Expectation Correction for Smoothed Inference in Switching Linear Dynamical Systems

David Barber DAVID.BARBER@IDIAP.CH
IDIAP Research Institute
Rue du Simplon 4
CH-1920 Martigny
Switzerland

Editor: Leslie Pack Kaelbling

Abstract

We introduce a method for approximate smoothed inference in a class of switching linear dynamical systems, based on a novel form of Gaussian Sum smoother. This class includes the switching Kalman Filter and the more general case of switch transitions dependent on the continuous latent state. The method improves on the standard Kim smoothing approach by dispensing with one of the key approximations, thus making fuller use of the available future information. Whilst the only central assumption required is projection to a mixture of Gaussians, we show that an additional conditional independence assumption results in a simpler but accurate alternative. Unlike the alternative Expectation Propagation procedure, our method consists only of a single forward and backward pass and is reminiscent of the standard smoothing 'correction' recursions in the simpler linear dynamical system. The method is stable and compares very favourably against alternative approximations, both in cases where a single mixture component provides a good approximation, and where a multimodal approximation of the posterior is required.

Keywords: Gaussian Sum Smoother, Switching Kalman Filter, Switching Linear Dynamical System, Expectation Propagation, Expectation Correction.

1. Switching Linear Dynamical System

The Linear Dynamical System (LDS) (Bar-Shalom and Li, 1998; West and Harrison, 1999) is a key temporal model in which a latent linear process generates the observed series. For complex time-series which are not well described globally by a single LDS, we may break the time-series into segments, each modelled by a potentially different LDS. This is the basis for the Switching LDS (SLDS) where, for each time t, a switch variable $s_t \in 1, \ldots, S$ describes which of the LDSs is to be used[1]. The observation (or 'visible') $v_t \in \mathcal{R}^V$ is linearly related to the hidden state $h_t \in \mathcal{R}^H$ by

$$v_t = B(s_t)h_t + \eta^v(s_t), \qquad \eta^v(s_t) \sim \mathcal{N}\left(\bar{v}(s_t), \Sigma^v(s_t)\right) \tag{1}$$

1. These systems also go under the names Jump Markov model/process, switching Kalman Filter, Switching Linear Gaussian State Space models, Conditional Linear Gaussian Models.

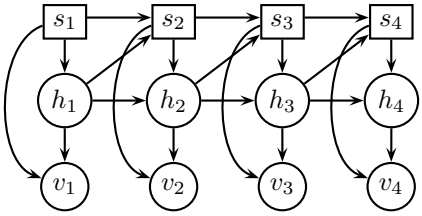


Figure 1: The independence structure of the aSLDS. Square nodes denote discrete variables, round nodes continuous variables. In the SLDS links from h to s are not normally considered.

where $\mathcal{N}(\mu, \Sigma)$ denotes a Gaussian distribution with mean μ and covariance Σ. The transition dynamics of the continuous hidden state h_t is linear,

$$h_t = A(s_t)h_{t-1} + \eta^h(s_t), \qquad \eta^h(s_t) \sim \mathcal{N}\left(\bar{h}(s_t), \Sigma^h(s_t)\right) \tag{2}$$

The switch variable s_t itself is Markovian, with transition $p(s_t|s_{t-1})$.

In this article, we will consider the more general model in which the switch s_t is dependent on both the previous s_{t-1} and h_{t-1}. We call this an augmented Switching Linear Dynamical System (aSLDS), in keeping with the terminology in Lerner (2002). An equivalent probabilistic model is[2] (see Figure(1))

$$p(v_{1:T}, h_{1:T}, s_{1:T}) = \prod_{t=1}^{T} p(v_t|h_t, s_t)p(h_t|h_{t-1}, s_t)p(s_t|h_{t-1}, s_{t-1})$$

with

$$p(v_t|h_t, s_t) = \mathcal{N}\left(\bar{v}(s_t) + B(s_t)h_t, \Sigma^v(s_t)\right), \qquad p(h_t|h_{t-1}, s_t) = \mathcal{N}\left(\bar{h}(s_t) + A(s_t)h_t, \Sigma^h(s_t)\right)$$

At time $t = 1$, $p(s_1|h_0, s_0)$ simply denotes the prior $p(s_1)$, and $p(h_1|h_0, s_1)$ denotes $p(h_1|s_1)$.

The SLDS is used in many disciplines, from econometrics to machine learning (Bar-Shalom and Li, 1998; Ghahramani and Hinton, 1998; Lerner et al., 2000; Kitagawa, 1994; Kim and Nelson, 1999; Pavlovic et al., 2001). The aSLDS has been used, for example, in state-duration modelling in acoustics (Cemgil et al., 2006) and econometrics (Chib and Dueker, 2004). See Lerner (2002) and Zoeter (2005) for recent reviews of work.

Inference

The aim of this article is to address how to perform inference in both the SLDS and aSLDS. In particular we desire the so-called *filtered* estimate $p(h_t, s_t|v_{1:t})$ and the *smoothed* estimate $p(h_t, s_t|v_{1:T})$, for any $1 \leq t \leq T$. Both filtered and smoothed inference in the SLDS is intractable, scaling exponentially with time (Lerner, 2002). To see this informally, consider the filtered posterior, which may be recursively computed using

$$p(s_t, h_t|v_{1:t}) = \sum_{s_{t-1}} \int_{h_{t-1}} p(s_t, h_t|s_{t-1}, h_{t-1}, v_t)p(s_{t-1}, h_{t-1}|v_{1:t-1}) \tag{3}$$

2. The notation $x_{1:T}$ is shorthand for $x_1, \ldots, x_T$.

At timestep 1, $p(s_1, h_1|v_1) = p(h_1|s_1, v_1)p(s_1|v_1)$ is an indexed set of Gaussians. At timestep 2, due to the summation over the states s_1, $p(s_2, h_2|v_{1:2})$ will be an indexed set of S Gaussians; similarly at timestep 3, it will be S^2 and, in general, gives rise to S^t Gaussians.

Our own interest in the SLDS stems from acoustic modelling, in which the time-series consists of many thousands of points (Mesot and Barber, 2006; Cemgil et al., 2006). For this, we require a stable and computationally feasible approximate inference, which is also able to deal with state-spaces of high dimension, H.

2. Expectation Correction

Our approach to approximate $p(h_t, s_t|v_{1:T})$ mirrors the Rauch-Tung-Striebel 'correction' smoother for the LDS (Rauch et al., 1965; Bar-Shalom and Li, 1998). Readers unfamiliar with this approach will find a short explanation in Appendix (A), which defines the important functions LDSFORWARD and LDSBACKWARD, which we shall make use of for inference in the aSLDS. Our correction approach consists of a single forward pass to recursively find the filtered posterior $p(h_t, s_t|v_{1:t})$, followed by a single backward pass to correct this into a smoothed posterior $p(h_t, s_t|v_{1:T})$. The forward pass we use is equivalent to standard Assumed Density Filtering (Minka, 2001). The main contribution of this paper is a novel form of backward pass, based only on collapsing the smoothed posterior to a mixture of Gaussians. However, we will discuss a simpler version of EC that makes an additional conditional independence assumption. This additional assumption is motivated by simplicity and also by the intuition that, in general, any deleterious effect on inference will be small.

2.1 Forward Pass (Filtering)

Readers familiar with Assumed Density Filtering may wish to continue directly to Section (2.2). Our aim is to form a recursion for $p(s_t, h_t|v_{1:t})$, based on a Gaussian mixture approximation[3] of $p(h_t|s_t, v_{1:t})$. Without loss of generality, we may decompose the filtered posterior as

$$p(h_t, s_t|v_{1:t}) = p(h_t|s_t, v_{1:t})p(s_t|v_{1:t}) \tag{4}$$

The exact representation of $p(h_t|s_t, v_{1:t})$ is a mixture with a $O(S^t)$ components. We therefore approximate this with a smaller I-component mixture

$$p(h_t|s_t, v_{1:t}) \approx \sum_{i_t=1}^{I} p(h_t|i_t, s_t, v_{1:t})p(i_t|s_t, v_{1:t})$$

where $p(h_t|i_t, s_t, v_{1:t})$ is a Gaussian parameterised with mean[4] $f(i_t, s_t)$ and covariance $F(i_t, s_t)$. To find a recursion for these parameters, consider

$$\begin{aligned} p(h_{t+1}|s_{t+1}, v_{1:t+1}) &= \sum_{s_t, i_t} p(h_{t+1}, s_t, i_t|s_{t+1}, v_{1:t+1}) \\ &= \sum_{s_t, i_t} p(h_{t+1}|s_t, i_t, s_{t+1}, v_{1:t+1})p(s_t, i_t|s_{t+1}, v_{1:t+1}) \end{aligned} \tag{5}$$

3. This derivation holds also for the aSLDS, unlike that presented in Alspach and Sorenson (1972).
4. Strictly speaking, we should use the notation $f_t(i_t, s_t)$ since, for each time t, we have a set of means indexed by i_t, s_t. This mild abuse of notation is used elsewhere in the paper.

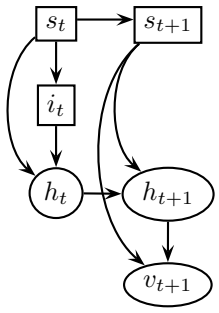


Figure 2: Structure of the mixture representation of the forward pass. Essentially, the forward pass defines a 'prior' distribution at time t which contains all the information from the variables $v_{1:t}$.

