IDIAP RESEARCH REPORT

# A stable Switching Kalman Smoother

David Barber [a]

IDIAP–RR 04-89

December 2004

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Report 04-89

# A stable Switching Kalman Smoother

David Barber

December2004

**Abstract.** We present a new method for approximate inference in Switching linear Gaussian State Space Models (also known as Switching Kalman Filters. The method is similar in spirit to the Rauch-Tung-Striebel smoother in the Kalman Filter case. Only a single Forward and Backward pass is required, both of which are numerically stable. The algorithm projects at each time, for both the Forward and Backward passes, the approximate Belief states onto either a single or a mixture of Gaussians. Unlike in Expectation Propagation, we find few difficulties with numerical stability.

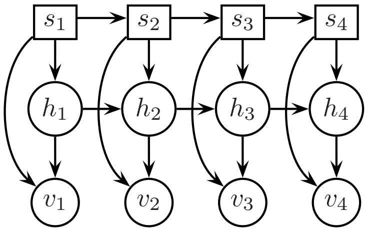


Figure 1: A Switching Kalman Filter. The variables **h** and **v** are Gaussian distributed. The Switch variables **s** are discrete, and control the means and variances of the Gaussian transitions, and possibly also the emissions. Links which are not allowed are those from continuous variables to discrete variables.

# 1 Introduction to the Switching Kalman Filter

The SKF is a popular Hybrid distribution. Its popularity stems from its powerful nature as being able to model both continuous process whilst switching between different Kalman Filter regimes.

Links which are not allowed are those from continuous variables to discrete variables. The reason for this is to explained below when we consider inference in this and related systems.

For example, we may have a standard Kalman Filter emission

$$\mathbf{v}(t) = A\mathbf{h}(t) + \boldsymbol{\eta}_V(t)$$

yet have the transition dynamics determined by the switch variables:

$$\mathbf{h}(t) = \left(I[s(t) = 1]B_1 + I[s(t) = 2]B_2\right)\mathbf{h}(t-1) + \boldsymbol{\eta}_H(t)$$

Here $I[x = y]$ has value 1 if the condition $x = y$ is satisfied, and is otherwise zero. In the above example, this has the effect that if $s(t) = 1$, we get the transition

$$\mathbf{h}(t) = B_1\mathbf{h}(t-1) + \boldsymbol{\eta}_H(t)$$

and if $s(t) = 2$, we have the transition

$$\mathbf{h}(t) = B_2\mathbf{h}(t-1) + \boldsymbol{\eta}_H(t)$$

This is an example of a form of Switching Dynamics. This means that a form of non-stationarity can be modelled. To complete the specification, we could set, say $p(s(t) = 1) = 0.1$, and $p(s(t) = 2) = 0.9$, so that the dynamics most of the time will use $B_2$, but will occasionally use $B_1$. A key question is, *is this model computationally tractable?* Imagine that we wish to compute the marginal distribution of the hidden variables $h$. Naively, we can integrate out immediately the switch variables to give the distribution

$$\prod_t p(\mathbf{v}(t)|\mathbf{h}(t)) \underbrace{\sum_{s(t)} p(\mathbf{h}(t)|\mathbf{h}(t-1), s(t))p(s(t))}_{\phi(\mathbf{h}(t-1),\mathbf{h}(t)}$$

with the usual conventions for the initial time $t = 1$. The factor $\phi(\mathbf{h}(t-1), \mathbf{h}(t)$ is a mixture of Gaussians[1]. In the specific example above, it is a mixture of two Gaussians, which we could therefore parameterise using $\theta_1(t)$ and $\theta_2(t)$ where $\theta$ represent mean and covariance parameters.

[1]For the case that the switch variables are functions of the continuous variables, this will typically result in a non-Gaussian contribution to the potential, taking us away from the mixture of Gaussians representation. This is problematic when we wish to calculate means and covariances of the mixture – it may not be analytically tractable.

Consider now using Belief Propagation (since the graph is non-loopy)[2, 3]. It's easy to see that at time $t$, the forward message will be a mixture of Gaussians with $2^t$ components. In practice, we can therefore cannot pass exact messages. In this sense, the intractability arises, not because the graph is loopy, but because we cannot make a simple description of the messages. In very special cases, this may not be quite as bad – for example it could be that the system is set up so that a limited number of Gaussian components is only possible, or that some of the switch variables have the effect of destroying past information.

A more natural way to pass messages (and is more useful for more general cases) is to consider a message which is a function of both the continuous and discrete switch state. For example, in the forward message, this would be

$$\rho_{t,t+1}\left(\mathbf{h}(t), s(t)\right)$$

Of course, many possible schemes could be envisaged as approximation schemes – sampling, such as simple particle filters, for example. However, in practice these are usually outclassed by 'analytic' approximation methods which try to approximate more directly the exact inference procedure.

