Switch-Reset Models : Exact and Approximate Inference

Chris Bracegirdle **David Barber**
University College London, Gower Street, London UK
{c.bracegirdle,d.barber}@cs.ucl.ac.uk

Abstract

Reset models are constrained switching latent Markov models in which the dynamics either continues according to a standard model, or the latent variable is resampled. We consider exact marginal inference in this class of models and their extension, the switch-reset models. A further convenient class of conjugate-exponential reset models is also discussed. For a length T time-series, exact filtering scales with T^2 and smoothing T^3. We discuss approximate filtering and smoothing routines that scale linearly with T. Applications are given to change-point models and reset linear dynamical systems.

1 LATENT MARKOV MODELS

For a time-series of observations $y_{1:T}$ and latent variables $x_{1:T}$, a latent Markov model defines a joint distribution

$$p(y_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(y_t|x_t)p(x_t|x_{t-1})$$

where $x_0 = \emptyset$. Due to the Markov structure, fig(1a), marginal inference in these well known models can be carried out using the classical $p(x_t, y_{1:t}) \equiv \alpha(x_t)$, $p(y_{t+1:T}|x_t) \equiv \beta(x_t)$, and $p(x_t|y_{1:T}) \equiv \gamma(x_t)$ message passing recursions (Barber, 2011)

$$\alpha(x_{t+1}) = p(y_{t+1}|x_{t+1}) \int_{x_t} p(x_{t+1}|x_t)\alpha(x_t) \quad (1.1)$$

$$\beta(x_{t-1}) = \int_{x_t} p(y_t|x_t)p(x_t|x_{t-1})\beta(x_t) \quad (1.2)$$

$$\gamma(x_t) = \int_{x_{t+1}} p(x_t|x_{t+1}, y_{1:t})\gamma(x_{t+1}) \quad (1.3)$$

where $\gamma(x_t) \propto \alpha(x_t)\beta(x_t)$ and $\gamma(x_T) \propto \alpha(x_T)$. These recursions hold more generally on replacing integration over any discrete elements of x by summation.

Writing $x_t = (h_t, s_t)$ for continuous h_t and discrete s_t, we identify a switching latent Markov model, fig(1b):

$$p(y_{1:T}, h_{1:T}, s_{1:T}) = \prod_{t=1}^{T} p(h_t|h_{t-1}, s_t)p(y_t|h_t, s_t)p(s_t|s_{t-1}) \quad (1.4)$$

in which we can deal with discontinuous jumps in the continuous latent state h_t by using a discrete 'switch' variable s_t. These models are also called 'conditional Markov models', 'jump Markov models', 'switching models', and 'changepoint models'.

Whilst these switching models are attractive and potentially powerful, they suffer from a well known computational difficulty: marginal inference of quantities such as $p(h_t|y_{1:t})$ scales with $O(S^t)$ due to the messages in the corresponding propagation algorithm (the analogue of the α-β recursions equation (1.3) applied to the variable $x_t \equiv (h_t, s_t)$) being mixtures with a number of components that grows exponentially with time t. A number of approximations to the exact posterior have been proposed to overcome the computational expense, including Particle Filtering (Doucet et al., 2000, 2001), Assumed Density Filtering (Alspach and Sorenson, 1972; Boyen and Koller, 1998), Expectation Propagation (Minka, 2001), Kim's Method (Kim, 1994; Kim and Nelson, 1999), Gibbs sampling (Carter and Kohn, 1996), and Expectation Correction (Barber, 2006). Such approximations work largely by approximating a mixture with fewer components, and were shown by Minka (2005) to correspond to various approaches to divergence minimisation.

An alternative, computationally simpler model is obtained by constraining the switch variable to have the effect of cutting temporal dependence[1]. We define a

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

[1] Some refer to equation (1.6) as a 'changepoint' model whilst others use this terminology for any switching latent Markov model of the form equation (1.4). Others refer

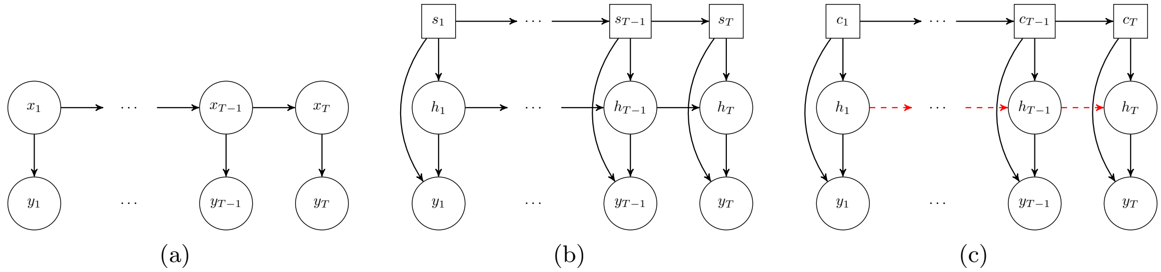

Figure 1: (a) Generic latent Markov model. (b) Conditional independence assumptions of a switching latent Markov model. The discrete switch $s_t \in \{1, \ldots, S\}$ selects from a set of S distinct transition and emission distributions. (c) Conditional independence assumptions of a reset latent Markov model. The binary reset variable c_t indicates whether the standard dynamics continues, $p^0(h_t|h_{t-1})$ $(c_t = 0)$ or whether the latent variable h_t is redrawn from the reset distribution $p^1(h_t)$ $(c_t = 1)$.

reset variable $c_t \in \{0, 1\}$ with Markov transition

$$p(c_t = j|c_{t-1} = i) = \tau_{j|i}, \qquad i, j \in \{0, 1\} \quad (1.5)$$

The continuous latent variable then transitions as

$$p(h_t|h_{t-1}, c_t) = \begin{cases} p^0(h_t|h_{t-1}) & c_t = 0 \\ p^1(h_t) & c_t = 1 \end{cases} \quad (1.6)$$

In this model, the latent binary variable c_t selects one of only two possible dynamics: either a continuation along the default dynamics $p^0(h_t|h_{t-1})$, or a 'reset' of the latent variable, drawing from the reset distribution $p^1(h_t)$. This reset process cuts the dependence on past states, see fig(1c). Finally, the reset model is completed by specifying an emission distribution[2]

$$p(y_t|h_t, c_t) = \begin{cases} p^0(y_t|h_t) & c_t = 0 \\ p^1(y_t|h_t) & c_t = 1 \end{cases} \quad (1.7)$$

For the reset model, it is well appreciated that filtered marginal inference $p(h_t, c_t|y_{1:t})$ scales as $O\left(t^2\right)$ (see for example Fearnhead and Liu (2007) and Barber and Cemgil (2010)), and smoothed marginal inference $p(h_t, c_t|y_{1:T})$ can be achieved in $O\left(T^3\right)$ time. Whilst this is a great saving from the exponential complexity of the switching model, cubic complexity is still prohibitive for large T and approximations may be required.

