Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures

François Caron, Manuel Davy, Arnaud Doucet, Emmanuel Duflos, Philippe Vanheeghe

Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures

F. Caron[1], M. Davy[1], A. Doucet[2], E. Duflos[1], P. Vanheeghe[1].

[1] CNRS/Ecole Centrale de Lille and INRIA-FUTURS SequeL team, Villeneuve d'Ascq, France

[2] University of British Columbia, Canada

Abstract

Using Kalman techniques, it is possible to perform optimal estimation in linear Gaussian state-space models. We address here the case where the noise probability density functions are of unknown functional form. A flexible Bayesian nonparametric noise model based on Dirichlet process mixtures is introduced. Efficient Markov chain Monte Carlo and Sequential Monte Carlo methods are then developed to perform optimal batch and sequential estimation in such contexts. The algorithms are applied to blind deconvolution and change point detection. Experimental results on synthetic and real data demonstrate the efficiency of this approach in various contexts.

Index Terms

Bayesian nonparametrics, Dirichlet Process Mixture, Markov Chain Monte Carlo, Rao-Blackwellization, Particle filter.

I. Introduction

Dynamic linear models are used in a variety of applications, ranging from target tracking, system identification, abrupt change detection, etc. The models are defined as follows :

$$\mathbf{x}_t = A_t\mathbf{x}_{t-1} + C_t\mathbf{u}_t + G_t\mathbf{v}_t \quad (1)$$

$$\mathbf{z}_t = H_t\mathbf{x}_t + \mathbf{w}_t \quad (2)$$

where $\mathbf{x}_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, $\mathbf{x}_t$ is the hidden state vector, $\mathbf{z}_t$ is the observation, $\mathbf{v}_t$ and $\mathbf{w}_t$ are sequences of mutually independent random variables such that $\mathbf{v}_t \overset{\text{i.i.d.}}{\sim} F^v$ and $\mathbf{w}_t \overset{\text{i.i.d.}}{\sim} F^w$. A_t and H_t are the known state and observation matrices, $\mathbf{u}_t$ is a known input, C_t the input transfer matrix and G_t is the state transfer matrix. Let us denote $\mathbf{a}_{i:j} = (\mathbf{a}_i, \mathbf{a}_{i+1}, ..., \mathbf{a}_j)$ for any sequence $\{\mathbf{a}_t\}$. The main use of

model (1)-(2) is to estimate the hidden state $\mathbf{x}_t$ given the observations $\mathbf{z}_{1:t}$ (filtering, with a forward recursion) or $\mathbf{z}_{1:T}$ for $t \leq T$ (smoothing, with a forward-backward recursion).

It is a very common choice to assume that the noise probability density functions (pdfs) F_v and F_w are Gaussian, with known parameters, as this enables the use of Kalman filtering/smoothing. In such a framework, Kalman techniques are optimal in the sense of minimizing the mean squared error. There are, however, a number of cases where the Gaussian assumption is inadequate, e.g. the actual observation noise distribution or the transition noise are multimodal (in Section VI, we provide several such examples). In this paper, we address the problem of *optimal state estimation when the probability density functions of the noise sequences are unknown and need to be estimated on-line or off-line from the data*. This problem takes place in the class of identification/estimation of linear models with unknown statistic noises.

A. Proposed approach

Our methodology[1] relies on the introduction of a Dirichlet Process Mixture (DPM), which is used to model the unknown pdfs of the state noise $\mathbf{v}_t$ and measurement noise $\mathbf{w}_t$. DPMs are flexible Bayesian nonparametric models which have become very popular in statistics over the last few years, to perform nonparametric density estimation [2–4]. Briefly, a realization of a DPM can be seen as an *infinite* mixture of pdfs with given parametric shape (e.g., Gaussian) where each pdf is denoted $f(\cdot|\theta)$. The parameters of the mixture (mixture weights and locations of the θ's) are given by the random mixture distribution $\mathbb{G}(\theta)$, which is sampled from a so-called *Dirichlet Process*. A prior distribution, denoted $\mathbb{G}_0(\theta)$ must be selected over the θ's (e.g., Normal-Inverse Wishart for the DPM of Gaussians case, where θ contains the mean vector and the covariance matrix), while the weights follow a distribution characterized by a positive real-valued parameter α. For small α, only a small fraction of the weights is significantly nonzero, whereas for large α, many weights are away from zero. Thus, the parameter α tunes the prior distribution of components in the mixture, without setting a precise number of components. Apart from this implicit, powerful clustering property, DPMs are computationally very attractive due to the so-called *Polya urn representation* which enables straightforward computation of the full conditional distributions associated to the latent variables θ.

B. Previous works

Several algorithms have been developed to estimate noise statistics in linear dynamic systems [5–8]. However, these algorithms assume Gaussian noise pdfs (with unknown mean and covariance matrix).

[1]Preliminary results were presented in Caron et al. [1].

As will be made clearer in the following, this is a special case of our framework: if the scaling coefficient α tends to 0, the realizations of the DPM of Gaussian pdfs converge in distribution to a single Gaussian with parameter prior distribution given by the base distribution $\mathbb{G}_0$. Algorithms have also been developed to deal with non-Gaussian noises distributions, such as student-t [9], α-stable [10] or mixture of Gaussians [11]. These works are based on a given prior parametric shape of the pdf which we do not assume in this paper.

Though many recent works have been devoted to DPMs in various contexts such as econometrics [12], geoscience [13] and biology [14, 15], this powerful class of models has never been used in the context of linear dynamic models (to the best of our knowledge). In this paper, we show that DPM-based dynamic models with unknown noise distributions can be defined easily. Moreover, we provide several efficient computational methods to perform Bayesian inference, ranging from Gibbs sampling (for offline estimation) to Rao-Blackwellized particle filtering for online estimation.

C. Paper organization

This paper is organized as follows. In Section II, we recall the basics of Bayesian nonparametric density estimation with DPMs. In Section III we present the dynamic model with unknown noise distributions. In Section IV we derive an efficient Markov chain Monte Carlo (MCMC) algorithm to perform optimal estimation in the batch (offline) case. In Section V, we develop a Sequential Monte Carlo (SMC) algorithm/Particle filter to perform optimal estimation in the sequential (online) case. All these algorithms can be interpreted as Rao-Blackwellized methods. In Section VII, we discuss some features of these algorithms, and we relate them to other existing approaches. Finally, in Section VI, we demonstrate our algorithms on two applications: blind deconvolution of impulse processes and a change point problem in biomedical time series. The last section is devoted to conclusions and future research directions.

II. Bayesian nonparametric density estimation

In this section, we review briefly Bayesian nonparametric density estimation[2]. We introduce Dirichlet processes as probabilistic measures on the space of probability measures, and we outline its discreteness. Then, the DPM model in presented.

[2]There are many ways to understand 'nonparametric'. In this paper, we follow many other papers in the same vein [2–4], where 'nonparametric' refers to the fact that the pdf of interest cannot be defined by a functional expansion with a finite-dimensional parameter space.

A. Density estimation

Let $\mathbf{y}_1, ..., \mathbf{y}_n$ be a statistically exchangeable sequence distributed with

$$\mathbf{y}_k \sim F(\cdot) \tag{3}$$

where $\sim$ means *distributed according to*. We are interested here in estimating $F(\cdot)$ and we consider the following nonparametric model

$$F(\mathbf{y}) = \int_{\Theta} f(\mathbf{y}|\theta) d\mathbb{G}(\theta) \tag{4}$$

where $\theta \in \Theta$ is called the latent variable or cluster variable, $f(\cdot|\theta)$ is the mixed pdf and $\mathbb{G}(\cdot)$ is the mixing distribution. Within the Bayesian framework, it is assumed that $\mathbb{G}(\cdot)$ is a Random Probability Measure (RPM) [4] distributed according to a prior distribution (i.e., a distribution over the set of probability distributions). We will select here the RPM to follow a Dirichlet Process (DP) prior.

B. Dirichlet Processes

Ferguson [16] introduced the Dirichlet Process (DP) as a probability measure on the space of probability measures. Given a probability measure $\mathbb{G}_0(\cdot)$ on a (measurable) space $(\mathcal{T}, \mathcal{A})$ and a positive real number α, a probability distribution $\mathbb{G}(\cdot)$ distributed according to a DP of base distribution $\mathbb{G}_0(\cdot)$ and scale factor α, denoted $\mathbb{G}(\cdot) \sim DP(\mathbb{G}_0(\cdot), \alpha)$, satisfies for any partition $A_1, ..., A_k$ of $\mathcal{T}$ and any k

$$(\mathbb{G}(A_1), ..., \mathbb{G}(A_k)) \sim \mathcal{D}\left(\alpha\mathbb{G}_0(A_1), ..., \alpha\mathbb{G}_0(A_k)\right) \tag{5}$$

where $\mathcal{D}$ is a standard Dirichlet distribution, classically defined for a set of random variables $(b_0, .., b_p) \sim \mathcal{D}(a_0, .., a_p)$ by

$$\mathcal{D}(a_0, .., a_p) = \frac{\Gamma(\sum_{l=0}^{p} a_l)}{\prod_{l=0}^{p} \Gamma(a_l)} \prod_{l=0}^{p} b_l^{a_l - 1} \delta_1(\sum_{l=0}^{p} b_l) \tag{6}$$

where Γ is the gamma function, and $\delta_u(v)$ is the Dirac delta function, which is zero whenever $v \neq u$. From the definition in Eq. (5), it is easy to show that for every $B \in \mathcal{T}$

$$\mathbb{E}\left[\mathbb{G}(B)\right] = \mathbb{G}_0(B) \tag{7}$$

$$\text{var}\left[\mathbb{G}(B)\right] = \frac{\mathbb{G}_0(B)\big(1 - \mathbb{G}_0(B)\big)}{1 + \alpha} \tag{8}$$

An important property is that the realizations of a Dirichlet process are *discrete*, with probability one. One can show that $\mathbb{G}$ admits the so-called *stick-breaking* representation, established by Sethuraman [17]:

$$\mathbb{G}(\cdot) \;=\; \sum_{j=1}^{\infty} \pi_j \delta_{U_j}(\cdot) \tag{9}$$

with $U_j \sim \mathbb{G}_0(\cdot)$, $\pi_j = \beta_j \prod_{l=1}^{j-1}(1-\beta_l)$ and $\beta_j \sim \mathcal{B}(1,\alpha)$ where $\mathcal{B}$ denotes the beta distribution. In the following, we omit $(\cdot)$ in $\mathbb{G}(\cdot)$ and other distributions, to simplify notations. Using Eq. (4), it comes that the following flexible prior model is adopted for the unknown distribution F

$$F(\mathbf{y}) = \sum_{j=1}^{\infty} \pi_j f(\mathbf{y}|U_j). \tag{10}$$

Apart from its flexibility, a fundamental motivation to use the DP model is the simplicity of the posterior update. Let $\theta_1, \ldots, \theta_n$ be n random samples from $\mathbb{G}$

$$\theta_k|\mathbb{G} \overset{\text{i.i.d.}}{\sim} \mathbb{G} \tag{11}$$

where $\mathbb{G} \sim DP(\mathbb{G}_0, \alpha)$ then the posterior distribution of $\mathbb{G}|\theta_{1:n}$ is also a DP

$$\mathbb{G}|\theta_{1:n} \sim DP(\frac{\alpha}{\alpha+n}\mathbb{G}_0 + \frac{1}{\alpha+n}\sum_{k=1}^{n}\delta_{\theta_k}\,,\, \alpha+n) \tag{12}$$

Moreover, it can be shown that the predictive distribution, computed by integrating out the RPM $\mathbb{G}$, admits the following *Polya urn* representation [18]

$$\theta_{n+1}|\theta_{1:n} \sim \frac{1}{\alpha+n}\sum_{k=1}^{n}\delta_{\theta_k} + \frac{\alpha}{\alpha+n}\mathbb{G}_0. \tag{13}$$

Therefore, conditionally on the latent variables $\theta_{1:n}$ sampled previously, the probability that a new sample is identical to an existing one is overall $\frac{n}{\alpha+n}$, whereas, with probability $\frac{\alpha}{\alpha+n}$, the new sample is distributed (independently) according to $\mathbb{G}_0$. It should be noted that several θ_k's might have the same value, thus the number of "alive" clusters (denoted M), that is, the number of distinct values of θ_k, is less than n.

