Stat Comput (2006) 16:279–292
DOI 10.1007/s11222-006-8770-8

Deterministic approximate inference techniques for conditionally Gaussian state space models

Onno Zoeter · Tom Heskes

Received: June 2004 / Accepted: March 2006

Abstract We describe a novel deterministic approximate inference technique for conditionally Gaussian state space models, i.e. state space models where the latent state consists of both multinomial and Gaussian distributed variables. The method can be interpreted as a smoothing pass and iteration scheme symmetric to an assumed density filter. It improves upon previously proposed smoothing passes by not making more approximations than implied by the projection onto the chosen parametric form, the assumed density. Experimental results show that the novel scheme outperforms these alternative deterministic smoothing passes. Comparisons with sampling methods suggest that the performance does not degrade with longer sequences.

Keywords State space models · Conditionally Gaussian state space models · Switching linear dynamical systems · Approximate inference · Expectation propagation · Assumed density filtering

1. Introduction

Many real-world problems can be described by models that extend the classical linear Gaussian dynamical system with (unobserved) discrete regime indicators. In such extended models the discrete indicators dictate what transition and observation model the process follows at a particular time point. The problems of tracking and estimation in models with maneuvering targets (Bar-Shalom and Li, 1993), multiple targets (Shumway and Stoffer, 1991), non-Gaussian disturbances (Kitagawa, 1996), unknown model parameters (Harrison and Stevens, 1976), failing sensors (Lerner *et al.*, 2000) and different trends (Hamilton, 1989) are all examples of problems that have been formulated as conditionally Gaussian state space models. Since the extented model is so general it has been invented and re-invented many times in multiple fields.

Although the extended model is very powerful, it is notorious for the fact that exact estimation of posteriors is intractable. In general, exact filtered, smoothed or predicted posteriors have a complexity exponential in the number of observations (Lerner and Parr, 2001). Section 2 describes the intractability in more detail.

In this article we introduce a deterministic approximation scheme that is particularly suited to find smoothed one and two-slice posteriors. It can be seen as a symmetric backward pass and iteration scheme for previously proposed assumed density filtering approaches (Harrison and Stevens, 1976).

The intractability problems are shared between all variants of the conditionally Gaussian state space model. In Section 2 we introduce the general model; variants where only the transition or only the observation model switch, or where states or observations are multi- or univariate can be treated as special cases. The assumed density filtering approach that forms the basis for our approximation scheme is described in Section 3. In Section 4 the symmetric backward pass is introduced. Since both the forward and the backward pass consist of local, greedy projections it makes sense to iterate them. Section 5 introduces the objective that is minimized by such an iteration scheme and gives an intuition how we should interpret fixed points. In Section 6 we describe two approximate smoothing passes that are often used for conditionally Gaussian state space models. In Section 7

O. Zoeter (✉) · T. Heskes
Biophysics, Radboud University Nijmegen, Geert Grooteplein 21,
6525 EZ, Nijmegen, The Netherlands
e-mail: o.zoeter@science.ru.nl

T. Heskes
e-mail: t.heskes@science.ru.nl

related deterministic approximation are discussed. Section 8 describes experiments that test the validity of the proposed method and compare it with the alternative backward pass and state-of-the-art sampling techniques.

2. Notation and problem description

In a *switching linear dynamical system* (SLDS) it is assumed that an observed sequence $\mathbf{y}_{1:T}$ of T, d-dimensional observations is generated as noisy observations from a first order Markov process. The latent space consists of a q-dimensional continuous state $\mathbf{x}_t$ and a discrete state s_t that can take on M states. Conditioned on s_{t-1} and s_t, $\mathbf{x}_t$ is a linear function of $\mathbf{x}_{t-1}$ subjected to Gaussian white noise

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_{t-1}, s_t, \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{x}_t; A_{s_{t-1},s_t}\mathbf{x}_t, Q_{s_{t-1},s_t}). \qquad (1)$$

In the above we have used $\boldsymbol{\theta}$ to denote the set of parameters in the model and $\mathcal{N}(.;.,.)$ is the Gaussian probability density function. The observation model is also linear-Gaussian and may differ between settings of s_{t-1} and s_t:

$$p(\mathbf{y}_t|\mathbf{x}_t, s_{t-1}, s_t, \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{y}_t; C_{s_{t-1},s_t}\mathbf{x}_t, R_{s_{t-1},s_t}). \qquad (2)$$

The discrete state follows a first order Markov chain in the discrete space:

$$p(s_t|s_{t-1}, \boldsymbol{\theta}) = \Pi_{s_{t-1}\to s_t}. \qquad (3)$$

At $t = 1$ we have $p(s_1|\boldsymbol{\theta}) = \pi_{s_1}$ and $p(\mathbf{x}_1|s_1, \boldsymbol{\theta})$ a Gaussian with known parameters. The graph that encodes the conditional independencies implied by these equations is presented in Fig. 1.