Our contribution is to introduce an exact, numerically stable correction smoothing method for reset models, in addition to demonstrating a fast and accurate linear-time approximation. We also consider an extension, the switch-reset model, which is able to model switching between a set of S continuous latent Markov models, but for which inference remains tractable.

2 RESET MODEL INFERENCE

A classical approach to deriving smoothing for the reset model is based on the α-β recursion. This has the advantage of being straightforward. However, for models such as the reset LDS, numerical stability issues are known to arise. In addition, it is unclear how best to form an approximation based on the α-β method. We first review the α-β approach.

2.1 α-β Smoothing

By writing

$$p(h_t, c_t, y_{1:T}) = \underbrace{p(h_t, c_t, y_{1:t})}_{\alpha(h_t, c_t)} \underbrace{p(y_{t+1:T}|h_t, c_t, \cancel{y_{1:t}})}_{\beta(h_t, c_t)}$$

we consider calculating the two components. The forward α message is standard and recursively calculated using equation (1.1) for the variable $x_t = (h_t, c_t)$. By defining[3]

$$\alpha(h_t, c_t) = \begin{cases} \alpha^0(h_t) & c_t = 0 \\ \alpha^1(h_t) & c_t = 1 \end{cases}$$

and $\alpha(c_t) = \int_{h_t} \alpha(h_t, c_t)$, we can identify two cases:

$$\alpha^0(h_{t+1}) = \tau_{0|0} p^0(y_{t+1}|h_{t+1}) \int_{h_t} p^0(h_{t+1}|h_t)\alpha^0(h_t)$$
$$+ \tau_{0|1} p^0(y_{t+1}|h_{t+1}) \int_{h_t} p^0(h_{t+1}|h_t)\alpha^1(h_t)$$
$$\alpha^1(h_{t+1}) = p^1(y_{t+1}|h_{t+1}) p^1(h_{t+1}) \sum_{c_t} \tau_{1|c_t} \alpha(c_t)$$

From these recursions, we see that the number of components in α grows linearly with time, making for an $O\left(t^2\right)$ computation for exact filtering.

to the piecewise reset model, section(5) as a 'changepoint' model. For this reason, in an attempt to avoid confusion, we refer to equation (1.6) as a reset model, a term which we feel also better reflects the assumptions of the model.

[2] Note that it is straightforward to include dependence on past observations $p(y_t|h_t, c_t, y_{1:t-1})$ if desired since these do not change the structure of the recursions.

[3] We will use component-conditional notation for these messages $\alpha(h_t, c_t) = \alpha(h_t|c_t)\alpha(c_t)$, which defines $\alpha(h_t|c_t) = p(h_t|c_t, y_{1:t})$ and $\alpha(c_t) = p(c_t, y_{1:t})$.

The backward β message $\beta(h_t, c_t) = p(y_{t+1:T}|h_t, c_t)$, is also calculated recursively using equation (1.2) as

$$\begin{aligned}
&\beta(h_{t-1}, c_{t-1}) \\
&\quad = \sum_{c_t} \tau_{c_t|c_{t-1}} \int_{h_t} p(y_t|h_t, c_t) p(h_t|h_{t-1}, c_t) \beta(h_t, c_t) \\
&\quad = \tau_{0|c_{t-1}} \underbrace{\int_{h_t} p^0(y_t|h_t) p^0(h_t|h_{t-1}) \beta(h_t, c_t = 0)}_{\beta^0(h_{t-1})} \\
&\qquad + \tau_{1|c_{t-1}} \underbrace{\int_{h_t} p^1(y_t|h_t) p^1(h_t) \beta(h_t, c_t = 1)}_{\beta^1_{t-1}}
\end{aligned}$$

where we have written

$$\beta(h_{t-1}, c_{t-1}) = \tau_{0|c_{t-1}} \beta^0(h_{t-1}) + \tau_{1|c_{t-1}} \beta^1_{t-1}$$

The recursions for these components are:

$$\begin{aligned}
\beta^0(h_{t-1}) &= \int_{h_t} p^0(y_t|h_t) p^0(h_t|h_{t-1}) \left[\tau_{0|0} \beta^0(h_t) + \tau_{1|0} \beta^1_t\right] \\
\beta^1_{t-1} &= \int_{h_t} p^1(y_t|h_t) p^1(h_t) \left[\tau_{0|1} \beta^0(h_t) + \tau_{1|1} \beta^1_t\right]
\end{aligned}$$

The posterior $p(h_t, c_t|y_{1:T}) \propto \alpha(h_t, c_t)\beta(h_t, c_t)$ is then a mixture of $(t+1) \times (T-t+1)$ components, and the algorithm scales as $O\left(T^3\right)$ to compute all the smoothed marginal posteriors. For large T, this can be expensive.

An obvious way to form an approximation is to drop components from either the α or β messages, or both. Dropping components from α is natural (since $\alpha(h_t, c_t)$ is a distribution in h_t, c_t). It is less natural to form an approximation by dropping β components since the β messages are not distributions—usually it is only their interaction with the α message that is of ultimate interest. We will discuss ways to achieve this in section(4). Use of the β message approach is also known to cause numerical instability in important models of interest, in particular the linear dynamical system (Verhaegen and Van Dooren, 2002). This motivates the desire to find a γ, 'correction smoother' recursion.