A natural way to proceed is therefore to make an approximation in some way to the messages. There have been many suggestions about how to do this. Currently, one of the most elegant approaches is Expectation Propagation[1], although this suffers from poor numerical stability.

In the following, we describe what a stable 'correction' SKF smoother. This will consist of a Forward and a Backward Pass. The Forward pass is relatively straightforward, and we consider this first. The novel contribution of this paper is in the Backpass, which is based on an approximate implementation of the analog of the Rung-Tauch-Striebel method in Kalman smoothing[4].

## 2 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

$$\underbrace{p(h_t|s_t, v_{1:t})}_{\approx N(f_t(s_t), F_t(s_t))} \underbrace{p(s_t|v_{1:t})}_{\equiv r_t(s_t)}$$

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 will be written as $r_t(s_t)$.

Our strategy will be to find first $p(h_t|s_t, v_{1:t})$. This can be obtained as follows:

$$p(h_t|s_t, v_{1:t}) = \sum_{s_{t-1}} p(h_t, s_{t-1}|s_t, v_{1:t}) \tag{2.1}$$

$$\propto \sum_{s_{t-1}} p(h_t, s_{t-1}, s_t, v_{1:t}) \tag{2.2}$$

$$\propto \sum_{s_{t-1}} p(h_t|s_{t-1}, s_t, v_{1:t}) p(s_{t-1}, s_t, v_{1:t}) \tag{2.3}$$

$$\propto \sum_{s_{t-1}} p(h_t|s_{t-1}, s_t, v_{1:t}) p(v_t|s_{t-1}, s_t, v_{1:t-1}) p(s_{t-1}, s_t|v_{1:t-1}) \tag{2.4}$$

$$\propto \sum_{s_{t-1}} p(h_t|s_{t-1}, s_t, v_{1:t}) p(v_t|s_{t-1}, s_t, v_{1:t-1}) p(s_t|s_{t-1}) p(s_{t-1}|v_{1:t-1}) \tag{2.5}$$

The only awkward term here is $p(h_t|s_{t-1}, s_t, v_{1:t})$. This can be found as follows

$$\begin{aligned} p(h_t|s_{t-1}, s_t, v_{1:t}) &\propto p(h_t, s_{t-1}, s_t, v_{1:t}) & (2.6)\\ &\propto p(h_t, v_t, s_{t-1}, s_t, v_{1:t-1}) & (2.7)\\ &\propto p(h_t, v_t|s_{t-1}, s_t, v_{1:t-1})p(s_{t-1}, s_t|v_{1:t-1}) & (2.8)\\ &\propto p(h_t, v_t|s_{t-1}, s_t, v_{1:t-1}) & (2.9) \end{aligned}$$

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

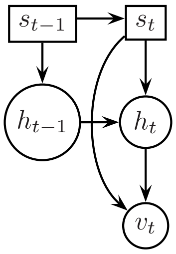


Figure 2: Structure of the forward pass. Essentially, the forward pass defines a 'prior' distribution at time $t-1$ which contains all the information from the variables $v_{1:t-1}$.

The term $p(h_t, v_t|s_{t-1}, s_t, v_{1:t-1})$ may be easily evaluated by realising that for each setting of the switch variables $s_{t-1}, s_t$ the distribution is a Gaussian. The means and covariances of this Gaussian are easily found from the relations

$$v_t = B(s_t)h_t + \eta_v(s_t)$$

$$h_t = A(s_t)h_{t-1} + \eta_h(s_t)$$

Using the above, we readily find

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

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

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

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

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

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

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

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

Using the above results, we are now in a position to calculate equation (2.5). For each setting of the variable $s_t$, we will therefore have a *mixture* of $S$ Gaussians. This is the exact calculation. Keeping all these Gaussians is expensive, since, at the next time step, we will have $S^2$ Gaussians, and in general, an exponential number of them as we progress through the forward recursion. There are many different strategies conceivable for approximating this mixture of Gaussians. Arguably the simplest is to replace the mixture with a single Gaussian which has the same mean and covariance as the mixture distribution. This is easy to do using the result in the appendix.

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

$$p(s_t|v_{1:t}) \propto \sum_{s_{t-1}} p(s_t, s_{t-1}, v_t, v_{1:t-1}) \tag{2.10}$$

$$= \sum_{s_{t-1}} p(v_t|s_t, s_{t-1}, v_{1:t-1}) p(s_t, s_{t-1}|v_{1:t-1}) \tag{2.11}$$

$$= \sum_{s_{t-1}} p(v_t|s_t, s_{t-1}, v_{1:t-1}) p(s_t|s_{t-1}) p(s_{t-1}|v_{1:t-1}) \tag{2.12}$$

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 $p(s_{t-1}|v_{1:t-1})$ is the previous prior.