EVALUATING $p(h_{t+1}|s_t, i_t, s_{t+1}, v_{1:t+1})$

We find $p(h_{t+1}|s_t, i_t, s_{t+1}, v_{1:t+1})$ from the joint distribution $p(h_{t+1}, v_{t+1}|s_t, i_t, s_{t+1}, v_{1:t})$, which is a Gaussian with covariance and mean elements[5]

$$\begin{aligned}
\Sigma_{hh} &= A(s_{t+1})F(i_t, s_t)A^{\mathsf{T}}(s_{t+1}) + \Sigma^h(s_{t+1}), \\
\Sigma_{vv} &= B(s_{t+1})\Sigma_{hh}B^{\mathsf{T}}(s_{t+1}) + \Sigma^v(s_{t+1}) \\
\Sigma_{vh} &= B(s_{t+1})F(i_t, s_t) \\
\mu_v &= B(s_{t+1})A(s_{t+1})f(i_t, s_t) \\
\mu_h &= A(s_{t+1})f(i_t, s_t)
\end{aligned} \tag{6}$$

These results are obtained from integrating the forward dynamics, Equations (1,2) over h_t, using the results in Appendix (B). To find $p(h_{t+1}|s_t, i_t, s_{t+1}, v_{1:t+1})$ we may then condition $p(h_{t+1}, v_{t+1}|s_t, i_t, s_{t+1}, v_{1:t})$ on v_{t+1} using the results in Appendix (C).

EVALUATING $p(s_t, i_t|s_{t+1}, v_{1:t+1})$

Up to a trivial normalisation constant the mixture weight in Equation (5) can be found from the decomposition

$$p(s_t, i_t|s_{t+1}, v_{1:t+1}) \propto p(v_{t+1}|i_t, s_t, s_{t+1}, v_{1:t})p(s_{t+1}|i_t, s_t, v_{1:t})p(i_t|s_t, v_{1:t})p(s_t|v_{1:t}) \tag{7}$$

The first factor in Equation (7), $p(v_{t+1}|i_t, s_t, s_{t+1}, v_{1:t})$ is given as a Gaussian with mean μ_v and covariance Σ_{vv}, as given in Equation (6). The last two factors $p(i_t|s_t, v_{1:t})$ and $p(s_t|v_{1:t})$ are given from the previous iteration. Finally, $p(s_{t+1}|i_t, s_t, v_{1:t})$ is found from

$$p(s_{t+1}|i_t, s_t, v_{1:t}) = \langle p(s_{t+1}|h_t, s_t)\rangle_{p(h_t|i_t, s_t, v_{1:t})} \tag{8}$$

where $\langle\cdot\rangle_p$ denotes expectation with respect to p. In the standard SLDS, Equation (8) is replaced by the Markov transition $p(s_{t+1}|s_t)$. In the aSLDS, however, Equation (8) will

5. We derive this for $\bar{h}_{t+1}, \bar{v}_{t+1} \equiv 0$, to ease notation.

generally need to be computed numerically. A simple approximation is to evaluate Equation (8) at the mean value of the distribution $p(h_t|i_t, s_t, v_{1:t})$. To take covariance information into account an alternative would be to draw samples from the Gaussian $p(h_t|i_t, s_t, v_{1:t})$ and thus approximate the average of $p(s_{t+1}|h_t, s_t)$ by sampling[6].

Closing the recursion

We are now in a position to calculate Equation (5). For each setting of the variable s_{t+1}, we have a mixture of $I \times S$ Gaussians which we numerically collapse back to I Gaussians to form

$$p(h_{t+1}|s_{t+1}, v_{1:t+1}) \approx \sum_{i_{t+1}=1}^{I} p(h_{t+1}|i_{t+1}, s_{t+1}, v_{1:t+1})p(i_{t+1}|s_{t+1}, v_{1:t+1})$$

Any method of choice may be supplied to collapse a mixture to a smaller mixture. A straightforward approach that we use in our code is based on repeatedly merging low-weight components, as explained in Appendix (D). In this way the new mixture coefficients $p(i_{t+1}|s_{t+1}, v_{1:t+1})$, $i_{t+1} \in 1, \ldots, I$ are defined.

The above completes the description of how to form a recursion for $p(h_{t+1}|s_{t+1}, v_{1:t+1})$ in Equation (4). A recursion for the switch variable is given by

$$p(s_{t+1}|v_{1:t+1}) \propto \sum_{i_t, s_t} p(s_{t+1}, i_t, s_t, v_{t+1}, v_{1:t})$$

The r.h.s. of the above equation is proportional to

$$\sum_{s_t, i_t} p(v_{t+1}|s_{t+1}, i_t, s_t, v_{1:t})p(s_{t+1}|i_t, s_t, v_{1:t})p(i_t|s_t, v_{1:t})p(s_t|v_{1:t})$$

where all terms have been computed during the recursion for $p(h_{t+1}|s_{t+1}, v_{1:t+1})$.

The Likelihood $p(v_{1:T})$

The likelihood $p(v_{1:T})$ may be found by recursing $p(v_{1:t+1}) = p(v_{t+1}|v_{1:t})p(v_{1:t})$, where

$$p(v_{t+1}|v_t) = \sum_{i_t, s_t, s_{t+1}} p(v_{t+1}|i_t, s_t, s_{t+1}, v_{1:t})p(s_{t+1}|i_t, s_t, v_{1:t})p(i_t|s_t, v_{1:t})p(s_t|v_{1:t})$$

In the above expression, all terms have been computed in forming the recursion for the filtered posterior $p(h_{t+1}, s_{t+1}|v_{1:t+1})$.

The procedure for computing the filtered posterior is presented in Algorithm (1).

6. Whilst we suggest sampling as part of the aSLDS update procedure, this does not equate this with a sequential sampling procedure, such as Particle Filtering. The sampling here is a form of exact sampling, for which no convergence issues arise, being used only to numerically compute Equation (8).

Algorithm 1 aSLDS Forward Pass. Approximate the filtered posterior $p(s_t|v_{1:t}) \equiv \rho_t$, $p(h_t|s_t, v_{1:t}) \equiv \sum_{i_t} w_t(i_t, s_t)\mathcal{N}(f_t(i_t, s_t), F_t(i_t, s_t))$. Also we return the approximate log-likelihood $\log p(v_{1:T})$. We require $I_1 = 1, I_2 \leq S, I_t \leq S \times I_{t-1}$. $\theta(s) = A(s), B(s), \Sigma^h(s), \Sigma^v(s), \bar{h}(s), \bar{v}(s)$.

for $s_1 \leftarrow 1$ **to** S **do**
  $\{f_1(1, s_1), F_1(1, s_1), \hat{p}\} = \text{LDSFORWARD}(0, 0, v_1; \theta(s_1))$
  $\rho_1 \leftarrow p(s_1)\hat{p}$
end for

for $t \leftarrow 2$ **to** T **do**
  for $s_t \leftarrow 1$ **to** S **do**
    for $i \leftarrow 1$ **to** I_{t-1}, **and** $s \leftarrow 1$ **to** S **do**
      $\{\mu_{x|y}(i, s), \Sigma_{x|y}(i, s), \hat{p}\} = \text{LDSFORWARD}(f_{t-1}(i, s), F_{t-1}(i, s), v_t; \theta(s_t))$
      $p^*(s_t|i, s) \equiv \langle p(s_t|h_{t-1}, s_{t-1} = s)\rangle_{p(h_{t-1}|i_{t-1}=i, s_{t-1}=s, v_{1:t-1})}$
      $p'(s_t, i, s) \leftarrow w_{t-1}(i, s)p^*(s_t|i, s)\rho_{t-1}(s)\hat{p}$
    end for
    Collapse the $I_{t-1} \times S$ mixture of Gaussians defined by $\mu_{x|y}$,$\Sigma_{x|y}$, and weights $p(i, s|s_t) \propto p'(s_t, i, s)$ to a Gaussian with I_t components, $p(h_t|s_t, v_{1:t}) \approx \sum_{i_t=1}^{I_t} p(i_t|s_t, v_{1:t})p(h_t|s_t, i_t, v_{1:t})$. This defines the new means $f_t(i_t, s_t)$, covariances $F_t(i_t, s_t)$ and mixture weights $w_t(i_t, s_t) \equiv p(i_t|s_t, v_{1:t})$.
    Compute $\rho_t(s_t) \propto \sum_{i,s} p'(s_t, i, s)$
  end for
  normalise ρ_t
  $L \leftarrow L + \log \sum_{s_t,i,s} p'(s_t, i, s)$
end for

2.2 Backward Pass (Smoothing)

The main contribution of this paper is to find a suitable way to 'correct' the filtered posterior $p(s_t, h_t|v_{1:t})$ obtained from the forward pass into a smoothed posterior $p(s_t, h_t|v_{1:T})$. We initially derive this for the case of a single Gaussian representation. The extension to the mixture case is straightforward and is given in Section (2.4). Our derivation holds for both the SLDS and aSLDS. We approximate the smoothed posterior $p(h_t|s_t, v_{1:T})$ by a Gaussian with mean $g(s_t)$ and covariance $G(s_t)$, and our aim is to find a recursion for these parameters. A useful starting point for a recursion is:

$$p(h_t, s_t|v_{1:T}) = \sum_{s_{t+1}} p(s_{t+1}|v_{1:T})p(h_t|s_t, s_{t+1}, v_{1:T})p(s_t|s_{t+1}, v_{1:T})$$

The term $p(h_t|s_t, s_{t+1}, v_{1:T})$ may be computed as

$$\begin{aligned} p(h_t|s_t, s_{t+1}, v_{1:T}) &= \int_{h_{t+1}} p(h_t, h_{t+1}|s_t, s_{t+1}, v_{1:T}) \\ &= \int_{h_{t+1}} p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:T})p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \\ &= \int_{h_{t+1}} p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \end{aligned} \tag{9}$$

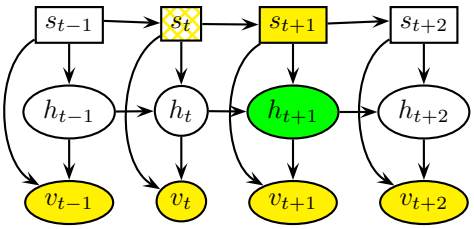


Figure 3: Our backpass approximates $p(h_{t+1}|s_{t+1}, s_t, v_{1:T})$ by $p(h_{t+1}|s_{t+1}, v_{1:T})$. Motivation for this is that s_t only influences h_{t+1} through h_t. However, h_t will most likely be heavily influenced by $v_{1:t}$, so that not knowing the state of s_t is likely to be of secondary importance. The green (darker) node is the variable we wish to find the posterior state of. The yellow (lighter shaded) nodes are variables in known states, and the hashed node a variable whose state is indeed known but assumed unknown for the approximation.