2.2 α-γ Smoothing

Considering the standard γ correction smoother derivation, equation (1.3), we may begin

$$\begin{aligned}
\gamma(h_{t-1}, c_{t-1}) &= p(h_{t-1}, c_{t-1}|y_{1:T}) \\
&= \sum_{c_t} \int_{h_t} p(h_{t-1}, c_{t-1}|h_t, c_t, y_{1:t-1}) \gamma(h_t, c_t)
\end{aligned}$$

The naïve approach is then to write

$$p(h_{t-1}, c_{t-1}|h_t, c_t, y_{1:t-1}, \cancel{y_t}) = \frac{p(h_t, c_t, h_{t-1}, c_{t-1}|y_{1:t})}{p(h_t, c_t|y_{1:t})}$$

However, the filtered distribution in the denominator $p(h_t, c_t|y_{1:t})$ is a mixture distribution. This is inconvenient since we cannot represent in closed form the result of the division of a mixture by another mixture. This means that using a γ recursion is not directly accessible for this model. However, by considering an equivalent model, it is possible to perform γ smoothing.

2.3 $\tilde{\alpha}$-$\tilde{\gamma}$ Run-Length Smoothing

We may define an equivalent reset model based on the 'run length', $\rho_t \geq 0$, which counts the number of steps since the last reset (Adams and MacKay, 2007; Fearnhead and Liu, 2007):

$$\rho_t = \begin{cases} 0 & c_t = 1 \\ \rho_{t-1} + 1 & c_t = 0 \end{cases} \tag{2.1}$$

and $c_t = \mathbb{I}\left[\rho_t = 0\right]$. Formally, one can write a Markov transition on the run-length defined by

$$p(\rho_t|\rho_{t-1}) = \begin{cases} \tau_{1|1} & \rho_{t-1} = 0, \rho_t = 0 \\ \tau_{1|0} & \rho_{t-1} > 0, \rho_t = 0 \\ \tau_{0|1} & \rho_{t-1} = 0, \rho_t = 1 \\ \tau_{0|0} & \rho_{t-1} > 0, \rho_t = \rho_{t-1} + 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

and a corresponding latent Markov model

$$p(h_t|h_{t-1}, \rho_t) = \begin{cases} p^0(h_t|h_{t-1}) & \rho_t > 0 \\ p^1(h_t) & \rho_t = 0 \end{cases} \tag{2.3}$$

Finally

$$p(y_t|h_t, \rho_t) = \begin{cases} p^0(y_t|h_t) & \rho_t > 0 \\ p^1(y_t|h_t) & \rho_t = 0 \end{cases}$$

This model is then formally equivalent to the reset model defined by equations (1.5, 1.6, 1.7). Since this is a latent Markov model, we can apply the standard filtering and smoothing recursions:

$$\begin{aligned}
\tilde{\alpha}(h_t, \rho_t) = p(y_t|h_t, \rho_t) &\sum_{\rho_{t-1}} p(\rho_t|\rho_{t-1}) \\
&\times \int_{h_{t-1}} p(h_t|h_{t-1}, \rho_t) \tilde{\alpha}(h_{t-1}, \rho_{t-1})
\end{aligned}$$

We can again distinguish two cases:

$$\begin{aligned}
\tilde{\alpha}(h_t, \rho_t = 0) &= p^1(y_t|h_t) p^1(h_t) \\
&\quad \times \sum_{\rho_{t-1}} p(\rho_t = 0|\rho_{t-1}) \tilde{\alpha}(\rho_{t-1})
\end{aligned} \tag{2.4}$$

$$\begin{aligned}
\tilde{\alpha}(h_t, \rho_t > 0) &= p^0(y_t|h_t) p(\rho_t|\rho_{t-1} = \rho_t - 1) \\
&\times \int_{h_{t-1}} p^0(h_t|h_{t-1}) \tilde{\alpha}(h_{t-1}, \rho_{t-1} = \rho_t - 1)
\end{aligned} \tag{2.5}$$

In this case the $\tilde{\alpha}$ messages are therefore not mixtures, but single-component distributions. The filtered posterior in the original reset model is obtained from

$$\alpha(h_t, c_t) = \begin{cases} \tilde{\alpha}(h_t, \rho_t = 0) & c_t = 1 \\ \sum_{\rho_t > 0} \tilde{\alpha}(h_t, \rho_t) & c_t = 0 \end{cases}$$

The run-length gives a natural interpretation of the components in the α message, namely that the components of the $\alpha(h_t, c_t)$ message are in fact simply the run-length components themselves.

Since the $\tilde{\alpha}$ messages are single components, one may implement the standard correction approach for smoothing on this redefined model:

$$\tilde{\gamma}(h_{t-1}, \rho_{t-1}) = \sum_{\rho_t} \int_{h_t} p(h_{t-1}, \rho_{t-1} | h_t, \rho_t, y_{1:t-1}) \tilde{\gamma}(h_t, \rho_t)$$

$$= \sum_{\rho_t} p(\rho_{t-1}|\rho_t, y_{1:t-1}) \underbrace{\int_{h_t} \frac{p(h_t|h_{t-1}, \rho_t)\tilde{\alpha}(h_{t-1}|\rho_{t-1})}{p(h_t|\rho_t, y_{1:t-1})}}_{\text{dynamics reversal}} \tilde{\gamma}(h_t, \rho_t)$$

where $p(\rho_{t-1}|\rho_t, y_{1:t-1}) \propto p(\rho_t|\rho_{t-1})\tilde{\alpha}(\rho_{t-1})$. Since $\tilde{\alpha}(h_{t-1}|\rho_{t-1})$ is a single component, the 'dynamics reversal' is a single component, and no numerical difficulties arise. Similarly to the above,

$$\gamma(h_t, c_t) = \begin{cases} \tilde{\gamma}(h_t, \rho_t = 0) & c_t = 1 \\ \sum_{\rho_t > 0} \tilde{\gamma}(h_t, \rho_t) & c_t = 0 \end{cases} \qquad (2.6)$$

The resulting $\tilde{\alpha}$-$\tilde{\gamma}$ recursion provides a numerically stable way to perform smoothed inference in reset models since both the $\tilde{\alpha}$ and $\tilde{\gamma}$ messages are distributions.