The *scaling coefficient* α tunes the number of "alive" clusters M. For large n, Antoniak [19] showed that $\mathbb{E}[M|\alpha,n] \simeq \alpha\log(1+\frac{n}{\alpha})$. As α tends to zero, most of the samples θ_k share the same value, whereas when α tends to infinity, the θ_k are almost i.i.d. samples from $\mathbb{G}_0$.

C. Dirichlet Process Mixtures

Using these modeling tools, it is now possible to reformulate the density estimation problem using the following hierarchical model known as DPM [19]:

$$\begin{aligned} \mathbb{G} &\sim DP(\mathbb{G}_0, \alpha,), \quad \text{and, for } k = 1, \ldots, n \\ \theta_k|\mathbb{G} &\sim \mathbb{G}, \\ \mathbf{y}_k|\theta_k &\sim f(\cdot|\theta_k) \end{aligned} \tag{14}$$

It should be noted that DPMs can model a wide variety of pdfs. In particular, assuming Gaussian $f(\cdot|\theta_k)$, the parameter contains both the mean and the covariance, and, depending on $\mathbb{G}_0$, the corresponding DPM may have components with large/small variances.

D. Estimation objectives

The objective of DPM-based density estimation boils down to estimating the posterior distribution $p(\theta_{1:n}|\mathbf{y}_{1:n})$, because the probability $\mathbb{G}$ can be integrated out analytically by using the Polya urn representation. Although DPMs were introduced in the 70's, these models were too complex to handle numerically before the introduction of Monte Carlo simulation based methods. Efficient MCMC algorithms [2, 3, 20–22] as well as Sequential Importance Sampling [23, 24] enable to sample from $p(\theta_{1:n}|\mathbf{y}_{1:n})$. However, these algorithms cannot be applied to our class of models, which is presented below, because the noise sequences $\mathbf{v}_t$ and $\mathbf{w}_t$ are not observed directly.

III. Dynamic Linear Model with Unknown Noise Distribution

The linear dynamic model defined in Eq.'s (1)-(2) relies on the unknown noises $\{\mathbf{v}_t\}$ and $\{\mathbf{w}_t\}$ distributions, which are assumed to be DPMs in this paper.

A. DPM noise models

For both $\{\mathbf{v}_t\}$ and $\{\mathbf{w}_t\}$, the pdf $f(\cdot|\theta)$ is assumed here to be a Gaussian, denoted $\mathcal{N}(\mu_t^v, \Sigma_t^v)$ and $\mathcal{N}(\mu_t^w, \Sigma_t^w)$ respectively. The base distributions $\mathbb{G}_0^v$ and $\mathbb{G}_0^w$ are assumed to be normal inverse Wishart distributions [25] denoted $\mathbb{G}_0^v = \mathcal{NIW}(\mu_0^v, \kappa_0^v, \nu_0^v, \Lambda_0^v)$ and $\mathbb{G}_0^w = \mathcal{NIW}(\mu_0^w, \kappa_0^w, \nu_0^w, \Lambda_0^w)$. The hyperparameters $\psi^v = \{\mu_0^v, \kappa_0^v, \nu_0^v, \Lambda_0^v\}$ and $\psi^w = \{\mu_0^w, \kappa_0^w, \nu_0^w, \Lambda_0^w\}$ are assumed fixed but unknown. Finally, the scale parameters α^v and α^w are also assumed fixed and unknown. Overall, the sets of hyperparameters are denoted $\phi^v = \{\alpha^v, \psi^v\}$, $\phi^w = \{\alpha^w, \psi^w\}$ and $\phi = \{\phi^v, \phi^w\}$. For the sake of presentation clarity, we assume that these hyperparameters are known, but in Subsection IV-B, we address the case of unknown hyperparameters by defining priors and a specific estimation procedure.

To summarize, we have the following models

$$\mathbb{G}^v|\phi^v \sim DP(\mathbb{G}_0^v, \alpha^v), \qquad \mathbb{G}^w|\phi^w \sim DP(\mathbb{G}_0^w, \alpha^w), \tag{15}$$

and for $t = 1, 2, \ldots$

$$\begin{aligned} &\theta_t^v|\mathbb{G}^v \overset{\text{i.i.d.}}{\sim} \mathbb{G}^v, && \theta_t^w|\mathbb{G}^w \overset{\text{i.i.d.}}{\sim} \mathbb{G}^w, \\ &\mathbf{v}_t|\theta_t^v \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_t^v, \Sigma_t^v). && \mathbf{w}_t|\theta_t^w \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_t^w, \Sigma_t^w). \end{aligned} \tag{16}$$

where $\theta_t^v = \{\mu_t^v, \Sigma_t^v\}$ (resp. $\theta_t^w = \{\mu_t^w, \Sigma_t^w\}$) is the latent cluster variable giving the mean and covariance matrix for that cluster, and $\theta_t = \{\theta_t^v, \theta_t^w\}$. This model is written equivalently as $\mathbf{v}_t \sim F^v(\mathbf{v}_t)$ and $\mathbf{w}_t \sim F^w(\mathbf{w}_t)$ where F^v and F^w are fixed but unknown distributions written as

$$F^v(\mathbf{v}_t) = \int \mathcal{N}(\mathbf{v}_t; \mu, \Sigma) d\mathbb{G}^v(\mu, \Sigma), \tag{17}$$

$$F^w(\mathbf{w}_t) = \int \mathcal{N}(\mathbf{w}_t; \mu, \Sigma) d\mathbb{G}^w(\mu, \Sigma) \tag{18}$$

In other words, F^v and F^w are countable infinite mixtures of Gaussian pdfs of unknown parameters, and the mixing distributions $\mathbb{G}^v$ and $\mathbb{G}^w$ are sampled from Dirichlet processes.

B. Estimation of the state parameters

In this work, our objective is to estimate $\mathbb{G}^v$ and $\mathbb{G}^w$ as well as the latent variables $\{\theta_t\}$ and state variable $\{\mathbf{x}_t\}$ at each time t, conditional on the observations $\{\mathbf{z}_t\}$. In practice, only the state variable is of interest – $\mathbb{G}^v$, $\mathbb{G}^w$ and $\{\theta_t\}$ are *nuisance* parameters. Ideally, one would like to estimate online the sequence of posterior distributions $p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t},\phi)$ as t increases or the offline posterior $p(\mathbf{x}_{0:T}|\mathbf{z}_{1:T},\phi)$, where T is the fixed length of the observation sequence $\mathbf{z}_{1:T}$. Thanks to the Polya urn representation, it is possible to integrate out analytically $\mathbb{G}^v$ and $\mathbb{G}^w$ from these posteriors. The parameters $\theta_{1:t}$ and $\theta_{1:T}$ remain and the inference is based upon $p(\mathbf{x}_{0:t},\theta_{1:t}|\mathbf{z}_{1:t},\phi)$ or $p(\mathbf{x}_{0:T},\theta_{1:T}|\mathbf{z}_{1:T},\phi)$. The posterior $p(\mathbf{x}_{0:t},\theta_{1:t}|\mathbf{z}_{1:t},\phi)$ satisfies for any t

$$p(\mathbf{x}_{0:t},\theta_{1:t}|\mathbf{z}_{1:t},\phi) = p(\mathbf{x}_{0:t}|\theta_{1:t},\mathbf{z}_{1:t},\phi)p(\theta_{1:t}|\mathbf{z}_{1:t},\phi). \tag{19}$$

Conditional upon θ_t, Eq.'s (1)-(2) may be rewritten as

$$\mathbf{x}_t = F_t\mathbf{x}_{t-1} + \mathbf{u}'_t(\theta_t) + G_t\mathbf{v}'_t(\theta_t) \tag{20}$$

$$\mathbf{z}_t = H_t\mathbf{x}_t + \mu^w_t + \mathbf{w}'_t(\theta_t) \tag{21}$$

where $\mathbf{u}'_t(\theta_t) = C_t\mathbf{u}_t + G_t\mu^v_t$ and μ^w_t are known inputs, $\mathbf{v}'_t(\theta_t)$ and $\mathbf{w}'_t(\theta_t)$ are centered white Gaussian noise of known covariance matrices Σ^v_t and Σ^w_t, respectively. Thus $p(\mathbf{x}_{0:t}|\theta_{1:t},\mathbf{z}_{1:t},\phi)$ (resp. $p(\mathbf{x}_{0:T}|\theta_{1:T},\mathbf{z}_{1:T},\phi)$) is a Gaussian distribution whose parameters can be computed using a Kalman filter (resp. smoother) [26] for given $\theta_{1:t}$ (resp.$\theta_{1:T}$).

One is generally interested in computing the marginal MMSE state estimate $\widehat{\mathbf{x}}^{\text{MMSE}}_{t|t'} = \mathbb{E}\left[\mathbf{x}_t|\mathbf{z}_{1:t'}\right]$ (with $t'=t$ or $t'=T$)

$$\begin{aligned}
\widehat{\mathbf{x}}^{\text{MMSE}}_{t|t'} &= \int \mathbf{x}_t p(\mathbf{x}_t,\theta_{1:t'}|\mathbf{z}_{1:t'},\phi)d(\mathbf{x}_t,\theta_{1:t'}) \\
&= \int \mathbf{x}_t p(\mathbf{x}_t|\theta_{1:t'},\mathbf{z}_{1:t'},\phi)p(\theta_{1:t'}|\mathbf{z}_{1:t'},\phi)d(\mathbf{x}_t,\theta_{1:t'}) \\
&= \int \widehat{\mathbf{x}}_{t|t'}(\theta_{1:t'})p(\theta_{1:t'}|\mathbf{z}_{1:t'},\phi)d\theta_{1:t'}
\end{aligned} \tag{22}$$

where $\widehat{\mathbf{x}}_{t|t}(\theta_{1:t})$ (resp. $\widehat{\mathbf{x}}_{t|T}(\theta_{1:T})$) is the mean of the Gaussian $p(\mathbf{x}_t|\theta_{1:t},\mathbf{z}_{1:t},\phi)$ (resp. $p(\mathbf{x}_t|\theta_{1:T},\mathbf{z}_{1:T},\phi)$). Both $\widehat{\mathbf{x}}_{t|t}(\theta_{1:t})$ and $\widehat{\mathbf{x}}_{t|T}(\theta_{1:t})$ are computed by the Kalman filter/smoother, see Sections IV and V below.