## 2.1 Extension to Mixture of Gaussians

Here I want to extend the Forward Pass so that the collapse has, for each state $s_t$, not just a single Gaussian, but a set of Gaussians (somewhat akin to particle filtering). We use $i_t \in 1 : I$ to represent the Gaussian mixture component.

$$p(s_t, h_t|v_{1:t}) = p(h_t|s_t, v_{1:t}) p(s_t|v_{1:t}) \tag{2.13}$$

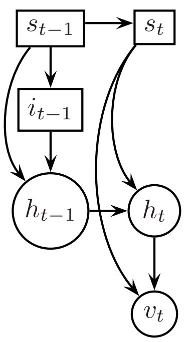


Figure 3: Structure of the forward pass. Essentially, the forward pass defines a 'prior' distribution at time $t-1$ which contains all the information from the variables $v_{1:t-1}$.

As before, 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:

$$p(h_t|s_t, v_{1:t}) = \sum_{i_{t-1}, s_{t-1}} p(h_t, i_{t-1}, s_{t-1}|s_t, v_{1:t}) \tag{2.14}$$

$$\propto \sum_{i_{t-1}, s_{t-1}} p(h_t, i_{t-1}, s_{t-1}, s_t, v_{1:t}) \tag{2.15}$$

$$\propto \sum_{i_{t-1}, s_{t-1}} p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t}) p(i_{t-1}, s_{t-1}, s_t, v_{1:t}) \tag{2.16}$$

$$\propto \sum_{i_{t-1}, s_{t-1}} p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t}) p(v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1}) p(i_{t-1}, s_{t-1}, s_t|v_{1:t-1}) \tag{2.17}$$

$$\propto \sum_{i_{t-1}, s_{t-1}} p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t}) \tag{2.18}$$

$$\times\, p(v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1}) p(s_t|s_{t-1}) p(i_{t-1}|s_{t-1}, v_{1:t-1}) p(s_{t-1}|v_{1:t-1}) \tag{2.19}$$

The only awkward term here is $p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t})$. This can be found as follows

$$p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t}) \propto p(h_t, i_{t-1}, s_{t-1}, s_t, v_{1:t}) \tag{2.20}$$

$$\propto p(h_t, v_t, i_{t-1}, s_{t-1}, s_t, v_{1:t-1}) \tag{2.21}$$

$$\propto p(h_t, v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1})p(i_{t-1}, s_{t-1}, s_t|v_{1:t-1}) \tag{2.22}$$

$$\propto p(h_t, v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1}) \tag{2.23}$$

From this joint distribution, conditioning on $v_t$ gives us $p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t})$. 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

$$v_t = B(s_t)h_t + \eta_v(s_t)$$

$$h_t = A(s_t)h_{t-1} + \eta_h(s_t)$$

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 (**??**). For each setting of the variable $s_t$, we will therefore have a *mixture* of $I \times S$ Gaussians. This is the exact calculation. Keeping all these Gaussians is expensive, since, at the next time step, we will have $S^2$ Gaussians, and in general, an exponential number of them as we progress through the forward recursion.

There are many different strategies conceivable for approximating this mixture of Gaussians. Previously we replaced the mixture with a single Gaussian which has the same mean and covariance as the mixture distribution. Here, we will replace it rather with another MOGs, albeit with (usually) a smaller number of components.

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

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

What about $p(s_t|v_{1:t})$?

$$p(s_t|v_{1:t}) \propto \sum_{s_{t-1}} p(s_t, s_{t-1}, v_t, v_{1:t-1}) \tag{2.24}$$

$$= \sum_{s_{t-1}} p(v_t|s_t, s_{t-1}, v_{1:t-1})p(s_t, s_{t-1}|v_{1:t-1}) \tag{2.25}$$

$$= \sum_{s_{t-1}} p(v_t|s_t, s_{t-1}, v_{1:t-1})p(s_t|s_{t-1})p(s_{t-1}|v_{1:t-1}) \tag{2.26}$$

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 $p(s_{t-1}|v_{1:t-1})$ is the previous prior.

# 3 The Backward Pass

Before we discuss our method for the SKF Backpass, we will look at the simpler case of the Kalman smoother – we will use the results from this to inspire our method.