Furthermore, a simple approximate smoothing algorithm is available based on dropping components from $\tilde{\alpha}$ and subsequently from $\tilde{\gamma}$. Simple schemes such as dropping low weight components can be very effective in this case since the weight of the component is directly related to its contribution to the posterior distribution.

2.4 Bracket Smoothing

Insight into the above $\tilde{\gamma}$ recursion can be obtained by introducing the index ς to correspond to the number of observation points until the next reset. We will characterise this index as $\varsigma_t \in \{1, \ldots, T-t+1\}$, where $\varsigma_t = T - t + 1$ corresponds to there being no reset in the sequence after observation point t. The forward run-length $\rho_t \in \{0, \ldots, t\}$ at observation t corresponds to the number of observation points since the last reset. We then index the smoothed posterior[4] $p(h_t, \rho_t, \varsigma_t|y_{1:T}) = p(h_t|\rho_t, \varsigma_t, y_{1:T})p(\rho_t, \varsigma_t|y_{1:T})$.

The smoothed partition posterior $p(\rho_t, \varsigma_t|y_{1:T})$ can then be calculated by a simple backward recursion, noting that in the no-reset case $\varsigma_t > 1$,

$$p(\rho_t, \varsigma_t|y_{1:T}) = p(\rho_{t+1} = \rho_t + 1, \varsigma_{t+1} = \varsigma_t - 1|y_{1:T}) \qquad (2.7)$$

In the reset case $\varsigma_t = 1 \Leftrightarrow \rho_{t+1} = 0$, so

$$p(\rho_t, \varsigma_t = 1|y_{1:T}) = p(\rho_t, \rho_{t+1} = 0|y_{1:T})$$

$$= p(\rho_t|\rho_{t+1} = 0, y_{1:t}) \sum_{\varsigma_{t+1}} p(\rho_{t+1} = 0, \varsigma_{t+1}|y_{1:T}) \qquad (2.8)$$

since $\rho_t \perp\!\!\!\perp y_{t+1:T}|\,\rho_{t+1} = 0$. Then

$$p(\rho_t|\rho_{t+1} = 0, y_{1:t}) \propto p(\rho_{t+1} = 0|\rho_t)p(\rho_t|y_{1:t})$$

and $p(\rho_t|y_{1:t}) \propto \tilde{\alpha}(\rho_t)$. These recursions enable one to fully compute the discrete component $p(\rho_t, \varsigma_t|y_{1:T})$.

Reset points partition the sequence, so conditioning on ρ_t, ς_t simplifies the model to use only standard dynamics p^0 on the 'bracket' $y_{\rho_t,\varsigma_t} \equiv y_{t-\rho_t:t+\varsigma_t-1}$. Smoothing for the joint is then obtained using

$$\underbrace{p(h_t, \rho_t, \varsigma_t|y_{1:T})}_{\tilde{\gamma}(h_t, \rho_t, \varsigma_t)} = \underbrace{p(h_t|\rho_t, \varsigma_t, y_{\rho_t,\varsigma_t})}_{\tilde{\gamma}(h_t|\rho_t, \varsigma_t)} \underbrace{p(\rho_t, \varsigma_t|y_{1:T})}_{\tilde{\gamma}(\rho_t, \varsigma_t)}$$

For $p(h_t|\rho_t, \varsigma_t, y_{\rho_t,\varsigma_t})$ we may run any smoothing routine with the dynamics p^0 on the bracket y_{ρ_t,ς_t}, with

$$\tilde{\gamma}(h_t|\rho_t, \varsigma_t > 1) = \int_{h_{t+1}} p(h_t|h_{t+1}, c_{t+1} = 0)$$

$$\times\, \tilde{\gamma}(h_{t+1}|\rho_{t+1} = \rho_t + 1, \varsigma_{t+1} = \varsigma_t - 1) \qquad (2.9)$$

noting that $\tilde{\gamma}(h_t|\rho_t, \varsigma_t = 1) = \tilde{\alpha}(h_t|\rho_t)$.

Finally, these messages enable us to perform smoothing for the original reset model by appealing to equation (2.6) after marginalising the future reset index ς_t.

2.5 $\tilde{\beta}$ Recursion

It is also useful to index the β recursion with the ς indexing variable, and the recursions become

$$\tilde{\beta}(h_{t-1}, c_{t-1}, \varsigma_{t-1}) = \begin{cases} \tau_{1|c_{t-1}} \tilde{\beta}^1_{t-1} & \varsigma_{t-1} = 1 \\ \tau_{0|c_{t-1}} \tilde{\beta}^0(h_{t-1}, \varsigma_{t-1}) & \varsigma_{t-1} > 1 \end{cases}$$

$$\tilde{\beta}^1_{t-1} = \int_{h_t} p^1(y_t|h_t)p^1(h_t) \sum_{\varsigma_t} \tilde{\beta}(h_t, \rho_t = 1, \varsigma_t)$$

$$\tilde{\beta}^0(h_{t-1}, \varsigma_{t-1}) = \int_{h_t} \begin{array}{l} p^0(y_t|h_t)p^0(h_t|h_{t-1}) \\ \times \tilde{\beta}(h_t, \rho_t = 0, \varsigma_t = \varsigma_{t-1} - 1) \end{array}$$

We can then combine any combination of α or $\tilde{\alpha}$ with β or $\tilde{\beta}$; since each such message features a single component, this is useful for motivating approximations.

[4] The component $p(h_t|\rho_t, \varsigma_t, y_{1:T})$ describes the distribution of h_t given that (i) the previous reset occurred ρ_t time-steps ago (or there has not been a reset prior to time t if $\rho_t = t$); and (ii) the next reset occurs ς_t time-steps in the future (or there is no reset after time t if $\varsigma_t = T - t + 1$). This is equivalent to asserting that the previous reset occurred at time $t - \rho_t$ (or there was no previous reset if $t - \rho_t < 1$) and that the next reset occurs at time $t + \varsigma_t$ (or there is no future reset if $t + \varsigma_t > T$).