Computing these estimates still requires integration w.r.t. the θ's, see Eq. (22). This kind of integral is not feasible in closed-form, but it can be computed numerically by using Monte Carlo

integration [27]. Briefly, assume that a set of N weighted samples $\{\theta_{1:t}^{(i)}\}_{i=1,...,N}$ with weights $w_t^{(i)}$ are distributed according to $p(\theta_{1:t}|\mathbf{z}_{1:t},\phi)$, then e.g., $\widehat{\mathbf{x}}_{t|t}^{\text{MMSE}}$ is computed as

$$\widehat{\mathbf{x}}_{t|t}^{\text{MMSE}} \approx \sum_{i=1}^{N} w_t^{(i)} \widehat{\mathbf{x}}_{t|t}(\theta_{1:t}^{(i)}) \tag{23}$$

In Eq. (23), the main difficulty consists of generating the weighted samples $\{\theta_{1:t}^{(i)}\}_{i=1,...,N}$ from the marginal posterior $p(\theta_{1:t}|\mathbf{z}_{1:t},\phi)$ (and similarly, from $p(\theta_{1:T}|\mathbf{z}_{1:T},\phi)$ in the offline case).

- For **offline (batch) estimation** $(t = T)$, this can be done by MCMC by building a Markov chain of samples $\{\theta_{1:T}^{(i)}\}_{i=1,...,N}$ with target distribution $p(\theta_{1:T}|\mathbf{z}_{1:T},\phi)$ (in that case, $w_t^{(i)} = 1/N$). The MCMC algorithms available in the literature to estimate these Bayesian nonparametric models – e.g. [3, 21] – are devoted to density estimation in cases where the data are observed directly. They do not apply to our case because here, the sequences $\{\mathbf{v}_t\}$ and $\{\mathbf{w}_t\}$ are not observed directly. One only observes $\{\mathbf{z}_t\}$, assumed to be generated by the dynamic model (1)-(2). Section IV proposes an MCMC algorithm dedicated to this model.
- For **online (sequential) estimation**, samples can be generated by sequential importance sampling, as detailed in Section V.

IV. MCMC ALGORITHM FOR OFF-LINE STATE ESTIMATION

In this Section, we consider the offline state estimation. As outlined above, this requires to compute estimates from the posterior $p(\mathbf{x}_{0:T}, \theta_{1:T}|\mathbf{z}_{1:T})$, where we recall that $\theta_t = \{\theta_t^v, \theta_t^w\} = \{\mu_t^v, \Sigma_t^v, \mu_t^w, \Sigma_t^w\}$ is the latent variable as defined above. We first assume that the hyperparameters are fixed and known (Subsection IV-A), then we let them be unknown, with given prior distributions (Subsection IV-B).

A. Fixed and known hyperparameters

In this subsection, the hyperparameter vector ϕ is assumed fixed and known. The marginal posterior $p(\theta_{1:T}|\mathbf{z}_{1:T},\phi)$ can be approximated through MCMC using the Gibbs sampler [27] presented in Algorithm 1 below.

Algorithm 1: Gibbs sampler to sample from $p(\theta_{1:T}|\mathbf{z}_{1:T},\phi)$

- Initialization: For $t = 1, ..., T$, sample $\theta_t^{(1)}$ from an arbitrary initial distribution, e.g. the prior.
- Iteration i, $i = 2, \ldots, N' + N$:
  - For $t = 1, \ldots, T$, sample $\theta_t^{(i)} \sim p(\theta_t|\mathbf{z}_{1:T}, \theta_{-t}^{(i)}, \phi)$ where $\theta_{-t}^{(i)} = \{\theta_1^{(i)}, .., \theta_{t-1}^{(i)}, \theta_{t+1}^{(i-1)}, .., \theta_T^{(i-1)}\}$

To implement Algorithm 1, one needs to sample from the conditional pdf $p(\theta_t|\mathbf{z}_{1:T}, \theta_{-t}, \phi)$ for each of the $N' + N$ iterations (including N' burn-in iterations). From Bayes' rule, we have

$$p(\theta_t|\mathbf{z}_{1:T}, \theta_{-t}, \phi) \propto p(\mathbf{z}_{1:T}|\theta_{1:T}) p(\theta_t|\theta_{-t}, \phi). \tag{24}$$

where $p(\theta_t|\theta_{-t},\phi) = p(\theta_t^v|\theta_{-t}^v,\phi^v)p(\theta_t^w|\theta_{-t}^w,\phi^w)$. From the Polya urn representation, these two terms are written as (for w, replace v with w below):

$$p(\theta_t^v|\theta_{-t}^v,\phi^v) \;=\; \frac{1}{\alpha^v+T-1}\sum_{k=1,k\neq t}^{T}\delta_{\theta_k^v}(\theta_t^v) + \frac{\alpha^v}{\alpha^v+T-1}\mathbb{G}_0^v(\theta_t^v|\psi^v), \tag{25}$$

Thus $p(\theta_t|\mathbf{z}_{1:T},\theta_{-t},\phi)$ can be sampled from with a Metropolis-Hastings (MH) step, where the candidate pdf is the conditional prior $p(\theta_t|\theta_{-t},\phi)$. The acceptance probability is thus given by

$$\rho(\theta_t^{(i)},\theta_t^{(i)*}) = \min\left(1, \frac{p(\mathbf{z}_{1:T}|\theta_t^{(i)*},\theta_{-t}^{(i)})}{p(\mathbf{z}_{1:T}|\theta_t^{(i)},\theta_{-t}^{(i)})}\right) \tag{26}$$

where $\theta_t^{(i)*}$ is the candidate cluster sampled from $p(\theta_t|\theta_{-t},\phi)$.

The computation of the acceptance probability requires to compute the likelihood $p(\mathbf{z}_{1:T}|\theta_t^{(i)},\theta_{-t}^{(i)})$. This can be done in $O(T)$ operations using a Kalman filter. However, this has to be done for $t = 1,\ldots,T$ and one finally obtains an algorithm of computational complexity $O(T^2)$. Here, we propose to use instead the backward-forward recursion developed in [28], to obtain an algorithm of overall complexity $O(T)$. This algorithm uses the following likelihood decomposition obtained by applying conditional probability rules to $p(\mathbf{z}_{1:t-1},\mathbf{z}_t,\mathbf{z}_{t+1:T}|\theta_{1:T})$

$$p(\mathbf{z}_{1:T}|\theta_{1:T}) = p(\mathbf{z}_{1:t-1}|\theta_{1:t-1})p(\mathbf{z}_t|\theta_{1:t},\mathbf{z}_{1:t-1})\int_{\mathcal{X}} p(\mathbf{z}_{t+1:T}|\mathbf{x}_t,\theta_{t+1:T})p(\mathbf{x}_t|\mathbf{z}_{1:t},\theta_{1:t})d\mathbf{x}_t \tag{27}$$

with

$$p(\mathbf{z}_{t:T}|\mathbf{x}_{t-1},\theta_{t:T}) \;=\; \int_{\mathcal{X}} p(\mathbf{z}_{t+1:T}|\mathbf{x}_{t-1},\theta_{t:T})p(\mathbf{z}_t,\mathbf{x}_t|\theta_t,\mathbf{x}_{t-1})d\mathbf{x}_t \tag{28}$$

The first two terms of the r.h.s. in Eq. (27) are computed by a forward recursion based on the Kalman filter [28]. The third term can be evaluated by a backward recursion according to Eq. (28). It is shown in [28] that if $\int_{\mathcal{X}} p(\mathbf{z}_{t:T}|\mathbf{x}_{t-1},\theta_{t:T})d\mathbf{x}_{t-1} < \infty$ then $\frac{p(\mathbf{z}_{t:T}|\mathbf{x}_{t-1},\theta_{t:T})}{\int_{\mathcal{X}} p(\mathbf{z}_{t:T}|\mathbf{x}_{t-1},\theta_{t:T})d\mathbf{x}_{t-1}}$ is a Gaussian distribution w.r.t. $\mathbf{x}_{t-1}$, of mean $m'_{t-1|t}(\theta_{t:T})$ and covariance $P'_{t-1|t}(\theta_{t:T})$. Even if $p(\mathbf{z}_{t:T}|\mathbf{x}_{t-1},\theta_{t:T})$ is not integrable in $\mathbf{x}_{t-1}$, the quantities $P'^{-1}_{t-1|t}(\theta_{t:T})$ and $P'^{-1}_{t-1|t}(\theta_{t:T})m'_{t-1|t}(\theta_{t:T})$ satisfy the backward information filter recursion (see Appendix). Based on Eq. (27), the density $p(\theta_t|\mathbf{z}_{1:T},\theta_{-t},\phi)$ is expressed by

$$p(\theta_t|\mathbf{z}_{1:T},\theta_{-t}) \;\propto\; p(\theta_t|\theta_{-t},\phi)p(\mathbf{z}_t|\theta_{1:t},\mathbf{z}_{1:t-1})\int_{\mathcal{X}} p(\mathbf{z}_{t+1:T}|\mathbf{x}_t,\theta_{t+1:T})p(\mathbf{x}_t|\mathbf{z}_{1:t},\theta_{1:t})d\mathbf{x}_t \tag{29}$$

Algorithm 2 summarizes the full posterior sampling procedure. It is the step-by-step description of Algorithm 1 that accounts for the factorization of the likelihood given by Eq. (27).

Algorithm 2: MCMC algorithm to sample from $p(\theta_{1:T}|\mathbf{z}_{1:T},\phi)$

Initialization $i = 1$

- For $t = 1,...,T$, sample $\theta_t^{(1)}$.

Iteration i, $i = 2,\ldots,N'+N$

- Backward recursion: For $t = T, .., 1$, compute and store $P'^{-1}_{t|t+1}(\theta^{(i-1)}_{t+1:T})$ and $P'^{-1}_{t|t+1}(\theta^{(i-1)}_{t+1:T})m'_{t|t+1}(\theta^{(i-1)}_{t+1:T})$
- Forward recursion: For $t = 1, .., T$
  - Perform a Kalman filter step with $\theta_t = \theta^{(i-1)}_t$, store $\widehat{\mathbf{x}}_{t|t}(\theta^{(i)}_{1:t-1}, \theta^{(i-1)}_t)$ and $\Sigma_{t|t}(\theta^{(i)}_{1:t-1}, \theta^{(i-1)}_t)$.
  - Metropolis-Hastings step :
    - Sample a candidate cluster
$$\theta^{(i)*}_t \sim p(\theta_t|\theta^{(i)}_{-t}, \phi) \tag{30}$$
    - Perform a Kalman filter step with $\theta_t = \theta^{(i)*}_t$, store $\widehat{\mathbf{x}}_{t|t}(\theta^{(i)}_{1:t-1}, \theta^{(i)*}_t)$ and $\Sigma_{t|t}(\theta^{(i)}_{1:t-1}, \theta^{(i)*}_t)$
    - Compute
$$\rho(\theta^{(i)}_t, \theta^{(i)*}_t) = \min\left(1, \frac{p(\mathbf{z}_{1:T}|\theta^{(i)*}_t, \theta^{(i)}_{-t})}{p(\mathbf{z}_{1:T}|\theta^{(i)}_t, \theta^{(i)}_{-t})}\right) \tag{31}$$
    - With probability $\rho(\theta^{(i)}_t, \theta^{(i)*}_t)$, set $\theta^{(i)}_t = \theta^{(i)*}_t$, otherwise $\theta^{(i)}_t = \theta^{(i-1)}_t$.

