IDIAP RESEARCH REPORT

Construction and comparison of approximations for switching linear gaussian state space models

David Barber and Bertrand Mesot

IDIAP–RR 05-06

February 2005

IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-06

Construction and comparison of approximations for switching linear gaussian state space models

David Barber and Bertrand Mesot

February 2005

Abstract. We introduce a new method for approximate inference in Hybrid Dynamical Graphical models, in particular, for switching dynamical networks. For the important special case of switching linear Gaussian state space models (switching Kalman Filters), our method is a novel form of Gaussian sum smoother, consisting of a single forward and backward pass. Our method is particularly well suited to switching observation models, since one of the key approximations is obviated. We compare our method very favourably against a range of competing techniques, including sequential Monte Carlo and Expectation Propagation, for which we also derive a novel numerically more stable implementation using the 'auxiliary variable trick'. We show that the use of mixture representations for both filtering and smoothing can dramatically improve the quality of the approximation.

1 Introduction

Hybrid graphical models are stochastic systems which contain both continuous and discrete hidden/latent variables. Such models appear naturally in applications where a continuous state space dynamics can switch between different behavioural regimes, which affects useful form of non-stationarity. Here we will be interested in temporal models of the form

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

where t is a discrete time index, h_t is the state of the continuous hidden vector, s_t is a discrete switch variable, and v_t is the visible/observed vector (usually this is continuous). The notation $x_{1:T}$ is shorthand for $x_1, \ldots, x_T$. At time $t = 1$, $p(s_1|s_0)$ simply denotes the prior $p(s_1)$. The graphical model corresponding to this distribution is depicted in fig(1), which represents a temporal chain; extending our considerations in this work to tree structures and higher order processes is straightforward.

1.0.1 Switching Linear State Space models

A special case of the above framework is the switching linear Gaussian state space model, also known as the Switching Kalman Filter/Smoother (SKF), which is well known in many different disciplines [3, 9, 21, 19, 12]. It is useful to describe these models as stochastic linear recursions with additive Gaussian noise. The observation or visible variable v_t is linearly related to the hidden state h_t by

$$v_t = B(s_t)h_t + \eta^v(s_t), \qquad \eta^v(s_t) \sim \mathcal{N}\left(\bar{v}(s_t), \Sigma^v(s_t)\right) \tag{2}$$

where $\bar{v}(s_t)$ is the mean of the switch dependent observation (emission) noise at time t. Similarly, $\Sigma^v(s_t)$ is the covariance. The transition dynamics is linear,

$$h_t = A(s_t)h_{t-1} + \eta^h(s_t), \qquad \eta^h(s_t) \sim \mathcal{N}\left(\bar{h}(s_t), \Sigma^h(s_t)\right) \tag{3}$$

where $\bar{h}(s_t)$ is the mean of the transition noise, and $\Sigma^h(s_t)$ the corresponding covariance. This is an example of a form of switching dynamics since the variable s_t controls which of a discrete set of linear hidden dynamics and emissions will be used. The discrete switch variable s_t itself is Markovian, with transition $p(s_t|s_{t-1})$. For notational simplicity, we dropped time suffices, and the reader should bear in mind that all that follows holds also when the SKF parameters are time dependent. Usually, we will set the means to zero, but they may be used, for example, to model time dependent external inputs. An equivalent probabilistic formulation of the above equations is given by equation (1) with

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

1.1 Forward-Backward Equations (Belief Propagation)

We are interested in how to perform inference in models of the form equation (1), both for the SKF and more general scenarios. 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 $1 \leq t \leq T$. From the general theory of graphical models, since the distribution has the form of a chain, straightforward belief propagation (a special case of the Junction Tree algorithm) over the pair h_t, s_t will produce the exact inferences [11]. Normally, therefore, such chain like structures would not cause much anxiety. However, the hybrid networks we consider are formally intractable since the *representation* of the messages in belief propagation (or any other exact approach) is exponentially complex in time.

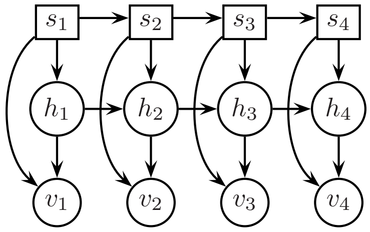


Figure 1: A Switching State-Space model. Square nodes denote discrete variables, round nodes continuous variables. The variables h_t are continuous. Most commonly the visible output variables v_t are continuous. The switch variables s_t are discrete, and control the hidden dynamics, and possibly also the emissions. Links which are not allowed are those from continuous hidden variables to discrete variables.

For readers less familiar with the probabilistic approach, we'll briefly describe here how to perform inference on chains [11, 20]. The presentation here follows [10]. First, let's simplify the notation, and write the distribution as

$$p = \prod_t \phi\left(x_{t-1}, v_{t-1}, x_t, v_t\right)$$

where $x_t \equiv h_t \otimes s_t$, and $\phi\left(x_{t-1}, v_{t-1}, x_t, v_t\right) \equiv p(x_t|x_{t-1})p(v_t|x_t)$. Our aim is to define 'messages' ρ, λ (these correspond to the α and β messages in the Hidden Markov Model framework [16]) which contain information from past observations and future observations respectively. 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 information from the future 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

$$p(x_t|v_{1:T}) \propto \rho\left(x_t\right) \lambda\left(x_t\right) \tag{4}$$

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

$$p(x_{t-1}, x_t|v_{1:T}) \propto \rho\left(x_{t-1}\right) \phi\left(x_{t-1}, v_{t-1}, x_t, v_t\right) \lambda\left(x_t\right) \tag{5}$$

Taking the above equation as a starting point, we have

$$p(x_t|v_{1:T}) \propto \int_{x_{t-1}} \rho\left(x_{t-1}\right) \phi\left(x_{t-1}, v_{t-1}, x_t, v_t\right) \lambda\left(x_t\right) \tag{6}$$

Consistency with equation (4) requires (neglecting irrelevant scalings)

$$\rho\left(x_t\right) \lambda\left(x_t\right) \propto \int_{x_{t-1}} \rho\left(x_{t-1}\right) \phi\left(x_{t-1}, v_{t-1}, x_t, v_t\right) \lambda\left(x_t\right) \tag{7}$$

Similarly, we can integrate equation (5) over x_t to get the marginal at time x_{t-1} which by consistency should be proportional to $\rho\left(x_{t-1}\right) \lambda\left(x_{t-1}\right)$. Hence

$$\rho\left(x_t\right) \propto \frac{\int_{x_{t-1}} \rho\left(x_{t-1}\right) \phi\left(x_{t-1}, x_t\right) \lambda\left(x_t\right)}{\lambda\left(x_t\right)}, \qquad \lambda\left(x_{t-1}\right) \propto \frac{\int_{x_t} \rho\left(x_{t-1}\right) \phi\left(x_{t-1}, x_t\right) \lambda\left(x_t\right)}{\rho\left(x_{t-1}\right)} \tag{8}$$

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

$$\text{Forward Recursion: } \rho\left(x_t\right) \propto \int_{x_{t-1}} \rho\left(x_{t-1}\right) \phi\left(x_{t-1}, v_{t-1}, x_t, v_t\right) \tag{9}$$

$$\text{Backward Recursion: } \lambda\left(x_{t-1}\right) \propto \int_{x_t} \phi\left(x_{t-1}, v_{t-1}, x_t, v_t\right) \lambda\left(x_t\right) \tag{10}$$

which are the usual definitions of the messages, defined as a set of independent recursions. The extension to more general singly connected structures is straightforward and results in partially independent recursions which communicate only at branch points of the tree [11].

Why is inference intractable?

Without loss of generality, let's write the forward message $\rho(h_t, s_t)$ as $\rho\left(s_t\right)\rho(h_t|s_t)$. Then the forward recursion equation (9) is

$$\rho\left(s_t\right)\rho(h_t|s_t) \propto \sum_{s_{t-1}} \rho\left(s_{t-1}\right) p(s_t|s_{t-1}) \int_{h_{t-1}} p(h_t|h_{t-1}, s_{t-1}) p(v_t|h_t, s_t) \rho(h_{t-1}|s_{t-1})$$

Even if the integral is tractable, due to the summation over s_{t-1}, for each state s_t, the distribution $\rho(h_t|s_t)$ is a mixture. At each time step the number of mixture component parameters that need to be passed increases by a factor of S, resulting in an exponential increase in the complexity of representing the messages with time. Hence, the intractability here arises not because of the structure of the graph, but because the messages cannot be represented in a compact way.

1.2 Approximation Strategies

Whilst much of what we consider in this article extends readily to more general noise models, by far the most popular scenario studied in the literature are SKFs, and most reported approximation strategies are specific to this case.

For the SKF, the filtered estimate $\rho(h_t|s_t)$ will be a mixture of Gaussians, with an exponential explosion of components with time. One useful strategy for filtering, therefore, is to represent $\rho(h_{t-1}|s_{t-1})$ by a set of K Gaussians, and project the set of $K \times S$ Gaussians that represent $\rho(h_t|s_t)$ back to a mixture of K Gaussians. This is the so-called Gaussian Sum approximation [2], and is a form of Assumed Density Filtering (ADF) [14]. The fact that $\lambda\left(x_t\right)$ is not a distribution in x_t has important consequences for any approximation method since it is not clear how to perform such collapse or projection methods on non-distributions. For this reason, approximate filtering is 'easier' than smoothing.

To make a Gaussian Sum approximation suitable for smoothing, [13] used a two-filter method in which the dynamics of the chain are reversed, and the Gaussian Sum approximation on the forward and reversed dynamics combined. In principle, such an approach is potentially attractive, although the implementation in [13] is somewhat inelegant in the specific approach used to form the dynamics reversal, appealing to unnecessary heuristics. Our approach may be viewed as somewhat similar to a form of dynamics reversal, but is formulated consistently within the given constraints of the probabilistic framework.

Expectation Propagation (EP) [14] addresses the fact that $\lambda(x_t)$ is not a distribution by using equation (8) to form the projection (or 'collapse'). In the numerator, the terms $\int_{x_{t-1}} \rho\left(x_{t-1}\right) \phi\left(x_{t-1}, x_t\right) \lambda\left(x_t\right)$ and $\int_{x_t} \rho\left(x_{t-1}\right) \phi\left(x_{t-1}, x_t\right) \lambda\left(x_t\right)$ represent the joint distributions $p(x_t|v_{1:T})$ and $p(x_{t-1}|v_{1:T})$. Since these *are* distributions (albeit a mixture in the hybrid case), they may be projected/collapsed to a single or smaller number of components. The update for the ρ message is then given by division by the λ potential, and vice versa. The resulting recursions, due to the approximation, are no longer independent and [10] show that using more than a single forward sweep and backward sweep often improves on the quality of the approximation. However, Expectation Propagation is notoriously unstable. In order to eliminate numerical instabilities in our experimental comparisons, we have used our own implementation of EP, see section (4), which is numerically relatively stable and based on the work in [4]. A difficulty with EP is that division of potentials only makes sense for members of the exponential family. More complex methods could be envisaged in which rather than an explicit

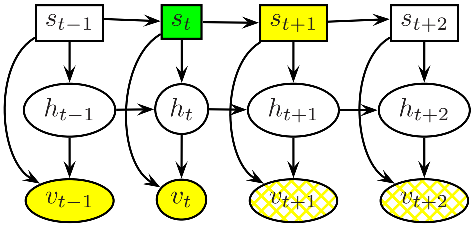


Figure 2: The approximation used in the GPB2 method. One approximates $p(s_t|s_{t+1}, v_{1:T})$ by $p(s_t|s_{t+1}, v_{1:t})$. In general, one wouldn't expect this approximation to improve much upon the filtered estimate, since all the future observations are discarded. The only additional information, beyond that used by the filtered estimate, is the state of s_{t+1}. 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 nodes are variables whose states are indeed known, but assumed unknown for the approximation.

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.

Variational methods [9] are interesting since they are able to exploit the structure in the hidden space. For example, if the switch s_t has a factorial structure, the variational methods can still be tractably implemented. Many other methods are more difficult since they scale exponentially with the number of hidden factors. Whilst the variational methods are therefore potentially useful, they suffer in the sense that their goal is not to approximate the marginal inference $p(h_t|v_{1:T})$, but rather the joint distribution $p(h_{1:T}|v_{1:T})$. This puts them at a disadvantage when compared to other methods that more directly approximate the marginal [18]. In this work, we consider only the case where s_t has a tractably small number of states, and therefore we will not consider them further in this article.