3 THE RESET LDS

The latent linear dynamical system (LDS) is defined by a latent Markov model on vectors which update according to a linear Gaussian transition and emission. For this model, the well known Kalman α filtering (Kalman, 1960) and γ smoothing (Rauch, Tung, and Striebel, 1965) are available. The corresponding α update LDSForward and γ update LDSBackward are provided in the supplementary material. The reset LDS is defined by:

$$p(\mathbf{h}_t|\mathbf{h}_{t-1},c_t) = \begin{cases} \mathcal{N}\left(\mathbf{h}_t|\mathbf{A}^0\mathbf{h}_{t-1}+\bar{\mathbf{h}}^0,\mathbf{Q}^0\right) & c_t=0 \\ \mathcal{N}\left(\mathbf{h}_t|\bar{\mathbf{h}}^1,\mathbf{Q}^1\right) & c_t=1 \end{cases}$$

$$p(\mathbf{y}_t|\mathbf{h}_t,c_t) = \begin{cases} \mathcal{N}\left(\mathbf{y}_t|\mathbf{B}^0\mathbf{h}_t+\bar{\mathbf{y}}^0,\mathbf{R}^0\right) & c_t=0 \\ \mathcal{N}\left(\mathbf{y}_t|\mathbf{B}^1\mathbf{h}_t+\bar{\mathbf{y}}^1,\mathbf{R}^1\right) & c_t=1 \end{cases}$$

3.1 Marginal Inference

We explain briefly how to implement filtering based on the run-length formalism for the reset LDS. In this case the filtered distribution is represented by

$$\tilde{\alpha}(\mathbf{h}_t|\rho_t) = \mathcal{N}\left(\mathbf{h}_t|\mathbf{f}_t(\rho_t),\mathbf{F}_t(\rho_t)\right)$$

and since $\tilde{\alpha}(\rho_t)=p(\rho_t|\mathbf{y}_{1:t})p(\mathbf{y}_{1:t})$, we take $p(\rho_t|\mathbf{y}_{1:t})\equiv w_t(\rho_t)$ and $p(\mathbf{y}_{1:t})\equiv l_t$. Filtering then corresponds to sequentially updating mean parameter $\mathbf{f}_t(\rho_t)$, covariance $\mathbf{F}_t(\rho_t)$, mixture weights $w_t(\rho_t)$, and likelihood l_t—see supplementary material for the algorithm. This form of filtering is particularly useful since, as explained in section(2.3), an exact correction smoother follows.

In order to compare approximate methods based on neglecting message components, we also implement the $\beta(\mathbf{h}_t,c_t)$ messages. Each $\beta^0(\mathbf{h}_{t-1})$ is calculated in canonical form using the standard information filter (see for example Cappé et al. (2005)). Each β^0 message is of the form $\sum_j k_{tj}\exp -\frac{1}{2}\left(\mathbf{h}_t^\top\mathbf{G}_{tj}\mathbf{h}_t-2\mathbf{h}_t^\top\mathbf{g}_{tj}\right)$, where the constant terms k_{tj} are necessary to compute the weights in the full posterior. The constant β_{t-1}^1 is easily found using a similar calculation. Formally, one carries out the β recursion under the assumption of a mixture of canonical forms. The resulting lengthy expressions are given in the supplementary material.

The result is that $\alpha(\mathbf{h}_t,c_t)$ is represented as a mixture of Gaussians in moment form, whereas $\beta(\mathbf{h}_t,c_t)$ is a mixture of squared exponentials in canonical form. To compute the smoothed posterior $p(\mathbf{h}_t,c_t|\mathbf{y}_{1:T})\propto\alpha(\mathbf{h}_t,c_t)\beta(\mathbf{h}_t,c_t)$, we need to multiply out both mixtures, converting the resulting mixture of moment-canonical interactions to a mixture of moments. Note that the frequent moment-to-canonical conversions required for the β message and forming the smoothed posterior mean that this procedure is computationally less stable and more expensive than correction based smoothing (Verhaegen and Van Dooren, 2002).

Since numerical stability is of such concern in the LDS, it is imperative to have a correction-based smoother for the reset LDS. There are two routes to achieve this: either we can use the run-length $\tilde{\alpha}$-$\tilde{\gamma}$ formalism, section(2.3), or apply the bracket smoother from section(2.4). Both are essentially equivalent and require that we have first computed the $\tilde{\alpha}$ messages. Deriving these smoothers is straightforward—the final bracket smoother algorithm is in the supplementary material.

4 APPROXIMATE INFERENCE

Filtering has overall $O\left(T^2\right)$ complexity, meanwhile smoothing has $O\left(T^3\right)$ complexity. For long time-series $T\gg 1$, this can be prohibitively expensive, motivating a consideration of approximations.

4.1 Approximate Filtering

The α message (or equivalently, the $\tilde{\alpha}$ message) is constructed as a mixture of distributions; it is therefore easy to motivate any reduced component mixture approximation technique (Titterington et al., 1985). Our implementation uses the $\tilde{\alpha}$ formalism, and to approximate we simply retain the M components with largest weight, reducing the forward pass to $O\left(MT\right)$. That is, we rank the $\tilde{\alpha}(h_t,\rho_t)$ components by the weight $\tilde{\alpha}(\rho_t)$, and retain only the ρ_t with largest weight.

4.2 Approximate $\tilde{\alpha}$-$\tilde{\beta}$

A naïve approximate algorithm is to drop components from the β message (or equivalently, the $\tilde{\beta}$ message) according to the weights of the components in the β message mixture. However, the β message components in themselves are not of interest and dropping components based on low β weight gives generally poor performance. When the α (or $\tilde{\alpha}$) and β (or $\tilde{\beta}$) messages are combined, the result is the smoothed posterior. The weights of these smoothed components are a function not just of the weights in the α and β messages, but of all parameters in the messages. The relationship between those parameters and the resulting component weights can be complex (see supplementary material for the case of the reset LDS).