State post-Sampling (for non-burn-in iterations only)

- For $i = N' + 1, ..., N' + N$, compute $\widehat{\mathbf{x}}_{t|T}(\theta^{(i)}_{1:T}) = \mathbb{E}\left(\mathbf{x}_t|\theta^{(i)}_{1:T}, \mathbf{z}_{1:T}\right)$ for all t with a Kalman smoother.

It can be easily established that the simulated Markov chain $\left\{\theta^{(i)}_{1:T}\right\}$ is ergodic with limiting distribution $p(\theta_{1:T}|\mathbf{z}_{1:T})$. After N' burn-in, the N last iterations of the algorithm are kept, and the MMSE estimates of θ_t and $\mathbf{x}_t$ for all $t = 0, \ldots, T$ are computed as explained in Subsection III-B, using

$$\widehat{\theta}^{\text{MMSE}}_{t|T} = \frac{1}{N}\sum_{i=N'+1}^{N'+N} \theta^{(i)}_t \quad \widehat{\mathbf{x}}^{\text{MMSE}}_{t|T} = \frac{1}{N}\sum_{i=N'+1}^{N'+N} \widehat{\mathbf{x}}_{t|T}(\theta^{(i)}_{1:T}) \tag{32}$$

B. Unknown hyperparameters

The hyperparameters in vector ϕ have some influence on the correct estimation of the DPMs F^v and F^w. In this subsection, we include them in the inference by considering them as unknowns with prior distributions:

$$\alpha^v \sim \mathcal{G}(\frac{\eta}{2}, \frac{\nu}{2}),\ \alpha^w \sim \mathcal{G}(\frac{\eta}{2}, \frac{\nu}{2}), \tag{33}$$

$$\psi^v \sim p_0(\psi^v),\ \psi^w \sim p_0(\psi^w) \tag{34}$$

where η and ν are known constants and p_0 is a pdf with fixed and known parameters. The posterior probability $p(\alpha^v|\mathbf{x}_{1:T}, \theta_{1:T}, \mathbf{z}_{1:T}, \psi^v, \phi^w)$ reduces to $p(\alpha^v|M^v, T)$ where M^v is the number of distinct values taken by the clusters $\theta^v_{1:T}$. As shown in [19], this pdf can be expressed by

$$p(\alpha^v|M^v, T) \propto \frac{s(T, M^v)(\alpha^v)^{M^v}}{\sum_{k=1}^{T} s(T, k)(\alpha^v)^k} p(\alpha^v) \tag{35}$$

where the $s(T,k)$ are the absolute values of Stirling numbers of the first kind. We can sample from the above pdf with a Metropolis-Hasting step using the prior Gamma pdf $p(\alpha^v) = \mathcal{G}(\frac{\eta}{2}, \frac{\nu}{2})$ as proposal (and similarly for α^w). Other methods have been proposed that allow direct sampling, see for example West [29], and Escobar and West [21].

The posterior probability $p(\psi^v|\mathbf{x}_{1:T}, \theta_{1:T}, \mathbf{z}_{1:T}, \alpha^v, \phi^w)$ reduces to $p(\psi^v|\theta^{v\,\prime}_{1:M^v})$ where $\theta^{v\,\prime}_{1:M^v}$ is the set of distinct values taken by the clusters $\theta^v_{1:T}$. It is expressed by

$$p(\psi^v|\mathbf{x}_{1:T}, \theta_{1:T}, \mathbf{z}_{1:T}, \alpha^v, \phi^w) \propto p_0(\psi^v) \prod_{k=1}^{M^v} \mathbb{G}_0^v(\theta_k^{v\,\prime}|\psi^v) \tag{36}$$

We can sample from this pdf with a Metropolis-Hasting step using the prior Gamma pdf $p_0(\psi^v)$ as proposal whenever direct sampling is not possible.

V. Rao-Blackwellized Particle Filter algorithm for online state estimation

Many applications, such as target tracking, require *online* state estimation. In this case, the MCMC approach is inadequate as it requires availability of the entire dataset to perform state estimation. In this section, we develop the online counterpart to the MCMC procedure presented in Section IV: a sequential Monte Carlo method (also known as particle filter) is implemented, to sample on-line from the sequence of probability distributions $\{p(\mathbf{x}_{0:t}, \theta_{1:t}|\mathbf{z}_{1:t}),\ t = 1, 2, \ldots\}$. Here, the hyperparameter vector ϕ is assumed to be known, therefore it is omitted in the following. Online hyperparameter estimation is discussed in Section VII.

As explained in Subsection III-B, we need to sample from $p(\theta_{1:t}|\mathbf{z}_{1:t})$, because $p(\mathbf{x}_{0:t}|\theta_{1:t}, \mathbf{z}_{1:t})$ can be computed using Kalman techniques. (The sampling procedure is indeed a generalization of the Rao-Blackwellized particle filter [30] to DPMs.) At time t, $p(\mathbf{x}_t, \theta_{1:t}|\mathbf{z}_{1:t})$ is approximated through a set of N particles $\theta^{(1)}_{1:t}, \ldots, \theta^{(N)}_{1:t}$ by the following empirical distribution

$$P_N(\mathbf{x}_t, \theta_{1:t}|\mathbf{z}_{1:t}) = \sum_{i=1}^{N} w_t^{(i)} \mathcal{N}(\mathbf{x}_t; \widehat{\mathbf{x}}_{t|t}(\theta^{(i)}_{1:t}), \Sigma_{t|t}(\theta^{(i)}_{1:t})) \tag{37}$$

The parameters $\widehat{\mathbf{x}}_{t|t}(\theta^{(i)}_{1:t})$ and $\Sigma_{t|t}(\theta^{(i)}_{1:t})$ are computed recursively for each particle i using the Kalman filter [26]. In order to build the algorithm, we note that

$$p(\theta^{(i)}_{1:t}|\mathbf{z}_{1:t}) \propto p(\theta^{(i)}_{1:t-1}|\mathbf{z}_{1:t-1}) p(\mathbf{z}_t|\theta^{(i)}_{1:t}, \mathbf{z}_{1:t-1}) p(\theta^{(i)}_t|\theta^{(i)}_{1:t-1}) \tag{38}$$

where

$$\begin{aligned} p(\mathbf{z}_t|\theta^{(i)}_{1:t}, \mathbf{z}_{1:t-1}) &= p(\mathbf{z}_t|\theta^{(i)}_t, \theta^{(i)}_{1:t-1}, \mathbf{z}_{1:t-1}) \\ &= \mathcal{N}(\mathbf{z}_t; \widehat{\mathbf{z}}_{t|t-1}(\theta^{(i)}_{1:t}), S_{t|t-1}(\theta^{(i)}_{1:t})) \end{aligned}$$

and

$$\widehat{\mathbf{z}}_{t|t-1}(\theta_{1:t}^{(i)}) = H_t \left[F_t \, \widehat{\mathbf{x}}_{t-1|t-1}(\theta_{1:t-1}^{(i)}) + C_t \mathbf{u}_t + G_t \mu_t^{v\ (i)} \right] + \mu_t^{w\ (i)} \tag{39}$$
$$S_{t|t-1}(\theta_{1:t}^{(i)}) = H_t \left[F_t \, \Sigma_{t-1|t-1}(\theta_{1:t-1}^{(i)}) \, F_t^{\mathsf{T}} + G_t \Sigma_t^{v\ (i)} G_t^{\mathsf{T}} \right] H_t^{\mathsf{T}} + \Sigma_t^{w\ (i)}$$

The Rao-Blackwellized Particle Filter (RBPF) algorithm proceeds as follows.

Algorithm 3: Rao-Blackwellized Particle Filter to sample from $p(\theta_{1:t}|\mathbf{z}_{1:t})$

At time 0.

- For $i = 1, .., N$, sample $\left(\widehat{\mathbf{x}}_{0|0}^{(i)}, \Sigma_{0|0}^{(i)} \right) \sim p_0(\mathbf{x}_{0|0}, \Sigma_{0|0})$.
- Set $w_0^{(i)} \leftarrow \frac{1}{N}$

At each time t ($t \geq 1$), do for $i = 1, \ldots, N$

- Sample $\widetilde{\theta}_t^{(i)} \sim q(\theta_t | \theta_{1:t-1}^{(i)}, \mathbf{z}_{1:t})$
- Compute $\{\widehat{\mathbf{x}}_{t|t-1}(\theta_{1:t-1}^{(i)}, \widetilde{\theta}_t^{(i)}), \Sigma_{t|t-1}(\theta_{1:t-1}^{(i)}, \widetilde{\theta}_t^{(i)}), \widehat{\mathbf{x}}_{t|t}(\theta_{1:t-1}^{(i)}, \widetilde{\theta}_t^{(i)}), \Sigma_{t|t}(\theta_{1:t-1}^{(i)}, \widetilde{\theta}_t^{(i)})\}$ by using a Kalman filter step from $\{\widehat{\mathbf{x}}_{t-1|t-1}(\theta_{1:t-1}^{(i)}), \Sigma_{t-1|t-1}(\theta_{1:t-1}^{(i)}), \widetilde{\theta}_t^{(i)}, \mathbf{z}_t)\}$
- For $i = 1, \ldots, N$, update the weights according to

$$\widetilde{w}_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(\mathbf{z}_t | \theta_{1:t-1}^{(i)}, \widetilde{\theta}_t^{(i)}, \mathbf{z}_{1:t-1}) p(\widetilde{\theta}_t^{(i)} | \theta_{1:t-1}^{(i)})}{q(\widetilde{\theta}_t^{(i)} | \theta_{1:t-1}^{(i)}, \mathbf{z}_{1:t})} \tag{40}$$

- Compute $S = \sum_{i=1}^N \widetilde{w}_t^{(i)}$ and for $i = 1, \ldots, N$, set $\widetilde{w}_t^{(i)} \leftarrow \frac{\widetilde{w}_t^{(i)}}{S}$
- Compute $N_{\text{eff}} = \left[\sum_{i=1}^N \left(\widetilde{w}_t^{(i)} \right)^2 \right]^{-1}$
- If $N_{\text{eff}} \leq \eta$, then resample the particles – that is, duplicate the particles with large weights are remove the particles with small weights. This results in a new set of particles denoted $\theta_t^{(i)}$ with weights $w_t^{(i)} = \frac{1}{N}$
- Otherwise, rename the particles and weights by removing the $\widetilde{\cdot}$'s.

Particle filtering convergence results indicate that the variance of the Monte Carlo estimates depends highly on the importance distribution selected. Here, the conditionally optimal importance distribution is $q(\theta_t | \theta_{1:t-1}^{(i)}, \mathbf{z}_{1:t}) = p(\theta_t | \theta_{1:t-1}^{(i)}, \mathbf{z}_{1:t})$, see [30]. However, it cannot be used, as the associated importance weights do not admit a closed-form expression[3]. In practice, the evolution pdf $p(\theta_t | \theta_{1:t-1})$ was used as the importance distribution.