### The Kalman Filter

Imagine that we have completed a forward pass, so that we have, for the KF, the filtered distributions $p(h_t|v_{1:t})$. We'll discuss how to calculate the smoother posteriors $p(h_t|v_{1:T})$, without using $\lambda$ recursions. (This is important, since the lambda messages are difficult to approximate in the SKF case).

$$p(h_t|v_{1:T}) \propto \sum_{h_{t+1}} p(v_{1:T}, h_t, h_{t+1}) \tag{3.1}$$

$$\propto \sum_{h_{t+1}} p(h_t|v_{1:T}, h_{t+1})p(h_{t+1}|v_{1:T}) \tag{3.2}$$

$$\propto \sum_{h_{t+1}} p(h_t|v_{1:t}, h_{t+1})p(h_{t+1}|v_{1:T}) \tag{3.3}$$

Let's look at the term

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

The easy way to find this distribution is to consider

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

We can work out this joint distribution in the usual manner by finding the joint mean and joint covariance. The term $p(h_t|v_{1:t})$ is known from the Forward Pass. To find the conditional distribution $p(h_t|v_{1:t}, h_{t+1})$, we use the results in section (4). From this we can easily write

$$p(h_t|v_{1:t}, h_{t+1}) \equiv h_t = \hat{A}_t h_{t+1} + \hat{m}_t + \hat{\eta}_t$$

for appropriately defined $\hat{A}_t$, $\hat{m}_t$ and $\hat{\eta}_t \sim N(0, \hat{\Sigma}_t)$. Then

$$p(h_t|v_{1:T}) = N(g_t, cov = G_t)$$

is a Gaussian distribution with mean

$$g_t \equiv \langle h_t|v_{1:T}\rangle = \hat{A}_t \langle h_{t+1}|v_{1:T}\rangle$$

and covariance

$$\left\langle \Delta h_t \Delta h_t^T|v_{1:T}\right\rangle = \hat{A}_t \left\langle \Delta h_{t+1} \Delta h_{t+1}^T|v_{1:T}\right\rangle \hat{A}_t^T + \hat{\Sigma}_t$$

Or

$$G_t = \hat{A}_t G_{t+1} \hat{A}_t^T + \hat{\Sigma}_t$$

In this way, we directly find the smoothed posterior without defining $\lambda$ messages. This will be useful when we attempt to extend this approach to SKF's, since in the above procedure, we are working only with distributions, and not conditional distributions. This procedure is equivalent to the Rauch-Tung-Striebel Kalman smoother[4].

## 3.1 The SKF case

Let's try to write a backward recursion for the (smoothed) posteriors, in a similar way to that in the Kalman Filter:

$$p(h_t, s_t|v_{1:T}) \tag{3.4}$$
$$\propto \sum_{s_{t+1}} \int_{h_{t+1}} p(h_t, h_{t+1}, s_{t+1}, s_t, v_{1:T}) \tag{3.5}$$
$$\propto \sum_{s_{t+1}} \int_{h_{t+1}} p(h_t, s_t|h_{t+1}, s_{t+1}, v_{1:T}) p(h_{t+1}, s_{t+1}, v_{1:T}) \tag{3.6}$$
$$\propto \sum_{s_{t+1}} \int_{h_{t+1}} p(h_t, s_t|h_{t+1}, s_{t+1}, v_{1:t}) p(h_{t+1}, s_{t+1}|v_{1:T}) \tag{3.7}$$

The first factor may be written

$$p(h_t, s_t|h_{t+1}, s_{t+1}, v_{1:t}) \tag{3.8}$$
$$\propto p(h_t, s_t, h_{t+1}, s_{t+1}, v_{1:t}) \tag{3.9}$$
$$\propto p(h_{t+1}, h_t, s_{t+1}, s_t|v_{1:t}) \tag{3.10}$$
$$\propto p(h_{t+1}, h_t|s_{t+1}, s_t, v_{1:t}) p(s_{t+1}, s_t|v_{1:t}) \tag{3.11}$$
$$\propto p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t}) \underbrace{p(h_{t+1}|s_{t+1}, s_t, v_{1:t}) p(s_t|s_{t+1}, v_{1:t})}_{\propto p(s_t|h_{t+1}, s_{t+1}, v_{1:t})} \tag{3.12}$$
$$= p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t}) \frac{p(h_{t+1}|s_{t+1}, s_t, v_{1:t}) p(s_t|s_{t+1}, v_{1:t})}{\sum_{s'_t} p(h_{t+1}|s_{t+1}, s'_t, v_{1:t}) p(s'_t|s_{t+1}, v_{1:t})} \tag{3.13}$$

We then make the observation that

$$p(h_{t+1}|s_{t+1}, v_{1:T}) p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t}) \tag{3.14}$$
$$= p(h_t, h_{t+1}|s_t, s_{t+1}, v_{1:T}) \tag{3.15}$$
$$= p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:T}) p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \tag{3.16}$$

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

$$p(h_t, s_t|v_{1:T}) = \sum_{s_{t+1}} p(s_{t+1}|v_{1:T}) \int_{h_{t+1}} p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t}) \frac{p(h_{t+1}|s_{t+1}, s_t, v_{1:t}) p(s_t|s_{t+1}, v_{1:t})}{\sum_{s'_t} p(h_{t+1}|s_{t+1}, s'_t, v_{1:t}) p(s'_t|s_{t+1}, v_{1:t})} p(h_{t+1}|s_{t+1}, v_{1:T}) \tag{3.17}$$