We can, however, motivate an approximation by observing the bracket smoothing results of section(2.4). First, we note that whatever algorithm we choose to implement (α-β, $\tilde{\alpha}$-$\tilde{\beta}$, α-γ, or $\tilde{\alpha}$-$\tilde{\gamma}$), the resulting exact posterior has identical structure. In the bracket smoother, the pair (ρ_t,ς_t) specifies exactly when the previous and next resets occur, so this intuition can be applied to each smoothing algorithm. In equation (2.7), we observed that the posterior mass transfers directly through the backward recursion in the no-reset case.

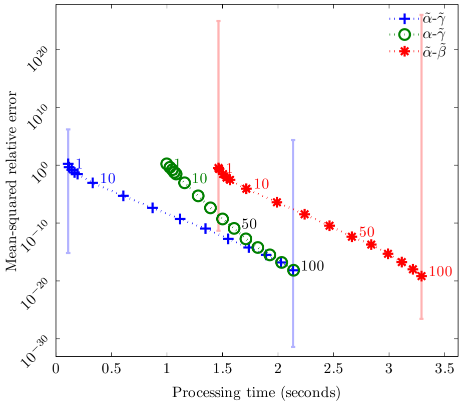

Figure 2: Comparison of approximation accuracy for the reset LDS. 1000 time-series ($T = 100$) were randomly generated using a single dimension for y_t and h_t. We show the median error (compared with the exact correction-smoothed posterior) of the linear-time smoother based on approximate (blue) and exact filtering (green), and the quadratic-complexity $\tilde{\beta}$ smoother with approximate filtering (red), versus the mean running time. Error bars show max and min values. In each case, the points on the curve correspond to $N = 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$ components in each approximate message. The error is calculated as $\text{mean}_t \left[\left(\langle h_t \rangle - \langle h'_t \rangle \right) / \langle h_t \rangle \right]^2$.

After calculating a full (exact or approximate) $\tilde{\alpha}$ recursion, we can approximate the $\tilde{\alpha}$-$\tilde{\beta}$ algorithm as follows. First, calculate a full $\tilde{\beta}$ message. Second, calculate the components corresponding to $\varsigma_t = 1$ given as $\tau_{1|c_t(\rho_t)} \tilde{\beta}_t^1 \tilde{\alpha}(h_t, \rho_t)$. Third, combine the $\tilde{\alpha}$ and $\tilde{\beta}$ messages for those components we know to have significant mass according to equation(2.7), corresponding to $\varsigma_t > 1$. Finally, renormalise the weights to form the approximate posterior. In this way, we limit the number of posterior components to N. The requirement of a full $\tilde{\beta}$ message, however, means the algorithm has $O\left(T^2\right)$ complexity.

4.3 Approximate $\tilde{\alpha}$-$\tilde{\gamma}$

By doing something similar with the $\tilde{\alpha}$-$\tilde{\gamma}$ recursion, we derive a linear-time algorithm. First calculate a full approximate $\tilde{\alpha}$ message. Second, calculate the components corresponding to $\varsigma_t = 1$: the weights are given by equation (2.8), and the moments by the $\tilde{\alpha}$ message. Third, calculate the moments for those components we know to have significant mass according to equation (2.7) corresponding to $\varsigma_t > 1$, with the correction smoother update. Finally, renormalise the weights to form the approximate posterior. This is equivalent to dropping components from $\tilde{\gamma}$ and limits the number of components calculated at each point to N.

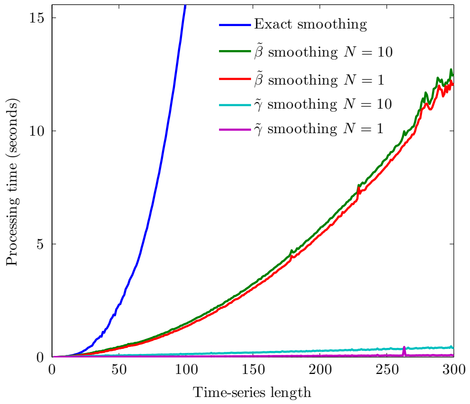

Figure 3: Median running time (10 iterations) for the reset LDS with variable time-series length.

4.4 Example: Reset LDS

We implemented a linear-time algorithm by limiting the number of $\tilde{\alpha}$ and $\tilde{\gamma}$ components, dropping lowest-weight components when the limit is exceeded, and compared the results with the quadratic-complexity $\tilde{\alpha}$-$\tilde{\beta}$ approximate implementation. To aid a direct comparison of methods, we also ran approximate $\tilde{\gamma}$ smoothing based on the exact filtered posterior since this has overall quadratic complexity comparable with the $\tilde{\beta}$ routine. Results are shown in fig(2), in which we show how the runtimes and relative errors in the smoothed posteriors compare for different numbers of components.

We demonstrate the run-time reduction for different time-series lengths in fig(3).

4.5 Discussion

Various approximation schemes are available for Gaussian mixtures. Here, we simply dropped posterior components. This motivates a discussion of whether such scheme provides a stable approximation, and how to select the number of components to retain. Each retained posterior component corresponds, according to the bracket smoother, to a unique local partition of the time-series; in the worst case, each of the posterior components has equal mass. In this case, the discrete components of the filtering and smoothing messages correspond to little or no posterior belief about the probability of a reset at each point. Hence it may be fair to say that the model is not well suited to the data: a reparameterisation or different model may be appropriate. When considering the number of message components to retain, however, the 'cut-off' weight of the dropped components is known in each routine and can be used to conclude whether retaining more components may be worth the computational expense.

The approximation routines are structured in a flexible

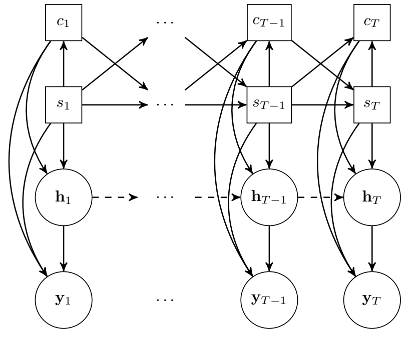

Figure 4: Switch-Reset model structure.

way so as to allow different schemes to that used in our implementation. One example would be to only drop posterior components from messages when the mass of such components falls below a predetermined threshold, though this has the effect of increasing worst-case computational complexity. Finally, the smoothed posterior weights, calculated according to the bracket smoother and used in the $\tilde{\gamma}$ approximation, are calculated only from the filtered weights; so it is possible to conclude something about the number of smoothed components that may be reasonably dropped by filtering only.