Throughout this article the parameters $\boldsymbol{\theta}$ are assumed to be known. The interest is in filtered and smoothed one and two-slice posteriors. If we treat $\mathbf{z}_t \equiv \{s_t, \mathbf{x}_t\}$ as a single *conditionally Gaussian* (CG) distributed random variable we obtain an independence structure identical to the basic linear dynamical system. (Appendix A introduces the CG distribution and defines standard operations.) This might lead us to assume that inference is easy. This however, is *not* the case. One way to see this is by looking at the posterior

$$p(s_t, \mathbf{x}_t|\mathbf{y}_{1:T}, \boldsymbol{\theta}) = \sum_{s_{1:T\backslash t}} p(\mathbf{x}_t|s_{1:T}, \mathbf{y}_{1:T}, \boldsymbol{\theta}) p(s_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\theta}). \qquad (4)$$

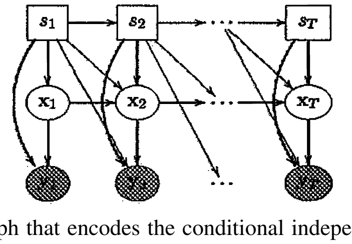

Fig. 1 The graph that encodes the conditional independencies in the SLDS. Ellipses denote Gaussian distributed variables and rectangles denote multinomial distributed variables. Shading emphasizes the fact that a particular variable is observed

Since the history of regime changes is unknown, we have to take into account all possible sequences of indicator variables $s_{1:T}$. The CG family is not closed under marginalization, so the result of summing over all possible sequences in (4) is of a more complex form than a simple CG distribution: conditioned on s_t the posterior is a Gaussian mixture with M^{T-1} modes.

A second way of interpreting the exponential growth is by inspecting a recursive filtering algorithm. At every timeslice the number of modes in the exact posterior increases by a factor M, since all the modes considered in the previous slice are propagated to and updated in the next slice by M possible models. In the next section we describe an approximate inference algorithm where this growth is avoided by a projection at every time step.

3. Assumed density filtering

3.1. Local approximations

In the previous section we have seen that the number of modes in the exact filtered posteriors increases M-fold with every new observation. An obvious, and in practice very powerful, approximation is to first incorporate evidence $\mathbf{y}_t$ and then to approximate the resulting posterior over $\mathbf{z}_t$ by a "best fitting" conditional Gaussian distribution. Here "best fitting" is defined as the CG distribution that minimizes the *Kullback-Leibler (KL) divergence* between the original and approximate distribution. The KL-divergence between distributions $\hat{p}(\mathbf{z}_t)$ and $q(\mathbf{z}_t)$ is defined as

$$\mathrm{KL}(\hat{p}(\mathbf{z}_t)||q(\mathbf{z}_t)) \equiv \sum_{\mathbf{z}_t} \hat{p}(\mathbf{z}_t) \log \frac{\hat{p}(\mathbf{z}_t)}{q(\mathbf{z}_t)} \qquad (5)$$

and is not symmetric in $\hat{p}$ and q (see e.g. Cover and Thomas, 1991, for an introduction to the KL divergence). In (5) we have used the somewhat sloppy shorthand notation of $\sum_{\mathbf{z}_t}$ for the operation of integrating over $\mathbf{x}_t$ and summing over s_t, a shorthand that we will use in the remainder of this article. The CG distribution $\hat{q}(\mathbf{z}_t)$ closest to $\hat{p}(\mathbf{z}_t)$ in the sense of KL-divergence is the CG distribution that has the same moments as $\hat{p}$ (Whittaker, 1989). That is, for each value of s_t, the mixture $\hat{p}(\mathbf{x}_t|s_t)$ is approximated in $\hat{q}(\mathbf{x}_t|s_t)$ by a single Gaussian with the same mean and covariance as $\hat{p}(\mathbf{x}_t|s_t)$. Motivated by these "collapses" of mixtures onto single Gaussians, we introduce the notation

$$\hat{q}(\mathbf{z}_t) = \mathrm{Collapse}\,(\hat{p}(\mathbf{z}_t)) \equiv \underset{q\in CG}{\mathrm{argmin}}\, \mathrm{KL}(\hat{p}(\mathbf{z}_t)||q(\mathbf{z}_t)).$$

A precise definition is given in Appendix A.

3.2. The sum-product algorithm

If the growth of complexity is avoided by a local projection in every recursion step, the computational requirements of an approximate filter are linear in the number of observations instead of exponential. This resulting approximate forward pass is referred to as *assumed density filtering*, or *generalized pseudo Bayes 2 (GPB2)* (Bar-Shalom and Li, 1993) amongst others and has been proposed independently many times. The oldest reference we are aware of is Harrison and Stevens (1976).

