Expectation Correction for an augmented class of Switching Linear Gaussian Models

David Barber
IDIAP Research Institute
CH-1920 Martigny, Switzerland

IDIAP–RR 05-71

March 2005

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 05-71

Expectation Correction for an augmented class of Switching Linear Gaussian Models

David Barber
IDIAP Research Institute
CH-1920 Martigny, Switzerland

March 2005

Abstract. We consider approximate inference in a class of switching linear Gaussian State Space models which includes the switching Kalman Filter and the more general case of switch transitions dependent on the continuous hidden state. The method is a novel form of Gaussian sum smoother consisting of a single forward and backward pass, and compares favourably against a range of competing techniques, including sequential Monte Carlo and Expectation Propagation.

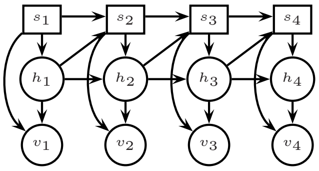

Figure 1: An augmented Switching State-Space model. Square nodes denote discrete variables, round nodes continuous variables. In switching linear Gaussian models (SKFs) links from h to s are not normally considered.

1 Switching Linear models

Switching linear Gaussian state space models are well known in many different disciplines [1, 2, 3, 4]. In these models, the observation or visible vector variable $v_t \in \mathcal{R}^V$ is linearly related to the hidden state vector $h_t \in \mathcal{R}^H$ by

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

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,

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

where $\bar{h}(s_t)$ is the mean of the transition noise, and $\Sigma^h(s_t)$ the corresponding covariance. This is a form of switching dynamics since the variable $s_t \in 1, \ldots, S$ 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 $p(s_t|h_{t-1}, s_{t-1})$. The potential dependence of s_t on the previous continuous value h_{t-1} differentiates this model from the simpler Switching Kalman Filter (SKF). An equivalent probabilistic model is (see fig(1))

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

with $p(v_t|h_t, s_t) = \mathcal{N}\left(\bar{v}(s_t) + B(s_t)h_t, \Sigma^v(s_t)\right)$ and $p(h_t|h_{t-1}, s_t) = \mathcal{N}\left(\bar{h}(s_t) + A(s_t)h_t, \Sigma^h(s_t)\right)$. The notation $x_{1:T}$ is shorthand for $x_1, \ldots, x_T$. At time $t = 1$, $p(s_1|h_0, s_0)$ simply denotes the prior $p(s_1)$. Here we address approximate inference in such models; 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 $1 \leq t \leq T$. For the SKF, the exact filtered estimate $\rho(h_t|s_t) \equiv p(h_t|s_t, v_{1:t})$ will be a mixture of Gaussians, with an exponential explosion of components with time. One useful strategy for filtering, therefore, is to project the exact $\rho(h_t|s_t)$ to a set of K Gaussians. This is a Gaussian Sum approximation [5], and is a form of Assumed Density Filtering (ADF) [6]. This approach is also taken in [3] which deals only with the special case of the SKF, and not the more general class considered here. Smoothing in [3] is approximated only for the discrete switch variables and is not based on approximating directly a principled backwards recursion. To make a Gaussian Sum approximation suitable for SKF smoothing, [4] used a two-filter method in which the dynamics of the chain are heuristically reversed. Generalised Pseudo Bayes2 (GPB2) [1] 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. This corresponds to a potentially severe loss of future information and, in general, GPB2 cannot be expected to improve much on ADF. Expectation Propagation (EP) [6, 7] corresponds to an approximate implementation of Belief Propagation and the assumption that the posterior $p(h_t|s_t, v_{1:T})$ is well approximated by a single Gaussian. Whilst a mixture approximation is possible, this would be expensive; in addition numerical instabilities need to be treated with some care[8]. EP may also in principle be applied to the more general class of models considered here, although we are unaware of any attempts to do so.