The recursion therefore requires $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$, which we can write as

$$p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \propto p(h_{t+1}|s_{t+1}, v_{1:T})p(s_t|s_{t+1}, h_{t+1}, v_{1:t}) \tag{10}$$

The difficulty here is that the functional form of $p(s_t|s_{t+1}, h_{t+1}, v_{1:t})$ is not squared exponential in h_{t+1}, so that $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$ will not be Gaussian. One possibility would be to approximate the non-Gaussian $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$ by a Gaussian (or mixture thereof) by minimising the Kullback-Leilbler divergence between the two, or performing moment matching in the case of a single Gaussian. A simpler alternative is to make the assumption $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$, see Figure(3). This makes life easy since $p(h_{t+1}|s_{t+1}, v_{1:T})$ is already known from the previous backward recursion. Under this assumption, the recursion becomes

$$p(h_t, s_t|v_{1:T}) \approx \sum_{s_{t+1}} p(s_{t+1}|v_{1:T})p(s_t|s_{t+1}, v_{1:T}) \left\langle p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})\right\rangle_{p(h_{t+1}|s_{t+1}, v_{1:T})} \tag{11}$$

We call the procedure resulting from the conditional independence assumption 'standard' EC. In Appendix (E) we show how standard EC is equivalent to a partial Discrete-Continuous factorisation approximation. Equation (11) forms the basis of the standard EC backward pass. How we implement the recursion for the continuous and discrete factors is detailed below[7].

7. Equation (11) has the pleasing form of an RTS backpass for the continuous part (analogous to LDS case), and a discrete smoother (analogous to a smoother recursion for the HMM). In the standard Forward-Backward algorithm for the HMM (Rabiner, 1989), the posterior $\gamma_t \equiv p(s_t|v_{1:T})$ is formed from the product of $\alpha_t \equiv p(s_t|v_{1:t})$ and $\beta_t \equiv p(v_{t+1:T}|s_t)$. This approach is also analogous to EP (Heskes and Zoeter, 2002). In the correction approach, a direct recursion for γ_t in terms of γ_{t+1} and α_t is formed, without explicitly defining β_t. The two approaches to inference are known as $\alpha - \beta$ and $\alpha - \gamma$ recursions.

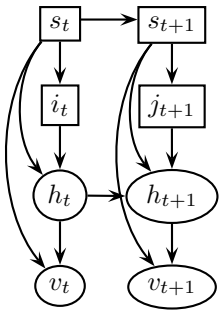


Figure 4: Structure of the backward pass for mixtures. Given the smoothed information at timestep $t+1$, we need to work backwards to integrate the filtered information from time t to 'correct' the filtered estimate at time t.

EVALUATING $\langle p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|s_{t+1},v_{1:T})}$

$\langle p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|s_{t+1},v_{1:T})}$ is a Gaussian in h_t, whose statistics we will now compute. First we find $p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})$ which may be obtained from the joint distribution

$$p(h_t, h_{t+1}|s_t, s_{t+1}, v_{1:t}) = p(h_{t+1}|h_t, s_{t+1})p(h_t|s_t, v_{1:t}) \tag{12}$$

which itself can be found from a forward dynamics from the filtered estimate $p(h_t|s_t, v_{1:t})$. The statistics for the marginal $p(h_t|s_t, s_{t+1}, v_{1:t})$ are simply those of $p(h_t|s_t, v_{1:t})$, since s_{t+1} carries no extra information about h_t[8]. The only remaining uncomputed statistics are the mean of h_{t+1}, the covariance of h_{t+1} and cross-variance between h_t and h_{t+1}, which are given by

$$\begin{aligned}
&\langle h_{t+1}\rangle = A(s_{t+1})f_t(s_t)\\
&\Sigma_{t+1,t+1} = A(s_{t+1})F_t(s_t)A^{\mathsf{T}}(s_{t+1}) + \Sigma^h(s_{t+1}), \qquad \Sigma_{t+1,t} = A(s_{t+1})F_t(s_t)
\end{aligned}$$

Given the statistics of Equation (12), we may now condition on h_{t+1} to find $p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})$. Doing so effectively constitutes a reversal of the dynamics,

$$h_t = \overleftarrow{A}(s_t, s_{t+1})h_{t+1} + \overleftarrow{\eta}(s_t, s_{t+1})$$

where $\overleftarrow{A}$ and $\overleftarrow{\eta}(s_t, s_{t+1}) \sim \mathcal{N}(\overleftarrow{m}(s_t, s_{t+1}), \overleftarrow{\Sigma}(s_t, s_{t+1}))$ are easily found using the conditioned Gaussian results in Appendix (C). Averaging the above reversed dynamics over $p(h_{t+1}|s_{t+1}, v_{1:T})$, we find that $\langle p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|s_{t+1},v_{1:T})}$ is a Gaussian with statistics

$$\mu_t = \overleftarrow{A}(s_t, s_{t+1})g(s_{t+1}) + \overleftarrow{m}(s_t, s_{t+1}), \quad \Sigma_{t,t} = \overleftarrow{A}(s_t, s_{t+1})G(s_{t+1})\overleftarrow{A}^{\mathsf{T}}(s_t, s_{t+1}) + \overleftarrow{\Sigma}(s_t, s_{t+1})$$

These equations directly mirror the standard RTS backward pass, see Algorithm (4).

8. Integrating over h_{t+1} means that the information from s_{t+1} passing through h_{t+1} via the term $p(h_{t+1}|s_{t+1}, h_t)$ vanishes. Also, since s_t is known, no information from s_{t+1} passes through s_t to h_t.

Evaluating $p(s_t|s_{t+1}, v_{1:T})$

The main departure of EC from previous methods is in treating the term

$$p(s_t|s_{t+1}, v_{1:T}) = \langle p(s_t|h_{t+1}, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|s_{t+1},v_{1:T})} \tag{13}$$

The term $p(s_t|h_{t+1}, s_{t+1}, v_{1:t})$ is given by

$$p(s_t|h_{t+1}, s_{t+1}, v_{1:t}) = \frac{p(h_{t+1}|s_{t+1}, s_t, v_{1:t})p(s_t, s_{t+1}|v_{1:t})}{\sum_{s'_t} p(h_{t+1}|s_{t+1}, s'_t, v_{1:t})p(s'_t, s_{t+1}|v_{1:t})} \tag{14}$$

Here $p(s_t, s_{t+1}|v_{1:t}) = p(s_{t+1}|s_t, v_{1:t})p(s_t|v_{1:t})$, where $p(s_{t+1}|s_t, v_{1:t})$ occurs in the forward pass, Equation (8). In Equation (14), $p(h_{t+1}|s_{t+1}, s_t, v_{1:t})$ is found by marginalising Equation (12).

Computing the average of Equation (14) with respect to $p(h_{t+1}|s_{t+1}, v_{1:T})$ may be achieved by any numerical integration method desired. The simplest approximation is to evaluate the integrand at the mean value of the averaging distribution[9] $p(h_{t+1}|s_{t+1}, v_{1:T})$. Otherwise, sampling from the Gaussian $p(h_{t+1}|s_{t+1}, v_{1:T})$, has the advantage that covariance information is used[10].

Closing the Recursion

We have now computed both the continuous and discrete factors in Equation (21), which we wish to use to write the smoothed estimate in the form $p(h_t, s_t|v_{1:T}) = p(s_t|v_{1:T})p(h_t|s_t, v_{1:T})$. The distribution $p(h_t|s_t, v_{1:T})$ is readily obtained from the joint Equation (21) by conditioning on s_t to form the mixture

$$p(h_t|s_t, v_{1:T}) = \sum_{s_{t+1}} p(s_{t+1}|s_t, v_{1:T})p(h_t|s_t, s_{t+1}, v_{1:T})$$

which may be collapsed to a single Gaussian (or mixture if desired). The smoothed posterior $p(s_t|v_{1:T})$ is given by

$$\begin{aligned} p(s_t|v_{1:T}) &= \sum_{s_{t+1}} p(s_{t+1}|v_{1:T})p(s_t|s_{t+1}, v_{1:T}) \\ &= \sum_{s_{t+1}} p(s_{t+1}|v_{1:T}) \left\langle p(s_t|h_{t+1}, s_{t+1}, v_{1:t})\right\rangle_{p(h_{t+1}|s_{t+1},v_{1:T})} . \end{aligned} \tag{15}$$

9. Replacing h_{t+1} by its mean gives the simple approximation

$$\langle p(s_t|h_{t+1}, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|s_{t+1},v_{1:T})} \approx \frac{1}{Z}\frac{e^{-\frac{1}{2}z_{t+1}^{\mathsf{T}}(s_t,s_{t+1})\Sigma^{-1}(s_t,s_{t+1}|v_{1:t})z_{t+1}(s_t,s_{t+1})}}{\sqrt{\det\Sigma(s_t, s_{t+1}|v_{1:t})}}p(s_t|s_{t+1}, v_{1:t})$$

where $z_{t+1}(s_t, s_{t+1}) \equiv \langle h_{t+1}|s_{t+1}, v_{1:T}\rangle - \langle h_{t+1}|s_t, s_{t+1}, v_{1:t}\rangle$ and Z ensures normalisation over s_t. $\Sigma(s_t, s_{t+1}|v_{1:t})$ is the filtered covariance of h_{t+1} given s_t, s_{t+1} and the observations $v_{1:t}$, which may be taken from Σ_{hh} in Equation (6).