In Algorithm 1 assumed density filtering is presented in the spirit of the *sum-product algorithm* (Kschischang *et al.*, 2001). The model Eqs. (1), (2), and (3) enter as *factors* ψ_t, potentials defined as

$$\begin{aligned}\psi_t(\mathbf{z}_{t-1},\mathbf{z}_t) &\equiv p(\mathbf{y}_t,\mathbf{z}_t|\mathbf{z}_{t-1},\boldsymbol{\theta})\\ &= p(\mathbf{y}_t|\mathbf{x}_t,s_{t-1},s_t,\boldsymbol{\theta})p(\mathbf{x}_t|\mathbf{x}_{t-1},s_{t-1},s_t,\boldsymbol{\theta})\\ &\quad\times p(s_t|s_{t-1},\boldsymbol{\theta}),\end{aligned} \tag{6}$$

with

$$\begin{aligned}\psi_1(\mathbf{z}_0,\mathbf{z}_1) &\equiv p(\mathbf{y}_1,\mathbf{z}_1|\boldsymbol{\theta})\\ &= p(\mathbf{y}_1|\mathbf{x}_1,s_1,\boldsymbol{\theta})p(\mathbf{x}_1|s_1,\boldsymbol{\theta})p(s_1|\boldsymbol{\theta})\ ,\end{aligned} \tag{7}$$

a convenient definition at the boundary.

To make the similarity between the filtering and smoothing pass more clear we introduce a distinct notation for approximate one-slice marginals, $\hat{q}_t(\mathbf{z}_t) \approx p(\mathbf{z}_t|\mathbf{y}_{1:T},\boldsymbol{\theta})$, and *forward messages*, $\alpha_t(\mathbf{z}_t)$. The messages fulfill a similar role as in the regular Kalman filter. In an exact filter, the forward messages would satisfy $\alpha_t(\mathbf{z}_t) \propto p(\mathbf{z}_t|\mathbf{y}_{1:t},\boldsymbol{\theta})$, here they are approximations of these quantities.

Figure 2 represents the *factor graph* that corresponds to Algorithm 1 and 2 that will be presented in Section 4. In Fig. 2 variables are represented by ovals, factors by solid squares and messages by arcs. Note that the variables $\mathbf{y}_{1:T}$ are not depicted, they are always observed and part of the factors.

The assumed density filter starts at $t = 1$. The approximation $\hat{q}_1(\mathbf{z}_1)$ is obtained by normalizing ψ_1:

$$\hat{q}_1(\mathbf{z}_1) \propto \psi_1(\mathbf{z}_1).$$

The posterior $\hat{q}_1(\mathbf{z}_1)$ is then a CG distribution with $\hat{q}_1(\mathbf{x}_1|s_1 = m)$ corresponding to the prior $p(\mathbf{x}_1|s_1 = m,\boldsymbol{\theta})$ updated in the light of observation $\mathbf{y}_1$ and observation model m. Similarly $\hat{q}_1(s_1 = m)$ is the prior probability that the system starts in regime m appropriately weighted by the likelihood that $\mathbf{y}_1$ was generated by model m. Since $\hat{q}_1(\mathbf{z}_1)$ is still CG, there is no need for an approximation at this point. Since the current approximation of the belief state of $\mathbf{z}_1$ is only based on $\mathbf{y}_1$ we set $\alpha_1(\mathbf{z}_1) = \hat{q}_1(\mathbf{z}_1)$. The need for messages will become clear in the next sections.

A recursive filtering step is done by making a prediction and measurement update step. The message α_{t-1} is multiplied by ψ_t and normalized to get a belief over the states $\mathbf{z}_{t-1,t}$:

$$\hat{p}_t(\mathbf{z}_{t-1,t}) \propto \alpha_{t-1}(\mathbf{z}_{t-1})\psi_t(\mathbf{z}_{t-1,t}).$$

At the start of recursion, $\alpha_{t-1}(\mathbf{z}_{t-1})$ is a conditional Gaussian potential with M modes. The belief $\hat{p}_t(\mathbf{z}_{t-1,t})$ is a conditional Gaussian distribution with M^2 components: the M components from $\alpha_{t-1}(\mathbf{z}_{t-1})$ propagated and updated according to M different models.

The marginal $\hat{p}_t(\mathbf{z}_t) = \sum_{\mathbf{z}_{t-1}} \hat{p}_t(\mathbf{z}_{t-1},\mathbf{z}_t)$ is not CG. Instead $\hat{p}_t(\mathbf{x}_t|s_t)$ is a *mixture* of Gaussians with M components. If we would use $\hat{p}_t(\mathbf{z}_t)$ as the new forward message, the filtering pass would give exact results, but, at the next step in the recursion in step 1 in Algorithm 1, the number of components in the joint would increase by a factor M, implying an exponential number of components in t.