Variational methods [2] approximate the joint distribution $p(h_{1:T}, s_{1:T}|v_{1:T})$ rather than the marginal inference $p(h_t, s_t|v_{1:T})$. This puts them at a disadvantage when compared to other methods that directly approximate the marginal. Popular alternative approaches are based on sequential Monte Carlo. Whilst potentially powerful, these non-analytic methods typically suffer in high-dimensional hidden spaces since they are often based on naive importance sampling, which restricts their practical use. Implementations of Rao-Blackwellisation (see for example [9]) may not help in difficult problems where the continuous posterior is highly non-Gaussian, and we are unaware of methods that have addressed this. An important point is that in the class of models we consider, the observation distribution $p(v|h)$ is Gaussian – this greatly facilitates analytic approximation schemes, as opposed to sampling methods. Exact inference consists of an exponential number of components in time. However, in practice, due to the Markovian nature of the dynamics, we expect that the effective correlation time of the posterior will be very much shorter than the total length of the time series. Hence a much smaller effective number of components may produce a reasonable approximation, and this motivates our method.

2 Expectation Correction

Our approach to compute $p(h_t, s_t|v_{1:T})$ mirrors the Rauch-Tung-Striebel 'correction' smoother for Kalman Filters [1] which consists of a single forward pass to recursively find $p(h_t, s_t|v_{1:t})$ and a single backward recursion to correct these filtered posteriors into smoothed posteriors $p(h_t, s_t|v_{1:T})$. Specifically, our method contains three main approximations : (a) collapse of $p(h_t|s_t, s_{t-1}, v_{1:T})$ to a mixture; (b) dropping of a dependence $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$, (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 accurate when compared with the popular GPB2 method, which discards all future information; step (c) is fairly benign, and step (a) can be made more accurate by increasing the number of Gaussians.

2.1 Forward Pass (Filtering)

Our aim is to form a recursion for $p(s_t, h_t|v_{1:t})$, based on a Gaussian mixture approximation of $p(h_t|s_t, v_{1:t})$. The approach is fairly standard, and as such our description here is brief. The starting point is the exact recursion

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

where we can write the last factor as $p(h_{t-1}, |s_{t-1}, v_{1:t-1})p(s_{t-1}|v_{1:t-1})$. For each state s_{t-1}, we approximate the filtered estimate by a set of I Gaussian components:

$$p(h_{t-1}|s_{t-1}, v_{1:t-1}) \approx \sum_{i_{t-1}=1}^{I} p(h_{t-1}|i_{t-1}, s_{t-1}, v_{1:t-1}) p(i_{t-1}|s_{t-1}, v_{1:t-1}) \tag{2}$$

$p(h_{t-1}|i_{t-1}, s_{t-1}, v_{1:t-1})$ will be parameterised by mean $f(i_{t-1}, s_{t-1})$ and covariance $F(i_{t-1}, s_{t-1})$, and our recursion will take the form of update equations for these parameters. Using the approximation eq(2) in eq(1) we can form a recursion first for $p(h_t|s_t, v_{1:t})$,

$$p(h_t|s_t, v_{1:t}) \propto \sum_{s_{t-1}, i_{t-1}} p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t}) c(i_{t-1}, s_{t-1}) \tag{3}$$

where the mixture components $c(i_{t-1}, s_{t-1})$ are

$$p(v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1}) p(i_{t-1}|s_{t-1}, v_{1:t-1}) p(s_{t-1}|v_{1:t-1}) p(s_t|i_{t-1}, s_{t-1}, v_{1:t-1}) \tag{4}$$

Once computed, we collapse this mixture of $I \times S$ components to a smaller mixture, as explained later. In computing eq(4), the first three factors of are unproblematic. The final term $p(s_t|i_{t-1}, s_{t-1}, v_{1:t-1})$ is formally found from

$$p(s_t|i_{t-1}, s_{t-1}, v_{1:t-1}) = \langle p(s_t|h_{t-1}, s_{t-1}) \rangle_{p(h_{t-1}|i_{t-1}, s_{t-1}, v_{1:t-1})} \tag{5}$$

A simple approximation of the above would be to evaluate the integrand at the mean value $\overline{h_{t-1}} \equiv \langle h_{t-1}|i_{t-1}, s_{t-1}, v_{1:t-1} \rangle$. Otherwise, drawing samples from the Gaussian $p(h_{t-1}|i_{t-1}, s_{t-1}, v_{1:t-1})$ to find the average of $p(s_t|s_{t-1}, h_{t-1})$ is straightforward.

