A Simple Alternative Derivation of the Expectation Correction Algorithm

Bertrand Mesot and David Barber

***Abstract*— The Switching Linear Dynamical System (SLDS) is a popular model in time-series analysis. However, the complexity of inferring the state of the latent variables scales exponentially with the length of the time-series, resulting in many approximation strategies in the literature. We focus on the recently devised Expectation Correction (EC) approximation which can be considered a form of Gaussian Sum Smoother. The algorithm has excellent numerical performance compared to a wide range of competing techniques, exploiting more fully the available information than, for example, Generalised Pseudo Bayes. We show that EC can be seen as an extension to the SLDS of the Rauch, Tung, Striebel inference algorithm for the Linear Dynamical System. This yields a simpler derivation of the EC algorithm and facilitates comparison with existing, similar approaches.**

***Index Terms*— Switching Linear Dynamical Systems, Approximate Inference, Expectation Correction.**

I. Introduction

The Linear Dynamical System (LDS) [1] is a key temporal model in which a latent linear process generates the observed time-series; see Fig. 1. For time-series which are not well described by a single LDS, we may model each observation by a potentially different LDS. This is the basis for the Switching LDS (SLDS) where, for each time step t, a *discrete* switch variable $s_t \in \{1, \ldots, S\}$ describes which of the LDSs is to be used; see Fig. 2. The observation (or 'visible' variable) $\mathbf{v}_t \in \mathbb{R}^V$ is linearly related to the hidden state $\mathbf{h}_t \in \mathbb{R}^H$ by

$$\mathbf{v}_t = \mathbf{B}_{s_t}\mathbf{h}_t + \boldsymbol{\eta}^{\mathcal{V}}_{s_t}, \quad \boldsymbol{\eta}^{\mathcal{V}}_{s_t} \sim \mathcal{N}\big(\boldsymbol{\mu}_{\mathcal{V}}(s_t), \boldsymbol{\Sigma}_{\mathcal{V}}(s_t)\big) \tag{1}$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Normal (Gaussian) distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The hidden state $\mathbf{h}_t$ at the t-th time step is linearly related to the state at the previous time step by

$$\mathbf{h}_t = \mathbf{A}_{s_t}\mathbf{h}_{t-1} + \boldsymbol{\eta}^{\mathcal{H}}_{s_t}, \quad \boldsymbol{\eta}^{\mathcal{H}}_{s_t} \sim \mathcal{N}\big(\boldsymbol{\mu}_{\mathcal{H}}(s_t), \boldsymbol{\Sigma}_{\mathcal{H}}(s_t)\big). \tag{2}$$

Eqs. 1 and 2 define the projection and transition probabilities $p(\mathbf{v}_t \,|\, \mathbf{h}_t, s_t)$ and $p(\mathbf{h}_t \,|\, \mathbf{h}_{t-1}, s_t)$, respectively[1]. The dynamics of the switch variables is assumed Markovian, with transition $p(s_t \,|\, s_{t-1})$. The SLDS is used in many disciplines, from econometrics to machine learning [1], [2], [3], [4]. See also [5] and [6] for recent reviews of work.

A quantity which is often required is the marginal (smoothed) posterior probability $p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:T})$ of the hidden variables $\mathbf{h}_t$ and s_t, given a sequence of T observations $\mathbf{v}_{1:T}$. For the SLDS, inferring the posterior distribution is computationally intractable since the exact posterior is an exponentially large mixture of Gaussians; see for example [5]. Various algorithms have been devised to address this problem; see [5], [6] for a review. We focus on the recently devised Expectation Correction (EC) algorithm [7] which has excellent comparative performance. Here we emphasise a reformulation of EC that simplifies the exposition and has the additional benefit of clarifying the relationship between EC and other approximation algorithms. EC is motivated by the Rauch, Tung, Striebel (RTS) smoother [8] which, for the simpler LDS, corrects the filtered posterior into its smoothed form. Before presenting our extension of the RTS strategy to the switching model, we first review RTS inference in the more straightforward LDS.

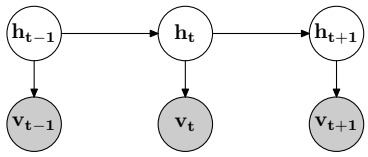

Fig. 1. Dynamic Bayesian network representation of the LDS; $\mathbf{h}_t$ represents the continuous hidden variable and $\mathbf{v}_t$ the observation.