Or,

$$p(h_t, s_t|v_{1:T}) = \sum_{s_{t+1}} p(s_{t+1}|v_{1:T}) \int_{h_{t+1}} p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t}) p(s_t|h_{t+1}, s_{t+1}, v_{1:t}) p(h_{t+1}|s_{t+1}, v_{1:T}) \tag{3.18}$$

Hence

$$p(h_t, s_t|v_{1:T}) = \sum_{s_{t+1}} p(s_{t+1}|v_{1:T}) p(h_t|s_{t+1}, s_t, v_{1:T}) \int_{h_{t+1}} \frac{p(h_{t+1}|s_{t+1}, s_t, v_{1:t}) p(s_t|s_{t+1}, v_{1:t})}{\sum_{s'_t} p(h_{t+1}|s_{t+1}, s'_t, v_{1:t}) p(s'_t|s_{t+1}, v_{1:t})} p(h_{t+1}|h_t, s_t, s_{t+1}, v_{1:T}) \tag{3.19}$$

$$p(h_t, s_t|v_{1:T}) = \sum_{s_{t+1}} p(s_{t+1}|v_{1:T})p(h_t|s_{t+1}, s_t, v_{1:T})$$
$$\int_{h_{t+1}} p(s_t|h_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}|h_t, s_t, s_{t+1}, v_{1:T}) \quad (3.20)$$

What's nice about the above formula, is that we can see the role of the switch variables, and their interaction with the continuous variables. When there are no switch variables, the integral term is unity, and the method is the same as for the KF. It is potentially advantageous to work with this form since then any approximation of the integral will still give the exact results in the KF case, when there are no switch variables. The fast way to find $p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:T})$ and $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$ is to find the joint distribution equation (3.15), and then condition. We need to do this now in two stages. First we need to find the distribution

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

This is found from conditioning the joint distribution

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

This is used to define the backward equation

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

Then the joint distribution has the following mean and covariances

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

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

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

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

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

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

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

Using the conditioning method, we can then write

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

This can be used in approximation methods (see below), most obviously as a fluctuation expansion.

The only term we haven't discussed is

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

The final expression contains easily computable terms.

## 3.2 Approximating the Integral

Perhaps the simplest approximation is to replace $h_{t+1}$ by it's mean value, not just as a function of $h_t$, but as the mean of $h_t$. That is,

$$h_{t+1} \rightarrow \overrightarrow{A}(s_t, s_{t+1}) \langle h_t | s_t, s_{t+1}, v_{1:T} \rangle + \overrightarrow{m}(s_t, s_{t+1})$$

More explicitly, this is

$$h_{t+1} \rightarrow \overrightarrow{A}(s_t, s_{t+1}) \left( \overleftarrow{A}(s_t, s_{t+1}) g_{t+1}(s_{t+1}) + \overleftarrow{m}(s_t, s_{t+1}) \right) + \overrightarrow{m}(s_t, s_{t+1})$$

This procedure automatically produces a correctly normalised distribution. This mixture of Gaussians and can be easily collapsed to a single Gaussian for each state $s_t$. (Also, for future work, every thing should extend fairly easily to the case of using more than a single Gaussian to represent the mixtures). Note that this simple idea is equivalent to replacing the average over $p(h_{t+1}|h_t, s_t, s_{t+1}, v_{1:T})$ with an average with respect to $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$. I'm not sure this is a great idea. May be better to use this as just the first term in a fluctuation expansion. Of course, all kinds of other approximations may be considered. I think the variational ones would be too expensive.

A couple of comments about the above procedure: Replacing $h_{t+1}$ with $\langle h_{t+1} | s_t, s_{t+1}, v_{1:T} \rangle$ means that the integral is approximated by

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

where $z_{t+1}(s_t, s_{t+1}) \equiv \langle h_{t+1}|s_t, s_{t+1}, v_{1:T} \rangle - \langle h_{t+1}|s_t, s_{t+1}, v_{1:t} \rangle$ and $Z$ is a constant to ensure normalistion over $s_t$.

An interesting point is that, whereas as in EP, we have to divide potentials (which corresponds to subtracting the canonical parameters), here we subtract *moments*, if only the first moment in this simple approximation. This is what makes this method numerically stable. If one looked at a more complex integral approximation, such as a higher-order fluctuation expansion, then we might have the subtraction of covariance matrices.

An even simpler method than the above would have been to, in equation (3.33), approximate $p(s_t|h_{t+1}, s_{t+1}, v_{1:t})$ by $p(s_t|s_{t+1}, v_{1:t})$. This would have had the effect that the integral equation (3.20) is approximated by $p(s_t|s_{t+1}, v_{1:t})$. Our procedure above in which we replace approximate $p(h_{t+1}|h_t, s_t, s_{t+1}, v_{1:T})$ is potentially more accurate since it takes into account future information, which is neglected in the simpler approximation. Also, it is a useful starting point for more complex analytic approximations.