10. This is a form of exact sampling since drawing samples from a Gaussian is easy. This should not be confused with meaning that this use of sampling renders EC a sequential Monte-Carlo sampling scheme.

Numerical Stability

Numerical stability is a concern even in the LDS, and the same is to be expected for the aSLDS. Since the standard LDS recursions LDSFORWARD and LDSBACKWARD are embedded within the EC algorithm, we may immediately take advantage of the large body of work on stabilizing the LDS recursions, such as the Joseph form (which is implemented in our code for both the forward and backward passes), or the square root form (Verhaegen and Van Dooren, 1986).

2.3 Remarks

The standard-EC Backpass procedure is closely related to Kim's method (Kim, 1994; Kim and Nelson, 1999). In both standard-EC and Kim's method, the approximation $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$, is used to form a numerically simple backward pass. The other 'approximation' in EC is to numerically compute the average in Equation (15). In Kim's method, however, an update for the discrete variables is formed by replacing the required term in Equation (15) by

$$\langle p(s_t|h_{t+1}, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|s_{t+1},v_{1:T})} \approx p(s_t|s_{t+1}, v_{1:t}) \tag{16}$$

This approximation[11] decouples the discrete backward pass in Kim's method from the continuous dynamics, since $p(s_t|s_{t+1}, v_{1:t}) \propto p(s_{t+1}|s_t)p(s_t|v_{1:t})/p(s_{t+1}|v_{1:t})$ can be computed simply from the filtered results alone. The fundamental difference therefore between EC and Kim's method is that the approximation, Equation (16), is not required by EC. The EC backward pass therefore makes fuller use of the future information, resulting in a recursion which intimately couples the continuous and discrete variables. Unlike Kim (1994) and Lerner et al. (2000), where $g_t, G_t \equiv f_t, F_t$ and only the backward pass mixture weights are updated from the forward pass, EC actually changes the Gaussian parameters g_t, G_t in a non-trivial way. The resulting effect on the quality of the approximation can be profound, as we will see in the experiments.

The Expectation Propagation algorithm, discussed in more detail in Section (3), makes the central assumption, as in EC, of collapsing the posteriors to a Gaussian family (Zoeter, 2005). However, in EP, collapsing to a mixture of Gaussians is difficult – indeed, even working with a single Gaussian may be numerically unstable. In contrast, EC works largely with moment parameterisations of Gaussians, for which relatively few numerical difficulties arise. As explained in the derivation of Equation (11), the conditional independence assumption $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$ is not strictly necessary in EC. We motivate it by computational simplicity, since finding an appropriate moment matching approximation of $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$ in Equation (10) requires a relatively expensive non-Gaussian integration. The important point here is that, if we did treat $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$ more correctly, the only assumption in EC would be a collapse to a mixture of Gaussians, as in EP. As a point of interest, as in EC, the exact computation requires only a single forward and backward pass, whilst EP is an 'open' procedure requiring iteration to convergence.

11. In the HMM, this is exact, but in the SLDS the future observations carry information about s_t.

Algorithm 2 aSLDS: EC Backward Pass. Approximates $p(s_t|v_{1:T})$ and $p(h_t|s_t, v_{1:T}) \equiv \sum_{j_t=1}^{J_t} u_t(j_t, s_t)\mathcal{N}(g_t(j_t, s_t), G_t(j_t, s_t))$ using a mixture of Gaussians. $J_T = I_T, J_t \leq S \times I_t \times J_{t+1}$. This routine needs the results from Algorithm (1).

$G_T \leftarrow F_T$, $g_T \leftarrow f_T$, $u_T \leftarrow w_T$
for $t \leftarrow T-1$ **to** 1 **do**
  for $s \leftarrow 1$ **to** S, $s' \leftarrow 1$ **to** S, $i \leftarrow 1$ **to** I_t, $j' \leftarrow 1$ **to** J_{t+1} **do**
    $(\mu, \Sigma)(i, s, j', s') = \text{LDSBACKWARD}(g_{t+1}(j', s'), G_{t+1}(j', s'), f_t(i, s), F_t(i, s), \theta(s'))$
    $p(i, s|j', s') = \langle p(s_t = s, i_t = i|h_{t+1}, s_{t+1} = s', j_{t+1} = j', v_{1:t})\rangle_{p(h_{t+1}|s_{t+1}=s', j_{t+1}=j', v_{1:T})}$
    $p(i, s, j', s'|v_{1:T}) \leftarrow p(s_{t+1} = s'|v_{1:T})u_{t+1}(j', s')p(i, s|j', s')$
  end for
  for $s_t \leftarrow 1$ **to** S **do**
    Collapse the mixture defined by weights $p(i_t = i, s_{t+1} = s', j_{t+1} = j'|s_t, v_{1_T}) \propto p(i, s_t, j', s'|v_{1:T})$, means $\mu(i_t, s_t, j_{t+1}, s_{t+1})$ and covariances $\Sigma(i_t, s_t, j_{t+1}, s_{t+1})$ to a mixture with J_t components. This defines the new means $g_t(j_t, s_t)$, covariances $G_t(j_t, s_t)$ and mixture weights $u_t(j_t, s_t)$.
    $p(s_t|v_{1:T}) \leftarrow \sum_{i_t, j', s'} p(i_t, s_t, j', s'|v_{1:T})$
  end for
end for

2.4 Using Mixtures in the Backward Pass

The extension to the mixture case is straightforward, based on the representation

$$p(h_t|s_t, v_{1:T}) \approx \sum_{j_t=1}^{J} p(j_t|s_t, v_{1:T})p(h_t|j_t, v_{1:T}).$$

Analogously to the case with a single component,

$$p(h_t, s_t|v_{1:T}) = \sum_{i_t, j_{t+1}, s_{t+1}} p(s_{t+1}|v_{1:T})p(j_{t+1}|s_{t+1}, v_{1:T})p(h_t|j_{t+1}, s_{t+1}, i_t, s_t, v_{1:T}) \cdot \langle p(i_t, s_t|h_{t+1}, j_{t+1}, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|j_{t+1}, s_{t+1}, v_{1:T})}$$

The average in the last line of the above equation can be tackled using the same techniques as outlined in the single Gaussian case. To approximate $p(h_t|j_{t+1}, s_{t+1}, i_t, s_t, v_{1:T})$ we consider this as the marginal of the joint distribution

$$p(h_t, h_{t+1}|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}) = p(h_t|h_{t+1}, i_t, s_t, j_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$$

As in the case of a single mixture, the problematic term is $p(h_{t+1}|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$. Analogously to before, we may make the assumption

$$p(h_{t+1}|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|j_{t+1}, s_{t+1}, v_{1:T})$$

meaning that information about the current switch state s_t, i_t is ignored. As in the single component case, in principle, this assumption may be relaxed and a moment matching approximation be performed instead. We can then form

$$p(h_t|s_t, v_{1:T}) = \sum_{i_t, j_{t+1}, s_{t+1}} p(i_t, j_{t+1}, s_{t+1}|s_t, v_{1:T})p(h_t|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$$

This mixture can then be collapsed to smaller mixture using any method of choice, to give

$$p(h_t|s_t, v_{1:T}) \approx \sum_{j_t} p(j_t|s_t, v_{1:T}) p(h_t|j_t, v_{1:T})$$

The resulting algorithm is presented in Algorithm (2), which includes using mixtures in both forward and backward passes.

3. Relation to other methods

Approximate inference in the SLDS has been a long-standing research topic, generating an extensive literature, to which it is difficult to serve justice. See Lerner (2002) and Zoeter (2005) for good reviews of previous work. A brief summary of some of the major existing approaches follows.

Assumed Density Filtering Since the exact filtered estimate $p(h_t|s_t, v_{1:t})$ is an (exponentially large) mixture of Gaussians a useful remedy is to project at each stage of the recursion Equation (3) back to a limited set of K Gaussians. This is a *Gaussian Sum Approximation* (Alspach and Sorenson, 1972), and is a form of *Assumed Density Filtering* (ADF) (Minka, 2001). Similarly, Generalised Pseudo Bayes2 (GPB2) (Bar-Shalom and Li, 1998; Bar-Shalom and Fortmann, 1988) also performs filtering by collapsing to a mixture of Gaussians. This approach to filtering is also taken in Lerner et al. (2000) which performs the collapse by removing spatially similar Gaussians, thereby retaining diversity.

Several smoothing approaches directly use the results from ADF. The most popular is Kim's method, which updates the filtered posterior weights to form the smoother. As discussed in Section (2.3), Kim's smoother corresponds to a potentially severe loss of future information and, in general, cannot be expected to improve much on the filtered results from ADF. The more recent work of Lerner et al. (2000) is similar in spirit to Kim's method, whereby the contribution from the continuous variables is ignored in forming an approximate recursion for the smoothed $p(s_t|v_{1:T})$. The main difference is that for the discrete variables, Kim's method is based on a correction smoother, (Rauch et al., 1965), whereas Lerner's method uses a Belief Propagation style backward pass (Jordan, 1998). Neither method correctly integrates information from the continuous variables. How to form a recursion for a mixture approximation, which does not ignore information coming through the continuous hidden variables is a central contribution of our work.

Kitagawa (1994) used a two-filter method in which the dynamics of the chain are reversed. Essentially, this corresponds to a Belief Propagation method which defines a Gaussian sum approximation for $p(v_{t+1:T}|h_t, s_t)$. However, since this is not a density in h_t, s_t, but rather a conditional likelihood, formally one cannot treat this using density propagation methods. In Kitagawa (1994), the singularities resulting from incorrectly treating $p(v_{t+1:T}|h_t, s_t)$ as a density are heuristically finessed.