II. The RTS Algorithm

The RTS algorithm performs smoothed inference in the LDS, which admits exact linear-time computation. It uses a forward-backward approach where the forward pass computes the filtered posterior $p(\mathbf{h}_t \,|\, \mathbf{v}_{1:t})$, and the backward pass corrects this to form the desired smoothed posterior $p(\mathbf{h}_t \,|\, \mathbf{v}_{1:T})$. Since only Gaussian distributions are involved, conditioning and marginalisation are straightforward.

A. Forward Pass

The filtered posterior $p(\mathbf{h}_t \,|\, \mathbf{v}_{1:t})$ is obtained by conditioning on $\mathbf{v}_t$ the joint distribution $p(\mathbf{v}_t, \mathbf{h}_t \,|\, \mathbf{v}_{1:t-1})$. For a given time step, it can be computed by means of the *forward* recursion

$$p(\mathbf{h}_t \,|\, \mathbf{v}_{1:t}) \propto p(\mathbf{v}_t \,|\, \mathbf{h}_t) \left\langle p(\mathbf{h}_t \,|\, \mathbf{h}_{t-1}) \right\rangle_{p(\mathbf{h}_{t-1} \,|\, \mathbf{v}_{1:t-1})} \tag{3}$$

where $\langle \cdot \rangle_p$ denotes the average with respect to the distribution p and $p(\mathbf{h}_{t-1} \,|\, \mathbf{v}_{1:t-1})$ is the filtered posterior at the previous time step. The recursion is initialised with $p(\mathbf{h}_1 \,|\, \mathbf{v}_1) \propto p(\mathbf{v}_1 \,|\, \mathbf{h}_1)\, p(\mathbf{h}_1)$, where $p(\mathbf{h}_1)$ is a given prior distribution.

B. Backward Pass

The smoothed posterior at the t-th time step is obtained from the *backward* recursion

$$p(\mathbf{h}_t \,|\, \mathbf{v}_{1:T}) = \left\langle p(\mathbf{h}_t \,|\, \mathbf{h}_{t+1}, \mathbf{v}_{1:T}) \right\rangle_{p(\mathbf{h}_{t+1} \,|\, \mathbf{v}_{1:T})} \tag{4}$$

Bertrand Mesot is with the IDIAP Research Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. David Barber is with the Department of Computer Science, University College London, U.K.

[1]The $\mathcal{H}$ and $\mathcal{V}$ symbols are used to indicate whether a parameter is associated with the hidden or visible variable, respectively.

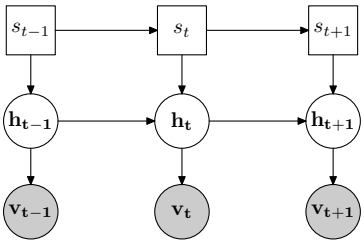

Fig. 2. Dynamic Bayesian network representation of the SLDS; s_t and $\mathbf{h}_t$ represent the discrete and continuous hidden variables and $\mathbf{v}_t$ the observation.

where $p(\mathbf{h}_{t+1} \,|\, \mathbf{v}_{1:T})$ is the smoothed posterior at the next time step. Since $\mathbf{h}_t$ is independent of any future observations once $\mathbf{h}_{t+1}$ is known, the *backward* transition probability $p(\mathbf{h}_t \,|\, \mathbf{h}_{t+1}, \mathbf{v}_{1:T})$ is given by

$$p(\mathbf{h}_t \,|\, \mathbf{h}_{t+1}, \mathbf{v}_{1:t}) \propto p(\mathbf{h}_{t+1} \,|\, \mathbf{h}_t)\, p(\mathbf{h}_t \,|\, \mathbf{v}_{1:t}). \tag{5}$$

which only involves the forward transition probability and the filtered posterior at time t. The backward pass is initialised with the filtered posterior obtained at the T-th step, since both filtered and smoothed posteriors match at that point.

C. Implementation

The pseudo-codes for computing the filtered and smoothed posteriors with the RTS method are given in Algorithms 1 and 2, respectively. In Algorithm 1, $\mathbf{x}$ and $\mathbf{X}$ correspond to the mean and covariance of $\mathbf{h}_t$ under $p(\mathbf{h}_t \,|\, \mathbf{v}_{1:t-1})$.

Algorithm 1 RTS forward pass. This function computes the mean $\mathbf{f}_t$ and covariance $\mathbf{F}_t$ of $\mathbf{h}_t$ under $p(\mathbf{h}_t \,|\, \mathbf{v}_{1:t})$ for $t \in [0, T]$, as well as the log-likelihood $l \equiv \log p(\mathbf{v}_{1:T})$. The prior mean and covariance are denoted by $\boldsymbol{\mu}_\mathcal{P}$ and $\boldsymbol{\Sigma}_\mathcal{P}$.