From the particles, the MMSE estimate and posterior covariance matrix of $\mathbf{x}_t$ are given by

$$\widehat{\mathbf{x}}_{t|t}^{\text{MMSE}} = \sum_{i=1}^N w_t^{(i)} \widehat{\mathbf{x}}_{t|t}(\theta_{1:t}^{(i)}) \tag{41}$$

[3]When using the optimal importance distribution, the weights computation requires the evaluation of an integral with respect to θ_t. It is possible to integrate analytically w.r.t. the cluster means μ^v and μ^w, but not w.r.t. the covariances.

$$\widehat{\Sigma}_{t|t} = \sum_{i=1}^{N} w_t^{(i)} \left[\Sigma_{t|t}(\theta_{1:t}^{(i)}) + (\widehat{\mathbf{x}}_{t|t}(\theta_{1:t}^{(i)}) - \widehat{\mathbf{x}}_{t|t}^{\text{MMSE}})(\widehat{\mathbf{x}}_{t|t}(\theta_{1:t}^{(i)}) - \widehat{\mathbf{x}}_{t|t}^{\text{MMSE}})^T \right] \tag{42}$$

VI. Applications

In this section, we present two applications of the above model and algorithms[4]. We address, first, blind deconvolution, second, change point detection in biomedical time series. In each case, we assume that the statistics of the state noise are unknown, and modelled as a DPM.

A. Blind deconvolution of impulse processes

Various fields of Engineering and Physics, such as image de-blurring, spectroscopic data analysis, audio source restoration, etc. require blind deconvolution. We follow here the model presented in [31] for blind deconvolution of Bernoulli-Gaussian processes, which is recalled below.

1) Statistical Model: Let $H = \begin{pmatrix} 1 & h_1 & .. & h_L \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{h} \end{pmatrix}$ and $\mathbf{x}_t = \begin{pmatrix} v_t & v_{t-1} & ... & v_{t-L} \end{pmatrix}^T$. The observed signal z_t is the convolution of the sequence $\mathbf{x}_t$ with a finite impulse response filter H, observed in additive white Gaussian noise w_t. The observation model is then

$$z_t = H\mathbf{x}_t + w_t \tag{43}$$

where $w_t \sim \mathcal{N}(0, \sigma_w^2)$ with σ_w^2 is the assumed known variance of w_t. The state space model can be written as follows:

$$\mathbf{x}_t = F\mathbf{x}_{t-1} + Gv_t \tag{44}$$

where $F = \begin{pmatrix} 0 & 0_{1\times L} \\ 0_{L\times 1} & I_L \end{pmatrix}$, $G = \begin{pmatrix} 1 \\ 0_{L\times 1} \end{pmatrix}$, $0_{m\times n}$ is the zero matrix of size $m \times n$ and I_m is the identity matrix of size $m \times m$. The state transition noise v_t is supposed to be independent from w_t, and distributed according to the mixture

$$v_t \ \sim \ \lambda F^v + (1-\lambda)\delta_0 \tag{45}$$

where δ_0 is the Dirac delta function at 0 and F^v is a DPM of Gaussians defined in Eq. (17). In other words, the noise is alternatively zero, or distributed according to a DPM of Gaussians.

For simplicity reasons, we introduce latent Bernoulli variables $r_t \in \{0, 1\}$ such that $\Pr(r_t = 1) = \lambda$ and $v_t|(r_t = 1) \sim f(\cdot|\theta_t^v),\ v_t|(r_t = 0) \sim \delta_0$. Consider the cluster variable φ_t^v defined by $\varphi_t^v = \theta_t^v$ if $r_t = 1$ and $\varphi_t^v = (0, 0)$ (i.e. parameters corresponding to the delta-mass) if $r_t = 0$, that is, $\varphi_t^v \sim \lambda F^v + (1-\lambda)\delta_{(0,0)}$. By integrating out F^v, one has

$$\varphi_t^v|\varphi_{-t}^v \sim \lambda p(\varphi_t^v|\varphi_{-t}^v, r_t = 1) + (1-\lambda)\delta_{(0,0)} \tag{46}$$

[4]See Caron et al. [1] for an application on a regression problem.

where $p(\varphi_t^v|\varphi_{-t}^v, r_t = 1)$ is the Polya urn representation on the set $\widetilde{\varphi}_{-t}^v = \{\varphi \in \varphi_{-t}^v | \varphi \neq \delta_{(0,0)}\}$ of size T' given by

$$\varphi_t^v|(\varphi_{-t}^v, r_t = 1) \sim \frac{\sum_{k=1,k\neq t}^{T'} \delta_{\varphi_k^v} + \alpha^v \mathbb{G}_0^v}{\alpha^v + T'} \tag{47}$$

The probability λ is considered as a random variable with a beta prior density $p(\lambda) = \mathcal{B}(\zeta, \tau)$ where ζ and τ are known parameters. The random variable λ can be marginalized out in Eq. (46)

$$\varphi_t^v|\varphi_{-t}^v \sim \frac{a(\varphi_{-t}^v)}{a(\varphi_{-t}^v) + b(\varphi_{-t}^v)} p(\varphi_t^v|\varphi_{-t}^v, r_t = 1) + \frac{b(\varphi_{-t}^v)}{a(\varphi_{-t}^v) + b(\varphi_{-t}^v)} \delta_{(0,0)} \tag{48}$$

where

$$a(\varphi_{-t}^v) = \zeta + \sum_{k=1,k\neq t}^{T} r_k \tag{49}$$

$$b(\varphi_{-t}^v) = \tau + \sum_{k=1,k\neq t}^{T} (1 - r_k) \tag{50}$$

where $r_t = 0$ if $\varphi_t^v = (0, 0)$ and $r_t = 1$ otherwise.

The hyperparameters are $\phi = (\alpha^v, \mathbf{h})$ (the hyperparameters of the base distribution $\mathbb{G}_0^v$ are assumed fixed and known). These hyperparameters are assumed random with prior distribution $p(\phi) = p(\alpha^v)p(\mathbf{h})$, where

$$p(\alpha^v) = \mathcal{G}(\frac{\eta}{2}, \frac{\nu}{2}),\ p(\mathbf{h}) = \mathcal{N}(0, \sigma_w^2 \Sigma_{\mathbf{h}}) \tag{51}$$

where η, ν and $\Sigma_{\mathbf{h}}$ are known. Conditional on $\mathbf{x}_{0:t}$, the following conditional posterior is obtained straighforwardly

$$p(\mathbf{h}|\mathbf{x}_{0:t}, \mathbf{z}_{1:T}) = \mathcal{N}(\mathbf{m}, \sigma_w^2 \Sigma'_{\mathbf{h}}) \tag{52}$$

where

$$\Sigma'^{-1}_{\mathbf{h}} = \Sigma_{\mathbf{h}}^{-1} + \sum_{t=1}^{T} \mathbf{v}_{t-1:t-L} \mathbf{v}'_{t-1:t-L}$$

$$\mathbf{m} = \Sigma'_{\mathbf{h}} \sum_{t=1}^{T} \mathbf{v}_{t-1:t-L} \left(z_t - v_t\right)$$

Samples $\mathbf{x}_{0:t}^{(i)}$ can be generated from the Gaussian posterior $p(\mathbf{x}_{0:t}|\varphi_{1:T}^{v\ (i)}, \mathbf{z}_{1:T}, \phi^{(i-1)})$ with the simulation smoother [32]. This algorithm complexity is $O(T)$.

The aim is to approximate by MCMC the joint posterior pdf $p(\mathbf{v}_{1:T}, \varphi_{1:T}, \phi|\mathbf{z}_{1:T})$. This is done by implementing Algorithm 3 for the cluster variable, whereas the other variables are sampled by Metropolis-Hastings or direct sampling w.r.t their conditional posterior.

2) Simulation results: This model has been simulated with the following parameters: $T = 120$, $L = 3$, $\mathbf{h} = \begin{pmatrix} -1.5 & 0.5 & -0.2 \end{pmatrix}$, $\lambda = 0.4$, $\sigma_w^2 = 0.1$, $F^v = 0.7\mathcal{N}(2, .5) + 0.3\mathcal{N}(-1, .1)$, $\Sigma_{\mathbf{h}} = 100$, $\eta = 3$, $\nu = 3$, $\zeta = 1$, $\tau = 1$. The hyperparameters of the base distribution are $\mu_0 = 0$, $\kappa_0 = 0.1$, $\nu_0 = 4, \Lambda_0 = 1$. For the estimation, 10,000 MCMC iterations are performed, with 7,500 burn-in iterations. Fig. 1 (top) displays the MMSE estimate of $\mathbf{v}_{1:T}$ together with its true value. As can be seen in Fig. 1 (bottom), the signal is correctly estimated and the residual is quite small. Also, as can be seen in Fig. 2, the estimated pdf F^v is quite close to the true one. In particular, the estimated pdf matches the two modes of the true pdf. Multiple simulations with different starting values were runned, and the results appeared insensitive to initialization. This suggest that the MCMC sampler explores properly the posterior.

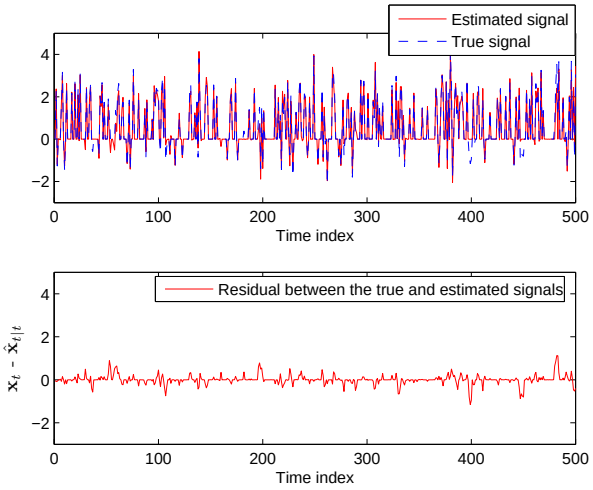

Fig. 1. Top picture: True (dashed line) and MMSE estimated (solid line) signal $\mathbf{v}_{1:T}$ after 10,000 MCMC iterations (7,500 burn-in). v_t is supposed to be either 0 with probability λ, or to be distributed from an unknown pdf F^v with probability $(1 - \lambda)$. Bottom picture: residual $e_t = v_t - E[v_t|\mathbf{z}_{1:T}]$ between the true and estimated signals. Although the distribution F^v is unknown, the state v_t is almost correctly estimated.