Generalised Pseudo Bayes2 (GPB2) [3, 12, 15] is a popular approximation method for smoothed inference. In order to form a tractable recursion for the smoothed switch variables, the approximation $p(s_t|s_{t+1}, v_{1:T}) \approx p(s_t|s_{t+1}, v_{1:t})$ is used, see fig(2). This corresponds to a potentially severe loss of future information and, in general, GPB2 cannot be expected to improve much on ADF.

Some of the most popular approaches to filtering and smoothing are based on sequential Monte Carlo [8]. Whilst potentially powerful, these non-analytic methods typically suffer in high-dimensional latent/hidden spaces since they are often based on naive importance sampling, which restricts their practical use. Implementations of Rao-Blackwellisation (see for example [7]) may not help in difficult problems where the continuous posterior is highly non-Gaussian, and we are unaware of methods that have addressed this.

2 Expectation Correction

Our aim is to introduce a method for computing the smoothed estimate $p(h_t|v_{1:T})$ that is numerically stable and extendable. Essentially, we replace the λ message with a recursion that works directly on distributions. The approach is analogous to the Rauch-Tung-Striebel 'correction' smoother for Kalman Filters [17], although its application in the Hybrid framework requires some care. Our approach will be essentially a Gaussian sum approximation for a single forward and backward pass. We will show that this gives similar performance to competing methods for inference in SKFs when a single Gaussian is used in the approximation. More importantly, we will show how to extend our method to use mixture representations which can lead to dramatic improvements in difficult cases where the posterior is strongly multimodal. We'll do this in some generality, and then in section (3) apply this to the special

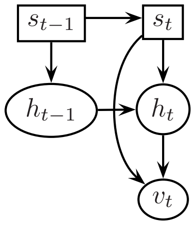


Figure 3: Structure of the forward pass. Essentially, the forward pass defines a 'prior' distribution at time $t-1$ which contains all the information from the variables $v_{1:t-1}$.

case of SKFs. In this general derivation, we'll assume that the forward propagation of posterior estimates from one time step to the next is unproblematic. First we'll describe the forward pass, since this is required to be precomputed for our backpass.

2.1 Forward Pass (Filtering)

Our approach for the forward pass is fairly standard, and is essentially Assumed Density Filtering. The forward pass is a distribution $p(s_t, h_t|v_{1:t})$ and it is convenient to write this in the form $p(s_t, h_t|v_{1:t}) = p(h_t|s_t, v_{1:t})p(s_t|v_{1:t})$. Then, rather than using the single forward recursion, equation (9), we can form a separate recursion for $p(s_t|v_{1:t})$ and $p(h_t|s_t, v_{1:t})$.

A recursion for $p(h_t|s_t, v_{1:t})$

This can be obtained as follows:

$$\begin{aligned} p(h_t|s_t, v_{1:t}) &= \sum_{s_{t-1}} p(h_t, s_{t-1}|s_t, v_{1:t}) \\ &\propto \sum_{s_{t-1}} p(h_t|s_{t-1}, s_t, v_{1:t})p(v_t|s_{t-1}, s_t, v_{1:t-1})p(s_t|s_{t-1})p(s_{t-1}|v_{1:t-1}) \end{aligned} \tag{11}$$

The factor $p(s_t|s_{t-1})$ is given by the model, and $p(s_{t-1}|v_{1:t-1})$ comes from recursion at the previous timestep. The term $p(h_t|s_{t-1}, s_t, v_{1:t})$ can be found using the dynamics as follows

$$p(h_t|s_{t-1}, s_t, v_{1:t}) \propto p(h_t, s_{t-1}, s_t, v_{1:t}) \propto p(h_t, v_t|s_{t-1}, s_t, v_{1:t-1}) \tag{12}$$

We can find the joint distribution $p(h_t, v_t|s_{t-1}, s_t, v_{1:t-1})$, and then condition on v_t to easily find the distribution $p(h_t|s_{t-1}, s_t, v_{1:t})$. Similarly, the factor $p(v_t|s_{t-1}, s_t, v_{1:t-1})$ in equation (11) is straightforward since this corresponds to a single forward iteration of the dynamics with known switch states, integrated over all h_{t-1}, h_t. We assume that this, in general, would cause little difficulty.

Using the above results, we are now in a position to calculate equation (11). For each setting of the variable s_t, we will therefore have a mixture of S distributions $p(h_t|s_t, s_{t-1}, v_{1:t})$, with a suitable mixture coefficient. Since, with time, this will entail an exponential growth S^t of mixture components, we need to approximate the mixture $p(h_t|s_t, v_{1:t})$ with a simpler representation $q(h_t|s_t, v_{1:t})$. Perhaps the simplest idea is to project/collapse each mixture back to a single component at each timestep. Many different collapse methods can be envisaged. Arguably the most natural is to impose that that the sufficient statistics of the projection $q(h_t|s_t, v_{1:t})$ match those of $p(h_t|s_t, v_{1:t})$. For example, if $p(h_t|s_t, v_{1:t})$ is a mixture of Gaussians, and $q(h_t|s_t, v_{1:t})$ is a single Gaussian, it would be reasonable that the mean and covariance of q should be set to the mean and covariance of the mixture of Gaussians – this is straightforward to do, as shown in appendix (B).

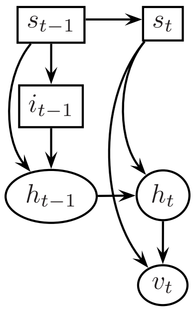


Figure 4: Structure of the mixture representation of the forward pass. Essentially, the forward pass defines a 'prior' distribution at time $t-1$ which contains all the information from the variables $v_{1:t-1}$.

A recursion for $p(s_t|v_{1:t})$

$$p(s_t|v_{1:t}) \propto \sum_{s_{t-1}} p(s_t, s_{t-1}, v_t, v_{1:t-1}) \qquad = \sum_{s_{t-1}} p(v_t|s_t, s_{t-1}, v_{1:t-1}) p(s_t|s_{t-1}) p(s_{t-1}|v_{1:t-1}) \tag{13}$$

The factor $p(v_t|s_t, s_{t-1}, v_{1:t-1})$ is straightforward to calculate, since this just requires forward propagation with known switch states. The factor $p(s_t|s_{t-1})$ is trivial, whilst the factor $p(s_{t-1}|v_{1:t-1})$ comes from the previous timestep.

2.2 Forward Pass : mixture representation

Here we extend the forward pass so that the collapse has, for each state s_t, not just a single component, but a set of I components.

$$q(h_t|s_t, v_{1:t}) = \sum_{i_t} p(h_t|i_t, s_t, v_{1:t}) p(i_t|s_t, v_{1:t}) \tag{14}$$

We use $i_t \in 1, \ldots, I$ to represent the mixture component[1].

A recursion for $p(h_t|s_t, v_{1:t})$

As in the single component case, our strategy will be to find first $p(h_t|s_t, v_{1:t})$. We will assume that the mixture coefficients $p(i_{t-1}|s_{t-1}, v_{1:t-1})$ have been given to us from a previous timestep. We will address how to set these for the current time step $p(i_t|s_t, v_{1:t})$ in due course. We may then proceed as follows:

$$p(h_t|s_t, v_{1:t}) \propto \sum_{i_{t-1}, s_{t-1}} p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t}) p(v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1}) \\ p(s_t|s_{t-1}) p(i_{t-1}|s_{t-1}, v_{1:t-1}) p(s_{t-1}|v_{1:t-1}) \tag{15}$$

The term $p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t})$ can be found from

$$p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t}) \propto p(h_t, v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1}) \tag{16}$$

The right hand side of the above equation is easy to find since it corresponds to a single forward propagation from the previous filtered state. Then conditioning the joint distribution $p(h_t, v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1})$

[1] In our code we include the possibility of using a time dependent number of components, I_t, since there may be some regions where a smaller or larger number is useful. Also, a little care is required at the beginning of the chain since at time $t = 1$, the exact filtered estimate $p(h_1|v_1, s_1)$ is not a mixture, and in general we require $I_t \le S \times I_{t-1}$.

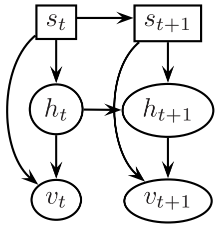


Figure 5: Structure of the backward pass.

on v_t gives $p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t})$. Using the above results, we are now in a position to calculate equation (15). For each setting of the variable s_t, we will therefore have a mixture of $I \times S$ distributions $p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t})$ which we can collapse back to a mixture of S distributions, which defines the mixture weights $p(i_t|s_t, v_{1:t})$. How to collapse a mixture to another mixture is partly a matter of taste. For computational expediency, we recommend either a simple merging of components that have low weight, or retention of the largest components. For the SKF case we describe explicitly the method we use in appendix section (B.1).

A recursion for $p(s_t|v_{1:t})$

By analogy with the single component case, we have:

$$
\begin{aligned}
p(s_t|v_{1:t}) &\propto \sum_{i_{t-1}, s_{t-1}} p(s_t, i_{t-1}, s_{t-1}, v_t, v_{1:t-1}) \\
&= \sum_{s_{t-1}} p(v_t|s_t, s_{t-1}, v_{1:t-1}) p(s_t|s_{t-1}) p(i_{t-1}|s_{t-1}, v_{1:t-1}) p(s_{t-1}|v_{1:t-1}) \qquad (17)
\end{aligned}
$$

where all factors in the final expression are known.

2.3 Backpass

In the following, we describe a general ‘correction’ smoother. This will consist of ‘correcting’ the filtered expected (or marginal) estimates $p(s_t, h_t|v_{1:t})$ obtained from the Forward Pass into smoothed estimates $p(s_t, h_t|v_{1:T})$. Let’s try to write a backward recursion for the (smoothed) posteriors, in a way analogous to the Rauch-Tung-Striebel (RTS) correction method for SKFs [17].

$$
\begin{aligned}
p(h_t, s_t|v_{1:T}) &\propto \sum_{s_{t+1}} \int_{h_{t+1}} p(h_t, s_t|h_{t+1}, s_{t+1}, v_{1:T}) p(h_{t+1}, s_{t+1}, v_{1:T}) \\
&\propto \sum_{s_{t+1}} \int_{h_{t+1}} p(h_t, s_t|h_{t+1}, s_{t+1}, v_{1:t}) p(h_{t+1}, s_{t+1}|v_{1:T}) \qquad (18)
\end{aligned}
$$

The first factor may be written

$$
p(h_t, s_t|h_{t+1}, s_{t+1}, v_{1:t}) \propto p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t}) p(s_t|h_{t+1}, s_{t+1}, v_{1:t}) \qquad (19)
$$

Using the above formula, we can write the backward recursion as

$$
p(h_t, s_t|v_{1:T}) = \sum_{s_{t+1}} \int_{h_{t+1}} p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t}) p(s_t|h_{t+1}, s_{t+1}, v_{1:t}) p(h_{t+1}, s_{t+1}|v_{1:T}) \qquad (20)
$$

Formally, this is sufficient to define a backwards recursion directly for the smoothed estimate $p(h_t, s_t|v_{1:T})$.

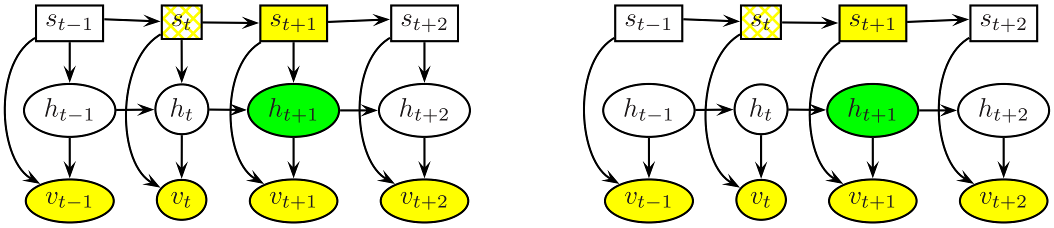


Figure 6: Backpass approximation. Left: Our approximation replaces $p(h_{t+1}|s_{t+1}, s_t, v_{1:T})$ by $p(h_{t+1}|s_{t+1}, v_{1:T})$. Motivation for this is that s_t only influences h_{t+1} through h_t. However, h_t will most likely be heavily influenced by $v_{1:t}$, so that not knowing the state of s_t is likely to be of secondary importance. Right : In the case that the switches affect only the observations, and not the dynamics, the 'approximation' is exact since, given the other evidence, s_t has no influence on h_{t+1}. The green (darker) node is variable we wish to find the posterior state of. The yellow(lighter shaded) nodes are variables in known states, and the hashed nodes are variables whose states are indeed known, but assumed unknown for the approximation.