$l \leftarrow 0$
for $t \leftarrow 1$ **to** T **do**
  if $t > 1$ **then**
    $\mathbf{x} \leftarrow \mathbf{A}\mathbf{f}_{t-1}$
    $\mathbf{X} \leftarrow \mathbf{A}\mathbf{F}_{t-1}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma}_\mathcal{H}$
  else
    $\mathbf{x} \leftarrow \boldsymbol{\mu}_\mathcal{P}$
    $\mathbf{X} \leftarrow \boldsymbol{\Sigma}_\mathcal{P}$
  end if
  $\{p(\mathbf{v}_t \,|\, \mathbf{v}_{1:t-1}), \mathbf{f}_t, \mathbf{F}_t\} \leftarrow \text{COND}\{\mathbf{x}, \mathbf{X}, \mathbf{B}, \boldsymbol{\Sigma}_\mathcal{V}, \mathbf{v}_t, \mathbf{0}\}$
  $l \leftarrow l + \log p(\mathbf{v}_t \,|\, \mathbf{v}_{1:t-1})$
end for

Algorithm 2 RTS backward pass. This function computes the mean $\mathbf{g}_t$ and covariance $\mathbf{G}_t$ of $\mathbf{h}_t$ under $p(\mathbf{h}_t \,|\, \mathbf{v}_{1:T})$ for $t \in [0, T]$.

$\mathbf{g}_t \leftarrow \mathbf{f}_t$
$\mathbf{G}_t \leftarrow \mathbf{F}_t$
for $t \leftarrow T - 1$ **to** 1 **do**
  $\{\alpha, \mathbf{g}_t, \mathbf{G}_t\} \leftarrow \text{COND}\{\mathbf{f}_t, \mathbf{F}_t, \mathbf{A}, \boldsymbol{\Sigma}_\mathcal{H}, \mathbf{g}_{t+1}, \mathbf{G}_{t+1}\}$
end for

The conditioning of $p(\mathbf{v}_t, \mathbf{h}_t \,|\, \mathbf{v}_{1:t-1})$ on $\mathbf{v}_t$ in the forward pass and the conditioning of $p(\mathbf{h}_t, \mathbf{h}_{t+1} \,|\, \mathbf{v}_{1:t})$ on $\mathbf{h}_{t+1}$ in the backward pass is performed by the COND function, whose pseudo-code is given in Algorithm 3. To improve numerical stability the conditioning is performed by means of Joseph's formula [1]. The first four arguments of the COND function are: the prior mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ of the hidden variable we are interested in, the matrix $\mathbf{C}$ which indicates how to transform the hidden variable into the conditioned one, and the prior covariance $\boldsymbol{\Sigma}_\mathcal{C}$ of the conditioned variable. The main difference between (3) and (4) is that the latter requires an averaging *after* the conditioning. This can be easily performed in the COND function by providing the mean $\mathbf{w}$ and covariance $\mathbf{W}$ of the variable we want to average on. In the forward pass, where no averaging is required, $\mathbf{w} = \mathbf{v}_t$ and $\mathbf{W} = \mathbf{0}$.

Algorithm 3 COND$\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \boldsymbol{\Sigma}_\mathcal{C}, \mathbf{w}, \mathbf{W}\}$. See text for the meaning of the arguments and [1] for a detailed explanation of the algorithm.

$\boldsymbol{\mu}_\mathcal{C} \leftarrow \mathbf{C}\boldsymbol{\mu}$
$\boldsymbol{\Sigma}_{\mathcal{C}\mathcal{H}} \leftarrow \mathbf{C}\boldsymbol{\Sigma}$
$\boldsymbol{\Sigma}_{\mathcal{C}\mathcal{C}} \leftarrow \boldsymbol{\Sigma}_{\mathcal{C}\mathcal{H}}\, \mathbf{C}^\mathsf{T} + \boldsymbol{\Sigma}_\mathcal{C}$
$\mathbf{K} \leftarrow \boldsymbol{\Sigma}_{\mathcal{C}\mathcal{H}}\, \boldsymbol{\Sigma}_{\mathcal{C}\mathcal{C}}^{-1}$
$\mathbf{X} \leftarrow \mathbf{I} - \mathbf{K}\mathbf{C}$
$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \mathbf{K}(\mathbf{w} - \boldsymbol{\mu}_\mathcal{C})$
$\boldsymbol{\Sigma} \leftarrow \mathbf{X}\,\boldsymbol{\Sigma}\,\mathbf{X}^\mathsf{T} + \mathbf{K}\,(\boldsymbol{\Sigma}_\mathcal{C} + \mathbf{W})\,\mathbf{K}^\mathsf{T}$
$p \leftarrow |\boldsymbol{\Sigma}_{\mathcal{C}\mathcal{C}}|^{-\frac{1}{2}}\, \exp\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_\mathcal{C})^\mathsf{T}\, \boldsymbol{\Sigma}_{\mathcal{C}\mathcal{C}}^{-1}\, (\mathbf{w} - \boldsymbol{\mu}_\mathcal{C})\}$
return $\{p, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$

