t+L}\middle|\,\mathbf{x}_{0:t+L-1}^{(i)},\mathbf{y}_{1:t+L}\right)$ and set $\widetilde{\boldsymbol{\gamma}}_{0:t+L}^{(i)}=\left(\boldsymbol{\gamma}_{0:t+L-1}^{(i)},\widetilde{\boldsymbol{\gamma}}_{t+L}^{(i)}\right)$.
- For $i=1,\ldots,N$, compute the normalized importance weights $\widetilde{w}_{t+L}^{(i)}$ using (34) and (37).

Selection Step

- Multiply / discard particles $\widetilde{\boldsymbol{\gamma}}_{0:t+L}^{(i)}$ w.r.t. high / low normalized importance weights to obtain N particles $\boldsymbol{\gamma}_{0:t+L}^{\prime(i)}$, $e.g.$using systematic sampling.

MCMC Step

- For $i=1,\ldots,N$, apply to $\boldsymbol{\gamma}_{0:t+L}^{\prime(i)}$ a Markov transition kernel $K\left(\boldsymbol{\gamma}_{0:t+L}^{(i)}\middle|\,d\boldsymbol{\gamma}_{0:t+L}^{\prime(i)}\right)$ with invariant distribution $p\left(d\boldsymbol{\gamma}_{0:t+L}\middle|\,\mathbf{y}_{1:t+L}\right)$ to obtain N particles $\boldsymbol{\gamma}_{0:t+L}^{(i)}$.

3.5.2 Implementation of the MCMC Steps

There is an unlimited number of choices for the MCMC transition kernel. Here a one-at-a-time MH algorithm is adopted that updates at time $t+L$ the values of the Markov process from time t to $t+L$. More specifically, $\boldsymbol{\gamma}_k^{(i)}$, $k = t, \ldots, t+L$, $i = 1, \ldots, N$, is sampled according to an MCMC with $p\left(d\boldsymbol{\gamma}_k \middle| \boldsymbol{\gamma}_{-k}^{(i)}, \mathbf{y}_{1:t+L}\right)$ as target distribution, where $\boldsymbol{\gamma}_{-k}^{(i)} \triangleq \left(\boldsymbol{\gamma}_{0:t-1}^{\prime(i)}, \boldsymbol{\gamma}_t^{(i)}, \ldots, \boldsymbol{\gamma}_{k-1}^{(i)}, \boldsymbol{\gamma}_{k+1}^{\prime(i)}, \ldots, \boldsymbol{\gamma}_{t+L}^{\prime(i)}\right)$. Evaluation of $p\left(d\boldsymbol{\gamma}_k \middle| \boldsymbol{\gamma}_{-k}^{(i)}, \mathbf{y}_{1:t+L}\right)$ can be done efficiently via a backward-forward algorithm of $O(L+1)$ complexity [7]. The algorithm is fully described in [2]. The computational complexity of the whole algorithm at each iteration is clearly $O(N)$. At first glance, it could appear necessary to keep in memory the paths of all the trajectories $\boldsymbol{\gamma}_{0:t}^{(i)}$, so that the storage requirements would increase linearly with time. In fact, the importance distribution π_1, $\pi\left(\boldsymbol{\gamma}_t \middle| \boldsymbol{\gamma}_{0:t-1}, \mathbf{y}_{1:t}\right)$ and the associated importance weights, depend on $\boldsymbol{\gamma}_{0:t-1}$ only via a set of low-dimensional sufficient statistics $\mathbf{m}_{t|t}^{\mathbf{a},\mathbf{h}(i)}$, $\mathbf{P}_{t|t}^{\mathbf{a},\mathbf{h}(i)}$, and only these values need to be kept in memory. Thus, the storage requirements are also $O(N)$ and do not increase over time.