However, it's clear that, in general, the representation $p(h_t, s_t|v_{1:T})$ will contain an exponential number of mixture components since the number of mixtures increases by a factor S at each iteration. Furthermore, the integral in this representation needs to be approximated, for which this form of the recursion is not particularly suited since the variable h_{t+1} is entwined in different places. A simple approximation would be to replace $p(s_t|h_{t+1}, s_{t+1}, v_{1:t})$ by $p(s_t|s_{t+1}, v_{1:t})$ and there may be some merit in this. However, we prefer to replace h_{t+1} by some kind of average value although, in the above form, it is not clear what the distribution of h_{t+1} is, and therefore what average value it should take. For these reasons, we prefer to find an alternative where we have more confidence that we can replace problematic terms with reasonable approximations. Consider therefore

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

This form of the recursion is potentially more useful since it is clearly a mixture of the distributions $p(h_t|s_t, s_{t+1}, v_{1:T})$ with an associated set of mixture weights $p(s_{t+1}|s_t, v_{1:T})$. Usually, both the distributions $p(h_t|s_t, s_{t+1}, v_{1:T})$ and weights $p(s_{t+1}|s_t, v_{1:T})$ will be difficult to obtain exactly. We'll consider both of these terms now separately.

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

We can write $p(h_t|s_t, s_{t+1}, v_{1:T})$ as the marginal of the joint distribution $p(h_t, h_{t+1}|s_t, s_{t+1}, v_{1:T})$. This joint distribution is somewhat difficult to find exactly, and we seek an approximation that is in keeping with the RTS spirit. This motivates the following factorisation

$$\begin{aligned} p(h_t, h_{t+1}|s_t, s_{t+1}, v_{1:T}) &= p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:T}) p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \\ &= p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t}) p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \end{aligned} \tag{22}$$

The first factor in equation (22) may be found from considering the joint distribution

$$p(h_t, h_{t+1}|s_t, s_{t+1}, v_{1:t}) = p(h_{t+1}|h_t, s_t, s_{t+1}, v_{1:t}) p(h_t|s_t, v_{1:t}) \tag{23}$$

which itself can be found from a simple forward dynamics from the filtered estimate $p(h_t|s_t, v_{1:t})$. Then conditioning equation (23) to find $p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})$ effectively constitutes a reversal of the forward dynamics.

The second factor $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$ in equation (22) may cause some difficulty, and is depicted in fig(6). When the switch variables affect only the observations and not the dynamics, then $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \equiv p(h_{t+1}|s_{t+1}, v_{1:T})$. Since we know $p(h_{t+1}|s_{t+1}, v_{1:T})$ from the previous smoothed estimate, for the case of a *switching observation model*, no additional approximations are required. Otherwise, we make the simple approximation $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$. Other approximations may also be suitable, but we have found that, at least in the context of SKFs, this often produces a reasonable approximation. Compared with the GPB2 method, fig(2), the dropping of this dependence is rather delicate and, reiterating, introduces *no* approximation in the case of switching observation models.

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

We also need to consider

$$p(s_t|h_{t+1}, s_{t+1}, v_{1:t}) = \frac{p(h_{t+1}|s_{t+1}, s_t, v_{1:t})p(s_t|s_{t+1}, v_{1:t})}{\sum_{s'_t} p(h_{t+1}|s_{t+1}, s'_t, v_{1:t})p(s'_t|s_{t+1}, v_{1:t})} \tag{24}$$

The term $p(h_{t+1}|s_{t+1}, s_t, v_{1:t})$ is readily found by marginalising equation (23). The only term we haven't discussed is

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

which is straightforward.

2.3.3 Approximating the Average $\langle p(s_t|h_{t+1}, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|s_{t+1}, v_{1:T})}$

We now have all the ingredients required to look at the integral in equation (21), namely

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

which we recognise as the average $\langle p(s_t|h_{t+1}, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|s_{t+1}, v_{1:T})}$. Perhaps the simplest approximation is to replace h_{t+1} by it's mean value. Replacing h_{t+1} with $\langle h_{t+1}|s_{t+1}, v_{1:T}\rangle$ means that the integral is approximated by

$$\langle p(s_t|h_{t+1}, s_{t+1}, v_{1:t})\rangle \approx \frac{1}{Z} p(h_{t+1} = \langle h_{t+1}|s_{t+1}, v_{1:T}\rangle \,|s_{t+1}, s_t, v_{1:t})p(s_t|s_{t+1}, v_{1:t}) \tag{25}$$

where Z is a constant to ensure normalization over s_t.

More sophisticated approximations of the average would correspond to the 'correction' of higher order moments. Fluctuation expansions $h_{t+1} \approx \langle h_{t+1}|s_{t+1}, v_{1:T}\rangle + \eta_{t+1}$ spring to mind as an obvious extension.

Forming the Recursion

It is useful to put 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

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

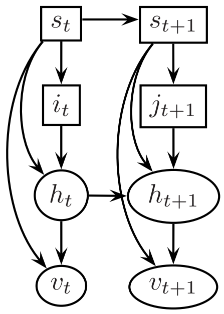


Figure 7: Structure of the backward pass for mixtures.

which may be collapsed to a single distribution using a standard approach.

The term $p(s_t|v_{1:T})$ is given by (using our approximation for the average)

$$p(s_t|v_{1:T}) = \sum_{s_{t+1}} p(s_{t+1}|v_{1:T}) \frac{p(h_{t+1} = \langle h_{t+1}\rangle |s_{t+1}, s_t, v_{1:t})p(s_t|s_{t+1}, v_{1:t})}{\sum_{s'_t} p(h_{t+1} = \langle h_{t+1}\rangle |s_{t+1}, s'_t, v_{1:t})p(s'_t|s_{t+1}, v_{1:t})} \quad (26)$$

In many applications it is quite likely that a single component collapse would be sufficient since the forward pass $p(h_t|s_t, v_{1:t})$, which *is* a mixture, is simply being corrected to form the smoothed estimate. Indeed, in most cases, the smoothed estimate is indeed 'smoother', so that we might not anticipate much need for a backpass using mixtures. Nevertheless, we describe below how to do this in the contingency that such an extension might be useful, depending on the application at hand.

2.3.4 Backward pass : mixture representation

Here we show how to collapse $p(h_t|s_t, v_{1:T})$ to a mixture $\sum_{j_t} p(j_t|s_t, v_{1:T})p(h_t|j_t, s_t, v_{1:T})$. This will make use of the mixture representation of our forward messages. Analogously to the case with a single component, we can write

$$\begin{aligned} p(h_t, s_t|v_{1:T}) = &\sum_{i_t, j_{t+1}, s_{t+1}} p(s_{t+1}|v_{1:T})p(j_{t+1}|s_{t+1}, v_{1:T})p(h_t|j_{t+1}, s_{t+1}, i_t, s_t, v_{1:T}) \\ &\times \int_{h_{t+1}} p(i_t, s_t|h_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}|j_{t+1}, s_{t+1}, v_{1:T}) \end{aligned} \quad (27)$$

As in the case of a single component, we need to approximate the average in the last line of the above equation. As before, a simple minded approximation is to replace the average of $p(i_t, s_t|h_{t+1}, s_{t+1}, v_{1:t})$ over h_{t+1} with $p(i_t, s_t|h_{t+1}, s_{t+1}, v_{1:t})$ evaluated with h_{t+1} set to its average value $\langle h_{t+1}|j_{t+1}, s_{t+1}, v_{1:T}\rangle$. Again more sophisticated approximations may readily be considered. In the above,

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

where $p(h_{t+1}|i_t, s_t, s_{t+1}, v_{1:t})$ is found from marginalising the joint distribution

$$p(h_{t+1}, h_t|s_{t+1}, i_t, s_t, v_{1:t}) = p(h_{t+1}|h_t, s_{t+1})p(h_t|i_t, s_t, v_{1:t}) \quad (28)$$

Again, we need to approximate $p(h_t|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$. We can use the same method as before by considering this as the marginal of the joint distribution

$$\begin{aligned} p(h_t, h_{t+1}|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}) &= p(h_t|h_{t+1}, i_t, s_t, j_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}) \\ &\approx p(h_t|h_{t+1}, i_t, s_t, j_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}|j_{t+1}, s_{t+1}, v_{1:T}) \end{aligned} \quad (29)$$

Integrating equation (27) over h_t, we have

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

where $\overline{h_{t+1}} \equiv \langle h_{t+1}|j_{t+1}, s_{t+1}, v_{1:T}\rangle$. Using the above, we can form the distribution

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

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

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

3 Switching Linear State Space models

We apply the previous general framework to the linear switching model (Switching Kalman Filter) of section (3). Our method will then consist of a standard forward pass using a Gaussian Sum approximation to calculate the filtered estimate, following by a correction backpass to form the smoothed estimate. In the following, for notational clarity, we derive the associated recursions for the case that the mean of the Gaussian noise is zero, although their inclusion is given in the algorithm in the appendices.

Here we will just derive Expectation Correction using only a *single* Gaussian for both the forward and backward passes. The straightforward extension to the mixture case is given in appendix (D), with corresponding pseudo-code and the required initialisations for the recursions. We also give in appendix (C) the likelihood $p(v_{1:T})$ approximation using mixtures of Gaussians.

3.1 The Forward Pass

The forward pass is a distribution $p(s_t, h_t|v_{1:t})$. It is convenient to write this is the form

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

where the continuous message will be approximated by a Gaussian with mean $f_t(s_t)$ and covariance $F_t(s_t)$. The discrete message $p(s_t|v_{1:t})$ will be written as $r_t(s_t)$. Expectation Correction will therefore produce a recursion for f_t, F_t and r_t.

A recursion for $p(h_t|s_t, v_{1:t})$

We will use the approach outlined in section (2.1), in which we will first find equation (12). That is, we find the joint distribution $p(h_t, v_t|s_{t-1}, s_t, v_{1:t-1})$, and condition on v_t to find the distribution $p(h_t|s_{t-1}, s_t, v_{1:t})$. The joint distribution $p(h_t, v_t|s_{t-1}, s_t, v_{1:t-1})$ is easily evaluated by realising that for each setting of the switch variables s_{t-1}, s_t the distribution over h_t, v_t is jointly Gaussian. In the sequel we use $\langle \cdot|c\rangle$ to denote averages conditional on the switch states expressed by c. Δ denotes a fluctuation, namely the deviation from the average, $\Delta x \equiv x - \langle x\rangle$. Hence the covariance matrix between v_t and h_t, knowing the switch states s_t, s_{t-1} is denoted $\left\langle \Delta v_t \Delta h_t^T|s_t, s_{t-1}\right\rangle$. The means and covariances are easily found from the relations

$$v_t = B(s_t)h_t + \eta^v(s_t), \qquad h_t = A(s_t)h_{t-1} + \eta^h(s_t)$$

Using the above, the covariance elements of the joint distribution are given by

$$\left\langle \Delta v_t \Delta v_t^T|s_t, s_{t-1}\right\rangle = B(s_t)\left\langle \Delta h_t \Delta h_t^T|s_t, s_{t-1}\right\rangle B^T(s_t) + \Sigma^v(s_t)$$

$$\left\langle \Delta h_t \Delta h_t^T | s_t, s_{t-1} \right\rangle = A(s_t) \left\langle \Delta h_{t-1} \Delta h_{t-1}^T | s_{t-1} \right\rangle A^T(s_t) + \Sigma^h(s_t) \tag{30}$$
$$\left\langle \Delta v_t \Delta h_t^T | s_t, s_{t-1} \right\rangle = B(s_t) \left\langle \Delta h_t \Delta h_t^T | s_t, s_{t-1} \right\rangle$$

whilst the means of the two variables are given by

$$\langle v_t | s_t, s_{t-1} \rangle = B(s_t) A(s_t) \langle h_{t-1} | s_{t-1} \rangle, \qquad \langle h_t | s_t, s_{t-1} \rangle = A(s_t) \langle h_{t-1} | s_{t-1} \rangle$$

In the above, using our moment representation of the forward messages

$$\langle h_{t-1} | s_{t-1} \rangle \equiv f_{t-1}(s_{t-1}), \qquad \left\langle \Delta h_{t-1} \Delta h_{t-1}^T | s_{t-1} \right\rangle \equiv F_{t-1}(s_{t-1})$$

Using the above results, we are now in a position to calculate equation (11). We can find $p(h_t|s_{t-1}, s_t, v_{1:t})$ by conditioning the joint Gaussian, using the results in the appendix (B). Similarly, $p(v_t|s_{t-1}, s_t, v_{1:t-1})$ is a Gaussian with mean and covariance given by $\langle v_t|s_t, s_{t-1}\rangle$ and $\left\langle \Delta v_t \Delta v_t^T | s_t, s_{t-1} \right\rangle$ above. For each setting of the variable s_t, we will therefore have a mixture of S Gaussians, which can easily be collapsed to a single Gaussian using the results in the appendix (B).