To avoid this growth $\hat{p}_t(\mathbf{z}_t)$ is approximated by the CG distribution closest to $\hat{p}_t(\mathbf{z}_t)$ in KL-sense:

$$\hat{q}_t(\mathbf{z}_t) = \text{Collapse}\,(\hat{p}_t(\mathbf{z}_t))\,.$$

As for $\hat{q}_1(\mathbf{z}_1)$, if we only perform a single forward pass, the approximate beliefs $\hat{q}_t(\mathbf{z}_t)$ are only based on $\mathbf{y}_{1:t}$. Therefore in Algorithm 1, we set $\alpha_t(\mathbf{z}_t) = \hat{q}_t(\mathbf{z}_t)$.

Since the growth of complexity is prevented by the projection in step 3 of Algorithm 1 the running time of assumed density filtering is linear in T, the number of observations.

4. Expectation propagation

4.1. Backward pass

After establishing a deterministic approximation for the forward pass, it is natural to look for an analogous backward, or smoothing, pass. Several attempts have been made in the literature (Kim and Nelson, 1999; Shumway and Stoffer, 1991).

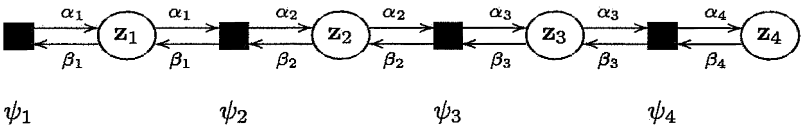

Fig. 2 The factor graph corresponding to an SLDS with four observations and with factors ψ_t defined by (6) and (7)

Algorithm 1 Assumed density filtering

At time step $t = 1$ there is no need yet for an approximation, we start the algorithm with:

$$\hat{q}_1(\mathbf{z}_1) = \alpha_1(\mathbf{z}_1) \propto \psi_1(\mathbf{z}_1) \; .$$

Then, for $t = 2, 3, \ldots T$, compute approximate filtered posteriors $\hat{q}_t(\mathbf{z}_t)$ as follows:

1. Construct a two-slice joint:

$$\hat{p}_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \propto \alpha_{t-1}(\mathbf{z}_{t-1})\psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \; .$$

2. Marginalize to obtain a one-slice filtered marginal

$$\hat{p}_t(\mathbf{z}_t) = \sum_{\mathbf{z}_{t-1}} \hat{p}_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \; .$$

3. Approximate $\hat{p}_t(\mathbf{z}_t)$ by $\hat{q}_t(\mathbf{z}_t)$, the CG distribution closest to $\hat{p}_t(\mathbf{z}_t)$ in KL-sense:

$$\hat{q}_t(\mathbf{z}_t) = \text{Collapse}\left(\hat{p}_t(\mathbf{z}_t)\right) \; .$$

4. Set $\alpha_t(\mathbf{z}_t) = \hat{q}_t(\mathbf{z}_t)$.

These have all included extra approximations beyond the projections onto the conditional Gaussian family; we will briefly review these in Section 6. Other approximations such as Helmick *et al.* (1995) are restricted to models with invertible dynamics.

The smoothing pass depends on *backward messages* $\beta_t(\mathbf{z}_t)$, with $t = 1, 2, \ldots, T$. These messages are analogous to the backward messages in the *hidden Markov model (HMM)* smoother or the smoother from the two-filter algorithm for the *linear dynamical system (LDS)*. In the exact case we would have $\beta_t(\mathbf{z}_t) \propto p(\mathbf{y}_{t+1:T}|\mathbf{z}_t, \boldsymbol{\theta})$, such that $\alpha_t(\mathbf{z}_t)\beta_t(\mathbf{z}_t) \propto p(\mathbf{z}_t|\mathbf{y}_{1:T}, \boldsymbol{\theta})$, which is of the form (4). In the current approximation scheme we have

$$\alpha_t(\mathbf{z}_t)\beta_t(\mathbf{z}_t) \propto \hat{q}_t(\mathbf{z}_t) \approx p(\mathbf{z}_t|\mathbf{y}_{1:T}, \boldsymbol{\theta}) \; , \tag{8}$$

where $\hat{q}_t(\mathbf{z}_t)$ is CG. In the factor graph in Fig. 2 this definition is depicted as follows: an approximate posterior belief over a variable can be constructed by multiplying all messages coming into that variable node.

A potential problem with an approximate backward pass based on β_t's is that, whereas forward messages α_t can always be normalized, the backward messages β_t in general cannot. The KL-divergence is only defined on proper distributions. We therefore cannot define a backward pass by approximating messages directly.

We propose to use an approximation which can be viewed as a special case of Expectation Propagation (Minka, 2001). A key difference with the previous approaches mentioned above is that instead of approximating messages, first *beliefs* are constructed which are then approximated. The new messages are then deduced from the approximated beliefs.