In eq(3) the term $p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t})$ can be found by conditioning the joint $p(h_t, v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1})$. For each setting of the switch variables i_{t-1}, s_{t-1}, s_t this joint distribution is a Gaussian with mean and covariance[1]

$$\Sigma_{hh} = A(s_t)F(i_{t-1}, s_{t-1})A^T(s_t) + \Sigma^h(s_t), \qquad \Sigma_{vv} = B(s_t)\Sigma_{hh}B^T(s_t) + \Sigma^v(s_t)$$
$$\Sigma_{vh} = B(s_t)F(i_{t-1}, s_{t-1}), \quad \mu_v = B(s_t)A(s_t)f(i_{t-1}, s_{t-1}), \quad \mu_h = A(s_t)f(i_{t-1}, s_{t-1})$$

It is straightforward to find from this joint distribution the conditional and marginal factors

$$p(h_t, v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1}) \qquad = p(h_t|i_{t-1}, s_{t-1}, s_t, v_{1:t})p(v_t|i_{t-1}, s_{t-1}, s_t, v_{1:t-1})$$

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

We are now in a position to calculate eq(3). For each setting of the variable s_t, we will therefore have a mixture of $I \times S$ Gaussians which we collapse to form[2]

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

In this way the new mixture coefficients $p(i_t|s_t, v_{1:t})$, $i_t \in 1, \ldots, I$ are defined. Finally, we can make a recursion for $p(s_t|v_{1:t})$ based on

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

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

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

All factors in this expression have already been computed in the recursion for $p(h_t|s_t, v_{1:t})$.

2.2 Backpass (Smoothing)

The main contribution of this paper is to find an accurate way to 'correct' the filtered expected (or marginal) estimates $p(s_t, h_t|v_{1:t})$ obtained from the Forward Pass into smoothed estimates $p(s_t, h_t|v_{1:T})$. For reasons of space we'll derive this for the case of a single Gaussian representation for both the Forward and Backpass. The smoothed posterior $p(h_t|s_t, v_{1:T})$ is in this case approximated by a Gaussian

[1] We derive this for $\bar{h}_t, \bar{v}_t \equiv 0$, to ease notation.

[2] We may collapse a mixture of Gaussians $p(x) = \sum_i p_i \mathcal{N}(x|\mu_i, \Sigma_i)$ to a single Gaussian with mean $\sum_i p_i \mu_i$ and covariance $\sum_i p_i \left(\Sigma_i + \mu_i \mu_i^T\right) - \mu\mu^T$. To collapse a mixture to a K-component mixture we retain the $K - 1$ Gaussians with the largest mixture weights – the remaining $N - K$ Gaussians are simply merged to a single Gaussian using the above method. Recursively merging the two Gaussians with the lowest mixture weights gave similar experimental performance.

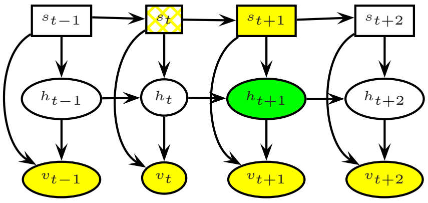

Figure 2: Our backpass approximates $p(h_{t+1}|s_{t+1}, s_t, v_{1:T})$ by $p(h_{t+1}|s_{t+1}, v_{1:T})$. Motivation for this is that s_t only influences h_{t+1} through h_t. However, h_t will most likely be heavily influenced by $v_{1:t}$, so that not knowing the state of s_t is likely to be of secondary importance. The green (darker) node is the variable we wish to find the posterior state of. The yellow (lighter shaded) nodes are variables in known states, and the hashed node a variable whose states are indeed known but assumed unknown for the approximation.

with mean $g(s_t)$ and covariance $G(s_t)$. 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 Kalman Filters [1].

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

The first factor may be decomposed $p(h_t|h_{t+1}, s_{t+1}, s_t, v_{1:t})p(s_t|h_{t+1}, s_{t+1}, v_{1:t})$. Formally, this is sufficient to define a backwards recursion directly for the smoothed estimate $p(h_t, s_t|v_{1:T})$, which contains an exponential number of mixture components since the number of mixtures increases by a factor S at each iteration. The integral in this representation needs to be approximated, for which this form of the recursion is not well suited since the variable h_{t+1} is entwined in different places. A simple, but key insight into forming a more useful recursion is the equation

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

where the r.h.s. may be written

$$\sum_{s_{t+1}} p(h_t|s_t, s_{t+1}, v_{1:T})p(s_{t+1}|v_{1:T}) \underbrace{\int_{h_{t+1}} p(s_t|h_{t+1}, s_{t+1}, v_{1:t})p(h_{t+1}|s_{t+1}, v_{1:T})}_{p(s_t|s_{t+1},v_{1:T})} \tag{6}$$