5 PIECEWISE RESET MODELS

The recursions for the reset LDS are straightforward since the messages are closed within the space of the mixture of Gaussians. Other classes of model admit similar closure properties, and we briefly describe two such here based on the piecewise-constant assumption:

$$p(h_t|h_{t-1},c_t) = \begin{cases} \delta(h_t - h_{t-1}) & c_t = 0 \\ p^1(h_t) & c_t = 1 \end{cases}$$

for which equation (2.5) is trivially rewritten

$$\tilde{\alpha}(h_t, \rho_t > 0) = p^0(y_t|h_t) p(\rho_t|\rho_{t-1} = \rho_t - 1) \\ \times \tilde{\alpha}(h_{t-1} = h_t, \rho_{t-1} = \rho_t - 1) \quad (5.1)$$

and similarly for equation (2.9)

$$\tilde{\gamma}(h_t|\rho_t, \varsigma_t > 1) = \tilde{\gamma}(h_{t+1} = h_t|\rho_{t+1} = \rho_t + 1, \varsigma_{t+1} = \varsigma_t - 1)$$

Any model can be considered in which $p(y|h_t, c_t)$ and $p^1(h_t)$ are conjugate-exponential pairs. For example if we have an inverse Gamma reset distribution $p^1(h_t) = \Gamma^{-1}(h_t)$ and a Gaussian emission $p(y_t|h_t, c_t) = \mathcal{N}(y_t|0, h_t)$, then the filtered and smoothed posteriors are mixtures of inverse Gamma terms. Similarly, one can consider a piecewise-constant Poisson reset model in which the rate h_t is constant until reset from a Gamma distribution. The resulting posterior is a mixture of Gamma distributions (Barber and Cemgil, 2010). Bayesian priors over Gaussian mean and precision (for conjugacy, usually Gaussian and Gamma/Wishart respectively) fit readily into the piecewise-constant framework.

Example problems are well known for piecewise reset models, including the coal-mine disaster data of Jarrett (1979) and the well-logging data of Ó Ruanaidh and Fitzgerald (1996). We provide an example of the latter in the supplementary material, using a Gaussian prior over the piecewise-constant mean of Gaussian data.

6 SWITCH-RESET MODELS

The reset model defined above is limited to two kinds of dynamics—either continuation along the standard dynamics p^0 or the reset p^1. The switch-reset model enriches this by defining a set of S dynamical models,

$$p(h_t|h_{t-1},c_t,s_t) = \begin{cases} p^0(h_t|h_{t-1},s_t) & c_t = 0 \\ p^1(h_t|s_t) & c_t = 1 \end{cases}$$

$$p(y_t|h_t,c_t,s_t) = \begin{cases} p^0(y_t|h_t,s_t) & c_t = 0 \\ p^1(y_t|h_t,s_t) & c_t = 1 \end{cases}$$

with a Markov transition on the switch variables $p(s_t|s_{t-1}, c_{t-1})$. The reset is deterministically defined by $c_t = \mathbb{I}[s_t \neq s_{t-1}]$, see fig(4).

The intuition is that after a reset the model chooses from an available set of S dynamical models p^1. Another reset occurs if the state s_t changes from s_{t-1}. At that point the latent variable is reset, after which the dynamics continues. This is therefore a switching model, but with resets[5]. Inference in the class of switch-reset models is straightforward—we give a derivation in the supplementary material—however in the LDS case, the naïve correction approach runs into the same analytical difficulty as in section(2.2). The intuitive interpretation of the posterior components that we observed for the basic reset model transfers to the switching model, and the approximation schemes described herein can be easily applied.

6.1 Example

We implemented a switch-reset LDS using the linear-time $\tilde{\alpha}$-$\tilde{\gamma}$ smoother. In fig(5) we apply the model to a short speech audio signal[6] of $10,000$ observations. For these data, the latent variable $\mathbf{h}_t$ is used to model the coefficients of the autoregressive lags, and we assume each observation $y_t = \sum_{m=1}^{6} h_t^m y_{t-m} + \epsilon$. Compared with a standard hidden Markov model in which a set of fixed autoregressive coefficients is used, this example provides a rich model in which the coefficients are free to evolve between state changes.

[5] Called a Reset-HMM in Barber and Cemgil (2010).

[6] These data are from the TIMIT database.