It is useful to put the message in the form

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

Clearly, the first term is given by (under our approximation for the integral)

$$p(s_t|v_{1:T}) = \sum_{s_{t+1}} p(s_{t+1}|v_{1:T}) \frac{p(h_{t+1} = \langle h_{t+1} \rangle | s_{t+1}, s_t, v_{1:t}) p(s_t|s_{t+1}, v_{1:t})}{\sum_{s'_t} p(h_{t+1} = \langle h_{t+1} \rangle | s_{t+1}, s'_t, v_{1:t}) p(s'_t|s_{t+1}, v_{1:t})} \tag{3.21}$$

Using the above, we can form the distribution

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

This can then be collapsed to a single Gaussian (if desired) using the usual approach.

## 3.3 Backpass using Mixtures

In the single Gaussian case, the backpass $p(h_t|s_t, v_{1:T})$ was approximated as a mixture of Gausssians, which was itself then collapsed to a single Gaussian. Here, we wish to do two things. Firstly, we will collapse $p(h_t|s_t, v_{1:T})$ to a mixture of Gaussians, $\sum_{j_t} p(j_t|s_t, v_{1:T})p(h_t|j_t, s_t, v_{1:T})$, and secondly, we will make use of the Mixture representation of our Forward messages to do this.

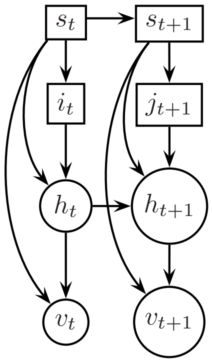


Figure 4: Structure of the backward pass for mixtures.

$$
\begin{aligned}
& p(h_t, s_t|v_{1:T}) && (3.22)\\
& \propto \sum_{j_{t+1}, s_{t+1}} \int_{h_{t+1}} p(h_t, h_{t+1}, j_{t+1}, s_{t+1}, s_t, v_{1:T}) && (3.23)\\
& \propto \sum_{j_{t+1}, s_{t+1}} \int_{h_{t+1}} p(h_t, s_t|h_{t+1}, j_{t+1}, s_{t+1}, v_{1:T})p(h_{t+1}, j_{t+1}, s_{t+1}, v_{1:T}) && (3.24)\\
& \propto \sum_{j_{t+1}, s_{t+1}} \int_{h_{t+1}} p(h_t, s_t|h_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}, j_{t+1}, s_{t+1}|v_{1:T}) && (3.25)\\
& \propto \sum_{j_{t+1}, s_{t+1}} \int_{h_{t+1}} \sum_{i_t} p(h_t, s_t, i_t|h_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}, j_{t+1}, s_{t+1}|v_{1:T}) && (3.26)\\
& \propto \sum_{j_{t+1}, s_{t+1}} \int_{h_{t+1}} \sum_{i_t} p(h_t, s_t|i_t, h_{t+1}, s_{t+1}, v_{1:t})p(i_t|h_{t+1}, s_{t+1}, v_{1:t}) && (3.27)\\
& \times p(h_{t+1}, j_{t+1}, s_{t+1}|v_{1:T}) && (3.28)
\end{aligned}
$$

The reason for introducing $i_t$ will (hopefully) become clear. The first factor may be written

$$
\begin{aligned}
& p(h_t, s_t|i_t, h_{t+1}, s_{t+1}, v_{1:t}) && (3.29)\\
& \propto p(i_t, h_t, s_t, h_{t+1}, s_{t+1}, v_{1:t}) && (3.30)\\
& \propto p(h_{t+1}, h_t, s_{t+1}, s_t|i_t, v_{1:t}) && (3.31)\\
& \propto p(h_{t+1}, h_t|s_{t+1}, s_t, i_t, v_{1:t})p(s_{t+1}, s_t|i_t, v_{1:t}) && (3.32)\\
& \propto p(h_t|h_{t+1}, s_{t+1}, s_t, i_t, v_{1:t}) \underbrace{p(h_{t+1}|s_{t+1}, s_t, i_t, v_{1:t})p(s_t|s_{t+1}, i_t, v_{1:t})}_{\propto p(s_t|h_{t+1}, s_{t+1}, i_t, v_{1:t})} && (3.33)\\
& = p(h_t|h_{t+1}, s_{t+1}, s_t, i_t, v_{1:t})p(s_t|h_{t+1}, s_{t+1}, i_t, v_{1:t}) && (3.34)
\end{aligned}
$$