Let e_{MSE} be the mean squared error (MSE), computed by

$$e_{MSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{v}_t - \mathbf{v}_{t|T}^{\text{MMSE}})^2} \tag{53}$$

To better highlight the performance of the proposed algorithm, we compared our model/algorithm (denoted M1) with the following models, denoted M2 to M8:

M2. In this model, the pdf is assumed known and set to the true value $F^v = 0.7\mathcal{N}(2, .5) + 0.3\mathcal{N}(-1, .1)$. The model is simply a Jump Linear Model that jumps between three modes

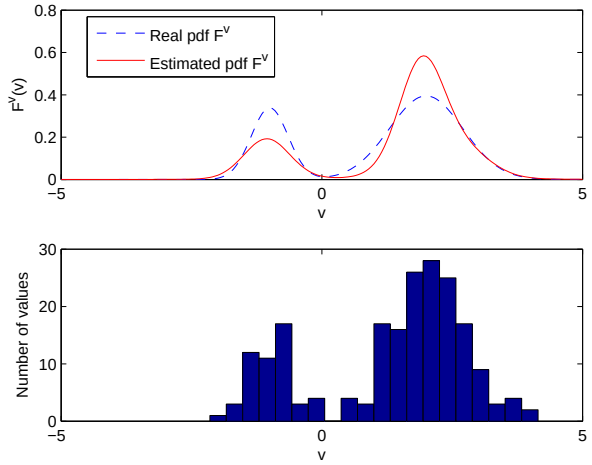

Fig. 2. (Top) True (dashed line) and estimated (solid line) pdf F^v. The true pdf F^v is a mixture of two Gaussians $0.7\mathcal{N}(2,.5)+0.3\mathcal{N}(-1,.1)$. It is supposed to be unknown and jointly estimated with the state vector with $10,000$ MCMC iterations (7,500 burn-in) given a vector of 120 observations $\mathbf{z}_{1:T}$. The estimated pdf matches correctly the two modes of the true distribution. (Bottom) Histogram of the simulated values v, sampled from F^v which a mixture of two Gaussians $0.7\mathcal{N}(2,.5)+0.3\mathcal{N}(-1,.1)$

of resp. mean/covariance $(0,0)$, $(2,.5)$ and $(-1,.1)$ with resp. prior probabilities $(1-\lambda)$, 0.7λ and 0.3λ.

M3. In this model, the pdf is assumed to be a Gaussian $\mathcal{N}(1.1,2.3)$. The first two moments of this Gaussian are the same as those of the true pdf F^v. The model is also a Jump Linear Model that jumps between two modes of resp. mean/covariance $(0,0)$ and $(1.1,2.3)$ with resp. prior probabilities $(1-\lambda)$ and λ.

M4-7. The model described in this article but with α^v fixed to 0.1 (M3), 1 (M4), 10 (M5) and 100 (M6).

M8. The model described in this article (M1) but with the observation noise variance σ_w^2 estimated with an inverse gamma prior $\sigma_w^2 \sim i\mathcal{G}(u,v)$ with $u=2$ and $v=0.1$. $\sigma_w^{2\ (i)}$ is sampled with Gibbs sampling with $\sigma_w^2|\mathbf{x}_{0:T},\mathbf{z}_{1:T},\mathbf{h} \sim i\mathcal{G}(u',v')$ and $u'=u+\frac{T}{2}$ and $v'=v+\frac{1}{2}\sum_{t=1}^{T}(\mathbf{z}_t-H\mathbf{x}_t)^2$.

The algorithm used for M2 and M3 is the Gibbs sampler with backward forward recursion given in [28]. For the same set of observations, each MCMC algorithm has been run with 10,000 iterations and 7,500 burn-in iterations. MMSE estimate $\mathbf{v}_{t|T}^{\text{MMSE}}$ and MSE e_{MSE} are computed for each model. 20 simulations have been performed; for each model, the mean and standard deviation of the MSE's over the 20 simulations are reported in Tab. I.

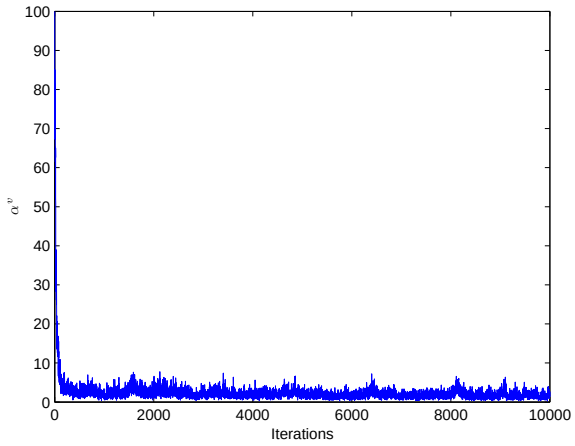

Fig. 3. Evolution of $\alpha^{v\ (i)}$ in function of Gibbs sampler iteration i. The value of α^v is initialized at 100.

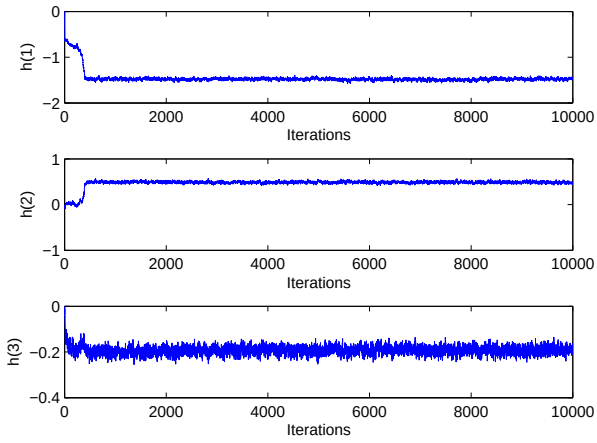

Fig. 4. Evolution of the three components of the vector $\mathbf{h}^{(i)}$ in function of Gibbs sampler iteration i. It is initialized at $[0\ 0\ 0]$. The value converges toward the true value $\mathbf{h} = [-1.5\ 0.5\ -0.2]$.

Tab. I. Comparison of our model/algorithm with other models

Simulation / Model	M1	M2	M3	M4	M5	M6	M7	M8
Mean	0.240	0.217	0.290	0.915	0.254	0.253	0.314	0.438
Standard deviation	0.067	0.058	0.085	0.818	0.062	0.086	0.222	0.421

Our model/algorithm (M1) gives MSE that is only 10% more than that of the model with fixed pdf (M2) even though the pdf is not exactly estimated. If the observation noise variance σ_w^2 is unknown and has to be estimated (M8), this has an impact on the estimation of the state vector still the sampler

converge more slowly to the true posterior. If the unknown pdf is set to be a Gaussian with large variance (M3), the MSE is 17% larger than with our approach. The estimation of α^v improves the estimation of the state vector: MSEs are higher for models M4-7 where α^v is set to a fixed value. This is especially true for $\alpha^v = 0.1$. With this small value, the sampler proposes new clusters very rarely and converges very slowly to the true posterior.

B. Change-point problems in biomedical time series

Let now consider a change-point problem in biomedical time series. The following problem has been discussed in [33] and [11]. Let consider patients who had recently undergone kidney transplant. The level of kidney function is given by the rate at which chemical substances are cleared from the blood, and the rate can be inferred indirectly from measurements on serum creatinine. If the kidney function is stable, the response series varies about a constant level. If the kidney function is improving (resp. decaying) at a constant level then the response series decays (resp. increases) linearly.

1) Statistical model: The linear model, formulated by Gordon and Smith [33] is given by

$$\mathbf{x}_t = F\mathbf{x}_{t-1} + G\mathbf{v}_t \tag{54}$$

$$z_t = H\mathbf{x}_t + w_t \tag{55}$$

where $\mathbf{x}_t = (m_t, \dot{m}_t)$, where m_t is the level and $\dot{m}_t$ the slope, $F = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, z_t is the measured creatinine and $H = \begin{pmatrix} 1 & 0 \end{pmatrix}$. Measurements are subject to errors due to mistakes in data transcription, equipment malfunction or blood contamination. w_t follows the following mixture model

$$w_t \sim \lambda^w \mathcal{N}(0, \sigma_1^w) + (1 - \lambda^w)\mathcal{N}(0, \sigma_2^w) \tag{56}$$

where $\lambda^w = 0.98$ is the probability that the measurements are correct, in that case the variance is $\sigma_1^w = 10^{-7}$ and $\sigma_2^w = 1$ otherwise. To capture the effects of jumps in the creatinine level, the state noise $\mathbf{v}_t$ is supposed to be distributed according to the following mixture model

$$\mathbf{v}_t \sim \lambda^v F^v + (1 - \lambda^v)\delta_{\theta_0^v} \tag{57}$$

where $\theta_0^v = \left\{ \begin{pmatrix} 0 & 0 \end{pmatrix}^T, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right\}$, $\lambda^v = 0.15$ is the probability of jump in the level and F^v is a DPM of Gaussians. Contrary to the model in [11], we do not define fixed jump levels. These levels, as well as their number, are estimated through the DPM.

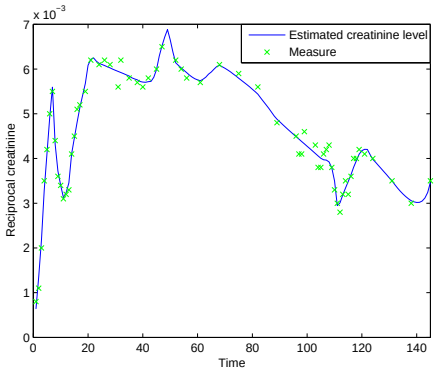

Fig. 5. Measured (cross) and estimated (solid line) creatinine level with 2000 MCMC iterations and 1000 burn-in iterations.

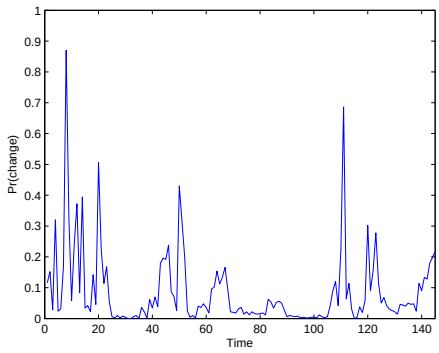

Fig. 6. Posterior probability of a jump in the creatinine level with 2000 MCMC iterations and 1000 burn-in iterations. For a threshold set to 0.5, the creatinine level experiences jumps at about times 8, 20 and 110.

2) Simulation results: The last model is applied to the data provided in Gordon and Smith [33] (and also exploited in [11]). The hyperparameters of the base distribution are $\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\kappa_0 = 10^6$, $\nu_0 = 4$, $\Lambda_0 = \frac{10^{-6}}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. For the estimation, 2,000 MCMC iterations (with 1,000 burn-in iterations) are performed. Fig. 5 presents the estimated creatinine level together with the measurements. Fig. 6 plots the posterior probability of a jump in the creatinine level. In particular, the estimated pdf matches the two modes of the true pdf. Multiple simulations with different starting values were runned, and the results appeared insensitive to initialization. This suggest that the MCMC sampler explores properly the posterior.

The estimation have also been made online with the Rao-Blackwellized algorithm with 1000 particles. We perform fixed-lag smoothing [34] to estimate $\mathbb{E}(\mathbf{x}_t|\mathbf{z}_{1:t+T})$, where T is set to 10. The

mean time per iteration is about 1s. The importance function used to sample the latent variables θ_t^v is prior pdf $p(\theta_t^v|\theta_{1:t-1}^v)$. For a detection threshold set at 0.5, the MCMC algorithm detects 3 peaks, while the RBPF only detects two peaks. The trade-off between false alarm and non detection may be tuned with the coefficient λ^v.