Calculating the filtered estimate $p(s_t|v_{1:t})$

$$p(s_t|v_{1:t}) \propto \sum_{s_{t-1}} p(s_t, s_{t-1}, v_t, v_{1:t-1}) \quad = \sum_{s_{t-1}} p(v_t|s_t, s_{t-1}, v_{1:t-1}) p(s_t|s_{t-1}) p(s_{t-1}|v_{1:t-1}) \tag{31}$$

The factor $p(v_t|s_t, s_{t-1}, v_{1:t-1})$ is straightforward to calculate, since this is just a mixture of Gaussians. The factor $p(s_t|s_{t-1})$ is trivial, whilst the factor $p(s_{t-1}|v_{1:t-1})$ comes from the previous timestep.

3.2 The Backward Pass

The backpass will directly yield an estimate of the smoothed posterior. Using a single Gaussian, our approximation to $p(h_t|s_t, v_{1:T})$ will have mean $g_t(s_t)$ and covariance $G_t(s_t)$. Expectation Correction will therefore produce a recursion for g_t, G_t and $l_t \equiv p(s_t|v_{1:T})$. The reader is directed to appendix (D) for details of the mixture of Gaussians calculation.

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

From section (2.3.1), we need $p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:T})$ and $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$ which can be found from the joint distribution $p(h_t, h_{t+1}|s_t, s_{t+1}, v_{1:T})$. Following the strategy presented in section (2.3.1), first we find the distribution $p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t})$. This is given by conditioning the joint distribution

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

which is used to define the backward equation

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

where $\overleftarrow{A}$ and $\overleftarrow{m}$ and $\overleftarrow{\eta}(s_t, s_{t+1}) \sim \mathcal{N}(0, \overleftarrow{\Sigma}_t(s_t, s_{t+1}))$ are easily found using the conditioned Gaussian results in appendix (A). Then the joint distribution $p(h_t, h_{t+1}|s_t, s_{t+1}, v_{1:T}) = p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t}) p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$ has the following mean and covariance

$$\langle h_t|s_t, s_{t+1}, v_{1:T}\rangle = \overleftarrow{A}(s_t, s_{t+1}) g_{t+1}(s_{t+1}) + \overleftarrow{m}(s_t, s_{t+1}) \tag{32}$$
$$\langle h_{t+1}|s_t, s_{t+1}, v_{1:T}\rangle = g_{t+1}(s_{t+1})$$
$$\left\langle \Delta h_{t+1} \Delta h_{t+1}^T | s_t, s_{t+1}, v_{1:T} \right\rangle = G_{t+1}(s_{t+1})$$
$$\left\langle \Delta h_t \Delta h_t^T | s_t, s_{t+1}, v_{1:T} \right\rangle = \overleftarrow{A}(s_t, s_{t+1}) G_{t+1}(s_{t+1}) \overleftarrow{A}^T(s_t, s_{t+1}) + \overleftarrow{\Sigma}_t(s_t, s_{t+1})$$
$$\left\langle \Delta h_t \Delta h_{t+1}^T | s_t, s_{t+1}, v_{1:T} \right\rangle = \overleftarrow{A}(s_t, s_{t+1}) G_{t+1}(s_{t+1})$$

From this, we can find easily the marginal $p(h_t|s_t, s_{t+1}, v_{1:T})$.

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

In the approximation, equation (25), we replace the average with respect to h_{t+1} by the integrand evaluated with h_{t+1} set to its mean value $g_{t+1}(s_{t+1})$. (A more refined approximation would also make use of the covariance $G_{t+1}(s_{t+1})$. Replacing h_{t+1} with $\langle h_{t+1}|s_{t+1}, v_{1:T}\rangle$ means that the integral is approximated by

$$\frac{1}{Z}\frac{e^{-\frac{1}{2}z^T_{t+1}(s_t,s_{t+1})\Sigma^{-1}(s_t,s_{t+1}|v_{1:t})z_{t+1}(s_t,s_{t+1})}}{\sqrt{\det \Sigma(s_t, s_{t+1}|v_{1:t})}}p(s_t|s_{t+1}, v_{1:t})$$

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 is a constant to ensure normalistion over s_t. The covariance $\Sigma(s_t, s_{t+1}|v_{1:t}) \equiv \langle \Delta h_{t+1}\Delta h_{t+1}|s_t, s_{t+1}, v_{1:t}\rangle$ is given by time shifting equation (30). Whereas as in Expectation Propagation we divide potentials (which corresponds to subtracting the canonical parameters), here we subtract *moments*, if only the first moment in this simple approximation. More complex approximations using fluctuation expansions or Gaussian Field approximations [5] immediately spring to mind. Variational approximations would likely prove too expensive.

The final mixture representation of the smoothed posterior can then be collapsed to a single Gaussian using the usual approach. The extension to the mixture collapse is given in appendix (D).

4 Expectation Propagation for SKF: the Auxiliary variable trick

Following along the same lines as Belief Propagation, we denote the Filtered posterior $p(h_t, s_t|v_{1:t})$ by $\rho(h_t, s_t)$ (up to a neglectable proportionality constant), which represents the state of h_t, s_t given all past observations $v_{1:t}$ up to the present t. Similarly, we all future information about h_t, s_t, is is contained in the quantity $p(v_{t+1:T}|h_t, s_t)$ which is denoted by $\lambda(h_t, s_t)$. Without loss of generality, we may factorise these functions as $\rho(h_t, s_t) \equiv \rho(h_t|s_t)\rho(s_t)$, and $\lambda(h_t, s_t) \equiv \lambda(h_t|s_t)\lambda(s_t)$

The two–times potential $p(h_{t-1}, s_{t-1}, h_t, s_t \mid v_{1:T})$ is then proportional to:

$$\rho_{t-2,t-1}(h_{t-1}, s_{t-1})p(v_t \mid h_t, s_t)p(h_t \mid h_{t-1}, s_t)p(s_t \mid s_{t-1})\lambda_{t+1,t}(h_t, s_t)$$

Hence

$$\begin{aligned}&p(h_{t-1}, h_t|s_{t-1}, s_t, v_{1:T})p(s_{t-1}, s_t|v_{1:T})\\ &\propto \rho(h_{t-1}|s_{t-1})p(v_t|h_t, s_t)p(h_t|h_{t-1}, s_t)\lambda(h_t|s_t)\rho(s_{t-1})p(s_t|s_{t-1})\lambda(s_t)\end{aligned}$$

Integrating the above, we find

$$\begin{aligned}&p(s_{t-1}, s_t|v_{1:T})\\ &\propto \rho(s_{t-1})p(s_t|s_{t-1})\lambda(s_t)\int_{h_{t-1},h_t}\rho(h_{t-1}|s_{t-1})p(v_t|h_t, s_t)p(h_t|h_{t-1}, s_t)\lambda(h_t|s_t)\end{aligned}$$

To find $p(h_{t-1}, h_t|s_{t-1}, s_t, v_{1:T})$, we only need to take into account those terms in the joint two time potential that depend on h_{t-1} and h_t for fixed s_t, s_{t-1}. Also, we know that this will be a Gaussian distribution. Therefore, we search for the Gaussian

$$p(h_{t-1}, h_t|s_{t-1}, s_t, v_{1:T}) \propto \rho(h_{t-1}|s_{t-1})p(v_t|h_t, s_t)p(h_t|h_{t-1}, s_t)\lambda(h_t|s_t)$$

Once we have found $p(h_{t-1}, h_t|s_{t-1}, s_t, v_{1:T})$ and $p(s_{t-1}, s_t|v_{1:T})$, we can use them do define the Belief Propagation recursions:

Forward Pass:

$$\begin{aligned}
\rho(h_t, s_t) &\propto \frac{\sum_{s_{t-1}} \int_{h_{t-1}} p(h_{t-1}, s_{t-1}, h_t, s_t|v_{1:T})}{\lambda(h_t, s_t)} \\
&\propto \frac{\sum_{s_{t-1}} p(s_{t-1}, s_t \mid v_{1:T}) \int_{h_{t-1}} p(h_{t-1}, h_t|s_{t-1}, s_t, v_{1:T})}{\lambda(s_t)\lambda(h_t|s_t)} \\
&\propto \frac{\sum_{s_{t-1}} p(s_{t-1}, s_t \mid v_{1:T}) p(h_t|s_{t-1}, s_t, v_{1:T})}{\lambda(s_t)\lambda(h_t|s_t)}
\end{aligned}$$

Backward Pass:

$$\begin{aligned}
\lambda(h_{t-1}, s_{t-1}) &\propto \frac{\sum_{s_t} \int_{h_t} p(h_{t-1}, s_{t-1}, h_t, s_t|v_{1:T})}{\rho(h_{t-1}, s_{t-1})} \\
&\propto \frac{\sum_{s_t} p(s_{t-1}, s_t|v_{1:T}) \int_{h_t} p(h_{t-1}, h_t|s_{t-1}, s_t, v_{1:T})}{\rho(s_{t-1})\rho(h_{t-1}|s_{t-1})} \\
&\propto \frac{\sum_{s_t} p(s_{t-1}, s_t|v_{1:T}) p(h_{t-1}|s_{t-1}, s_t, v_{1:T})}{\rho(s_{t-1})\rho(h_{t-1}|s_{t-1})}
\end{aligned}$$

We will use the following parameterisations of the messages:

$$\rho_{t-1,t}(h_t \mid s_t) \propto \frac{1}{\left|\tilde{F}(s_t)\right|^{1/2}} e^{-\frac{1}{2}(h_t - \tilde{f}(s_t))^T \tilde{F}^{-1}(s_t)(h_t - \tilde{f}(s_t))}$$

$$\lambda_{t+1,t}(h_t \mid s_t) = e^{-\frac{1}{2}\left(h_t G(s_t) h_t - 2h^t G(s_t) g_t(s_t)\right)}$$

The reader should bear in mind that where previous we used G to parameterise the smoothed estimate in Expectation Correction, here in Expectation Propagation, we use G to parameterise the λ message only. The smoothed posterior is given then by the product of the associated λ and ρ messages. In general, our notation in Expectation Propagation is to used tilded parameters such as $\tilde{F}$ to denote the moment representation, and untilded messages to denote the canonical representation, as will become clearer throughout the derivation.

4.0.1 The Auxiliary variable Trick

The auxiliary variable trick [1] is useful for deriving recursions involving λ messages simply, and avoids the explicit appearance of inverse noise covariance parameters, which improves numerical stability. Our use of the trick here is slightly different to that presented in [1] since here more care is needed with proportionality terms that in the simpler Kalman Filter may be neglected.

The potential $\lambda(h_t \mid s_t)$ can be expressed as a probability distribution of an auxiliary variable a_t which represents the amount of information coming from the future. If a_t has the following probability distribution:

$$p(a_t|h_t, s_t) = N(h_t, cov = G^{-1}(s_t)) \propto |G(s_t)|^{1/2} e^{-\frac{1}{2}(a_t - h_t)^T G_t(s_t)(a_t - h_t)}$$

then

$$\lambda(h_t|s_t) = \frac{e^{\frac{1}{2} g^T(s_t) G(s_t) g(s_t)}}{|G(s_t)|^{\frac{1}{2}}} p(a_t|h_t, s_t)|_{a_t = g(s_t)}$$

Note that the prefactor only plays a role in the discrete variable case $p(s_{t-1}, s_t|v_{1:T})$ and does not affect $p(h_{t-1}, h_t|s_{t-1}, s_t, v_{1:T})$.