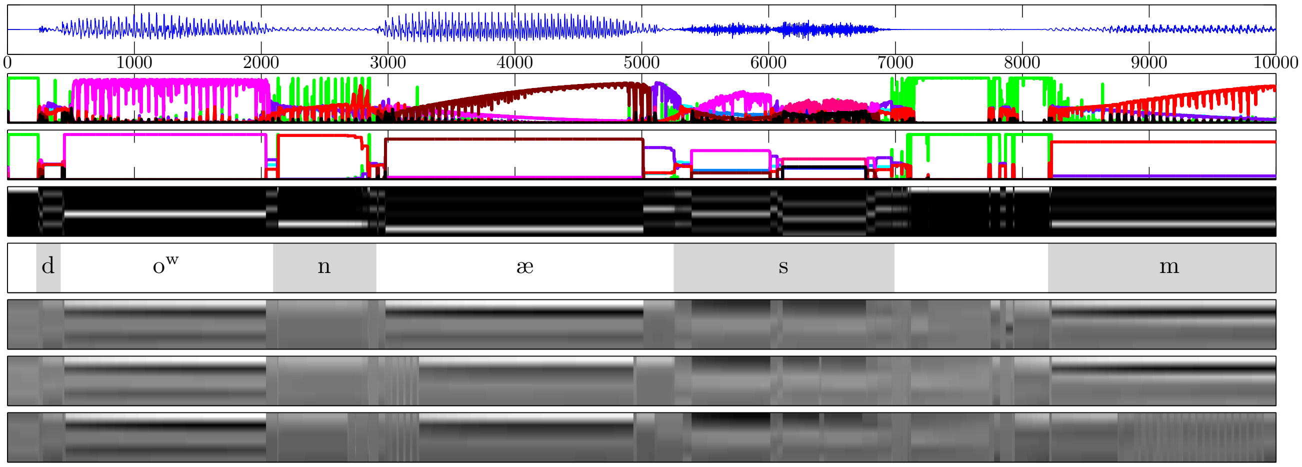

Figure 5: Switch-Reset LDS example with a short speech signal. We assumed that the signal is autoregressive of order 6, and used the switch-reset LDS to model the autoregressive coefficients with $S = 10$ different states, with $N = 10$ components retained in each message. From top to bottom, we show (i) the audio signal; (ii) the filtered posterior of each state; (iii)-(iv) the smoothed posterior state mass; (v) the main IPA phonemes comprising the sound; (vi) the mean value of the inferred autoregressive coefficients; (vii) also with $N = 5$; and (viii) with $N = 2$.

Using our MATLAB code on a 2.4GHz machine, filtering took less than 400 seconds and subsequent smoothing less than 200 further; in the exact case, however, the problem is intractable needing the moments of some $O\left(10^{11}\right)$ Gaussian components (each of dimension 6) for the smoothed posterior in total, for each state. The model is highly sensitive to the state parameters, and we performed a very simple manual search of the parameter space and considered the likelihood (l_T) surface[7], with states broadly corresponding to 'growing', 'steady', and 'reducing' signal magnitudes by considering the sum of the autoregressive coefficients. The results show clear state switches between different phonemes, and each phoneme corresponds to a different (combination of) states in the smoothed posterior.

A further example of the switch-reset LDS is given in the supplementary material.

7 SUMMARY AND CONCLUSION

We discussed probabilistic inference in reset models and switch-reset models. Such models have been used in the fields of bioinformatics (Boys and Henderson, 2004), finance (Davis et al., 2008) and signal processing (Fearnhead, 2005). The well-known β message passing algorithm is applicable and straightforward to derive in respect of reset models, but suffers some drawbacks. First, for certain classes of such model—notably the linear dynamical system—numerical stability is a concern. Second, it is difficult to contrive a linear-time algorithm for smoothed inference, due to the abstract nature of the β components.

To address these issues we went on to derive a correction smoother, based on a redefinition of the reset model in terms of run-length. We then contributed an interpretation of smoothing in terms of messages relating to future resets. The algorithms so defined overcome the numerical difficulties of the β approach and can be implemented with confidence using standard numerically-stable propagation routines in models such as the linear dynamical system. Moreover, the derivation is didactically useful when considering approximations to the smoothed posterior. The resulting approximations based on dropping low weight components in the filtered and smoothed posteriors give a linear-time algorithm that exhibits excellent performance, superior to previous approaches based on α-β smoothing. Further applications include piecewise reset models (widely known as changepoint models), for which the inference algorithms are readily transferred.

A switch-reset model was also discussed, motivated by a desire for multiple generating processes; the reset nature of the model significantly reduces the complexity in comparison with other switching systems, and the linear-time routines are applicable. The reset models are highly practical and do not suffer from the numerical difficulties apparent in the more general switching models. Furthermore, with robust and effective linear-time smoothing and filtering algorithms, they are inexpensive to deploy.

Demo code is available in the supplementary material.

[7] It is possible to use maximum likelihood learning techniques such as expectation maximisation.

References

R. P. Adams and D. J. C. MacKay. Bayesian online changepoint detection. Technical report, University of Cambridge, Cambridge, UK, 2007.

D. Alspach and H. Sorenson. Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972. ISSN 0018-9286.

D. Barber. Expectation correction for smoothed inference in switching linear dynamical systems. *The Journal of Machine Learning Research*, 7:2515–2540, 2006. ISSN 1532-4435.

D. Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2011. In press.

D. Barber and A. T. Cemgil. Graphical models for time series. *IEEE Signal Processing Magazine*, (18), November 2010.

X. Boyen and D. Koller. Tractable Inference for Complex Stochastic Processes. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998.

R. J. Boys and D. A. Henderson. A Bayesian Approach to DNA Sequence Segmentation. *Biometrics*, 60(3): 573–581, 2004. ISSN 1541-0420.

O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models.* Springer, New York, 2005.

C. K. Carter and R. Kohn. Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika*, 83(3):589, 1996. ISSN 0006-3444.

R. A. Davis, T. C. M. Lee, and G. A. Rodriguez-Yam. Break Detection for a Class of Nonlinear Time Series Models. *Journal of Time Series Analysis*, 29(5): 834–867, 2008. ISSN 1467-9892.

A. Doucet, N. De Freitas, K. Murphy, and S. Russell. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 176–183, 2000.

A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo methods in practice.* Springer Verlag, 2001. ISBN 0387951466.

P. Fearnhead. Exact Bayesian Curve Fitting and Signal Segmentation. *IEEE Transactions on Signal Processing*, 53(6):2160 – 2166, June 2005. ISSN 1053-587X.

P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (4):589–605, 2007. ISSN 1467-9868.

R. G. Jarrett. A Note on the Intervals Between Coal-Mining Disasters. *Biometrika*, 66(1):191 – 193, 1979.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

C. J. Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1-2):1–22, February 1994. ISSN 03044076.

C. J. Kim and C. R. Nelson. *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*, volume 1. MIT Press, 1999. ISBN 0262112388.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*, volume 17, pages 362–369, 2001.

T. P. Minka. Divergence measures and message passing. *Microsoft Research, Cambridge, UK, Tech. Rep. MSR-TR-2005-173*, 2005.

J. J. K. Ó Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian methods applied to signal processing.* Springer Verlag, 1996. ISBN 0387946292.

H. E. Rauch, F. Tung, and C. T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

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

M. Verhaegen and P. Van Dooren. Numerical aspects of different Kalman filter implementations. *Automatic Control, IEEE Transactions on*, 31(10):907–917, 2002. ISSN 0018-9286.