The message passing scheme is symmetric in the forward and backward pass. As in the previous section the presentation in Algorithm 2 is in the spirit of the sum-product algorithm.

It suffices to describe the backward recursion step. The start of the backward recursion follows by introducing a message $\beta_T = 1$. In the backward pass a new backward message β_{t-1} is computed as a function of the backward message β_t, the local potential ψ_t, *and* the forward message α_{t-1}. Given α_{t-1}, ψ_t and β_t an approximated two-slice posterior belief $\hat{p}_t(\mathbf{z}_{t-1,t})$ can be computed as

$$\hat{p}_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \propto \alpha_{t-1}(\mathbf{z}_{t-1})\psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t)\beta_t(\mathbf{z}_t).$$

In the factor graph in Fig. 2 this operation should be interpreted as follows: an approximate posterior belief over the domain of the factor t can be obtained by multiplying all incoming messages with factor ψ_t itself and normalizing.

As in the forward pass the marginal $\hat{p}_t(\mathbf{z}_{t-1})$ is a conditional mixture of Gaussians instead of CG. However since $\hat{p}_t(\mathbf{z}_{t-1})$ constitutes a proper distribution, the approximation

$$\hat{q}_{t-1}(\mathbf{z}_{t-1}) = \text{Collapse}\left(\hat{p}_t(\mathbf{z}_{t-1})\right)$$

is now well defined. From the definition in (8) we have $\hat{q}_{t-1}(\mathbf{z}_{t-1}) = \alpha_{t-1}(\mathbf{z}_{t-1})\beta_{t-1}(\mathbf{z}_{t-1})$. The message $\alpha_{t-1}(\mathbf{z}_{t-1})$ is kept fixed in this recursion step, so the new $\beta_{t-1}(\mathbf{z}_{t-1})$ follows by a division as in step 4 in Algorithm 2.

In the forward pass a new forward message α_t is computed analogously as a function of the old forward message α_{t-1}, the local potential ψ_t, and the backward message β_t, by constructing $\hat{q}_t(\mathbf{z}_t)$ and divididing by $\beta_t(\mathbf{z}_t)$. We initialize all messages with 1, so on the first forward pass the scheme is identical to the assumed density filtering algorithm discussed in Section 3.

A new two-slice marginal implies two new one-slice beliefs, so in step 4 we could compute two new messages instead of one. However, computing new backward (β_{t-1}) messages on a forward pass is redundant since these messages will be replaced on a backward pass before they would be used. A similar reasoning goes for the computation of new forward messages on a backward pass.

4.2. Iteration

If in step 3 of Algorithm 2 the new one-slice marginal is not approximated, i.e. if exact marginalization is performed, one forward, and one backward pass would suffice. This can easily be seen from the fact that the forward and backward messages can be computed independently: since the (exact) summation is a linear operation multiplying with $\beta_t(\mathbf{z}_t)$ in step 1 and dividing again in step 4 is redundant. In fact, the above scheme is identical to the two filter approach of finding smoothed estimates in a linear dynamical system if these redundant multiplication and division operations are left out. In the current setting however, with a "collapse" operation in step 3 that is not linear in α_{t-1} nor in β_t (see Appendix A), the forward and backward messages *do* interfere. Different backward messages β_t result in different forward messages α_t and vice versa.

So instead of performing one forward and backward pass, steps 1 to 4 in Algorithm 2 can be iterated to find local approximations that are as consistent as possible. In Section 5 we will introduce the objective that is minimized when such an iterative scheme is followed.

4.3. Supportiveness

One issue that is not discussed in Algorithm 2 is *supportiveness*. We say that step 4 in Algorithm 2 is supported, if all the beliefs that change because of the construction of the new messages remain normalizable. On the first forward pass this is automatically satisfied. Since α_1 is a proper distribution and $\beta_t = 1$ for all t and all ψ_t are proper conditional distributions, by induction, all α_t are proper distributions as well. A new message $\alpha_t^{\text{new}}(\mathbf{z}_t)$ changes belief $\hat{p}_{t+1}^{\text{new}}(\mathbf{z}_{t,t+1}) \propto \alpha_t^{\text{new}}(\mathbf{z}_t)\psi_t(\mathbf{z}_{t,t+1})\beta_{t+1}(\mathbf{z}_{t+1})$, which is normalizable by construction since $\beta_{t+1}(\mathbf{z}_{t+1})$ is 1 on the first forward pass. However, in general, due to the division in step 4, after a message is updated neighboring two-slice potentials are not guaranteed to be normalizable. For instance, on a backward pass, after replacing β_t with β_t^{new} (based on the two slice belief $\hat{p}_{t+1}^{\text{new}}(\mathbf{z}_{t,t+1})$), the neighboring belief

$$\hat{p}_t^{\text{new}}(\mathbf{z}_{t-1,t}) \propto \alpha_{t-1}(\mathbf{z}_{t-1})\psi_t(\mathbf{z}_{t-1,t})\beta_t^{\text{new}}(\mathbf{z}_t)$$