Expectation Propagation EP (Minka, 2001) corresponds to an approximate implementation of Belief Propagation[12] (Jordan, 1998; Heskes and Zoeter, 2002). Whilst EP may be applied to multiply-connected graphs, it does not fully exploit the numerical advantages present in the singly-connected aSLDS structure. Nevertheless, EP is the most sophisticated rival to Kim's method and EC, since it makes the least assumptions. For this reason, we'll explain briefly how EP works. First, let's simplify the notation, and write the distribution as $p = \prod_t \phi(x_{t-1}, v_{t-1}, x_t, v_t)$, where $x_t \equiv h_t \otimes s_t$, and $\phi(x_{t-1}, v_{t-1}, x_t, v_t) \equiv p(x_t|x_{t-1})p(v_t|x_t)$. EP defines 'messages' ρ, λ[13] which contain information from past and future observations respectively[14]. Explicitly, we define $\rho_t(x_t) \propto p(x_t|v_{1:t})$ to represent knowledge about x_t given all information from time 1 to t. Similarly, $\lambda_t(x_t)$ represents knowledge about state x_t given all observations from time T to time $t+1$. In the sequel, we drop the time suffix for notational clarity. We define $\lambda(x_t)$ implicitly through the requirement that the marginal smoothed inference is given by

$$p(x_t|v_{1:T}) \propto \rho(x_t)\lambda(x_t) \tag{17}$$

Hence $\lambda(x_t) \propto p(v_{t+1:T}|x_t, v_{1:t}) = p(v_{t+1:T}|x_t)$ and represents all future knowledge about $p(x_t|v_{1:T})$. From this

$$p(x_{t-1}, x_t|v_{1:T}) \propto \rho(x_{t-1})\phi(x_{t-1}, v_{t-1}, x_t, v_t)\lambda(x_t) \tag{18}$$

Taking the above equation as a starting point, we have

$$p(x_t|v_{1:T}) \propto \int_{x_{t-1}} \rho(x_{t-1})\phi(x_{t-1}, v_{t-1}, x_t, v_t)\lambda(x_t)$$

Consistency with Equation (17) requires (neglecting irrelevant scalings)

$$\rho(x_t)\lambda(x_t) \propto \int_{x_{t-1}} \rho(x_{t-1})\phi(x_{t-1}, v_{t-1}, x_t, v_t)\lambda(x_t)$$

Similarly, we can integrate Equation (18) over x_t to get the marginal at time x_{t-1} which, by consistency, should be proportional to $\rho(x_{t-1})\lambda(x_{t-1})$. Hence

$$\rho(x_t) \propto \frac{\int_{x_{t-1}} \rho(x_{t-1})\phi(x_{t-1}, x_t)\lambda(x_t)}{\lambda(x_t)}, \qquad \lambda(x_{t-1}) \propto \frac{\int_{x_t} \rho(x_{t-1})\phi(x_{t-1}, x_t)\lambda(x_t)}{\rho(x_{t-1})} \tag{19}$$

where the divisions can be interpreted as preventing overcounting of messages. In an exact implementation, the common factors in the numerator and denominator cancel.

12. Non-parametric belief propagation (Sudderth et al., 2003), which performs approximate inference in general continuous distributions, is also related to EP applied to the aSLDS, in the sense that the messages cannot be represented easily, and are approximated by mixtures of Gaussians.
13. These correspond to the α and β messages in the Hidden Markov Model framework (Rabiner, 1989).
14. In this Belief Propagation/EP viewpoint, the backward messages, traditionally labeled as β, correspond to conditional likelihoods, and not distributions. In contrast, in the EC approach, which is effectively a so-called $\alpha - \gamma$ recursion, the backward γ messages correspond to posterior distributions.

EP addresses the fact that $\lambda(x_t)$ is not a distribution by using Equation (19) to form the projection (or 'collapse'). In the numerator, the terms $\int_{x_{t-1}} \rho(x_{t-1})\,\phi(x_{t-1}, x_t)\,\lambda(x_t)$ and $\int_{x_t} \rho(x_{t-1})\,\phi(x_{t-1}, x_t)\,\lambda(x_t)$ represent $p(x_t|v_{1:T})$ and $p(x_{t-1}|v_{1:T})$. Since these *are* distributions (an indexed mixture of Gaussians in the SLDS), they may be projected/collapsed to a single indexed Gaussian. The update for the ρ message is then found from division by the λ potential, and vice versa[15]. To perform this division, the potentials in the numerator and denominator are converted to their canonical representations. To form the ρ update, the result of the division is then reconverted back to a moment representation. The collapse is nominally made to a single Gaussian since then explicit division is well defined. The resulting recursions, due to the approximation, are no longer independent and Heskes and Zoeter (2002) show that using more than a single forward sweep and backward sweep often improves on the quality of the approximation. This coupling is a departure from the exact recursions, which should remain independent, as in our EC approach.

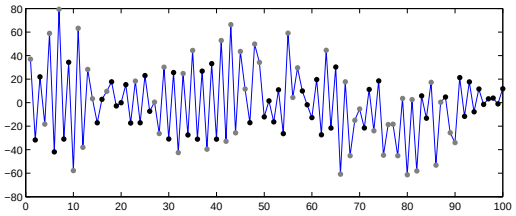
Applied to the SLDS, EP suffers from severe numerical instabilities (Heskes and Zoeter, 2002) and finding a way to minimize the corresponding EP free energy in an efficient, robust and guaranteed way remains an open problem. Damping the parameter updates is one suggested approach to heuristically improve convergence. The source of these numerical instabilities is not well understood since, even in cases when the posterior appears uni-modal, the method is problematic. The frequent conversions between moment and canonical parameterisation of Gaussians are most likely at the root of the difficulties. Our experience is that EP is currently unsuitable for large scale time series applications.

Variational Methods Ghahramani and Hinton (1998) used a variational method which approximates the joint distribution $p(h_{1:T}, s_{1:T}|v_{1:T})$ rather than the marginal inference $p(h_t, s_t|v_{1:T})$. This is a disadvantage when compared to other methods that directly approximate the marginal. The variational methods are nevertheless potentially attractive since they are able to exploit structural properties of the distribution, such as a factored discrete state-transition. In this article, we concentrate on the case of a small number of states S and hence will not consider variational methods further here[16].

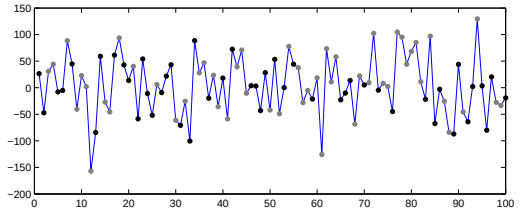
Sequential Monte Carlo (Particle Filtering) These methods form an approximate implementation of Equation (3), using a sum of delta functions to represent the posterior (see, for example, Doucet et al. (2001)). Whilst potentially powerful, these non-analytic methods typically suffer in high-dimensional hidden spaces since they are often based on naive importance sampling, which restricts their practical use. ADF is generally preferential to Particle Filtering since in ADF the approximation is a

15. In EP the explicit division of potentials only makes sense for members of the exponential family. More complex methods could be envisaged in which, rather than an explicit division, the new messages are defined by minimising some measure of divergence between $\rho(x_t)\lambda(x_t)$ and $\int_{x_{t-1}} \rho(x_{t-1})\,\phi(x_{t-1}, x_t)\,\lambda(x_t)$, such as the Kullback-Leibler divergence. Whilst this is certainly feasible, it is somewhat unattractive computationally since this would require for each timestep an expensive minimization.

16. Lerner (2002) discusses an approach in the case of a large structured discrete state transition. Related ideas could also be used in EC.



(a) Easy problem



(b) Hard problem

Figure 5: SLDS: Throughout, $S = 2$, $V = 1$ (scalar observations), $T = 100$, with zero output bias. $A(s) = 0.9999 * \text{orth}(\text{randn}(\text{H}, \text{H}))$, $B(s) = \text{randn}(\text{V}, \text{H})$, $\bar{v}_t \equiv 0$, $\bar{h}_1 = 10 * \text{randn}(\text{H}, 1)$, $\bar{h}_{t>1} = 0$, $\Sigma_1^h = I_H$, $p_1 = \text{uniform}$. The figures show typical examples for each of the two problems: (a) Easy problem. $H = 3$, $\Sigma^h(s) = I_H$, $\Sigma^v(s) = 0.1 I_V$, $p(s_{t+1}|s_t) \propto 1_{S\times S} + I_S$. (b) Hard problem. $H = 30$, $\Sigma^v = 30 I_V$,$\Sigma^h = 0.01 I_H$, $p(s_{t+1}|s_t) \propto 1_{S\times S}$.

mixture of non-trivial distributions, which is better at capturing the variability of the posterior. In addition, for applications where an accurate computation of the likelihood of the observations is required (see, for example Mesot and Barber (2006)), the inherent stochastic nature of sampling methods is undesirable.

4. Experiments

Our toy experiments examine the stability and accuracy of EC against several other methods on long time-series. In addition, we will compare the absolute accuracy of EC as a function of the number of mixture components on a short time-series, where exact inference may be explicitly evaluated. Only standard-EC is evaluated here, and evaluating EC with the relaxed conditional independence assumption is left for future work.

Testing EC in a problem with a reasonably long temporal sequence, T, is important since numerical stabilities may not be apparent in timeseries of just a few points. To do this, we sequentially generate hidden and visible states from a given model. Then, given only the parameters of the model and the visible observations (but not any of the hidden states $h_{1:T}, s_{1:T}$), the task is to infer $p(h_t|s_t, v_{1:T})$ and $p(s_t|v_{1:T})$. Since the exact computation is exponential in T, a formally exact evaluation of the method is infeasible. A simple alternative is to assume that the original sample states $s_{1:T}$ are the 'correct' inferences, and compare how our most probable posterior smoothed estimates $\arg\max_{s_t} p(s_t|v_{1:T})$ compare with the assumed correct sample s_t[17]. We look at two sets of experiments, one for the SLDS and one for the aSLDS. In both cases, scalar observations are used so that the complexity of the inference problem can be visually assessed.

17. We could also consider performance measures on the accuracy of $p(h_t|s_t, v_{1:T})$. However, we prefer to look at approximating $\arg\max_{s_t} p(s_t|v_{1:T})$ since the sampled discrete states are likely to correspond to the exact $\arg\max_{s_t} p(s_t|v_{1:T})$.