III. Expectation Correction

EC follows the same approach as the RTS algorithm. The forward pass computes the filtered posterior $p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:t})$ and the backward pass corrects this to form the smoothed posterior $p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:T})$. Without loss of generality, we write the filtered and smoothed posterior as a product of a continuous and a discrete distribution:

$$p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:t}) = p(\mathbf{h}_t \,|\, s_t, \mathbf{v}_{1:t})\, p(s_t \,|\, \mathbf{v}_{1:t})$$
$$p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:T}) = p(\mathbf{h}_t \,|\, s_t, \mathbf{v}_{1:T})\, p(s_t \,|\, \mathbf{v}_{1:T}).$$

Our approach will approximate both the filtered and smoothed posteriors as a finite mixture of Gaussians. Formally, this can be achieved using, for example, $p(\mathbf{h}_t \,|\, s_t, \mathbf{v}_{1:t}) \equiv \sum_i p(\mathbf{h}_t \,|\, i_t, s_t, \mathbf{v}_{1:t})\, p(i_t \,|\, s_t, \mathbf{v}_{1:t})$—see, for example, [7], [9]. Whereas in [7], mixtures of Gaussians are used, in our exposition we use only a *single* Gaussian—the extension to the mixture case is straightforward [7] and we prefer to present the central idea without the extra notational complexity of collapsing to mixtures.

A. Forward Pass

The filtered posterior $p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:T})$ is obtained by conditioning on $\mathbf{v}_t$ the joint distribution $p(\mathbf{v}_t, \mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:t-1})$. The equivalent of (3) for the SLDS reads

$$p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:t}) \propto \sum_{s_{t-1}} p(s_t \,|\, s_{t-1})\, p(s_{t-1} \,|\, \mathbf{v}_{1:t-1}) \times p(\mathbf{v}_t \,|\, \mathbf{h}_t, s_t)\, \big\langle p(\mathbf{h}_t \,|\, \mathbf{h}_{t-1}, s_t) \big\rangle_{p(\mathbf{h}_{t-1} \,|\, s_{t-1}, \mathbf{v}_{1:t-1})}$$

where $p(s_{t-1} \,|\, \mathbf{v}_{1:t-1})$ and $p(\mathbf{h}_{t-1} \,|\, s_{t-1}, \mathbf{v}_{1:t-1})$ are the discrete and continuous components of the filtered posterior at the

previous time step. After averaging over $\mathbf{h}_{t-1}$ and grouping similar factors, we obtain

$$\begin{aligned} p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:t}) \\ &\propto \sum_{s_{t-1}} p(s_{t-1}, s_t \,|\, \mathbf{v}_{1:t-1})\, p(\mathbf{v}_t, \mathbf{h}_t \,|\, s_{t-1}, s_t, \mathbf{v}_{1:t-1}) \\ &\propto \sum_{s_{t-1}} p(s_{t-1}, s_t \,|\, \mathbf{v}_{1:t})\, p(\mathbf{h}_t \,|\, s_{t-1}, s_t, \mathbf{v}_{1:t}). \end{aligned} \tag{6}$$

The continuous component $p(\mathbf{h}_t \,|\, s_{t-1}, s_t, \mathbf{v}_{1:t})$ corresponds to the filtered posterior of the LDS, as given by (3), and is proportional to

$$p(\mathbf{v}_t \,|\, \mathbf{h}_t, s_t) \left\langle p(\mathbf{h}_t \,|\, \mathbf{h}_{t-1}, s_t) \right\rangle_{p(\mathbf{h}_{t-1} \,|\, s_{t-1}, \mathbf{v}_{1:t-1})}. \tag{7}$$

The discrete component $p(s_{t-1}, s_t \,|\, \mathbf{v}_{1:t})$ is proportional to

$$p(\mathbf{v}_t \,|\, s_{t-1}, s_t, \mathbf{v}_{1:t-1})\, p(s_t \,|\, s_{t-1})\, p(s_{t-1} \,|\, \mathbf{v}_{1:t-1}) \tag{8}$$

where $p(\mathbf{v}_t \,|\, s_{t-1}, s_t, \mathbf{v}_{1:t-1})$ is obtained by integrating (7) over $\mathbf{h}_t$.