may not be normalizable. The requirement is that the sum of the respective inverse covariance matrices is positive definite. If a normalizability problem is detected, messages α_t^{new} and β_{t-1}^{new} can be computed as *damped* versions of the messages α_t^* and β_{t-1}^* suggested by step 4. We define damping as a convex combination of old and suggested messages in *canonical space* (see Appendix A), e.g. for a damped forward message:

$$\boldsymbol{\alpha}_t^{\text{new}} = \epsilon\boldsymbol{\alpha}_t^* + (1-\epsilon)\boldsymbol{\alpha}_t^{\mathbf{old}} . \tag{9}$$

In (9), $\boldsymbol{\alpha}_t^{\text{new}}$, $\boldsymbol{\alpha}_t^*$, and $\boldsymbol{\alpha}_t^{\mathbf{old}}$ (in boldface) are the canonical parameter vectors of their corresponding potentials. If a regular update ($\epsilon = 1$) results in a non-normalizable potential, a damping parameter ϵ with $0 \leq \epsilon < 1$ is chosen such that the resulting precision matrices for the neighboring two-slice belief are positive definite.

5. Free energy minimization

In this section we discuss the objective that is minimized when the steps in Algorithm 2 are iterated. We first show that in the exact case, the task of finding smoothed one and two-slice posteriors can be formulated as a minimization problem. Although this minimization problem remains intractable it will form the basis for an approximate objective as will be discussed in the second part of this section.

A variational distribution $p(\mathbf{z}_{1:T})$ can be introduced to get an objective $\mathcal{F}$ that is sometimes referred to as a *free-energy* (Yedidia *et al.*, 2001):

$$-\log Z = \min_p \mathcal{F}(p) \equiv \min_p \mathrm{KL}(p||p^*) - \log Z. \tag{10}$$

In the above p^* is a shorthand for the exact posterior $p(\mathbf{z}_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\theta})$, and the minimisation is over all proper distributions over the same domain as p^*. Since the KL-divergence is never negative and 0 if and only if $p = p^*$, $\mathcal{F}$ has a unique and correct minimum. In particular, if p forms a minimum of $\mathcal{F}$ its one and two slice marginals will be equal to those of p^*.

In terms of the potentials ψ_t, the exact posterior is written as:

$$p^*(\mathbf{z}_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\theta}) = \frac{1}{Z}\prod_{t=1}^{T}\psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t), \tag{11}$$

with $Z \equiv p^*(\mathbf{y}_{1:T}|\boldsymbol{\theta})$ the normalization constant. If we plug this into $\mathcal{F}$ we obtain

$$\mathcal{F}(p) = E(p) - S(p) \tag{12}$$

```
Algorithm 2 Expectation Propagation for an SLDS
Compute a forward pass by performing the following steps for $t = 1, 2, \ldots, T$, with $t' \equiv t$, and a
backward pass by performing the same steps for $t = T-1, T-2, \ldots, 1$, with $t' \equiv t-1$. Possibly
iterate forward-backward passes until convergence. At the boundaries keep $\alpha_0 = \beta_T = 1$.

  1. Construct a two-slice belief,
       $\hat{p}_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \propto \alpha_{t-1}(\mathbf{z}_{t-1}) \psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \beta_t(\mathbf{z}_t)$ .
  2. Marginalize to obtain a one-slice marginal
       $\hat{p}_t(\mathbf{z}_{t'}) = \sum_{\mathbf{z}_{t''}} \hat{p}_t(\mathbf{z}_{t-1}, \mathbf{z}_t)$ ,
     with $t'' \equiv \{t-1, t\} \backslash t'$.
  3. Find $\hat{q}_{t'}(\mathbf{z}_{t'})$ that approximates $\hat{p}_t(\mathbf{z}_{t'})$ best in Kullback-Leibler (KL) sense:
       $\hat{q}_{t'}(\mathbf{z}_{t'}) = \text{Collapse}\left(\hat{p}_t(\mathbf{z}'_t)\right)$ .
  4. Infer the new message by division.
       $\alpha_t(\mathbf{z}_t) = \frac{\hat{q}_t(\mathbf{z}_t)}{\beta_t(\mathbf{z}_t)}$ ,   $\beta_{t-1}(\mathbf{z}_{t-1}) = \frac{\hat{q}_{t-1}(\mathbf{z}_{t-1})}{\alpha_{t-1}(\mathbf{z}_{t-1})}$ .
```

$$E(p) \equiv -\sum_{t=1}^{T} \sum_{\mathbf{z}_{t-1,t}} p(\mathbf{z}_{t-1}, \mathbf{z}_t) \log \psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \tag{13}$$

$$S(p) \equiv -\sum_{\mathbf{z}_{1:T}} p(\mathbf{z}_{1:T}) \log p(\mathbf{z}_{1:T}). \tag{14}$$