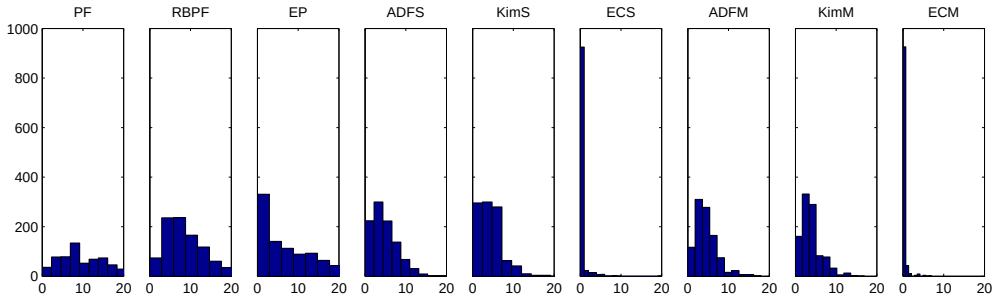


Figure 6: SLDS 'Easy' problem: The number of errors in estimating a binary switch $p(s_t|v_{1:T})$ over a time series of length $T = 100$. Hence 50 errors corresponds to random guessing. Plotted are histograms of the errors are over 1000 experiments. (PF) Particle Filter. (RBPF) Rao-Blackwellised PF. (EP) Expectation Propagation. (ADFS) Assumed Density Filtering using a Single Gaussian. (KimS) Kim's smoother using the results from ADFS. (ECS) Expectation Correction using a Single Gaussian ($I = J = 1$). (ADFM) ADF using a multiple of $I = 4$ Gaussians. (KimM) Kim's smoother using the results from ADFM. (ECM) Expectation Correction using a mixture with $I = J = 4$ components.

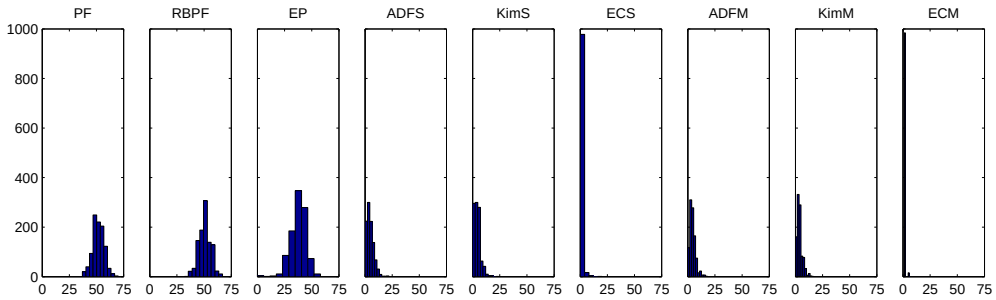


Figure 7: SLDS 'Hard' problem: The number of errors in estimating a binary switch $p(s_t|v_{1:T})$ over a time series of length $T = 100$. Hence 50 errors corresponds to random guessing. Plotted are histograms of the errors are over 1000 experiments.

SKF experiments

We chose experimental conditions that, from the viewpoint of classical signal processing, are difficult, with changes in the switches occurring at a much higher rate than the typical frequencies in the signal. We consider two different toy SLDS experiments. The 'easy' problem corresponds to a low dimensional state space, $H = 3$, with low visible noise. Conversely, the 'hard' problem corresponds to a high dimensional state space, $H = 30$, and high visible noise. See Figure(5) for full details of the experimental setup.

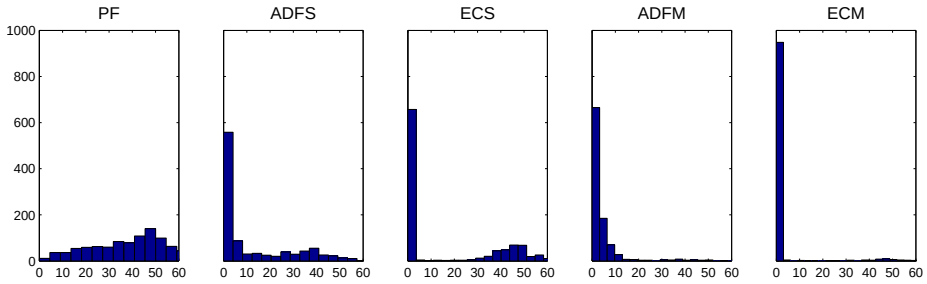


Figure 8: aSLDS: Histogram of the number of errors in estimating a binary switch $p(s_t|v_{1:T})$ over a time series of length $T = 100$. Hence 50 errors corresponds to random guessing. Plotted are histograms of the errors are over 1000 experiments. Augmented SKF results. ADFM used $I = 4$ Gaussians, and ECM used $I = J = 4$ Gaussians. We used 1000 samples to approximate Equation (13).

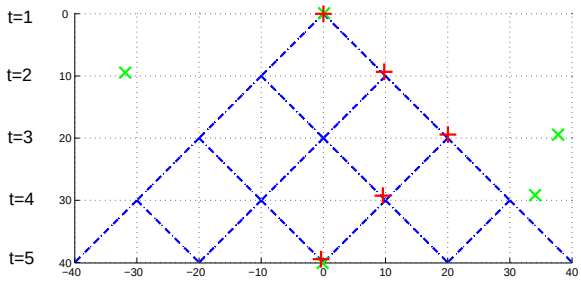
We compared methods using a single Gaussian, and methods using multiple Gaussians, see Figure(6) and Figure(7). For EC we use the mean approximation for the numerical integration of Equation (13). We included the Particle Filter merely for a point of comparison with ADF, since they are not designed to approximate the smoothed estimate, for which 1000 particles were used, with Kitagawa resampling, (Kitagawa, 1996). For the Rao-Blackwellised Particle Filter (Doucet et al., 2000), 500 particles were used, with Kitagawa resampling. We found that EP[18] was numerically unstable and often struggled to converge. To encourage convergence, we used the damping method in Heskes and Zoeter (2002), performing 20 iterations with a damping factor of 0.5. Nevertheless, the disappointing performance of EP is most likely due to conflicts resulting from numerical instabilities introduced by the frequent conversions between moment and canonical representations.

The various algorithms differ widely in performance, see Figure(6) and Figure(7). Not surprisingly, the best filtered results are given using ADF, since this is better able to represent the variance in the filtered posterior than the sampling methods. Unlike Kim's method, EC makes good use of the future information to clean up the filtered results considerably. One should bear in mind that both EC and Kim's method use the same ADF filtered results. These results show that EC may dramatically improve on Kim's method, so that the small amount of extra work in making a numerical approximation of $p(s_t|s_{t+1}, v_{1:T})$, Equation (13), may bring significant benefits.

Augmented switching model

In Figure(8), we chose a simple two state $S = 2$ transition distribution $p(s_{t+1} = 1|s_t, h_t) = \sigma\left(h_t^\mathsf{T} w(s_t)\right)$, where $\sigma(x) \equiv 1/(1 + e^{-x})$. Some care needs to be taken to make a model so that even exact inference would produce posterior switches close to the sampled switches. If the switch variables s_{t+1} can change wildly, which is possible given the above formula,

18. Generalised EP Zoeter (2005), which groups variables together improves on the results, but is still far inferior to the EC results presented here – Onno Zoeter personal communication.



(a)

	t=1	t=2	t=3	t=4	t=5
s=1	0.495	0.000	0.000	0.000	0.269
s=2	0.495	0.000	0.000	0.000	0.682
s=3	0.005	0.507	0.713	0.509	0.022
s=4	0.005	0.493	0.287	0.491	0.027

(b)

Figure 9: (a) The multipath problem. The particle starts from $(0, 0)$ at time $t = 1$. Subsequently, at each time point, either the vector $(10, 10)$ (corresponding to states $s = 1$ and $s = 3$) or $(-10, 10)$ (corresponding to states $s = 2$ and $s = 4$), is added to the hidden dynamics, perturbed by a small amount of noise, $\Sigma^h = 0.1$. The observations are $v = h + \eta^v(s)$. For states $s = 1, 2$ the observation noise is small, $\Sigma^v = 0.1I$, but for $s = 3, 4$ the noise in the horizontal direction has variance 1000. The visible observations are given by the green 'x'. The true hidden states are given by the red '+'.(b) The exact state smoothed state posteriors $p^{exact}(s_t|v_{1:T})$ computed by enumerating all paths (given by the blue dashed lines).

I	1	4	4	16	16	64	64	256	256
J	1	1	4	1	16	1	64	1	256
error	0.0989	0.0624	0.0365	0.0440	0.0130	0.0440	4.75e-4	0.0440	3.40e-8

Table 1: Errors in approximating the states for the multipath problem, see Figure(9). The mean absolute deviation $|p^{ec}(s_t|v_{1:T}) - p^{exact}(s_t|v_{1:T})|$ averaged over the $S = 4$ states of s_t and over the times $t = 1, \ldots, 5$, computed for different numbers of mixture component in EC. The 'mean' integral approximation, Equation (13), is used. The exact computation uses $S^{T-1} = 256$ mixtures.

essentially no information is left in the signal for any inference method to produce reasonable results. We therefore set $w(s_t)$ to a zero vector except for the first two components, which are independently sampled from a zero mean Gaussian with standard deviation 5. For each of the two switch states, s, we have a transition matrix $A(s)$, which we set to be block diagonal. The first 2×2 block is set to $0.9999R_\theta$ where R_θ is a 2×2 rotation matrix with angle θ chosen uniformly from 0 to 1 radians. This means that s_{t+1} is dependent on the first two components of h_t which are rotating at a restricted rate. The remaining $H - 2 \times H - 2$ block of $A(s)$ is chosen as (using MATLAB notation) $0.9999 * \text{orth}(\text{rand}(\text{H} - 2))$, which means a scaled randomly chosen orthogonal matrix. Throughout, $S = 2$, $V = 1$, $H = 30$, $T = 100$, with zero output bias. Using partly MATLAB notation, $B(s) = \text{randn}(\text{V}, \text{H})$, $\bar{v}_t \equiv 0$, $\bar{h}_1 = 10 * \text{randn}(\text{H}, 1)$, $\bar{h}_{t>1} = 0$, $\Sigma_1^h = I_H$, $p_1 = \text{uniform}$. $\Sigma^v = 30I_V$, $\Sigma^h = 0.1I_H$. We compare EC only against Particle Filters using 1000 particles, since other methods

would require specialised and novel implementations. In ADFM, $I = 4$ Gaussians were used, and for ECM, $I = J = 4$ Gaussians were used. Looking at the results in Figure(8), we see that EC performs very well, with some improvement in using the mixture representation $I, J = 4$ over a single Gaussian $I = J = 1$. The Particle Filter most likely failed since the hidden dimension is too high to be explored well with only 1000 particles.