The filtered posterior at time t, as given by (6), is a mixture of Gaussians. At each time step the number of mixture components is multiplied by S and thus grows exponentially with t. A simple approximate remedy is to collapse the mixture obtained to a mixture with fewer components. This corresponds to the so-called *Gaussian Sum Approximation* (GSA) [9] which is a form of *Assumed Density Filtering* [10]. It reduces the complexity of the forward pass to $\mathcal{O}(I \cdot S \cdot T)$, where I is the number of mixture components of the collapsed distribution. The recursion is initialised with $p(\mathbf{h}_1, s_1 \,|\, \mathbf{v}_1) \propto p(\mathbf{v}_1 \,|\, \mathbf{h}_1, s_1)\, p(\mathbf{h}_1 \,|\, s_1)\, p(s_1)$, where $p(\mathbf{h}_1 \,|\, s_1)$ and $p(s_1)$ are given prior distributions.

B. Backward Pass

The equivalent of (4) for the SLDS reads

$$\begin{aligned} p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:T}) = \sum_{s_{t+1}} p(s_{t+1} \,|\, \mathbf{v}_{1:T}) \\ \times \left\langle p(\mathbf{h}_t, s_t \,|\, \mathbf{h}_{t+1}, s_{t+1}, \mathbf{v}_{1:t}) \right\rangle_{p(\mathbf{h}_{t+1} \,|\, s_{t+1}, \mathbf{v}_{1:T})} \end{aligned} \tag{9}$$

where $p(s_{t+1} \,|\, \mathbf{v}_{1:T})$ and $p(\mathbf{h}_{t+1} \,|\, s_{t+1}, \mathbf{v}_{1:T})$ are the discrete and continuous components of the smoothed posterior at the next time step. The average in (9) can be written as[2]

$$\left\langle p(\mathbf{h}_t \,|\, \mathbf{h}_{t+1}, s_t, s_{t+1}, \mathbf{v}_{1:t})\, p(s_t \,|\, \mathbf{h}_{t+1}, s_{t+1}, \mathbf{v}_{1:t}) \right\rangle.$$

This is difficult to evaluate because of the dependency of s_t on $\mathbf{h}_{t+1}$. In its most simple form, EC approximates the average by

$$\underbrace{\left\langle p(\mathbf{h}_t \,|\, \mathbf{h}_{t+1}, s_t, s_{t+1}, \mathbf{v}_{1:t}) \right\rangle}_{p(\mathbf{h}_t \,|\, s_t, s_{t+1}, \mathbf{v}_{1:T})} \underbrace{\left\langle p(s_t \,|\, \mathbf{h}_{t+1}, s_{t+1}, \mathbf{v}_{1:t}) \right\rangle}_{p(s_t \,|\, s_{t+1}, \mathbf{v}_{1:T})}. \tag{10}$$

This is particularly appealing since the first factor corresponds to the smoothed posterior of the LDS, as given by (4), and can be evaluated by conditioning on $\mathbf{h}_{t+1}$ the joint distribution

$$p(\mathbf{h}_t, \mathbf{h}_{t+1} \,|\, s_t, s_{t+1}, \mathbf{v}_{1:t}) = p(\mathbf{h}_{t+1} \,|\, \mathbf{h}_t, s_t)\, p(\mathbf{h}_t \,|\, s_t, \mathbf{v}_{1:t}). \tag{11}$$

[2]To simplify notation, in the following we assume that the averages are taken with respect to $p(\mathbf{h}_{t+1} \,|\, s_{t+1}, \mathbf{v}_{1:T})$.