We then make the observation that

$$p(h_{t+1}|j_{t+1}, s_{t+1}, v_{1:T})p(h_t|h_{t+1}, s_{t+1}, i_t, s_t, v_{1:t}) \tag{3.35}$$
$$= p(h_t, h_{t+1}|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}) \tag{3.36}$$
$$= p(h_t|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})p(h_{t+1}|h_t, i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}) \tag{3.37}$$

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

$$p(h_t, s_t|v_{1:T}) = \sum_{i_t, j_{t+1}, s_{t+1}} p(s_{t+1}|v_{1:T})p(j_{t+1}|s_{t+1}, v_{1:T})p(h_t|j_{t+1}, s_{t+1}, i_t, s_t, v_{1:T}) \tag{3.38}$$
$$\times \int_{h_{t+1}} p(s_t|h_{t+1}, s_{t+1}, i_t, v_{1:t})p(i_t|h_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}|h_t, i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}) \tag{3.39}$$
$$= \sum_{i_t, j_{t+1}, s_{t+1}} p(s_{t+1}|v_{1:T})p(j_{t+1}|s_{t+1}, v_{1:T})p(h_t|j_{t+1}, s_{t+1}, i_t, s_t, v_{1:T}) \tag{3.40}$$
$$\times \int_{h_{t+1}} p(i_t, s_t|h_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}|h_t, i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}) \tag{3.41}$$

We can calculate

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

(Later I will approximate the average of the above by replacing $h_{t+1}$ with its average value.) The fast way to find $p(h_t|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$ and $p(h_{t+1}|h_t, i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$ is to use the same trick we used in the KF. However, we need to do this now in two stages. First we need to find the distribution

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

This is found from conditioning the joint distribution

$$p(h_{t+1}, h_t|s_{t+1}, i_t, s_t, v_{1:t}) = p(h_{t+1}|h_t, i_t, s_{t+1}, v_{1:t})p(h_t|i_t, s_t, v_{1:t}) \tag{3.42}$$
$$= p(h_{t+1}|h_t, s_{t+1})p(h_t|i_t, s_t, v_{1:t}) \tag{3.43}$$

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

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

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

$$\left\langle \Delta h_t\Delta h_t^T|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T}\right\rangle = \overleftarrow{A}(i_t, s_t, s_{t+1})G_{t+1}(j_{t+1}, s_{t+1})\overleftarrow{A}^T(i_t, s_t, s_{t+1}) + \overleftarrow{\Sigma}_t(i_t, s_t, s_{t+1}) \tag{3.44}$$

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

From this, we can find easily the marginal $p(h_t|i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T})$. Using Bayes, we can reexpress the joint as

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

Using the conditioning method, we can then write

$$h_{t+1}|h_t, i_t, s_t, j_{t+1}, s_{t+1}, v_{1:T} = \overrightarrow{A}(i_t, s_t, j_{t+1}, s_{t+1})h_t + \overrightarrow{m}(i_t, s_t, j_{t+1}, s_{t+1}) + \overrightarrow{\eta}(i_t, s_t, j_{t+1}, s_{t+1}) \tag{3.45}$$

This can be used in approximation methods (see below), most obviously as a fluctuation expansion.

I will use the same approximation as before for the backpass.

$$p(h_t, s_t|v_{1:T}) \approx \sum_{i_t, j_{t+1}, s_{t+1}} p(s_{t+1}|v_{1:T})p(j_{t+1}|s_{t+1}, v_{1:T})p(i_t, s_t|\overline{h_{t+1}}, s_{t+1}, j_{t+1}, v_{1:T}) \tag{3.46}$$

$$\times\, p(h_t|j_{t+1}, s_{t+1}, i_t, s_t, v_{1:T}) \tag{3.47}$$

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

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

# A toy experiment

A simple experiment is given in fig(5), in which we use the above approximate inference procedure for a one dimensional time series.