Effect of using mixtures

Our claim is that EC should cope in situations where, not just the filtered posterior $p(h_t|s_t, v_{1:t})$, but also the smoothed posterior $p(h_t|s_t, v_{1:T})$ is multimodal and, consequently, cannot be well represented by a single Gaussian[19]. We therefore constructed an SLDS which exhibits multimodality to see the effect of using EC with both I and J greater than 1. The 'multipath' scenario is described in Figure(9), where a particle traces a path through a two dimensional space. A small number of timesteps was chosen so that the exact $p(s_t|v_{1:T})$ could be computed by direction enumeration. The observation of the particle is at times extremely noisy in the horizontal direction. This induces multimodality of $p(h_t|s_t, v_{1:T})$ since there are several paths that might plausibly have been taken to give rise to the observations. The accuracy with which EC predicts the exact smoothed posterior is given in Table(1). For this problem we see that both the number of forward components I and the number of backward components J affects the accuracy of the approximation, generally with improved accuracy as the number of mixture components increases. For a 'perfect' approximation method, one would expect that when $I = J = S^{T-1} = 256$, then the approximation should be exact. The small error for this case in Table(1) may arise for several reasons : the collapse to a mixture of Gaussians (since the exact smoothed posterior is not a mixture of Gaussians), the extra independence assumption used in EC, or the simple mean approximation used to compute Equation (13). However, at least in this case, the effect of these assumptions on the performance is very small.

5. Discussion

Expectation Correction is a novel form of backward pass which makes less approximations than the standard approach from Kim (1994). In Kim's method, potentially important future information channeled through the continuous hidden variables is lost. Standard-EC, along with Kim's method, makes the additional assumption $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$. However, our experience is that this assumption is rather mild, since the state of h_{t+1} will be most heavily influenced by its immediate parent s_{t+1}. Knowing $v_{1:T}$ should in most cases give good information about the state of h_t, so that not knowing the state s_t will not cost much. However, of critical importance is the numerical stability of the method, particularly for long timeseries, where EC significantly out-benefits EP. In tracking situations where the visible information is (temporarily) not enough to specify accurately the hidden state, then representing the posterior $p(h_t|s_t, v_{1:T})$ using a mixture of Gaussians may improve results significantly. In EC, using a mixture of Gaussians[20] is fast and numerically stable, in contrast to EP. The conditional independence assumption

19. This should not be confused with the multimodality of $p(h_t|v_{1:T}) = \sum_{s_t} p(h_t|s_t, v_{1:T})p(s_t|v_{1:T})$.
20. Whilst we presented our work in terms of Gaussians, in principle the method should be applicable to more complex members of the exponential family.

of standard EC, namely $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$, may be relaxed, at the expense of requiring a moment matching Gaussian approximation of Equation (10). Whilst we did not do so, implementing this should not give rise to numerical instabilities since no potential divisions are required, merely the estimation of moments. In the experiments presented here, we did not pursue this option, since we believe that the effect of this conditional independence assumption is relatively weak. If this approach were taken, then the only remaining assumption of EC would be to project to a Gaussian mixture approximation of the filtered and smoothed posterior. This would make the assumptions essentially the same as that of the rival EP method. Interestingly, though, EC will only ever result in a single forward and backward pass, unlike EP which requires multiple forward and backward passes. This shows that the methods of computing the updates for the two approaches are fundamentally different, with EC carefully exploiting the chain structure of the distribution. EC has time complexity $O(S^2IJK)$ where S are the number of switch states, I and J are the number of Gaussians used in the Forward and Backward passes, and K is the time to compute the exact Kalman smoother for the system with a single switch state.

An interesting question is whether one could generalise EC to multiply connected structures. For cases amenable to cutset conditioning (Castillo et al., 1997), this is straightforward, though for more general cases, some care would be needed to avoid overcounting.

6. Conclusion

We have presented a method that can be used for approximate smoothed inference in an augmented class of switching linear dynamical systems with additive Gaussian noise. Our approximation is based on the idea that, although exact inference will consist of an exponentially large number of mixture components, due to the forgetting which commonly occurs in Markovian models, a finite number of mixture components may provide a reasonable approximation. Clearly, in systems with very long correlation times our method may require too many mixture components to produce a satisfactory result, although we are unaware of other techniques that would be able to cope well in that case. The main benefit of EC over the Kim smoothing approach is that future information is more accurately dealt with and, additionally, the method is relatively numerically stable compared to the alternative EP procedure. The relaxed version of EC makes the same basic assumptions as EP, but results only in a single forward and backward pass, each being based on a stable update procedure. In a related work, we have successfully applied EC to a problem in automatic speech recognition where we model a one dimensional speech signal using a SLDS (Mesot and Barber, 2006). The signal consists of many thousands of timepoints, and numerical stability is an important concern. The application also discusses parameter learning which can be achieved using the usual EM approach. We hope that the straightforward ideas presented here may help facilitate the practical application of dynamics hybrid networks to machine learning and related areas.

Software for Expectation Correction for this augmented class of Switching Linear Gaussian models is at `www.idiap.ch/~bmesot/ec`

Algorithm 3 LDS Forward Pass. Compute the filtered posteriors $p(h_t|v_{1:t}) \equiv \mathcal{N}(f_t, F_t)$ for a LDS with parameters $\theta = A, B, \Sigma^h, \Sigma^v, \bar{h}, \bar{v}$. The log-likelihood $L = \log p(v_{1:T})$ is also returned.

$F_0 \leftarrow 0$, $f_0 \leftarrow 0$, $L \leftarrow 0$
for $t \leftarrow 1, T$ **do**
  $\{f_t, F_t, p_t\} = \text{LDSFORWARD}(f_{t-1}, F_{t-1}, v_t; \theta)$
  $L \leftarrow L + \log p_t$
end for
function LDSFORWARD$(f, F, v; \theta)$
  Compute joint $p(h_t, v_t|v_{1:t-1})$:
  $\mu_h \leftarrow Af + \bar{h}, \qquad \mu_v \leftarrow B\mu_h + \bar{v}$
  $\Sigma_{hh} \leftarrow AFA^\mathsf{T} + \Sigma^h, \qquad \Sigma_{vv} \leftarrow B\Sigma_{hh} + \Sigma^v, \qquad \Sigma_{vh} \leftarrow B\Sigma_{hh}$
  Find $p(h_t|v_{1:t})$ by conditioning:
  $f' \leftarrow \mu_h + \Sigma_{vh}^\mathsf{T}\Sigma_{vv}^{-1}(v - \mu_v), \qquad F' \leftarrow \Sigma_{hh} - \Sigma_{vh}^\mathsf{T}\Sigma_{vv}^{-1}\Sigma_{vh}$
  Compute $p(v_t|v_{1:t-1})$:
  $p' \leftarrow \exp\left(-\frac{1}{2}(v - \mu_v)^\mathsf{T}\Sigma_{vv}^{-1}(v - \mu_v)\right)/\sqrt{\det 2\pi\Sigma_{vv}}$
  return f', F', p'
end function

Acknowledgements

I would like to thank Onno Zoeter and Tom Heskes for kindly providing their Expectation Propagation code, Silvia Chiappa for helpful discussions, and Bertrand Mesot for help with the simulations and for suggesting the relationship between the partial factorisation and independence viewpoints of standard-EC. I would also like to thank the reviewers for their many helpful comments.

Appendix A. Inference in the LDS

The LDS is defined by equations (1,2) in the case of a single switch $S = 1$. The LDS Forward and Backward passes define the important functions LDSFORWARD and LDSBACKWARD, which we shall make use of for inference in the aSLDS.

Forward Pass (Filtering)

The filtered posterior $p(h_t|v_{1:t})$ is a Gaussian which we parameterise with mean f_t and covariance F_t. These parameters can be updated recursively using $p(h_t|v_{1:t}) \propto p(h_t, v_t|v_{1:t-1})$, where the joint distribution $p(h_t, v_t|v_{1:t-1})$ has statistics (see Appendix (B))

$$\mu_h = Af_{t-1} + \bar{h}, \quad \mu_v = B\mu_h + \bar{v}$$

$$\Sigma_{hh} = AF_{t-1}A^\mathsf{T} + \Sigma^h, \quad \Sigma_{vv} = B\Sigma_{hh} + \Sigma^v, \quad \Sigma_{vh} = B\Sigma_{hh}$$

We may then find $p(h_t|v_{1:t})$ by conditioning $p(h_t, v_t|v_{1:t-1})$ on v_t, see Appendix (C). This gives rise to Algorithm (3).

Algorithm 4 LDS Backward Pass. Compute the smoothed posteriors $p(h_t|v_{1:T})$. This requires the filtered results from Algorithm (3).