We see that the first term $E(p)$, the *energy*, factors automatically since the posterior is a product of two-slice potentials ψ_t. For arbitrary $p(\mathbf{z}_{1:T})$ the second term $S(p)$, the *entropy*, does not factorize. However we can use the fact that the exact joint posterior on a chain is of the form (Whittaker, 1989)

$$p^*(\mathbf{z}_{1:T}) = \frac{\prod_{t=2}^{T} p^*(\mathbf{z}_{t-1}, \mathbf{z}_t)}{\prod_{t=2}^{T-1} p^*(\mathbf{z}_t)}. \tag{15}$$

We can therefore restrict the minimization problem to range over all possible *chains* parameterized by their one and two-slice marginals

$$p(\mathbf{z}_{1:T}) = \frac{\prod_{t=2}^{T} p_t(\mathbf{z}_{t-1}, \mathbf{z}_t)}{\prod_{t=2}^{T-1} q_t(\mathbf{z}_t)} . \tag{16}$$

In (16) the one and two slice marginals $q_t(\mathbf{z}_t)$ and $p_t(\mathbf{z}_{t-1}, \mathbf{z}_t)$ should be properly normalized distributions and such that they agree on their overlap:

$$\sum_{\mathbf{z}_{t-1}} p_t(\mathbf{z}_{t-1}, \mathbf{z}_t) = q_t(\mathbf{z}_t) = \sum_{\mathbf{z}_{t+1}} p_{t+1}(\mathbf{z}_t, \mathbf{z}_{t+1}). \tag{17}$$

Plugging (16) into (12) we get a minimization problem over a collection of one and two-slice marginals

$$\begin{aligned} -\log Z &= \min_{\{p_t, q_t\}} \mathcal{F}(\{p_t, q_t\}) \\ &\equiv \min_{\{p_t, q_t\}} \left\{ \sum_{t=1}^{T} \sum_{\mathbf{z}_{t-1,t}} p_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \log \frac{p_t(\mathbf{z}_{t-1}, \mathbf{z}_t)}{\psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t)} \right. \\ &\quad \left. - \sum_{t=2}^{T-1} \sum_{\mathbf{z}_t} q_t(\mathbf{z}_t) \log q_t(\mathbf{z}_t) \right\} , \end{aligned} \tag{18}$$

under the constraints that the marginals are proper distributions and consistent (17). So the exact posteriors can be found by a minimization procedure in terms of (constrained) one and two-slice marginals.

A problem remains of course that, as described in Section 2, the exact one and two-slice posteriors p_t and q_t are very complex, and in general will have exponentially many modes. So therefore, even if we would find a scheme that would minimise $\mathcal{F}(\{p_t, q_t\})$, the results could not be computed nor stored efficiently. So we will approximate (18) by restricting the one slice $q_t(\mathbf{z}_t)$ marginals to be conditionally Gaussian, i.e. the conditional posteriors $p(\mathbf{x}_t | s_t, \mathbf{y}_{1:T}, \boldsymbol{\theta})$ are approximated by a single Gaussian. The approximated, or 'pseudo' marginals are denoted by $\hat{p}_t$ and $\hat{q}_t$. The resulting free energy reads:

$$\mathcal{F}_{\text{EP}}(\hat{p}, \hat{q}) \equiv \sum_{t=1}^{T} \sum_{\mathbf{z}_{t-1,t}} \hat{p}_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \log \frac{\hat{p}_t(\mathbf{z}_{t-1}, \mathbf{z}_t)}{\psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t)}$$

$$-\sum_{t=2}^{T-1}\sum_{\mathbf{z}_t}\hat{q}_t(\mathbf{z}_t)\log\hat{q}_t(\mathbf{z}_t)\ . \tag{19}$$

Restricting the marginals makes exact agreement on overlaps possible only in trivial solutions. Instead the consistency requirements are weakened: overlapping two-slice marginals only need to agree on their *expectations*. That is, apart from being properly normalized, the marginals are required to have the same moments *after a collapse*:

$$\langle f(\mathbf{z}_t)\rangle_{\hat{p}_t} = \langle f(\mathbf{z}_t)\rangle_{\hat{q}_t} = \langle f(\mathbf{z}_t)\rangle_{\hat{p}_{t+1}}\ . \tag{20}$$

The intuition and motivation behind this approximate free energy is similar to the assumed density filtering approach. It is hoped that the collection of marginals $\hat{p}_t(\mathbf{z}_{t-1,t})$ and $\hat{q}_t(\mathbf{z}_t)$ are reasonable approximations to the exact marginals $p(\mathbf{z}_{t-1,t}|\mathbf{y}_{1:T},\boldsymbol{\theta})$ and $p(\mathbf{z}_t|\mathbf{y}_{1:T},\boldsymbol{\theta})$. The weak consistency constraints ensure that the two possible ways of computing the one-slice marginal $\hat{q}_t$ (based on $\hat{p}_t$ or $\hat{p}_{t+1}$) are identical *after* a collapse.