The second factor in (10) is still difficult to evaluate exactly. Formally, this term corresponds to

$$\left\langle p(s_t \,|\, \mathbf{h}_{t+1}, s_{t+1}, \mathbf{v}_{1:t}) \right\rangle \equiv p(s_t \,|\, s_{t+1}, \mathbf{v}_{1:T}). \tag{12}$$

The distinguishing feature of EC from other methods, such as *Generalised Pseudo Bayes* (GPB) [1], [2], [11] is in the approximation of $p(s_t \,|\, s_{t+1}, \mathbf{v}_{1:T})$. In GPB, $p(s_t \,|\, s_{t+1}, \mathbf{v}_{1:T}) \approx p(s_t \,|\, s_{t+1}, \mathbf{v}_{1:t})$, which depends only on the filtered posterior for s_t and does not include any information coming from the continuous variable $\mathbf{h}_{t+1}$. Since $p(s_t \,|\, s_{t+1}, \mathbf{v}_{1:t}) \propto p(s_{t+1} \,|\, s_t)\, p(s_t \,|\, \mathbf{v}_{1:t})$, computing the smoothed recursion for the switch states in GPB is equivalent to running the RTS backward pass on a Hidden Markov Model. This represents a potentially severe loss of information from the future and means any information from the continuous variables cannot be used when correcting the filtered results $p(s_t \,|\, \mathbf{v}_{1:t})$ into smoothed posteriors $p(s_t \,|\, \mathbf{v}_{1:T})$. In contrast, EC attempts to preserve future information passing through the continuous variables. The simplest approach within EC is to use the approximation

$$\begin{aligned} p(s_t \,|\, s_{t+1}, \mathbf{v}_{1:T}) &\equiv \left\langle p(s_t \,|\, \mathbf{h}_{t+1}, s_{t+1} \mathbf{v}_{1:t}) \right\rangle \\ &\approx p(s_t \,|\, \mathbf{h}_{t+1}, s_{t+1}, \mathbf{v}_{1:t})\big|_{\mathbf{h}_{t+1} = \langle \mathbf{h}_{t+1} \,|\, s_{t+1}, \mathbf{v}_{1:T} \rangle} \end{aligned} \tag{13}$$

where $\langle \mathbf{h}_{t+1} \,|\, s_{t+1}, \mathbf{v}_{1:T} \rangle$ is the mean of $\mathbf{h}_{t+1}$ with respect to $p(\mathbf{h}_{t+1} \,|\, s_{t+1}, \mathbf{v}_{1:T})$. Whereas in [7], other approximations are also considered, we only consider this simple (and fast) method because, in practice, it often suffices [7], [12], [13]. More sophisticated approximation schemes—which take into account the covariance of $\mathbf{h}_{t+1}$, for example—are straightforward to implement, if desired [7]. Finally, the right-hand-side of (13) can be evaluated by considering the joint distribution

$$\begin{aligned} p(\mathbf{h}_{t+1}, s_t \,|\, s_{t+1}, \mathbf{v}_{1:t}) \propto \\ p(\mathbf{h}_{t+1} \,|\, s_t, s_{t+1}, \mathbf{v}_{1:t})\, p(s_{t+1} \,|\, s_t)\, p(s_t \,|\, \mathbf{v}_{1:t}) \end{aligned} \tag{14}$$

where $p(\mathbf{h}_{t+1} \,|\, s_t, s_{t+1}, \mathbf{v}_{1:t})$ is obtained by marginalising (11) over $\mathbf{h}_t$.

In summary, the smoothed posterior, as given by (9), is a mixture of Gaussians of the form

$$\begin{aligned} p(\mathbf{h}_t, s_t \,|\, \mathbf{v}_{1:T}) = \\ \sum_{s_{t+1}} p(s_t, s_{t+1} \,|\, \mathbf{v}_{1:T})\, p(\mathbf{h}_t \,|\, s_t, s_{t+1}, \mathbf{v}_{1:T}). \end{aligned} \tag{15}$$

In its most generic form, EC approximates the discrete and continuous components by

$$\begin{aligned} p(s_t, s_{t+1} \,|\, \mathbf{v}_{1:T}) &\approx p(s_{t+1} \,|\, s_t) \left\langle p(s_t \,|\, \mathbf{h}_{t+1}, s_{t+1}, \mathbf{v}_{1:t}) \right\rangle \\ p(\mathbf{h}_t \,|\, s_t, s_{t+1}, \mathbf{v}_{1:T}) &\approx \left\langle p(\mathbf{h}_t \,|\, \mathbf{h}_{t+1}, s_t, s_{t+1}, \mathbf{v}_{1:t}) \right\rangle \end{aligned}$$

As for the forward pass, the number of mixture components is multiplied by S at each iteration. Hence, to retain tractability, the mixture in (15) is collapsed to a mixture with fewer components. The backward pass is initialised with the filtered posterior obtained at the T-th step, since both filtered and smoothed posteriors match at that point.