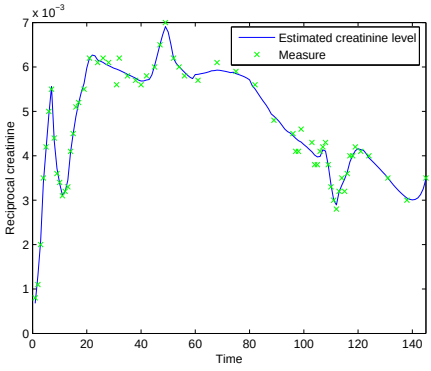

Fig. 7. Measured (cross) and estimated (solid line) creatinine level with a Rao-Blackwellized particle filter with 1000 particles.

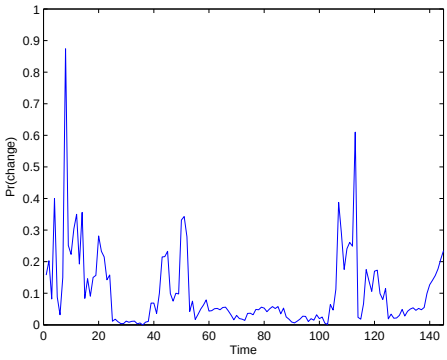

Fig. 8. Posterior probability of a jump in the creatinine level with the Rao-Blackwellized particle filter with 1000 particles. For a threshold set to 0.5, the creatinine level jumps are detected at about times 8 and 110.

VII. Discussion

In this section, we discuss several features of the approach proposed.

A. About Dirichlet Process-based modeling

DPMs have several main advantages. Firstly, sampling from the posterior distribution is made especially easy thanks to the Polya urn scheme. Second, the discreteness of the distribution $\mathbb{G}$ enables

straightforward estimation of the "number of components", without requiring reversible jump-like computational approaches. This discreteness has, however, some unexpected effects on inferences, which are reported in [35] and [36]. For example, the DP tends to favor a misbalance between the size of the groups of latent variables associated to the same cluster, and to concentrate the posterior distribution of the number of groups on a higher value. Dirichlet Processes realize nevertheless an attractive trade-off between versatile modeling properties and implementation advantages, which explain their success in various contexts – and our choice to use them in this paper.

B. About MCMC algorithms for DPMs

As stated in [3], the "single-site" marginal algorithm used in this paper may be stuck in a mode of the posterior: several noises samples $\mathbf{v}_t$ (resp. $\mathbf{w}_t$) are associated to the same cluster value U_j^v for some j in Eq. (9) (resp. $U_{j'}^w$) – in other words, there are many t's such that $\theta_t^v = U_j^v$ for some j (resp $\theta_t^w = U_{j'}^w$). Since the algorithm cannot change the value of θ_t^v for more than one $\mathbf{v}_t$ simultaneously, changes to θ_t^v occur rarely, as they require passage through a low-probability intermediate state in which noises $\mathbf{v}_t$ in the same group are not associated to the same cluster. In alternative algorithms, such as those given in [3], clusters are sampled in groups, which avoids this problem at the expense of an increased computational cost. Nevertheless, we have demonstrated empirically in Section VI that our MCMC scheme is indeed efficient in the applications presented.

C. About the hyperparameter estimation in the MCMC algorithm

As shown in the applications section, the estimation of the hyperparameter α improves the overall state estimation. It also makes the convergence of the Gibbs sampler faster. During the first iterations, the value of α is high, and the sampler proposes new clusters more easily. This enables efficient state space global exploration during the first iterations. When the "good" clusters have been found, the value of α decreases, and it eliminates useless clusters.

D. About the convergence of the Rao-Blackwellized particle filter

Because the DPMs F^v and F^w are static (infinite-dimensional) parameters, the Rao-Blackwellized particle filter suffers from an accumulation of errors over time. In other words, the particle filter is not able to move cluster values U_j^v's and U_j^w after they are initialized. This is a well known problem of *static parameter* estimation with particle filters. However, as the static component is not the estimated cluster θ_t but its prior distribution $\mathbb{G}$, this accumulation is less critical than with the estimation of true static parameters.

In Section V, the hyperparameter vector ϕ is assumed fixed, also because this is a static parameter. It could actually be estimated by implementing one of the particle filtering approaches to static parameter estimation. For example, the approaches in [37–40] are based on either kernel density methods, MCMC steps, or Maximum Likelihood. However, these algorithms also have important drawbacks (error accumulation with time in $O(t^2)$). An alternative solution consists of introducing an artificial dynamic on the hyperparameters [41] but it is not applicable to our problem: we would then loose the Polya urn structure given by Eq. (13).

E. About related approaches

Our model has some connections with Jump Linear Systems (JLS) [42, 43]. In JLS, a discrete indicator variable switches between a (known) fixed number of different (known) linear Gaussian models with some (known) prior probability. Our model may be interpreted as a JLS whose number of different models is unknown, mean vector and covariance matrix of the linear Gaussian models are unknowns as well as their prior probabilities. The model proposed in this paper can also be generalized in the following manner. Denote $\underline{\theta}_t = \{F_t, C_t, H_t, G_t, \mu_t^v, \Sigma_t^v, \mu_t^w, \Sigma_t^w\} = \{F_t, C_t, H_t, G_t, \theta_t\}$ and $\underline{\mathbb{G}}_0$ a prior distribution on $\underline{\theta}_t$. The following general hierarchical model

$$\begin{aligned}
\underline{\mathbb{G}} &\sim DP(\underline{\mathbb{G}}_0, \alpha), \\
\underline{\theta}_t|\underline{\mathbb{G}} &\sim \underline{\mathbb{G}}, \\
\mathbf{x}_t|\underline{\theta}_t, \mathbf{x}_{t-1} &\sim \mathcal{N}(F_t\mathbf{x}_{t-1} + C_t\mathbf{u}_t + G_t\mu_t^v, G_t\Sigma_t^v G_t'), \\
\mathbf{z}_t|\underline{\theta}_t, \mathbf{x}_t &\sim \mathcal{N}(H_t\mathbf{x}_t + \mu_t^w, \Sigma_t^w)
\end{aligned} \tag{58}$$

has more flexibility than common JLS: the number of different switching models is estimated, as well as the parameters of these models and their prior probabilities.

F. About observability

In order for the observation noise $\mathbf{w}_t$ pdf to be correctly estimated, some observability constraints must be ensured. Indeed, the pair $(\widetilde{F}, \widetilde{H})$ has to be fully observable, that is, the observability matrix

$$\begin{pmatrix} \widetilde{H} \\ \widetilde{H}\widetilde{F} \\ \cdots \\ \widetilde{H}\widetilde{F}^{n_x+n_z-1} \end{pmatrix} \tag{59}$$

must have rank $n_x + n_z$ (full rank), where $\widetilde{F} = \begin{pmatrix} F & 0_{n_x \times n_z} \\ 0_{n_z \times n_x} & I_{n_z} \end{pmatrix}$, $\widetilde{H} = \begin{pmatrix} H & I_{n_z} \end{pmatrix}$, n_x and n_z are resp. the length of the state and observation vectors.

VIII. Conclusion

In this paper, we have presented a Bayesian nonparametric model that enables state and observation noise pdfs estimation, in a linear dynamic model. The Dirichlet process mixture considered here is flexible and we have presented two simulation-based algorithms based on Rao-Blackwellization which allows us to perform efficiently inference. The approach has proven efficient in applications – in particular, we have shown that state estimation is possible even though the dynamic and observation noises are of unknown pdfs. We are currently investigating the following extensions of our methodology. First, it would be of interest to consider nonlinear dynamic models. Second, it would be important to develop time-varying Dirichlet process mixture models in cases where the noise statistics are assumed to evolve over time.

Appendix

A. Notations

μ and Σ are sampled from a Normal inverse Wishart distribution $\mathbb{G}_0$ of hyperparameters μ_0, κ_0, ν_0, Λ_0 if

$$\mu|\Sigma \sim \mathcal{N}(\mu_0, \frac{\Sigma}{\kappa_0})$$
$$\Sigma^{-1} \sim W(\nu_0, \Lambda_0^{-1})$$

where $W(\nu_0, \Lambda_0^{-1})$ is the standard Wishart distribution.

B. Backward forward recursion

The quantities $P'^{-1}_{t-1|t}(\theta_{t:T})$ and $P'^{-1}_{t-1|t}(\theta_{t:T})m'_{t-1|t}(\theta_{t:T})$ defined in Section IV-A always satisfy the following backward information filter recursion.

1) Initialization
   $P'^{-1}_{T|T}(\theta_T) = H_T^{\mathsf{T}}(\Sigma_T^w)^{-1}H_T$
   $P'^{-1}_{T|T}(\theta_T)m'_{T|T}(\theta_T) = H_T^{\mathsf{T}}(\Sigma_T^w)^{-1}(\mathbf{z}_T - \mu_T^w)$
2) Backward recursion. For $t = T-1..1$,

$$\Delta_{t+1} = \left[I_{n_v} + B^{\mathsf{T}}(\theta_{t+1})P'^{-1}_{t+1|t+1}(\theta_{t+1|T})B(\theta_{t+1})\right]^{-1} \tag{60}$$

$$P'^{-1}_{t|t+1}(\theta_{t+1:T}) = F_{t+1}^{\mathsf{T}}P'^{-1}_{t+1|t+1}(\theta_{t+1:T})(I_{n_x} - B(\theta_{t+1})\Delta_{t+1}(\theta_{t+1:T})B^{\mathsf{T}}(\theta_{t+1})P'^{-1}_{t+1|t+1}(\theta_{t+1:T}))F_{t+1}$$

$$\begin{aligned} P'^{-1}_{t|t+1}(\theta_{t+1:T})m'_{t|t+1}(\theta_{t+1:t}) = {} & F_{t+1}^{\mathsf{T}}(\theta_{t+1}) \times (I_{n_x} - P'^{-1}_{t+1|t+1}(\theta_{t+1:T})B(\theta_{t+1})\Delta_{t+1}(\theta_{t+1:T})B^{\mathsf{T}}(\theta_{t+1})) \\ & \times P'^{-1}_{t+1|t+1}(\theta_{t+1:T})\left(m'_{t+1|t+1}(\theta_{t+1:T}) - \mathbf{u}'_{t+1}(\theta_{t+1})\right) \end{aligned} \tag{61}$$

$$P'^{-1}_{t|t}(\theta_{t:T}) = P'^{-1}_{t|t+1}(\theta_{t+1:T}) + H_t^{\mathsf{T}}(\Sigma_t^w)^{-1}H_t \tag{62}$$

$$P'^{-1}_{t|t}(\theta_{t:T})m'_{t|t}(\theta_{t:T}) = P'^{-1}_{t|t+1}(\theta_{t+1:T})m'_{t|t+1}(\theta_{t+1:T}) + H_t^{\mathsf{T}}(\Sigma_t^w)^{-1}(\mathbf{z}_t - \mu_t^w) \tag{63}$$

where $B(\theta_t) = G_t \times \text{chol}(\Sigma_t^v)^{\mathsf{T}}$.