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

2.2.1 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 approximated by a method in keeping with the RTS spirit and motivates the following factorisation

$$p(h_t, h_{t+1}|s_t, s_{t+1}, v_{1:T}) = p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \tag{7}$$

The first factor on the r.h.s. of eq(7) may be found from the joint distribution

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

which itself can be found from a simple forward dynamics from the filtered estimate $p(h_t|s_t, v_{1:t})$. Then conditioning eq(8) to find $p(h_t|h_{t+1}, s_t, s_{t+1}, v_{1:t})$ effectively constitutes a reversal of the forward dynamics,

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

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 described before.

The second factor $p(h_{t+1}|s_t, s_{t+1}, v_{1:T})$ in eq(7) may cause some difficulty and is depicted in fig(2). We make the simple approximation $p(h_{t+1}|s_t, s_{t+1}, v_{1:T}) \approx p(h_{t+1}|s_{t+1}, v_{1:T})$. Crucially, contrary to the GPB2 method, we therefore retain all future and past visible information in the recursion. By combining the above results, we find that $p(h_t|s_t, s_{t+1}, v_{1:T})$ has the approximate statistics

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

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

$$p(s_t|s_{t+1}, v_{1:T}) = \langle p(s_t|h_{t+1}, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|s_{t+1}, v_{1:T})} \tag{9}$$

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

$$p(s_t|h_{t+1}, s_{t+1}, v_{1:t}) = \frac{p(h_{t+1}|s_{t+1}, s_t, v_{1:t})p(s_t, s_{t+1}|v_{1:t})}{\sum_{s'_t} p(h_{t+1}|s_{t+1}, s'_t, v_{1:t})p(s'_t, s_{t+1}|v_{1:t})} \tag{10}$$

where $p(s_t, s_{t+1}|v_{1:t}) = p(s_{t+1}|s_t, v_{1:t})p(s_t|v_{1:t})$, which is straightforward to find since $p(s_{t+1}|s_t, v_{1:t})$ occurs in the Forward Pass eq(5). In eq(9), $p(h_{t+1}|s_{t+1}, s_t, v_{1:t})$ is found by marginalising eq(8).

For the average eq(9), replacing h_{t+1} by its mean gives a fast approximation,

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

where $z_{t+1}(s_t, s_{t+1}) \equiv \langle h_{t+1}|s_{t+1}, v_{1:T}\rangle - \langle h_{t+1}|s_t, s_{t+1}, v_{1:t}\rangle$ and Z ensures normalisation over s_t. $\Sigma(s_t, s_{t+1}|v_{1:t})$ is the Forward pass filtered covariance of h_{t+1} given s_t, s_{t+1} and the observations $v_{1:t}$. We also approximate the average by sampling Gaussians from $p(h_{t+1}|s_{t+1}, v_{1:T})$, which has the advantage that covariance information is used.

Finally, we can compute the smoothed recursion 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 eq(6) by conditioning on s_t to form the mixture

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

which may be collapsed to a single Gaussian (or mixture if desired). The term $p(s_t|v_{1:T})$ is trivially given by integrating the approximation of eq(6) over h_t. Briefly, for the case of using mixtures in the Forward and Backpass, we begin with the central equation

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

$d(i_t, j_{t+1}, s_{t+1}) = p(s_{t+1}|v_{1:T})p(j_{t+1}|s_{t+1}, v_{1:T})\,\langle p(i_t, s_t|h_{t+1}, s_{t+1}, v_{1:t})\rangle_{p(h_{t+1}|j_{t+1}, s_{t+1}, v_{1:T})}$ One then proceeds in a similar way to the case of a single Gaussian[8].