IV. Implementation

Algorithms 4 and 5 give the pseudo-code of EC forward and backward passes. In Algorithm 5, the prefactor α in the expression $p_{s_t,s_{t+1}} \leftarrow \alpha\, p(s_{t+1} \mid s_t)\, p(s_t \mid \mathbf{v}_{1:t})$ differentiates EC from GPB. The COL routine collapses the mixture of Gaussians passed as arguments to a single Gaussian; see [7] for additional details and for an example of collapse to a mixture of Gaussians.

Algorithm 4 EC forward pass. This function computes the filtered posterior $p(s_t \mid \mathbf{v}_{1:t})$, the mean $\mathbf{f}_{s_t}$ and the covariance $\mathbf{F}_{s_t}$ of $\mathbf{h}_t$ under $p(\mathbf{h}_t \mid s_t, \mathbf{v}_{1:t})$, as well as the log-likelihood $l \equiv \log p(\mathbf{v}_{1:T})$. The prior mean and covariance are denoted by $\boldsymbol{\mu}_{\mathcal{P}}$ and $\boldsymbol{\Sigma}_{\mathcal{P}}$.

$l \leftarrow 0$
for $t \leftarrow 1$ **to** T **do**
  for all (s_{t-1}, s_t) **do**
    if $t > 1$ **then**
      $\mathbf{x} \leftarrow \mathbf{A}_{s_t} \mathbf{f}_{s_{t-1}}$
      $\mathbf{X} \leftarrow \mathbf{A}_{s_t} \mathbf{F}_{s_{t-1}} \mathbf{A}_{s_t}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\mathcal{H}}(s_t)$
    else
      $\mathbf{x} \leftarrow \boldsymbol{\mu}_{\mathcal{P}}(s_1)$
      $\mathbf{X} \leftarrow \boldsymbol{\Sigma}_{\mathcal{P}}(s_1)$
    end if
    $\{\alpha, \boldsymbol{\mu}_{s_{t-1},s_t}, \boldsymbol{\Sigma}_{s_{t-1},s_t}\}$
      $\leftarrow \text{COND}\{\mathbf{x}, \mathbf{X}, \mathbf{B}_{s_t}, \boldsymbol{\Sigma}_{\mathcal{V}}(s_t), \mathbf{v}_t, \mathbf{0}\}$
    $p_{s_{t-1},s_t} \leftarrow \alpha\, p(s_t \mid s_{t-1})\, p(s_{t-1} \mid \mathbf{v}_{1:t-1})$
  end for
  $p(\mathbf{v}_t \mid \mathbf{v}_{1:t-1}) \leftarrow \sum_{s_{t-1},s_t} p_{s_{t-1},s_t}$
  for all (s_{t-1}, s_t) **do**
    $p_{s_{t-1},s_t} \leftarrow p_{s_{t-1},s_t} \,/\, p(\mathbf{v}_t \mid \mathbf{v}_{1:t-1})$
  end for
  for all s_t **do**
    $p(s_t \mid \mathbf{v}_{1:t}) \leftarrow \sum_{s_{t-1}} p_{s_{t-1},s_t}$
    $p(s_{t-1} \mid s_t, \mathbf{v}_{1:t}) \leftarrow p_{s_{t-1},s_t} \,/\, p(s_t \mid \mathbf{v}_{1:t})$
    $\{\mathbf{f}_{s_t}, \mathbf{F}_{s_t}\} \leftarrow \text{COL}\{p(s_{t-1} \mid s_t, \mathbf{v}_{1:t}), \boldsymbol{\mu}_{s_{t-1},s_t}, \boldsymbol{\Sigma}_{s_{t-1},s_t}\}$
  end for
  $l \leftarrow l + \log p(\mathbf{v}_t \mid \mathbf{v}_{1:t-1})$
end for

V. Conclusion

We presented an alternative and simpler derivation of the EC algorithm which makes the relationship with the RTS algorithm more evident. EC is perhaps most naturally viewed as the extension of the time-honoured Gaussian Sum Filter [9] to the smoothing case. It is similar to GPB; both algorithms use the same forward pass, but the EC backward pass can be more accurate since it better preserves the information carried by the continuous variables. Furthermore, EC is not limited to the simple approximations (10) and (13), but can readily be extended to use more elaborate schemes [7]. In its most simple form—with collapse to a single Gaussian—it has been successfully used for inference on real-world time-series, including speech waveforms [12], [13] with $\mathcal{O}(10^5)$ time steps. In this case, EC proved to be more stable than EP while being more accurate and faster than Monte-Carlo approaches.