$G_T \leftarrow F_T$, $g_T \leftarrow f_T$
for $t \leftarrow T-1, 1$ **do**
  $\{g_t, G_t\} = \text{LDSBACKWARD}(g_{t+1}, G_{t+1}, f_t, F_t; \theta)$
end for
function LDSBACKWARD$(g, G, f, F; \theta)$
  $\mu_h \leftarrow Af + \bar{h}, \qquad \Sigma_{h'h'} \leftarrow AFA^\mathsf{T} + \Sigma^h, \qquad \Sigma_{h'h} \leftarrow AF$
  $\overleftarrow{\Sigma} \leftarrow F_t - \Sigma_{h'h}^\mathsf{T}\Sigma_{h'h'}^{-1}\Sigma_{h'h}, \quad \overleftarrow{A} \leftarrow \Sigma_{h'h}^\mathsf{T}\Sigma_{h'h'}^{-1}, \quad \overleftarrow{m} \leftarrow f - \overleftarrow{A}\mu_h$
  $g' \leftarrow \overleftarrow{A}g + \overleftarrow{m}, \quad G' \leftarrow \overleftarrow{A}G\overleftarrow{A}^\mathsf{T} + \overleftarrow{\Sigma}$
  return g', G'
end function

Backward Pass

The smoothed posterior $p(h_t|v_{1:T}) \equiv \mathcal{N}(g_t, G_t)$ can be computed by recursively using:

$$p(h_t|v_{1:T}) = \int_{h_{t+1}} p(h_t|h_{t+1}, v_{1:T})p(h_{t+1}|v_{1:T}) = \int_{h_{t+1}} p(h_t|h_{t+1}, v_{1:t})p(h_{t+1}|v_{1:T})$$

where $p(h_t|h_{t+1}, v_{1:t})$ may be obtained from the joint distribution

$$p(h_t, h_{t+1}|v_{1:t}) = p(h_{t+1}|h_t)p(h_t|v_{1:t}) \tag{20}$$

which itself can be obtained by forward propagation from $p(h_t|v_{1:t})$. Conditioning Equation (20) to find $p(h_t|h_{t+1}, v_{1:t})$ effectively reverses the dynamics,

$$h_t = \overleftarrow{A}_t h_{t+1} + \overleftarrow{\eta}_t$$

where $\overleftarrow{A}_t$ and $\overleftarrow{\eta}_t \sim \mathcal{N}(\overleftarrow{m}_t, \overleftarrow{\Sigma}_t)$ are found using the conditioned Gaussian results in Appendix (C). Then averaging the reversed dynamics over $p(h_{t+1}|v_{1:T})$ we find that $p(h_t|v_{1:T})$ is a Gaussian with statistics

$$g_t = \overleftarrow{A}_t g_{t+1} + \overleftarrow{m}_t, \quad G_t = \overleftarrow{A}_t G_{t+1} \overleftarrow{A}_t^\mathsf{T} + \overleftarrow{\Sigma}_t$$

This backward pass is given in Algorithm (4). For parameter learning of the A matrix, the smoothed statistic $\left\langle h_t h_{t+1}^\mathsf{T} \right\rangle$ is required. Using the above formulation, this is given by $\overleftarrow{A}_t G_{t+1} + \langle h_t \rangle \left\langle h_{t+1}^\mathsf{T} \right\rangle$. This is much simpler than the standard expressions cited in Shumway and Stoffer (2000) and Roweis and Ghahramani (1999).

Appendix B. Gaussian Propagation

Let y be linearly related to x through $y = Mx + \eta$, where $\eta \sim \mathcal{N}(\mu, \Sigma)$, and $x \sim \mathcal{N}(\mu_x, \Sigma_x)$. Then $p(y) = \int_x p(y|x)p(x)$ is a Gaussian with mean $M\mu_x + \mu$ and covariance $M\Sigma_x M^\mathsf{T} + \Sigma$.

Appendix C. Gaussian Conditioning

For a joint Gaussian distribution over the vectors x and y with means μ_x, μ_y and covariance elements Σ_{xx},Σ_{xy},Σ_{yy}, the conditional $p(x|y)$ is a Gaussian with mean $\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$ and covariance $\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$.

Appendix D. Collapsing Gaussians

The user may provide any algorithm of their choice for collapsing a set of Gaussians to a smaller set of Gaussians (Titterington et al., 1985). Here, to be explicit, we present a simple one which is fast, but has the disadvantage that no spatial information about the mixture is used.

First, we describe how to collapse a mixture to a *single* Gaussian: We may collapse a mixture of Gaussians $p(x) = \sum_i p_i \mathcal{N}(x|\mu_i, \Sigma_i)$ to a single Gaussian with mean $\sum_i p_i \mu_i$ and covariance $\sum_i p_i \left(\Sigma_i + \mu_i \mu_i^\mathsf{T}\right) - \mu\mu^\mathsf{T}$.

To collapse a mixture to a K-component *mixture* we retain the $K-1$ Gaussians with the largest mixture weights – the remaining $N-K$ Gaussians are simply merged to a single Gaussian using the above method. The alternative of recursively merging the two Gaussians with the lowest mixture weights gave similar experimental performance.

More sophisticated methods which retain some spatial information would clearly be potentially useful. The method presented in Lerner et al. (2000) is a suitable approach which considers removing Gaussians which are spatially similar (and not just low-weight components), thereby retaining a sense of diversity over the possible solutions.

Appendix E. The Discrete-Continuous factorisation Viewpoint

An alternative viewpoint is to proceed analogously to the Rauch-Tung-Striebel (RTS) correction method for the LDS (Bar-Shalom and Li, 1998):

$$\begin{aligned}
p(h_t, s_t|v_{1:T}) &= \sum_{s_{t+1}} \int_{h_{t+1}} p(s_t, h_t, h_{t+1}, s_{t+1}|v_{1:T}) \\
&= \sum_{s_{t+1}} p(s_{t+1}|v_{1:T}) \int_{h_{t+1}} p(h_t, s_t|h_{t+1}, s_{t+1}, v_{1:t}) p(h_{t+1}|s_{t+1}, v_{1:T}) \\
&= \sum_{s_{t+1}} p(s_{t+1}|v_{1:T}) \left\langle p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t}) p(s_t|h_{t+1}, s_{t+1}, v_{1:t}) \right\rangle \\
&\approx \sum_{s_{t+1}} p(s_{t+1}|v_{1:T}) \left\langle p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t}) \right\rangle \underbrace{\left\langle p(s_t|s_{t+1}, v_{1:T}) \right\rangle}_{p(s_t|s_{t+1}, v_{1:T})}
\end{aligned} \tag{21}$$

where angled brackets $\langle \cdot \rangle$ denote averages with respect to $p(h_{t+1}|s_{t+1}, v_{1:T})$. Whilst the factorised approximation in Equation (21) may seem severe, by comparing Equations (21) and (11) we see that it is equivalent to the apparently mild assumption $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$. Hence this factorised approximation is equivalent to the 'standard' EC approach in which the dependency on s_t is dropped.

References

D. L. Alspach and H. W. Sorenson. Nonlinear Bayesian Estimation Using Gaussian Sum Approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972.

Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.

Y. Bar-Shalom and Xiao-Rong Li. *Estimation and Tracking : Principles, Techniques and Software*. Artech House, Norwood, MA, 1998.

E. Castillo, J. M. Gutierrez, and A.S. Hadi. *Expert Systems and Probabilistic Network Models*. Springer, 1997.

A. T. Cemgil, B. Kappen, and D. Barber. A Generative Model for Music Transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 14(2):679 – 694, 2006.

S. Chib and M. Dueker. Non-markovian regime switching with endogenous states and time-varying state strengths. *Econometric Society 2004 North American Summer Meetings 600*, 2004.

A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Srpinger, 2001.

A. Doucet, N. de Freitas, K. Murphy, and S. Russell. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. *Uncertainty in Artificial Intelligence*, 2000.

Z. Ghahramani and G.E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):963–996, 1998.

T. Heskes and O. Zoeter. Expectation Propagation for approximate inference in dynamic Bayesian networks. In A. Darwiche and N. Friedman, editors, *Uncertainty in Artificial Intelligence*, pages 216–223, 2002.

M. I. Jordan. *Learning in Graphical Models*. MIT Press, 1998.

C-J. Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60: 1–22, 1994.

C-J. Kim and C. R. Nelson. *State-Space models with regime switching*. MIT Press, 1999.

G. Kitagawa. The Two-Filter Formula for Smoothing and an implementation of the Gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics*, 46(4):605–623, 1994.

G. Kitagawa. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

U. Lerner, R. Parr, D. Koller, and G. Biswas. Bayesian Fault Detection and Diagnosis in Dynamic Systems. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AIII-00)*, pages 531–537, 2000.

U. N. Lerner. *Hybrid Bayesian Networks for Reasoning about Complex Systems*. PhD thesis, Stanford University, 2002.

B. Mesot and D. Barber. Switching linear dynamical systems for noise robust speech recognition. IDIAP-RR 08, 2006.

T. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT Media Lab, 2001.

V. Pavlovic, J.M Rehg, and J. MacCormick. Learning switching linear models of human motion. In *Advances in Neural Information Processing systems (NIPS 13)*, pages 981–987, 2001.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2), 1989.

H. E. Rauch, G. Tung, and C. T. Striebel. Maximum Likelihood estimates of linear dynamic systems. *American Institute of Aeronautics and Astronautics Journal (AIAAJ)*, 3(8): 1445–1450, 1965.

S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*. Springer, 2000.

E. B. Sudderth, A. T. Ihler, and A. S. Freeman, W. T.and Willsky. Nonparametric belief propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 605–612, 2003.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Wiley, 1985.

M. Verhaegen and P. Van Dooren. Numerical Aspects of Different Kalman Filter Implementations. *IEEE Transactions of Automatic Control*, 31(10):907–917, 1986.

M. West and J. Harrison. *Bayesian forecasting and dynamic models*. Springer, 1999.

O. Zoeter. *Monitoring non-linear and switching dynamical systems*. PhD thesis, Radboud University Nijmegen, 2005.