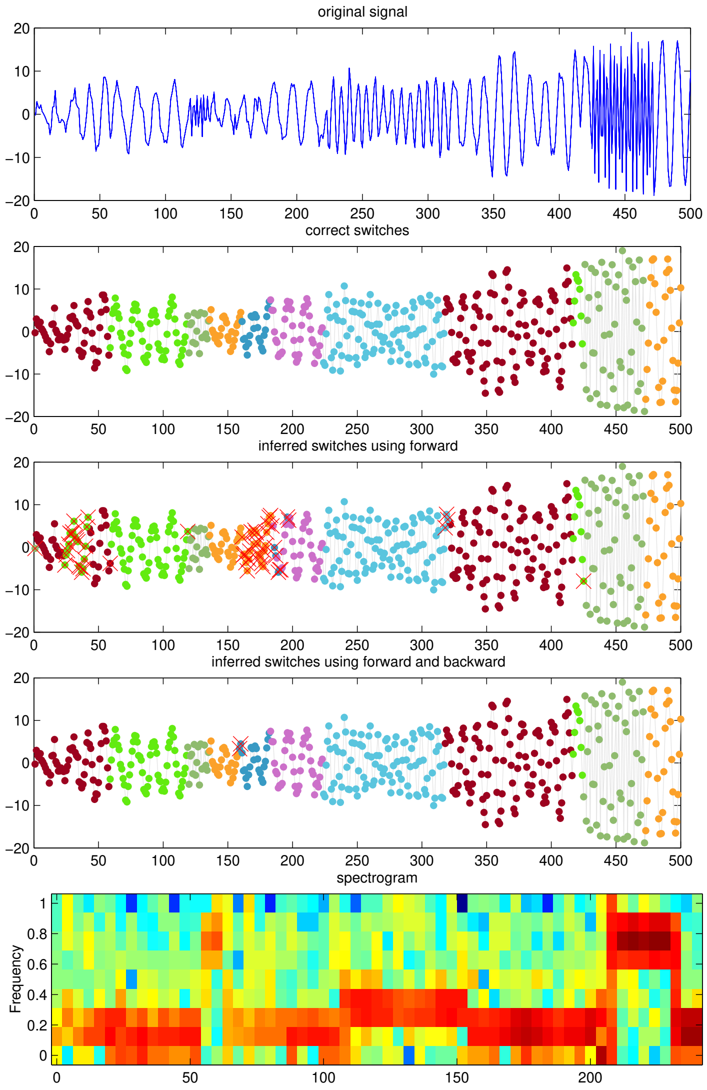


Figure 5: In the top figure is the original signal, below it the signal coloured according to the four switch states. This data was obtained by sampling from a model with four switch variables. Each dynamical regime corresponds to a noisy oscillation at a fixed frequency. We then try to infer the switch variables (using the known model that generated the data) to compute the posterior filtered estimates $p(s_t|v_{1:t})$, based on knowing the observation, but not the switch states that generated them. Estimated values for the switch variables that do not correspond to the 'correct' sample values are labelled with a $\times$. Then the posterior smoothed estimates $p(s_t|v_{1:T})$ are given. The bottom plot is the short window Fourier Transform with the frequency plotted vertically. White denotes a high energy, black low energy. The Fourier representation does not give a clear indication that at any time, a single noisy oscillator at a fixed frequency is responsible for the dynamics. However, the original correct switch variables used to generate the data are reasonably well inferred by the approximate Forward Pass method. The Backward Pass usually improves the situation considerably, although this cannot be guaranteed. The above results were obtained by projecting to a single Gaussian. No improvement using a projection to a mixture was observed.

# 4 Discussion

We have presented a new method for approximate inference in Switching linear Gaussian state space models, also known as Switching Kalman Filters. The forward pass in our method is nothing particularly original, and corresponds essentially to Assumed Density Filtering. Our new Backpass method was based on an analagous approach to 'correction' methods, such as the Rauch-Tung-Striebel method in Kalman smoothing. The method results from a simple first order fluctuation expansion, which results in the subtraction of moments. The method is easily extensible to a higher order perturbation, or other variational implementations. Furthermore, we can extend our SKF approximation method to retaining, not just a single Gaussian as an estimate of a Gaussian Mixture, but we can do this for may mixture components. This can also be done for the backpass. This avoids a major difficulty in Expectation Propagation where the division of a potential only works when we are using the exponential family. If the mixture distribution cannot be accurately represented by the exponential family, EP cannot be expected to produce reasonable results.

Unlike in Expectation Propagation, our method relies avoids message over counting by the subtraction of stabily-computed moments. In Expectation Propagation numerical instabilities arise due to avoiding over counting by dividing potentials. Curiously, our implementation using a projection to a mixture of Gaussians for both the forward and backward steps has not resulted in much improvement over projecting to a single Gaussian. Our method is relatively straightforward to implement and corresponds to a single forward and backward sweep.

# Appendix

## Finding the Conditional Gaussian from the joint

$$p(x,y) = \frac{1}{\sqrt{\det 2\pi\Sigma}} e^{-\frac{1}{2}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}^T \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{pmatrix}^{-1} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}} \tag{4.1}$$

Then

$$p(x|y) = N(mean = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y), cov = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}) \tag{4.2}$$

## Collapsing a Mixture of Gaussians to a single Gaussian

Consider a normalised ($\sum_i p_i = 1$) mixture of Gaussians distribution:

$$p(\mathbf{x}) = \sum_i p_i N(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

The mean and covariance of this distribution is

$$\boldsymbol{\mu} = \sum_i p_i \boldsymbol{\mu}_i$$

$$\boldsymbol{\Sigma} = \sum_i p_i \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T\right) - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

# References

[1] Tom Heskes and Onno 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.

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

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

[4] 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.