3 Experiments

We would like to test our Expectation Correction (EC) smoothing method in a problem with a reasonably long temporal sequence, T. We therefore sample a hidden state s_1 and h_1 from the prior, and then a visible observation v_1. Then, sequentially, we generate hidden and visible states for the next time step. The task for 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 $p(h_t|s_t, v_{1:T})$ and $p(s_t|v_{1:T})$. Since the exact computation is exponential in T, a formally exact evaluation of the method is infeasible. A simple performance measure therefore is to assume that the original sample states $s_{1:T}$ are the 'correct' inferences, and compare how our most probable posterior smoothed estimates $\arg\max_{s_t} p(s_t|v_{1:T})$ compare with the assumed correct sample ${s_t}^3$. We look at two sets of experiments, one for the standard switching linear Gaussian model (SKF) with transition $p(s_t|s_{t-1})$ and the other for the augmented model with transitions $p(s_t|h_{t-1}, s_{t-1})$. To make the experiments challenging, we chose large hidden dimension H, small hidden noise Σ^h and large visible noise Σ^v. Such settings encourage multi-modal filtered posteriors since the continuous hidden trajectories will most likely not overlap, and there is not enough visible information in previous time points to disambiguate the trajectories. EC gives good performance in 'easier' situations where existing methods also perform well[8].

SKF experiments

We chose experimental conditions that from the viewpoint of classical signal processing are difficult with changes in the switches occurring at a much higher rate than the typical frequencies in the signal. We compared methods using a single Gaussian, and methods using multiple Gaussians. The number of Gaussians used was set to $2 \times S$ throughout. In all experiments 500 particles were used for the Rao-Blackwellised Particle Filter. In fig(3)(left), we plot results on time series of length $T = 100$ with $S = 2$ switch states. Throughout, $V = 1$ (scalar observations) with zero output bias. The 'Smoothed Single' method is EC with a single Gaussian. For the multiple Gaussian methods, $I = J = 4$ Gaussians were used since this gave reasonable performance at modest computational expense. Our implementation of EP[8] was run to convergence. Using partly MATLAB notation, $A(s) = 0.9999 * \mathrm{orth}(\mathrm{randn}(\mathrm{H}, \mathrm{H}))$, $B(s) = \mathrm{randn}(\mathrm{V}, \mathrm{H})$, $\bar{v}_t \equiv 0$, $\bar{h}_1 = 10 * \mathrm{randn}(\mathrm{H}, 1)$, $\bar{h}_{t>1} = 0$, $\Sigma^h_1 = I_H$, $p_1 = \text{uniform}$, $H = 30$, $\Sigma^v = 30I_V$,$\Sigma^h = 0.01I_H$, $p(s_{t+1}|s_t) \propto 1_{S\times S}$. For EC we use the mean approximation for the intractable averages. Slightly better results are achieved using the sampling alternative.

EC with a small number of mixture components, apart from a small number of errors, performs very well. The Rao-Blackwellised PF (which should in principle cope well with large H) doesn't help much here since we used the implementation in [9] which makes the assumption that a single Gaussian adequately describes $p(h_t|s_t, v_{1:t})$. In our experiment, any method which does not deal with multi-modality of the posterior will perform poorly. Filtering can work reasonably well (so Particle Filtering should, in principle, be made to work better) provided they deal with multi-modality. However, our results suggest that smoothing can improve the performance significantly.

Augmented switching model

In fig(3)(right), we chose a simple two state $S = 2$ transition distribution $p(s_{t+1} = 1|s_t, h_t) = \sigma\left(h_t^T w(s_t)\right)$, where $\sigma(x) \equiv 1/(1 + e^{-x})$. Some care needs to be taken to make a model so that even exact inference would produce posterior switches close to the sampled switches. If the switch variables s_{t+1} can change wildly, which is possible given the above formula, essentially no information is left in the signal for any inference method to produce reasonable results. We therefore set $w(s_t)$ to a zero vector except for the first two components, which are independently sampled from a zero mean Gaussian with standard deviation 5. For each of the two switch states, s, we have a transition matrix

[3] We could also consider performance measures on the accuracy of $p(h_t|s_t, v_{1:T})$. However, we prefer to look at the switch variables alone since they should provide a more robust inference task.