Algorithm 5 EC backward pass. This function computes the smoothed posterior $p(s_t \mid \mathbf{v}_{1:T})$, the mean $\mathbf{g}_{s_t}$ and the covariance $\mathbf{G}_{s_t}$ of $\mathbf{h}_t$ under $p(\mathbf{h}_t \mid s_t, \mathbf{v}_{1:T})$.

for all s_T **do**
  $\mathbf{g}_{s_T} \leftarrow \mathbf{f}_{s_T}$
  $\mathbf{G}_{s_T} \leftarrow \mathbf{F}_{s_T}$
end for
for $t \leftarrow T-1$ **to** 1 **do**
  for all (s_t, s_{t+1}) **do**
    $\{\alpha, \boldsymbol{\mu}_{s_t,s_{t+1}}, \boldsymbol{\Sigma}_{s_t,s_{t+1}}\}$
      $\leftarrow \text{COND}\{\mathbf{f}_{s_t}, \mathbf{F}_{s_t}, \mathbf{A}_{s_{t+1}}, \boldsymbol{\Sigma}_{\mathcal{H}}(s_{t+1}), \mathbf{g}_{s_{t+1}}, \mathbf{G}_{s_{t+1}}\}$
    $p_{s_t,s_{t+1}} \leftarrow \alpha\, p(s_{t+1} \mid s_t)\, p(s_t \mid \mathbf{v}_{1:t})$
  end for
  for all (s_t, s_{t+1}) **do**
    $p_{s_t,s_{t+1}} \leftarrow p_{s_t,s_{t+1}} \,/\, \sum_{s_t} p_{s_t,s_{t+1}}$
  end for
  for all (s_t, s_{t+1}) **do**
    $p_{s_t,s_{t+1}} \leftarrow p_{s_t,s_{t+1}} \cdot p(s_{t+1} \mid \mathbf{v}_{1:T})$
  end for
  for all s_t **do**
    $p(s_t \mid \mathbf{v}_{1:T}) \leftarrow \sum_{s_{t+1}} p_{s_t,s_{t+1}}$
    $p(s_{t+1} \mid s_t, \mathbf{v}_{1:T}) \leftarrow p_{s_t,s_{t+1}} \,/\, p(s_t \mid \mathbf{v}_{1:T})$
    $\{\mathbf{g}_{s_t}, \mathbf{G}_{s_t}\} \leftarrow \text{COL}\{p(s_{t+1} \mid s_t, \mathbf{v}_{1:T}), \boldsymbol{\mu}_{s_t,s_{t+1}}, \boldsymbol{\Sigma}_{s_t,s_{t+1}}\}$
  end for
end for

Acknowledgements

This work was supported by the Swiss NSF MULTI project and the Swiss OFES through the PASCAL Network.

References

[1] Y. Bar-Shalom and X.-R. Li, *Estimation and tracking: principles, techniques and software*. Norwood, MA: Artech House, 1998.
[2] C.-J. Kim and C. R. Nelson, *State-Space Models with Regime Switching*. MIT Press, 1999.
[3] G. Kitagawa, "The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother," *Annals of the Institute of Statistical Mathematics*, vol. 46, no. 4, pp. 605–623, 1994.
[4] V. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *Advances in Neural Information Processing Systems (NIPS 13)*, 2001, pp. 981–987.
[5] U. N. Lerner, "Hybrid Bayesian networks for reasoning about complex systems," Ph.D. dissertation, Stanford University, 2002.
[6] O. Zoeter, "Monitoring non-linear and switching dynamical systems," Ph.D. dissertation, Radboud University, Nijmegen, 2005.
[7] D. Barber, "Expectation correction for smoothed inference in switching linear dynamical systems," *Journal of Machine Learning Research*, vol. 7, pp. 2515–2540, November 2006.
[8] H. E. Rauch, G. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *Journal of American Institute of Aeronautics and Astronautics*, vol. 3, no. 8, pp. 1445–1450, 1965.
[9] D. L. Alspach and H. W. Sorensen, "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE Transactions on Automatic Control*, vol. 17, no. 4, pp. 439–448, 1972.
[10] T. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, MIT Media Lab, 2001.
[11] C.-J. Kim, "Dynamic linear models with Markov-switching," *Journal of Econometrics*, vol. 60, no. 1–2, pp. 1–22, 1994.
[12] B. Mesot and D. Barber, "Switching linear dynamical systems for noise robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1850–1858, August 2007.
[13] B. Mesot, "Inference in switching linear dynamical systems applied to noise robust speech recognition of isolated digits," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne (EPFL), 2008, thesis 4059.