4.1 Forward Pass

$$p(h_t|s_{t-1}, s_t, v_{1:T}) \propto p(h_t, v_t, a_t|s_{t-1}, s_t, v_{1:t-1})|_{a_t=g(s_t)}$$

where we use the auxiliary variable to represent information coming from the future. This former expression is proportional to:

$$\exp\left\{-\frac{1}{2}\begin{pmatrix} h_t-\langle h_t\rangle \\ v_t-\langle v_t\rangle \\ a_t-\langle a_t\rangle \end{pmatrix}^T \begin{pmatrix} C_{hh} & C_{hv} & C_{ha} \\ C_{hv}^T & C_{vv} & C_{va} \\ C_{ha}^T & C_{va}^T & C_{aa} \end{pmatrix}^{-1} \begin{pmatrix} h_t-\langle h_t\rangle \\ v_t-\langle v_t\rangle \\ a_t-\langle a_t\rangle \end{pmatrix}\right\}$$

$$\langle h_t\rangle = A(s_t)\tilde{f}(s_{t-1})$$
$$\langle v_t\rangle = \langle B(s_t)h_t + \eta_v(s_t)\rangle = B(s_t)A(s_t)\tilde{f}(s_{t-1})$$
$$\langle a_t\rangle = \langle h_t + \eta_a(s_t)\rangle = \langle h_t\rangle = A(s_t)\tilde{f}(s_{t-1})$$
$$C_{hh} = \langle \Delta h_t \Delta h_t^T\rangle = A(s_t)\tilde{F}(s_{t-1})A^T(s_t) + \Sigma_h(s_t)$$
$$C_{hv} = \langle \Delta h_t \Delta v_t^T\rangle = C_{hh}B^T(s_t)$$
$$C_{ha} = C_{hh}$$
$$C_{vv} = \langle \Delta v_t \Delta v_t^T\rangle = B(s_t)C_{hh}B^T(s_t) + \Sigma_v(s_t)$$
$$C_{va} = B(s_t)C_{hh}$$
$$C_{aa} = \langle \Delta a_t \Delta a_t^T\rangle = C_{hh} + G^{-1}(s_t)$$
$$C_{aa}^{-1} = \left(G(s_t)C_{hh} + I\right)^{-1} G(s_t)$$

After conditioning on a_t and v_t, we get:

$$p(h_t|s_{t-1}, s_t, v_{1:T}) = p(h_t|a_t, s_{t-1}, s_t, v_{1:t})|_{a(s_t)=g(s_t)} \propto e^{-\frac{1}{2}(h_t-\tilde{q}(s_{t-1},s_t))^T\tilde{Q}^{-1}(s_{t-1},s_t)(h_t-\tilde{q}(s_{t-1},s_t))}$$

with:

$$\tilde{Q}(s_{t-1}, s_t) = C_{hh} - \begin{pmatrix} C_{hv} \\ C_{ha} \end{pmatrix}^T \begin{pmatrix} C_{vv} & C_{va} \\ C_{va}^T & C_{aa} \end{pmatrix}^{-1} \begin{pmatrix} C_{hv}^T \\ C_{ha}^T \end{pmatrix} \tag{33}$$

$$\tilde{q}(s_{t-1}, s_t) = \langle h_t\rangle + \begin{pmatrix} C_{hv} \\ C_{ha} \end{pmatrix}^T \begin{pmatrix} C_{vv} & C_{va} \\ C_{va}^T & C_{aa} \end{pmatrix}^{-1} \begin{pmatrix} v_t-\langle v_t\rangle \\ g(s_t)-\langle a_t\rangle \end{pmatrix} \tag{34}$$

A difficulty with these expressions is that $G(s_t)$ is not formally invertible. To avoid this difficulty, we need to reexpress the equations using C_{aa}^{-1}, which is well defined. We can do this by using partitioned matrix inverse results:

$$\begin{pmatrix} C_{vv} & C_{va} \\ C_{va}^T & C_{aa} \end{pmatrix}^{-1} = \begin{pmatrix} D_{vv} & D_{va} \\ D_{va}^T & D_{aa} \end{pmatrix}$$

$$D_{aa} = \left(C_{aa} - C_{va}^T C_{vv}^{-1} C_{va}\right)^{-1} = \left(I - C_{aa}^{-1} C_{va}^T C_{vv}^{-1} C_{va}\right)^{-1} C_{aa}^{-1}$$
$$D_{va} = -C_{vv}^{-1} C_{va} D_{aa}$$
$$D_{vv} = C_{vv}^{-1} + C_{vv}^{-1} C_{va} D_{aa} C_{va}^T C_{vv}^{-1}$$

To ensure that the above does not explicitly contain formally invertible terms, we can rewrite the contribution to $\tilde{q}(s_{t-1}, s_t)$ as

$$\begin{pmatrix} C_{hv} \\ C_{ha} \end{pmatrix}^T \begin{pmatrix} C_{vv} & C_{va} \\ C_{va}^T & C_{aa} \end{pmatrix}^{-1} \begin{pmatrix} v_t-\langle v_t\rangle \\ g(s_t)-\langle a_t\rangle \end{pmatrix}$$

$$\begin{pmatrix} C_{hv} \\ C_{ha} \end{pmatrix}^T \begin{pmatrix} D_{vv}(v_t-\langle v_t\rangle) + D_{va}(g(s_t)-\langle a_t\rangle) \\ D_{av}(v_t-\langle v_t\rangle) + D_{aa}(g(s_t)-\langle a_t\rangle) \end{pmatrix} \tag{35}$$

Since D_{aa} has always a postfactor $G(s_t)$, and $D_{va} = -C_{vv}^{-1}C_{va}D_{aa}$, the awkward terms $g(s_t)$ will always be accompanied by a prefactor $G(s_t)$. We can therefore replace each instance of $G(s_t)g(s_t)$ in the above by a redefined term $\hat{g}(s_t)$. As we will see, we can find a recursion for $\hat{g}(s_t)$, but not $g(s_t)$ alone.

Computing $p(s_{t-1}, s_t|v_{1:T})$

$$
\begin{aligned}
&p(s_{t-1}, s_t|v_{1:T}) \propto \rho(s_{t-1})p(s_t|s_{t-1})\lambda(s_t)\int_{h_{t-1},h_t}\rho(h_{t-1}|s_{t-1})p(v_t|h_t,s_t)p(h_t|h_{t-1},s_t)\lambda(h_t|s_t)\\
&\propto \frac{\rho(s_{t-1})p(s_t|s_{t-1})\lambda(s_t)e^{\frac{1}{2}g^T(s_t)G(s_t)g(s_t)}}{|G(s_t)|^{\frac{1}{2}}}\int_{h_{t-1},h_t}\rho(h_{t-1}|s_{t-1})p(v_t|h_t,s_t)p(h_t|h_{t-1},s_t)p(a_t = g_t|h_t,s_t)\\
&\propto \frac{\rho(s_{t-1})p(s_t|s_{t-1})\lambda(s_t)e^{\frac{1}{2}g^T(s_t)G(s_t)g(s_t)}}{|G(s_t)|^{\frac{1}{2}}}p(v_t, a_t = g_t|s_{t-1}, s_t, v_{1:t-1})
\end{aligned}
$$

$p(v_t, a_t \,|\, s_{t-1}, s_t, v_{1:t-1})$ has the following form:

$$
\frac{1}{z(s_{t-1}, s_t)}\exp\left\{-\frac{1}{2}\begin{pmatrix} v_t - \langle v_t\rangle \\ a_t - \langle a_t\rangle\end{pmatrix}^T\begin{pmatrix} C_{vv} & C_{va} \\ C_{va}^T & C_{aa}\end{pmatrix}^{-1}\begin{pmatrix} v_t - \langle v_t\rangle \\ a_t - \langle a_t\rangle\end{pmatrix}\right\}
$$

and can also be written as:

$$
p(v_t, a_t \,|\, s_{t-1}, s_t, v_{1:t-1}) = p(v_t \,|\, a_t, s_{t-1}, s_t, v_{1:t-1})p(a_t \,|\, s_{t-1}, s_t, v_{1:t-1}) \tag{36}
$$

By conditioning on a_t, we get:

$$
p(v_t \,|\, a_t, s_{t-1}, s_t, v_{1:t-1}) \propto \frac{1}{\left|\Sigma\right|^{\frac{1}{2}}}e^{-\frac{1}{2}\left(v_t-\mu\right)^T\Sigma^{-1}\left(v_t-\mu\right)}
$$

with:

$$
\begin{aligned}
&\Sigma = C_{vv} - C_{va}C_{aa}^{-1}C_{va}^T\\
&\mu = \langle v_t\rangle + C_{va}C_{aa}^{-1}(g(s_t) - \langle a_t\rangle)
\end{aligned} \tag{37}
$$

Note that, whilst $g(s_t)$ occurs in the expression for $\mu(s_t)$, it always occurs in combination with $G(s_t)$, which arises from $C_{aa}^{-1}g(s_t)$ in equation (37). The other factor is given by

$$
p(a_t \,|\, s_{t-1}, s_t, v_{1:t-1}) \propto \frac{1}{\left|C_{aa}\right|^{1/2}}\exp\left\{\left(a_t - \langle a_t\rangle\right)^T C_{aa}^{-1}\left(a_t - \langle a_t\rangle\right)\right\}
$$

Which, when $a_t = g(s_t)$, is proportional to:

$$
\frac{\exp\left\{-\frac{1}{2}\left(g(s_t) - A(s_t)\tilde{f}(s_{t-1})\right)^T\left(G(s_t)C_{hh} + I\right)^{-1}G(s_t)\left(g(s_t) - A(s_t)\tilde{f}(s_{t-1})\right)\right\}}{\left|G(s_t)C_{hh} + I\right|^{\frac{1}{2}}|G(s_t)|^{-\frac{1}{2}}}
$$

To cut down on the length of the expressions, we denote $G_t \equiv G_t(s_t)$, and similarly for other quantities. Putting this all together, we get

$$
\begin{aligned}
&p(s_{t-1}, s_t|v_{1:T})\\
&\propto \frac{\rho(s_{t-1})p(s_t|s_{t-1})\lambda(s_t)e^{\frac{1}{2}g_t^T G_t g_t}}{|\Sigma_{t-1,t}|^{\frac{1}{2}}|G_t C_{hh} + I|^{\frac{1}{2}}}e^{-\frac{1}{2}\left((v_t-\mu_t)^T\Sigma_{t-1,t}^{-1}(v_t-\mu_t)+(g_t-A_t\tilde{f}_{t-1})^T(G_tC_{hh}+I)^{-1}G_t(g_t-A_t\tilde{f}_{t-1})\right)}\\
&\propto \frac{\rho(s_{t-1})p(s_t|s_{t-1})\lambda(s_t)}{|\Sigma|^{\frac{1}{2}}|L_t|^{\frac{1}{2}}}e^{-\frac{1}{2}(v_t-\mu)^T\Sigma^{-1}(v_t-\mu)+\frac{1}{2}\hat{g}_t^T C_{hh}L_t^{-1}\hat{g}_t-\frac{1}{2}\tilde{f}_{t-1}^T A_t^T L_t^{-1}G_tA_t\tilde{f}_{t-1}+\tilde{f}_{t-1}A_t^T L_t^{-1}\hat{g}_t}
\end{aligned} \tag{38}
$$

where

$$L(s_t) = G(s_t)C_h h(s_t) + I, \qquad \hat{g}(s_t) = G(s_t)g(s_t)$$

The above equations are convenient since we do not need to assume invertibility of G. We then collapse the mixture of Gaussians defined by:

$$\sum_{s_{t-1}} p(s_{t-1} \mid s_t, v_{1:T})p(h_t \mid s_{t-1}, s_t, v_{1:T})$$

into a single Gaussian with mean $\tilde{x}(s_t)$ and covariance $\tilde{X}(s_t)$:

$$p(h_t \mid s_t, v_{1:T}) \propto \frac{1}{|\tilde{X}(s_t)|^{\frac{1}{2}}} e^{-\frac{1}{2}(h_t - \tilde{x}(s_t))^T \tilde{X}^{-1}(s_t)(h_t - \tilde{x}(s_t))} \tag{39}$$

with the following prefactor:

$$p(s_t \mid v_{1:T}) = \sum_{s_{t-1}} p(s_{t-1}, s_t \mid v_{1:T})$$

Calculating $p(s_{t-1}|s_t, v_{1:T})$ is straightforward by using equation (38). We now divide the Gaussian equation (39) by the corresponding backward message:

$$\begin{aligned} &\frac{p(s_t|v_{1:T})}{\lambda(s_t)} \frac{e^{-\frac{1}{2}(h_t - \tilde{x}(s_t))^T \tilde{X}^{-1}(s_t)(h_t - \tilde{x}(s_t))}}{|\tilde{X}(s_t)|^{\frac{1}{2}}} e^{\frac{1}{2}\left(h_t^T G(s_t) h_t - 2h_t G(s_t) g(s_t)\right)} \\ &= \frac{p(s_t|v_{1:T})}{\lambda(s_t)|\tilde{X}(s_t)|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(h_t F_t h_t - 2h_t^T f_t\right)} e^{-\frac{1}{2}\tilde{x}_t^T \tilde{X}^{-1} \tilde{x}_t} \end{aligned} \tag{40}$$

where $F_t = \tilde{X}^{-1} - G_t$ and $f_t = \tilde{X}^{-1}\tilde{x}_t - G(s_t)g_t$

$$\frac{p(s_t|v_{1:T})}{\lambda(s_t)|\tilde{X}(s_t)|^{\frac{1}{2}}} \frac{e^{\frac{1}{2} f_t^T F_t^{-1} f_t}}{e^{\frac{1}{2}\tilde{x}_t^T \tilde{X}^{-1} \tilde{x}_t}} e^{-\frac{1}{2}\left(h_t - F_t^{-1} f_t\right)^T F_t \left(h_t - F_t^{-1} f_t\right)} \tag{41}$$