4 Simulations

To illustrate the efficiency of our method we have applied the procedure to two scenarios. In the first case we generated two stationary time invariant order 2 autoregressive processes, mixed by two filters of length 5. In the second case we made the phase of the poles of the AR processes evolve from .1 to .4 from $t = 1, \ldots, 250$ and from .4 to .1 for $t = 251, \ldots, 500$. The algorithm was run with 100 particles with $L = 0$, the parameters of the system and the sources were initialized at random. The matrices $\mathbf{B}^{\mathbf{a}}$ and $\mathbf{B}^{\mathbf{h}}$ were set to $.01 \times \mathbf{I}$. The results for the two cases are presented on Fig. (1) and Fig. (2), where the two original sources and the sources estimated on-line using the particle filter are displayed. Note that for the first scenario convergence really occurs from iteration 150 approximately, and that source 2 was apparently mistaken with source 1 before. For the second scenario, as expected, the problem is much harder, especially when the two poles are close, or when zero/pole cancellation occurs. Application of the algorithm to real speech signals is currently under investigation.

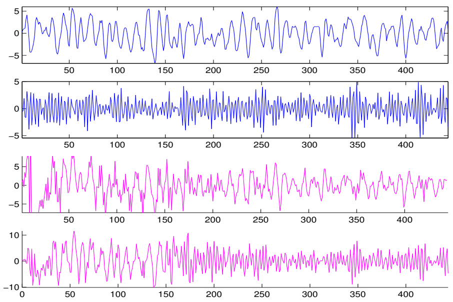

Figure 1: True (two top) and estimated (two bottom) sources.

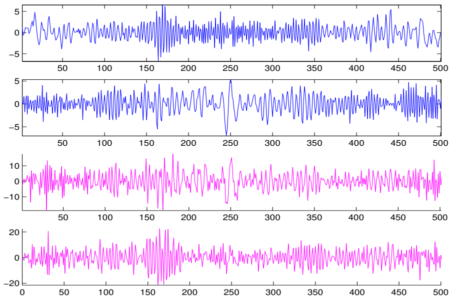

Figure 2: True (two top) and estimated (two bottom) sources.

5 REFERENCES

[1] B.D.O. Anderson and J.B. Moore, Optimal Filtering , Prentice-Hall, Englewood Cliffs, 1979.

[2] C. Andrieu, S.J. Godsill, “A Particle filter for audio source separation”, Tech. Rep. Cambridge University Engineering Department, 2000.

[3] D.L. Aspach and H.W. Sorenson, “Nonlinear Bayesian estimation using Gaussian sum approximations”, IEEE Trans. Auto. Cont., vol. 17, no. 4, pp. 439448, 1972.

[4] H. Broman, U. Lindgren, H. Sahlin and P. Stoica, “Source Separation: A TITO System Identification Approach”, Signal Processing 1998.

[5] R.S. Bucy and K.D. Senne, “Digital synthesis of nonlinear filters”, Automatica, vol. 7, 1971, pp. 287-298.

[6] D. Crisan, P. Del Moral and T. Lyons, “Discrete filtering using branching and interacting particle systems”, to appear Markov Processes and Related Fields, 2000.

[7] A. Doucet and C. Andrieu, “Iterative algorithms for optimal state estimation of jump Markov linear systems”, in Proc. Conf. IEEE ICASSP, 1999.

[8] A. Doucet, J.F.G. de Freitas and N.J. Gordon (eds.), Sequential Monte Carlo Methods in Practice, Springer-Verlag, to appear June 2000.

[9] A. Doucet, S.J. Godsill and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering”, to appear Statistics and Computing, 2000.

[10] N.J. Gordon, D.J. Salmond and A.F.M. Smith, “Novel approach to nonlinear/non-Gaussian Bayesian state estimation”, IEE Proceedings-F, vol. 140, no. 2, pp. 107-113, 1993.

[11] J.E. Handschin and D.Q. Mayne, “Monte Carlo techniques to estimate the conditional expectation in multi-stage nonlinear filtering”, Int. J. Cont., vol. 9, no. 5, 1969, pp. 547-559.

[12] C.P. Robert and G. Casella, Monte Carlo Statistical Methods, Springer-Verlag, 1999.

[13] E. Weinstein, A.V. Oppenheim, M. Feder and J.R. Buck, “Iterative and Sequential Algorithms for Multisensor Signal Enhancement”, IEEE. Trans. SP, Vol. 42, No. 4, April 1994.