For the Metropolis Hasting ratio, we need to compute the acceptance probability only with a probability constant

$$p(\mathbf{z}_{1:T}|\theta_{1:T}) \propto p(\mathbf{z}_t|\theta_{1:t}, \mathbf{z}_{1:t-1}) \int_{\mathcal{X}} p(\mathbf{z}_{t+1:T}|\mathbf{x}_t, \theta_{t+1:T}) p(\mathbf{x}_t|\mathbf{z}_{1:t}, \theta_{1:t}) d\mathbf{x}_t \tag{64}$$

If $\Sigma_{t|t}(\theta_{1:t}) \neq 0$ then it exists $\Pi_{t|t}(\theta_{1:t})$ and $Q_{t|t}(\theta_{1:t})$ such that $\Sigma_{t|t}(\theta_{1:t}) = Q_{t|t}(\theta_{1:t})\Pi_{t|t}(\theta_{1:t})Q^T_{t|t}(\theta_{1:t})$. The matrices $Q_{t|t}(\theta_{1:t})$ and $\Pi_{t|t}(\theta_{1:t})$ are straightforwardly obtained using the singular value decomposition of $\Sigma_{t|t}(\theta_{1:t})$. Matrix $\Pi_{t|t}(\theta_{1:t})$ is a $n_t \times n_t, 1 \leq n_t \leq n_x$ diagonal matrix with the nonzero eigenvalues of $\Sigma_{t|t}(\theta_{1:t})$ as elements. Then one has

$$\begin{aligned} p(\mathbf{z}_{1:T}|\theta_{1:T}) &\propto \mathcal{N}(\widehat{\mathbf{z}}_{t|t-1}(\theta_{1:t}), S_{t|t-1}(\theta_{1:t})) \left|\Pi_{t|t}(\theta_{1:t})Q^{\mathsf{T}}_{t|t}(\theta_{1:t})P'^{-1}_{t|t+1}(\theta_{t+1:T})Q_{t|t}(\theta_{1:t}) + I_{n_t}\right|^{-\frac{1}{2}} \\ &\times \exp(-\tfrac{1}{2}\widehat{\mathbf{x}}^{\mathsf{T}}_{t|t}(\theta_{1:t})P'^{-1}_{t|t+1}(\theta_{t+1:T})\widehat{\mathbf{x}}_{t|t}(\theta_{1:t}) - 2\widehat{\mathbf{x}}^{\mathsf{T}}_{t|t}(\theta_{1:t})P'^{-1}_{t|t+1}(\theta_{t+1:T})m'_{t|t+1}(\theta_{t+1:T}) \\ &-(m'_{t|t+1}(\theta_{t+1:T}) - \widehat{\mathbf{x}}_{t|t}(\theta_{1:t}))^{\mathsf{T}} \times P'^{-1}_{t|t+1}(\theta_{t+1:T})A_{t|t}(\theta_{1:t}) \times P'^{-1}_{t|t+1}(\theta_{t+1:T})(m'_{t|t+1}(\theta_{t+1:T}) - \widehat{\mathbf{x}}_{t|t}(\theta_{1:t}))) \end{aligned} \tag{65}$$

where

$$A_{t|t}(\theta_{1:t}) = Q_{t|t}(\theta_{1:t}) \left[\Pi^{-1}_{t|t}(\theta_{1:t}) + Q^{\mathsf{T}}_{t|t}(\theta_{1:t})P'^{-1}_{t|t+1}(\theta_{t+1:T})Q_{t|t}(\theta_{1:t})\right]^{-1} Q^{\mathsf{T}}_{t|t}(\theta_{1:t}) \tag{66}$$

The quantities $\widehat{\mathbf{x}}_{t|t}(\theta_{1:t})$, $\Sigma_{t|t}(\theta_{1:t})$, $\widehat{\mathbf{z}}_{t|t-1}(\theta_{1:t})$ and $S_{t|t-1}(\theta_{1:t})$ are, resp., the one-step ahead filtered estimate and covariance matrix of $\mathbf{x}_t$, the innovation at time t, and the covariance of this innovation. These quantities are provided by the Kalman filter, the system being linear Gaussian conditional upon $\theta_{1:t}$.

ACKNOWLEDGMENT

This work is partially supported by the Centre National de la Recherche Scientifique (CNRS) and the Région Nord-Pas de Calais.

REFERENCES

[1] F. Caron, M. Davy, A. Doucet, E. Duflos, and P. Vanheeghe, "Bayesian inference for dynamic models with dirichlet process mixtures," in *International Conference on Information Fusion, Florence, Italia*, 2006.

[2] S. Walker, P. Damien, P. Laud, and A. Smith, "Bayesian nonparametric inference for random distributions and related functions," *J. R. Statist. Soc. B*, vol. 61, no. 3, pp. 485–527, 1999.

[3] R. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, pp. 249–265, 2000.

[4] P. Muller and F. Quintana, "Nonparametric Bayesian data analysis," *Statistical science*, vol. 19, no. 1, pp. 95–110, 2004.

[5] R. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Transactions on Automatic Control*, vol. 15, no. 2, pp. 175–184, 1970.

[6] K. Myers and B. Tapley, "Adaptive sequential estimation with unknown noise statistics," *IEEE Transactions on Automatic Control*, vol. 21, no. 4, pp. 520–523, 1976.

[7] R. E. Maine and K. Iliff, "Formulation and implementation of a practical algorithm for parameter estimation with process and measurement noise," *SIAM Journal of Applied Mathematics*, vol. 41, no. 3, pp. 558–579, 1981.

[8] J. Maryak, J. Spall, and B. Heydon, "Use of the Kalman filter for inference in state-space models with unknown noise distributions," *IEEE Transactions on Automatic Control*, vol. 49, no. 1, 2004.

[9] N. Shephard, "Partial non-gaussian state space," *Biometrika*, vol. 81, pp. 115–131, 1994.

[10] M. J. Lombardi and S. J. Godsill, "On-line Bayesian estimation of signals in symmetric α-stable noise," *IEEE Transactions on Signal Processing*, vol. 53, pp. 1–6, 2005.

[11] C. Carter and R. Kohn, "Markov chain Monte Carlo in conditionally Gaussian state space models," *Biometrika*, vol. 83, no. 3, pp. 589–601, 1996.

[12] J. Griffin and M. Steel, "Semiparametric Bayesian inference for stochastic frontier models," *Journal of econometrics*, vol. 123, no. 1, pp. 121–152, 2004.

[13] A. Pievatolo and R. Rotondi, "Analysing the interevent time distribution to identify seismicity phases: a Bayesian nonparametric approach to the multiple-changepoint problem," *Applied statistics*, vol. 49, no. 4, pp. 543–562, 2000.

[14] K.-A. Do, P. Muller, and F. Tang, "A Bayesian mixture model for differential gene expression," *Journal of the Royal Statistical Society C*, vol. 54, no. 3, 2005.

[15] M. Medvedovic and S. Sivaganesan, "Bayesian infinite mixture model based clustering of gene expression profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1194–1206, 2002.

[16] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *The annals of statistics*, vol. 1, pp. 209–230, 1973.

[17] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[18] D. Blackwell and J. MacQueen, "Ferguson distributions via Polya urn schemes," *The annals of statistics*, vol. 1, pp. 353–355, 1973.

[19] C. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The annals of statistics*, vol. 2, pp. 1152–1174, 1974.

[20] M. Escobar and M. West, "Computing Bayesian nonparametric hierarchical models," Institute of statistics and decision sciences, Duke University, Durham, USA, Tech. Rep., 1992.

[21] ——, "Bayesian density estimation and inference using mixtures," *Journal of the american statistical association*, vol. 90, pp. 577–588, 1995.

[22] S. MacEachern and P. Muller, "Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models," in *Robust Bayesian Analysis*, F. Ruggeri and D. Rios-Insua, Eds. Springer-Verlag, 2000, pp. 295–316.

[23] S. MacEachern, M. Clyde, and J. Liu, "Sequential importance sampling for nonparametric Bayes models: the next generation," *The Canadian Journal of Statistics*, vol. 27, no. 2, pp. 251–267, 1999.

[24] P. Fearnhead, "Particle filters for mixture models with an unknown number of components," *Statistics and Computing*, vol. 14, pp. 11–21, 2004.

[25] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian data analysis*. Chapman and Hall, 1995.

[26] B. Anderson and J. Moore, *Optimal filtering*. Prentice-Hall, 1979.

[27] C. Robert and G. Casella, *Monte Carlo statistical methods*. Springer-Verlag, 1999.

[28] A. Doucet and C. Andrieu, "Iterative algorithms for state estimation of jump Markov linear systems," *IEEE transactions on signal processing*, vol. 49, no. 6, pp. 1216–1227, 2001.

[29] M. West, "Hyperparameter estimation in Dirichlet process mixture models," Institute of statistics and decision sciences, Duke University, Durham, USA, Tech. Rep., 1992.

[30] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in practice*. Springer-Verlag, 2001.

[31] A. Doucet and P. Duvaut, "Bayesian estimation of state-space models applied to deconvolution of Bernoulli-Gaussian processes," *Signal Processing*, vol. 57, pp. 147–161, 1997.

[32] J. Durbin and S. Koopman, "A simple and efficient simulation smoother for state space time series analysis," *Biometrika*, vol. 89, no. 3, pp. 603–615, 2002.

[33] K. Gordon and A. Smith, "Monitoring and modeling biomedical time series," *Journal of the American Statistical Association*, vol. 85, pp. 328–337, 1990.

[34] A. Doucet, N. Gordon, and V. Krishnamurthy, "Particle filters for state estimation of jump Markov linear systems," *IEEE Transactions on Signal Processing*, vol. 49, pp. 613–624, 2001.

[35] S. Petrone and A. Raftery, "A note on the Dirichlet prior in Bayesian nonparametric inference with partial exchangeability," *Statistics and probability letters*, vol. 36, pp. 69–83, 1997.

[36] P. Green and S. Richardson, "Modelling heterogeneity with and without the Dirichlet process," *Scandinavian journal of statistics*, vol. 28, no. 2, pp. 355–375, 2001.

[37] J. Liu and M. West, "Combined parameter and state estimation in simulation-based filtering," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. D. Freitas, and N. Gordon, Eds. Springer-Verlag, 2001.

[38] W. Gilks and C. Berzuini, "Following a moving target: Monte Carlo inference for dynamic Bayesian models," *Journal of the Royal Statistical Association B*, vol. 63, no. 1, pp. 127–146, 2001.

[39] A. Doucet and V. Tadic, "Parameter estimation in general state-space models using particle methods," *Ann. Inst. Statist. Math.*, vol. 55, no. 2, pp. 409–422, 2003.

[40] G. Poyiadjis, A. Doucet, and S. Singh, "Particle methods for optimal filter derivative: application to parameter estimation," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP'05*, 2005.

[41] C. Andrieu, M. Davy, and A. Doucet, "Efficient particle filtering for jump Markov systems. Application to time-varying autoregressions." *IEEE Transactions on signal processing*, vol. 51, no. 7, 2003.

[42] G. Ackerson and K. Fu, "On state estimation in switching environments," *IEEE Transactions on Automatic Control*, vol. 15, pp. 10–17, 1970.

[43] H. Akashi and H. Kumamoto, "Random sampling approach to state estimation in switching environments," *Automatica*, vol. 13, pp. 429–434, 1977.