In the moment representation of the forward message:

$$\begin{aligned} \tilde{F}(s_t) &= \left(\tilde{X}^{-1}(s_t) - G(s_t)\right)^{-1} \\ \tilde{f}(s_t) &= \tilde{F}(s_t)\left(\tilde{X}^{-1}\tilde{x}(s_t) - G(s_t)g(s_t)\right) \end{aligned}$$

with the following prefactor:

$$\rho(s_t) \propto \frac{p(s_t|v_{1:T})}{\lambda(s_t)} \frac{|\tilde{F}(s_t)|^{\frac{1}{2}}}{|\tilde{X}(s_t)|^{\frac{1}{2}}} \frac{e^{\frac{1}{2}\tilde{f}_t(s_t)^T \tilde{F}_t^{-1}(s_t)\tilde{f}_t(s_t)}}{e^{\frac{1}{2}\tilde{x}_t(s_t)^T \tilde{X}^{-1}(s_t)\tilde{x}_t(s_t)}}$$

4.2 Backward Pass

We restate the backward equations:

$$\lambda(h_{t-1}, s_{t-1}) \propto \frac{\sum_{s_t} p(s_{t-1}, s_t|v_{1:T})p(h_{t-1}|s_{t-1}, s_t, v_{1:T})}{\rho(s_{t-1})\rho(h_{t-1}|s_{t-1})}$$

The term $p(s_{t-1}, s_t|v_{1:T})$ has already been computed in the forward pass. Hence, we just need

$$p(h_{t-1}|s_{t-1}, s_t, v_{1:T}) \propto p(h_{t-1}, v_t, a_t \mid s_{t-1}, s_t, v_{1:t-1})|_{a_t = g(s_t)}$$

which is proportional to:

$$\exp\left\{-\frac{1}{2}\begin{pmatrix} h_{t-1}-\langle h_{t-1}\rangle \\ v_t-\langle v_t\rangle \\ a_t-\langle a_t\rangle \end{pmatrix}^T \begin{pmatrix} C_{hh} & C_{hv} & C_{ha} \\ C_{hv}^T & C_{vv} & C_{va} \\ C_{ha}^T & C_{va}^T & C_{aa} \end{pmatrix}^{-1} \begin{pmatrix} h_{t-1}-\langle h_{t-1}\rangle \\ v_t-\langle v_t\rangle \\ a_t-\langle a_t\rangle \end{pmatrix}\right\}$$

The only difference between this and the forward pass occurs in the time shift for h. Hence the statistics involving a_t and v_t alone are the same as for the forward pass. For the statistics involving h_{t-1}, we have

$$\langle h_{t-1}\rangle = \tilde{f}(s_{t-1})$$
$$C_{hh} = \langle \Delta h_{t-1}\Delta h_{t-1}^T\rangle = \tilde{F}(s_{t-1})$$
$$C_{hv} = \langle \Delta h_{t-1}\Delta v_t^T\rangle = \tilde{F}(s_{t-1})A^T(s_t)B^T(s_t)$$
$$C_{ha} = \langle \Delta h_{t-1}\Delta a_t^T\rangle = \tilde{F}(s_{t-1})A^T(s_t)$$

We need to compute

$$p(h_{t-1}|s_{t-1}, s_t, v_{1:T}) = p(h_{t-1}\,|\,a_t, s_{t-1}, s_t, v_{1:t})|_{a_t=g(s_t)}$$

This can be now done exactly as in the forward pass using equations (33,34), with just the definitions of the covariances changed to be those used in the backpass. We then collapse the mixture of Gaussians defined by:

$$\sum_s p(s_t\,|\,s_{t-1}, v_{1:T})p(h_t\,|\,s_{t-1}, s_t, v_{1:T})$$

into a single Gaussian with mean $\tilde{x}(s_{t-1})$ and covariance $\tilde{X}(s_{t-1})$:

$$p(h_{t-1}\,|\,s_t, v_{1:T}) \propto \frac{1}{|\tilde{X}(s_{t-1})|^{\frac{1}{2}}} e^{-\frac{1}{2}(h_{t-1}-\tilde{x}(s_{t-1}))^T\tilde{X}^{-1}(s_{t-1})(h_{t-1}-\tilde{x}(s_{t-1}))} \tag{42}$$

with the following pre–factor:

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

We now divide the Gaussian by the corresponding forward message:

$$\frac{p(s_{t-1}|v_{1:T})}{\rho(s_{t-1})} \frac{e^{-\frac{1}{2}(h_{t-1}-\tilde{x}(s_{t-1}))^T\tilde{X}^{-1}(s_{t-1})(h_{t-1}-\tilde{x}(s_{t-1}))}}{e^{-\frac{1}{2}(h_{t-1}-\tilde{f}(s_{t-1}))^T\tilde{F}^{-1}(s_{t-1})(h_{t-1}-(\mathbf{s_{t-1}}))}} \frac{|\tilde{F}(s_{t-1})|^{\frac{1}{2}}}{|\tilde{X}(s_{t-1})|^{\frac{1}{2}}} \tag{43}$$

In the canonical representation of the backward message:

$$\begin{aligned} G(s_{t-1}) &= \tilde{X}^{-1}(s_{t-1}) - \tilde{F}^{-1}(s_{t-1}) \\ \hat{g}(s_{t-1}) \equiv G(s_{t-1})g(s_{t-1}) &= \tilde{X}^{-1}(s_{t-1})\tilde{x}(s_{t-1}) - \tilde{F}^{-1}(s_{t-1})\tilde{f}(s_{t-1}) \end{aligned}$$

with the following prefactor:

$$\lambda(s_{t-1}) \propto \frac{p(s_{t-1}|v_{1:T})}{\rho(s_{t-1})} \frac{|\tilde{F}(s_{t-1})|^{\frac{1}{2}}}{|\tilde{X}(s_{t-1})|^{\frac{1}{2}}} \frac{e^{-\frac{1}{2}\tilde{x}(s_{t-1})^T\tilde{X}^{-1}(s_{t-1})\tilde{x}(s_{t-1})}}{e^{-\frac{1}{2}\tilde{f}(s_{t-1})^T\tilde{F}^{-1}(s_{t-1})\tilde{f}(s_{t-1})}}$$

Note that $g(s_t)$ only ever occurs in combination with $G(s_t)$. This means that, in both the forward and the backward passes, one can replace throughout $G(s_t)g(s_t)$ by a new variable $\hat{g}(s_t)$.

The reader may notice that in both the forward and backpasses, effectively, we multiply and divide by $\lambda(s_t)$ and $\rho(s_{t-1})$ respectively. In our implementation, therefore, to aid numerical stability, this unnecessary multiplication and division by the same value is removed.

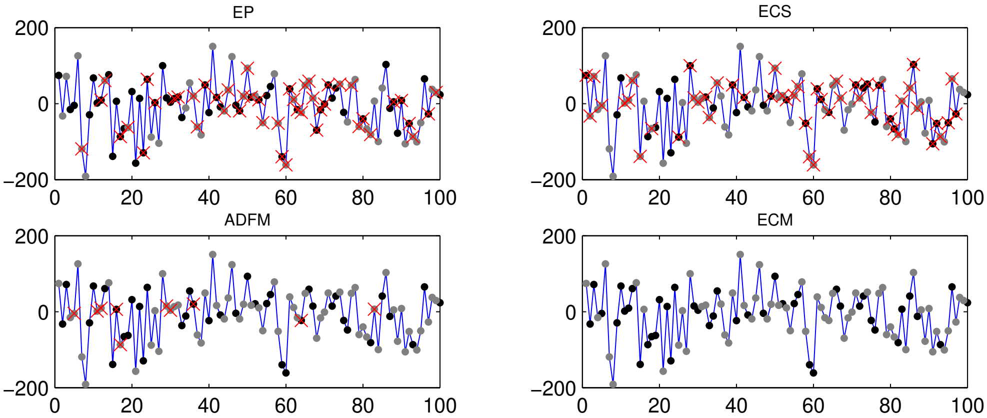
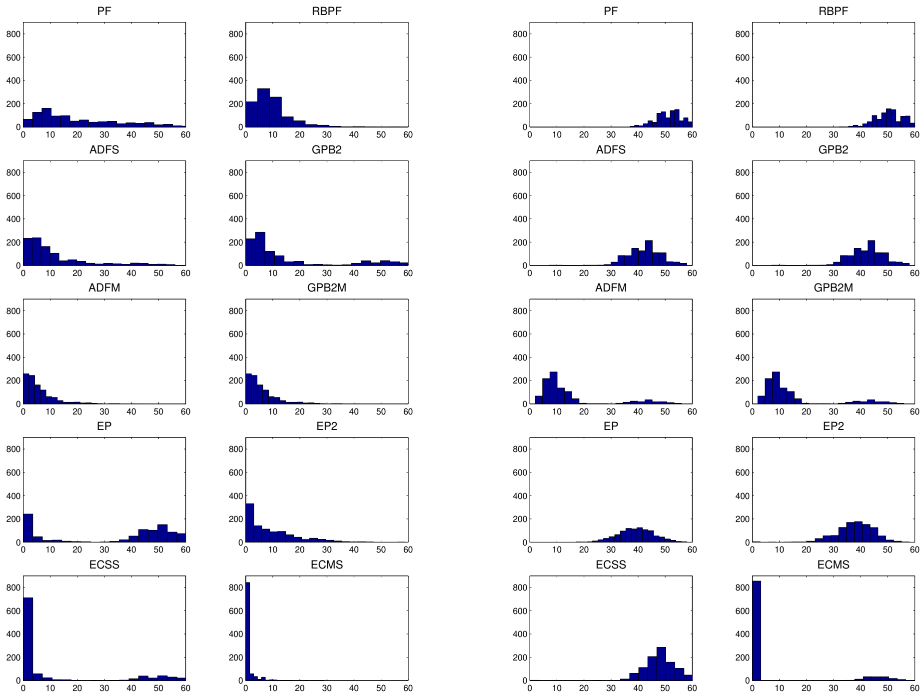


Figure 8: Results on a typical example from our 'hard' problem for the methods of Expectation Propagation (EP), Assumed Density Filtering using a mixture of 4 Gaussians (ADFM), Expectation Correction using a Single Gaussian (ECS), and Expectation Correction using a mixture of 4 Gaussians (ECM). Plotted is the one dimensional visible signal, with a marker coloured by the most probable posterior estimated switch variable, which can be one of two states. A cross indicates a switch variable inference 'error'. Only methods which have mixture representations of the posterior succeed – indeed, the ECM method gives no errors. See the caption of fig(9) for details of the experimental setup.

5 Experiments with Switching Linear Gaussian Dynamics

We would like to test our Expectation Correction smoothing method in a problem with a reasonably long temporal sequence, T. An obvious difficulty arises here in that, since the exact computation is exponential in T, a formally exact evaluation of the method is infeasible. A reasonable approach under these circumstances, is to suppose that generated switch variables will be close to the most probable state of the true posterior $p(s_t|v_{1:T})$. That is, we sample a hidden state s_1 and h_1 from the prior, and then a visible observation v_1. Then, sequentially, we generate hidden states and visible states for the next time steps. The task for smoothing inference is, given only the parameters of the model and the visible observations (but not any of the hidden states $h_{1:T}, s_{1:T}$), to infer $p(h_t|v_{1:T})$ and $p(s_t|v_{1:T})$. A simple performance measure 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 s_t. The reader should bear in mind, of course, that this is just a tractable surrogate for comparing our estimate of $p(s_t|v_{1:T})$ with the exact value of $p(s_t|v_{1:T})$.