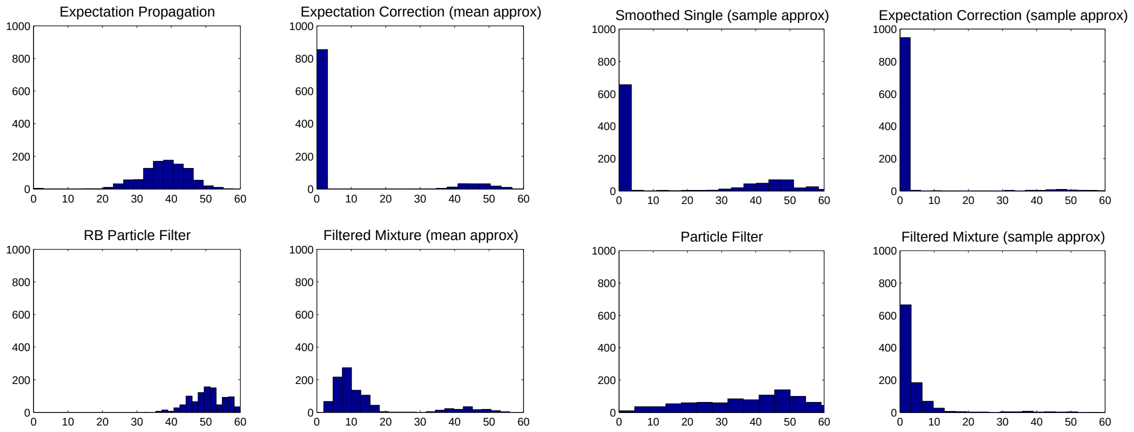

Figure 3: The number of errors in estimating a binary switch $p(s_t|v_{1:T})$ over a time series of length $T = 100$. Hence 50 errors corresponds to random guessing. Plotted are histograms of the errors are over 1000 experiments. Left: SKF results. Right: Augmented SKF results. Here we used 1000 samples to approximate equations (9,5).

$A(s)$, which we set to be block diagonal. The first 2×2 block is set to $0.9999 R_\theta$ where R_θ is a 2×2 rotation matrix with angle θ chosen uniformly from 0 to 1 radians. This means that s_{t+1} is dependent on the first two components of h_t which are rotating at a restricted rate. The remaining $H-2 \times H-2$ block of $A(s)$ is chosen as (using MATLAB notation) $0.9999{*}\text{orth}(\text{rand}(\text{H}-2))$, which means a scaled randomly chosen orthogonal matrix. Throughout, $S = 2$, $V = 1$, $H = 30$, $T = 100$, with zero output bias. For the multiple Gaussian methods, $I = J = 4$ Gaussians were used. Using partly MATLAB notation, $B(s) = \text{randn}(\text{V}, \text{H})$, $\bar{v}_t \equiv 0$, $\bar{h}_1 = 10 * \text{randn}(\text{H}, 1)$, $\bar{h}_{t>1} = 0$, $\Sigma^h_1 = I_H$, $p_1 = \text{uniform}$. $\Sigma^v = 30 I_V$,$\Sigma^h = 0.1 I_H$. In fig(3) we compare EC only against Particle Filters using 1000 particles since other methods would require specialised and novel implementations. The 'Smoothed Single' method is EC with a single Gaussian. In the Filtered method, $I = 4$ Gaussians were used, and for EC, $I = J = 4$ Gaussians were used. Again, EC performs very well. The Particle Filter most likely failed since the hidden dimension is too high to be explored well with only 1000 particles.

4 Discussion

We have presented a method that can be used for approximate inference in an augmented class of switching dynamical models with Gaussian noise. Our method is numerically stable and has time complexity $O(S^2 IJK)$ where S are the number of switch states, I and J are the number of Gaussians used in the Forward and Backward passes, and K is the time to compute the exact Kalman smoother for the system with a single switch state (roughly $T(V^2+H^3)$). Our approximation is based on the idea that the exact inference will consist of an exponentially large number of mixture components although, due to the exponential forgetting which commonly occurs in Markovian models, a finite number of mixture components may provide a good approximation. In systems with very long correlation times our method may require too many mixture components to be practical, although we are unaware of other techniques that would also cope well in this case. A key benefit of Expectation Correction is the use of mixtures which facilitates approximations when the switch depends on the continuous hidden variable in a non-trivial way. MATLAB software for Expectation Correction for this augmented class of Switching Linear Gaussian models is at `http://www.idiap.ch/∼barber/ecskf.zip` and fuller details including pseudocode are to be found in [8].

Acknowledgements: I would like to thank Bertrand Mesot for help with the simulations.

References

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

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

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

[4] 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.

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

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

[7] 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.

[8] D. Barber and B. Mesot. Construction and comparison of approximations for Switching Linear Gaussian State Space Models. Technical Report 05:06, IDIAP Research Institute, February 2005.

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