(a) Easy problem (b) Difficult problem

Figure 9: Histograms of the number of errors over 1000 experiments. Particle Filter(PF), Rao-Blackwellised Particle Filter(RBPF), Assumed Density Filtering, Single Gaussian (ADFS), Generalised Pseudo Bayes 2(GPB2), Assumed Density Filtering, Multiple Gaussians (ADFM), Generalised Pseudo Bayes 2 with Multiple Gaussians (GPB2M), Expectation Propagation (EP), Our Implementation of Expectation Propagation (EP2), Expectation Correction Smoothing with a Single Gaussian (ECSS), Expectation Correction Smoothing with Multiple Gaussians (ECSM). Throughout, $S = 2$, $V = 1$, $T = 100$, with zero output bias. For the multiple Gaussian methods, $I = J = 4$ Gaussians were used. For EP, 3 iterations were performed. Using partly MATLAB notation, $A(s) = 0.9999 * \mathrm{orth}(\mathrm{randn}(\mathrm{H},\mathrm{H}))$, $B(s) = \mathrm{randn}(\mathrm{V},\mathrm{H})$, $\bar{v}_t \equiv 0$, $\bar{h}_1 = 10 * \mathrm{randn}(\mathrm{H},1)$, $\bar{h}_{t>1} = 0$, $\Sigma_1^h = I_H$, $p_1 = \text{uniform}$. (a) Results on a relatively 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) Results on a relatively hard problem. $H = 30$, $\Sigma^v = 30 I_V$,$\Sigma^h = 0.01 I_H$, $p(s_{t+1}|s_t) \propto 1_{S\times S}$.

We look at two sets of experiments, fig(9), both on time series of length $T = 100$ with $S = 2$ switch states[2]. In both sets of experiments, we compared methods using a single Gaussian, and methods using multiple Gaussians. The number of Gaussians used was set to $2 \times S$ throughout. One is relatively ‘easy’ and the other relatively ‘hard’. From the viewpoint of classical signal processing, both experiments are extremely difficult in the sense that they cannot be solved by short time Fourier methods, since changes occur in the dynamics at a much higher rate than the typical frequencies in the signal, see fig(8).

In the easy experiments, we used a small hidden dimension $H = 3$, with a moderate amount of transition and observation noise. As can be seen from fig(9a), Particle Filtering performs reason-

[2]In fact we looked at time series of length $T = 105$, and computed the number of errors made on the time points 5 to 105. The reason for this is that we are restricted in the number of Gaussians that can be used in the first two time steps, and this means that the performance is atypically worse on the first couple of time points.

ably well, although its performance is enhanced by Rao-Blackwellisation (RBPF). Assumed Density Filtering using a single Gaussian and with a mixture performed roughly the same as RBPF, as did the methods based on Generalised Pseudo Bayes 2, using either the ADF single Gaussian results, or the Gaussian mixture results. A standard implementation of Expectation Propagation, even in this case, suffers from many numerical stability problems, but is improved somewhat by our own more stable implementation. The Expectation Correction method using a single Gaussian dramatically improves on the ADF single Gaussian filtered estimate. Using a small number of mixture components in Expectation Correction improves the situation further.

In the hard case, fig(9b) we used a larger hidden dimension, $H = 30$, with a small amount of transition noise, and a large amount of observation noise. We chose these parameters since this will most likely result in highly multi-modal posteriors. In this case, only those methods that used a mixture of Gaussians performed well – otherwise, the methods were little better than random guessing. Expectation Correction with a small number of mixture components, apart from a small number of errors, dramatically gives almost perfect performance. Readers interested in Particle Filters may wonder why Rao-Blackwellisation doesn't seem to perform well. Our explanation is that the standard implementation we used [7] still makes the assumption that a *single* Gaussian is adequate to describe the posterior filtered estimate $p(h_t|s_t, v_{1:t})$. In our 'hard' experiment, any method which does not deal with multi-modality of the posterior is doomed.

6 Discussion

We have presented a method that can be used in switching dynamical models, best suited to distributions conditional on the switch variable which are from the exponential family. We were motivated to use a mixture approximation since we know that the exact result for the posterior distribution is a mixture, albeit of a number of components exponential in time. However, in practice, due to the Markovian nature of the dynamics, we expect that the effective correlation length of the posterior will be very much shorter than the length of the time series. For this reason, a much smaller effective number of mixture components may be expected to produce a reasonable approximation.

Specifically, our method contains three approximations : (a) collapse of $p(h_t|s_t, s_{t-1}, v_{1:T})$ to a mixture; (b) dropping of a dependence $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$, which for switching observation models is exact; (c) approximation of the average of $p(s_t|h_{t+1}, s_{t+1}, v_{1:t})$ with respect to the distribution $p(h_{t+1}|s_{t+1}, v_{1:T})$. In (b), our approximation is relatively delicate, when compared with the popular Generalised Pseudo Bayes method, which discards all future information. A particularly appealing aspect of our approach is that steps (a) and (c) above can be made more accurate. Certainly in the case of Switching Linear Gaussian models, we have found experimental conditions where increasing the number of Gaussians in the approximation results in a dramatic improvement in performance. The complexity of each iteration of the Forward Pass scales linearly with the number of mixture components. However, the backpass is more complex, and scales with the number of forward mixtures used multiplied by the number of backward mixtures used. Our experience is that, in practice, a smaller number of backward mixtures may be sufficient since usually most of the improvement in performance comes from using mixtures in the forward pass. Indeed, often a single mixture is sufficient in the backward pass.

In the current work, only the simplest average approximation was considered, namely evaluation of the integrand at the mean, and more sophisticated methods may give better results. In this sense, the current work can be seen as a useful starting point.

Whilst the current method has many attractive properties, it cannot be feasibly applied to factorial representations of the switch variables, and this would require further approximations. Another area that we are currently investigating is dependencies of the discrete hidden variables on the continuous hidden variables, see for example [6], since this would enable us to have a powerful general purpose approximation method for a large class of practically useful models.

MATLAB software for Expectation Correction for Switching Linear Gaussian State Space models is at `http://www.idiap.ch/~barber/ecskf.zip`.

A Finding the Conditional Gaussian from the joint

For the joint Gaussian distribution over the vectors x and y

$$p(x,y) = \frac{1}{\sqrt{\det 2\pi\Sigma}} \exp\left(-\frac{1}{2}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}^T \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{pmatrix}^{-1} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}\right) \tag{44}$$

the conditional is given by

$$p(x|y) = N(mean = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y), cov = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}) \tag{45}$$

B Collapsing a Mixture of Gaussians

Consider a normalised ($\sum_i p_i = 1$) mixture of Gaussians distribution $p(x) = \sum_i p_i \mathcal{N}(x|\mu_i, \Sigma_i)$. The mean and covariance of this distribution is

$$\mu = \sum_i p_i \mu_i, \qquad \Sigma = \sum_i p_i\left(\Sigma_i + \mu_i\mu_i^T\right) - \mu\mu^T$$

B.1 Collapsing a Mixture of N Gaussians to a smaller Mixture of K Gaussians

There are many ways to do this. Ideally, one might use a method such as minimal Kullback-Leibler divergence between the large and the small mixture. Unfortunately, this is difficult and computationally expensive to approximate. The method that we use in the experiments is to simply retain the $K-1$ Gaussians with the largest mixture weights in the mixture we wish to approximate. The remaining $N-K$ Gaussians are simply merged to a single Gaussian using the above method. Of course, a disadvantage of such a simple approach is that no spatial information is taken into account in the approximation. A similarly non-spatial approach is to recursively merge the two Gaussians with the lowest mixture weights; this gave similar experimental performance.

C The Likelihood

One of the most elegant approaches is to use Bayes' rule recursively:

$$p(v_{1:T}) = p(v_1)p(v_2|v_1)p(v_3|v_{1:2})\dots p(v_T|v_{1:T-1}) \tag{46}$$

Consider

$$\begin{aligned}
p(v_t|v_{1:t-1}) &= \sum_{s_{t-1}} \int_{h_{t-1}} p(v_t, h_{t-1}, s_{t-1}|v_{1:t-1}) \\
&= \sum_{s_{t-1}} \int_{h_{t-1}} p(v_t|h_{t-1}, s_{t-1})p(h_{t-1}, s_{t-1}|v_{1:t-1}) \\
&= \sum_{s_{t-1},s_t} \int_{h_{t-1},h_t} p(v_t, h_t, s_t|h_{t-1}, s_{t-1})p(h_{t-1}, s_{t-1}|v_{1:t-1}) \\
&= \sum_{s_{t-1},s_t} p(s_t|s_{t-1})p(s_{t-1}|v_{1:t-1}) \underbrace{\int_{h_{t-1},h_t} p(v_t|h_t, s_t)p(h_t|h_{t-1}, s_t)p(h_{t-1}|s_{t-1}, v_{1:t-1})}_{p(v_t|s_t,s_{t-1},v_{1:t-1})}
\end{aligned} \tag{47}$$

Under our approximation scheme $p(h_{t-1}|s_{t-1}, v_{1:t-1})$ is a single Gaussian (the extension to the mixture case is trivial). Then $p(v_t|s_t, s_{t-1}, v_{1:t-1})$ is a Gaussian distribution with mean and covariance given by

$$c(s_t, s_{t-1}) = B(s_t)A(s_t)f(s_{t-1})$$

$$C(s_t, s_{t-1}) = B(s_t)\left(A(s_t)F(s_{t-1})A^T(s_t) + \Sigma^h(s_t)\right)B^T(s_t) + \Sigma^v(s_t)$$

Thus

$$p(v_t|s_t, s_{t-1}, v_{1:t-1}) = \frac{e^{-\frac{1}{2}(v(t)-c(s_t,s_{t-1}))^T C^{-1}(s_t,s_{t-1})(v(t)-c(s_t,s_{t-1}))}}{\sqrt{\det 2\pi C(s_t, s_{t-1})}}$$

which can be used directly in equation (47) and equation (46) to find the likelihood.

D Switching Linear Gaussian State Space Models : EC mixture approximation

D.1 Forward Pass

Here we extend the forward pass so that the collapse has, for each state s_t, not just a single Gaussian, but a set of Gaussians. We use $i_t \in 1 : I$ to represent the Gaussian mixture component. From the factorisation $p(s_t, h_t|v_{1:t}) = p(h_t|s_t, v_{1:t})p(s_t|v_{1:t})$, as for the single Gaussian case, our strategy will be to find, first $p(h_t|s_t, v_{1:t})$. The term $p(h_t, v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1})$ may be easily evaluated by realising that for each setting of the switch variables i_{t-1}, s_{t-1}, s_t the distribution is a Gaussian. The means and covariances of this Gaussian are easily found from the relations

$$v_t = B(s_t)h_t + \eta_v(s_t), \qquad h_t = A(s_t)h_{t-1} + \eta_h(s_t)$$

Using the above, we readily find

$$\left\langle \Delta v_t \Delta v_t^T | s_t, i_{t-1}, s_{t-1} \right\rangle = B(s_t)\left\langle \Delta h_t \Delta h_t^T | s_t, i_{t-1}, s_{t-1} \right\rangle B^T(s_t) + \Sigma^v(s_t)$$

$$\left\langle \Delta h_t \Delta h_t^T | s_t, i_{t-1}, s_{t-1} \right\rangle = A(s_t)\left\langle \Delta h_{t-1} \Delta h_{t-1}^T | i_{t-1}, s_{t-1} \right\rangle A^T(s_t) + \Sigma^h(s_t)$$

$$\left\langle \Delta v_t \Delta h_t^T | s_t, i_{t-1}, s_{t-1} \right\rangle = B(s_t)\left\langle \Delta h_t \Delta h_t^T | s_t, i_{t-1}, s_{t-1} \right\rangle$$

$$\langle v_t | s_t, i_{t-1}, s_{t-1} \rangle = B(s_t)A(s_t)\langle h_{t-1} | i_{t-1}, s_{t-1} \rangle$$

$$\langle h_t | s_t, i_{t-1}, s_{t-1} \rangle = A(s_t)\langle h_{t-1} | i_{t-1}, s_{t-1} \rangle$$

In the above, using our moment representation of the forward messages

$$\langle h_{t-1} | i_{t-1}, s_{t-1} \rangle \equiv f_{t-1}(i_{t-1}, s_{t-1})$$

$$\left\langle \Delta h_{t-1} \Delta h_{t-1}^T | s_{t-1} \right\rangle \equiv F_{t-1}(i_{t-1}, s_{t-1})$$

Using the above results, we are now in a position to calculate equation (15). For each setting of the variable s_t, we will therefore have a mixture of $I \times S$ Gaussians. There are many different strategies conceivable for approximating this mixture of Gaussians, and we use the one outlined in section (B.1), to give

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

In this way the new mixture coefficients $p(i_t|s_t, v_{1:t})$, $i_t \in 1, \ldots, I$ are defined.

Algorithm 1 The Switching Kalman Filter : Forward Pass using Mixtures. We require $I_1 = 1, I_2 \leq S, I_t \leq S \times I_{t-1}$.

procedure SwitchingKalmanForwardMixture
 $F_0 \leftarrow 0,\ f_0 \leftarrow 0,\ \rho_0 \leftarrow 1, w_0 \leftarrow 1$
 for $t \leftarrow 1, T$ **do**
  for $s_t \in S$ **do**
   for $s_{t-1} \in S$ **do**
    for $i_{t-1} \in I$ **do**
     $\tilde{h}_t \leftarrow A(s_t) f_{t-1}(i_{t-1}, s_{t-1}) + \bar{h}_t(s_t)$
     $\tilde{v}_t \leftarrow B(s_t)\tilde{h}_t + v_b(s_t) + \bar{v}_t(s_t)$
     $\Sigma_{hh} \leftarrow A(s_t) F_{t-1}(i_{t-1}, s_{t-1}) A^T(s_t) + \Sigma_t^h(s_t)$
     $\Sigma_{vv} \leftarrow B(s_t)\Sigma_{hh} + \Sigma^v(s_t)$
     $\Sigma_{vh} \leftarrow B(s_t)\Sigma_{hh}$
     $\Sigma_{x|y}(i_{t-1}, s_{t-1}) \leftarrow \Sigma_{hh} - \Sigma_{vh}^T \Sigma_{vv}^{-1} \Sigma_{vh}$
     $\mu_{x|y}(i_{t-1}, s_{t-1}) \leftarrow \tilde{h}_t + \Sigma_{vh}^T \Sigma_{vv}^{-1} (v_t - \tilde{v}_t)$
     $\hat{p} \leftarrow \frac{1}{\sqrt{\det \Sigma_{vv}}} \exp\left(-\frac{1}{2}(v_t - \tilde{v}_t)^T \Sigma_{vv}^{-1} (v_t - \tilde{v}_t)\right)$
     $(t = 1) : p'(i_{t-1}, s_{t-1}) \leftarrow p_{t=1}(s_t)\hat{p}$
     $(t = 2) : p'(i_{t-1}, s_{t-1}) \leftarrow p(s_t|s_{t-1})\rho_{t-1}(s_{t-1})\hat{p}$
     $(t > 2) : p'(i_{t-1}, s_{t-1}) \leftarrow w_{t-1}(i_{t-1}, s_{t-1}) p(s_t|s_{t-1})\rho_{t-1}(s_{t-1})\hat{p}$
    end for
   end for
   Normalise p' to a distribution over i_{t-1}, s_{t-1}
   $\rho_t(s_t) = \sum_{i_{t-1}, s_{t-1}} p'(i_{t-1}, s_{t-1})$
   $(t = 1)$ No Collapse
   (If $t > 1$) Collapse Gaussian with means $\mu_{x|y}$, covariances $\Sigma_{x|y}$ and mixture weights p' to a Gaussian with I components
   This defines the new means $f_t(i_t, s_t)$, covariances $F_t(i_t, s_t)$ and mixture weights $w_t(i_t, s_t)$
  end for
  normalise ρ_t
 end for
end procedure

D.2 Backward pass

Our aim is to make a recursion for a mixture representation of $p(h_t|s_t, v_{1:T})$. This is straightforward and follows from the approach given in section (2.3.4). We need to find $p(h_t|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$ and $p(h_{t+1}|h_t, i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$, which we do as follows. First we find the distribution $p(h_t|h_{t+1}, s_{t+1}, i_t, s_t, v_{1:t})$, which is itself found from conditioning the joint distribution

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

This is used to define the backward equation

$$h_t|h_{t+1}, i_t, s_t, s_{t+1}, v_{1:t} = \overleftarrow{A}(i_t, s_t, s_{t+1}) h_{t+1} + \overleftarrow{m}(i_t, s_t, s_{t+1}) + \overleftarrow{\eta}(i_t, s_t, s_{t+1})$$

Then the joint distribution $p(h_t, h_{t+1}|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$ has the following mean and covariances

$$\langle h_t|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}\rangle = \overleftarrow{A}(i_t, s_t, s_{t+1}) g_{t+1}(j_{t+1}, s_{t+1}) + \overleftarrow{m}(i_t, s_t, s_{t+1})$$

$$\langle h_{t+1}|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}\rangle = g_{t+1}(j_{t+1}, s_{t+1})$$

$$\left\langle \Delta h_{t+1} \Delta h_{t+1}^T | i_t, s_t, s_{t+1}, v_{1:T} \right\rangle = G_{t+1}(j_{t+1}, s_{t+1})$$

$$\left\langle \Delta h_t \Delta h_t^T | i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T} \right\rangle = \overleftarrow{A}(i_t, s_t, s_{t+1}) G_{t+1}(j_{t+1}, s_{t+1}) \overleftarrow{A}^T(i_t, s_t, s_{t+1}) + \overleftarrow{\Sigma}_t(i_t, s_t, s_{t+1})$$

$$\left\langle \Delta h_t \Delta h_{t+1}^T | i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T} \right\rangle = \overleftarrow{A}(i_t, s_t, s_{t+1}) G_{t+1}(j_{t+1}, s_{t+1})$$

From this, we can find the marginal $p(h_t|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$. Using again the simple approximation for the average in equation (27),

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

Integrating over h_t, we have

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

Using the above, we can form the distribution

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

This mixture can then be collapsed to another mixture of Gaussians using the usual approach to define

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

Algorithm 2 The Switching Kalman Filter : Expectation Correction using Mixtures. We require $I_1 = 1, I_2 \leq S, I_t \leq S \times I_{t-1}$. $J_T = I_T, J_t \leq S \times I_t \times J_{t+1}$

procedure SwitchingKalmanBackward
  $G_T \leftarrow F_T$, $g_T \leftarrow f_T$, $\lambda_T \leftarrow \rho_T$, $u_T \leftarrow w_T$
  for $t \leftarrow T, 1$ **do**
    for $s_t \in S$ **do**
      for $s_{t+1} \in S$ **do**
        for $i_t \in I_t$ **do**
          $\langle h_{t+1} | v_{1:t} \rangle (i_t) \leftarrow A(s_{t+1}) f_t(i_t, s_t) + \bar{h}_{t+1}(s_{t+1})$
          $\left\langle \Delta h_{t+1} \Delta h_{t+1}^T | v_{1:t} \right\rangle (i_t) \leftarrow A(s_{t+1}) F_t(i_t, s_t) A^T(s_{t+1}) + \Sigma^h(s_{t+1})$
          $\left\langle \Delta h_{t+1} \Delta h_t^T | v_{1:t} \right\rangle \leftarrow A(s_{t+1}) F_t(i_t, s_t)$
          $\overleftarrow{\Sigma}(i_t) \leftarrow F_t(i_t, s_t) - \left\langle \Delta h_{t+1} \Delta h_t^T | v_{1:t} \right\rangle^T \left\langle \Delta h_{t+1} \Delta h_{t+1}^T | v_{1:t} \right\rangle^{-1} \left\langle \Delta h_{t+1} \Delta h_t^T | v_{1:t} \right\rangle$
          $\overleftarrow{A}(i_t) \leftarrow \left\langle \Delta h_{t+1} \Delta h_t^T | v_{1:t} \right\rangle^T \left\langle \Delta h_{t+1} \Delta h_{t+1}^T | v_{1:t} \right\rangle^{-1}$
          $\overleftarrow{m}(i_t) \leftarrow f_t(i_t, s_t) - \overleftarrow{A}(i_t) \langle h_{t+1} | v_{1:t} \rangle (i_t)$
          for $j_{t+1} \in J_{t+1}$ **do**
            $\langle h_t | v_{1:T} \rangle (i_t, s_t, j_{t+1}, s_{t+1}) \leftarrow \overleftarrow{A}(i_t) g_{t+1}(j_{t+1}, s_{t+1}) + \overleftarrow{m}(i_t)$
            $\left\langle \Delta h_t \Delta h_t^T | v_{1:T} \right\rangle (i_t, s_t, j_{t+1}, s_{t+1}) \leftarrow \overleftarrow{A}(i_t) G_{t+1}(j_{t+1}, s_{t+1}) \overleftarrow{A}^T(i_t) + \overleftarrow{\Sigma}(i_t)$
            $\left\langle \Delta h_t \Delta h_{t+1}^T | v_{1:T} \right\rangle \leftarrow \overleftarrow{A}(i_t) G_{t+1}(j_{t+1}, s_{t+1})$
            $p(i_t, s_t, s_{t+1} | v_{1:t}) \leftarrow p(s_{t+1} | s_t) w_t(i_t, s_t) \rho_t(s_t)$
            $z = g_{t+1}(j_{t+1}, s_{t+1}) - \langle h_{t+1} | v_{1:t} \rangle (i_t)$
            $p(i_t, s_t | j_{t+1}, s_{t+1}, v_{1:T}) = \frac{p(i_t, s_t, s_{t+1} | v_{1:t})}{\sqrt{\det \left\langle \Delta h_{t+1} \Delta h_{t+1}^T | v_{1:t} \right\rangle (i_t)}} \exp\left(-\frac{1}{2} z^T \left\langle \Delta h_{t+1} \Delta h_{t+1}^T | v_{1:t} \right\rangle^{-1} (i_t) z \right)$
          end for
        end for
      end for
    end for
    Normalise $p(i_t, s_t | j_{t+1}, s_{t+1}, v_{1:T})$ to ensure a distribution over i_t, s_t
    for $s_t \in S$ **do**
      for $i_t \in I_t, s_{t+1} \in S, j_{t+1} \in J_{t+1}$ **do**
        $p(i_t, s_t, j_{t+1}, s_{t+1} | v_{1:T}) \leftarrow p(s_{t+1} | v_{1:T})(s_{t+1}) u_{t+1}(j_{t+1}, s_{t+1}) p(i_t, s_t | j_{t+1}, s_{t+1}, v_{1:T})$
      end for
      $p(s_t | v_{1:T}) \leftarrow \sum_{i_t, j_{t+1}, s_{t+1}} p(i_t, s_t, j_{t+1}, s_{t+1} | v_{1:T})$
      Collapse the mixture over the joint set of indices i_t, j_{t+1}, s_{t+1} defined by weights $p(i_t, s_{t+1}, j_{t+1} | s_t, v_{1_T})$, means $\langle h_t | v_{1:T} \rangle (i_t, s_t, j_{t+1}, s_{t+1})$ and covariances $\left\langle \Delta h_t \Delta h_t^T | v_{1:T} \right\rangle (i_t, s_t, j_{t+1}, s_{t+1})$. This defines the new means $g_t(j_t, s_t)$, covariances $G_t(j_t, s_t)$ and mixture weights $u_t(j_t, s_t)$
    end for
  end for
end procedure

We would like to thank Onno Zoeter and Tom Heskes for kindly providing their Expectation Propagation code.

References

[1] Felix Agakov and David Barber. An auxiliary variational method. IDIAP-RR 86, IDIAP, Rue de Simplon 4, Martigny, CH-1920, Switerland, June 2004. IDIAP-RR 04-86.

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

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

[4] D. Barber. The auxiliary variable trick for deriving kalman smoothers. IDIAP-RR 87, IDIAP, Rue de Simplon 4, Martigny, CH-1920, Switerland, December 2004. Submitted to IEEE trans. Automatic Control.

[5] D. Barber and P. Sollich. Gaussian fields for approximate inference in layered sigmoid belief networks. In S. A. Solla, T. K. Leen, and K. R Muller, editors, *Advances in Neural Information Processing Systems NIPS 12*. MIT Press, 2000. ISSN 1049 5258 ISBN 0 262 19450 3.

[6] A. T. Cemgil, B. Kappen, and D. Barber. A Generative Model for Music Transcription. *IEEE Transactions on Speech and Audio Processing*, 2004. In press : see www.idiap.ch/∼barber/publications/cemgil-pianoroll-submit.pdf.

[7] N. de Freitas. Rao-blackwellised particle filtering for fault diagnosis. In *IEEE Aerospace Conference Proceedings*, volume 4, pages 1767–1772, March 2002.

[8] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[9] Z. Ghahramani and G. Hinton. Switching state-space models. Technical Report CRG-TR-96-3, 1996.

[10] 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.

[11] F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer Verlag, 2001.

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

[13] 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.

[14] T. Minka. A family of algorithms for approximate Bayesian inference, 2001.

[15] K. P. Murphy. Switching Kalman Filters. Technical Report U. C. Berkeley, 1998.

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

[17] 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.

[18] D. Saad and M. Opper. *Advanced Mean Field Methods Theory and Practice*. MIT Press, 2001.

[19] R. H. Shumway and D. S. Stoffer. Dynamic Linear Models with Switching. *Journal of the American Statistical Association*, 86(415):763–769, 1991.

[20] M. A. Srinvas and R. J. McEliece. The Generalised Distributive Law. *IEEE Transactions of Information Theory*, 46(2):325–343, 2000.

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