Iterating the forward and backward pass as described in Section 4 can be interpreted as a heuristic to find a minimum of $\mathcal{F}_{\text{EP}}$ under constraints (20). The following theorem describes this relationship.

Theorem 1. *The collection of beliefs $\hat{p}_t(\mathbf{z}_{t-1,t})$ and $\hat{q}_t(\mathbf{z}_t)$ form fixed points of Algorithm 2 if and only if they are zero gradient points of $\mathcal{F}_{\text{EP}}$ under the appropriate constraints.*

The proof is presented in Appendix B.

The relationship between the algorithm and the objective $\mathcal{F}_{\text{EP}}$ is in fact somewhat stronger. It can be shown that if the algorithm converges, the collection of beliefs correspond to a local *minimum* of $\mathcal{F}_{\text{EP}}$ under the constraints. The proof is somewhat more involved and is given in Heskes and Zoeter (2002).

It must be stressed that Theorem 1 does not claim that the algorithm always converges. In hard problems the algorithm might get trapped in cycles or diverge. For such hard problems it often helps to make smaller or *damped* updates (Eq.(9)). In practice we observe that Algorithm 2 nearly always converges to a very reasonable approximation and for more 'harder' problems damping resolves convergence problems. However in a thorough implementation direct minimization of $\mathcal{F}_{\text{EP}}$ may be used when Algorithm 2 fails to converge. A direct minimization procedure is presented in Heskes and Zoeter (2002). This procedure is a lot slower than Algorithm 2, so therefore the latter remains the method of choice for a practical application.

6. Alternative backward passes

6.1. Approximated backward messages

Now that the symmetric backward pass is introduced, we briefly describe previous approaches and show how they differ from the method proposed in Algorithm 2.

The forward pass in Kim and Nelson (1999) is identical to the assumed density filtering discussed in Section 3. The backward pass differs from the one proposed in Algorithm 2. We will refer to it as *alternative backward pass (ABP)* in the remainder of this article. The ABP is based on the traditional Kalman smoother form (as opposed to the two-filter approach to smoothing in EP). Instead of $\beta_t(\mathbf{z}_t)\approx p(\mathbf{y}_{t+1:T}|\mathbf{z}_t)$ messages, approximations to the smoothed posteriors $p(\mathbf{x}_t|s_t,\mathbf{y}_{1:T})$ and $p(s_t|\mathbf{y}_{1:T})$ form the basis for recursion. The smoother treats the discrete and continuous latent states separately and differently. This forces us to adapt our notation slightly. In this section we use $p(\cdot)$ for (uncollapsed) distributions over two-slices, $\psi(\cdot)$ for the model equations (to emphasize the similarities with the factors from Section 3), and $q(\cdot)$ for (collapsed) one-slice marginals. For compactness we do not explicitly write down the dependence on $\boldsymbol{\theta}$.

As in the forward pass M^2 modes are computed ($p(\mathbf{x}_t, s_t, s_{t+1}|\mathbf{y}_{1:T})$ for all instantiations of s_t and s_{t+1}) and subsequently collapsed:

$$\begin{aligned}&q(\mathbf{x}_t, s_t|\mathbf{y}_{1:T})\\&= \text{Collapse}\left(\sum_{s_{t+1},\mathbf{x}_{t+1}} p(\mathbf{x}_t,\mathbf{x}_{t+1},s_t,s_{t+1}|\mathbf{y}_{1:T})\right).\end{aligned} \tag{21}$$

It is in the construction of $p(\mathbf{x}_t,\mathbf{x}_{t+1},s_t,s_{t+1}|\mathbf{y}_{1:T})$ that the ABP differs from Algorithm . The conditional posterior over $\mathbf{x}_t$ is computed as follows (Kim and Nelson, 1999):

$$\begin{aligned}&p(\mathbf{x}_t,\mathbf{x}_{t+1}|s_t,s_{t+1},\mathbf{y}_{1:T})\\&\quad= p(\mathbf{x}_t|\mathbf{x}_{t+1},s_t,s_{t+1},\mathbf{y}_{1:t})\,p(\mathbf{x}_{t+1}|s_t,s_{t+1},\mathbf{y}_{1:T})\\&\quad\approx p(\mathbf{x}_t|\mathbf{x}_{t+1},s_t,s_{t+1},\mathbf{y}_{1:t})\,q(\mathbf{x}_{t+1}|s_{t+1},\mathbf{y}_{1:T}) \qquad (22)\\&\quad= \frac{p(\mathbf{x}_t,\mathbf{x}_{t+1}|s_t,s_{t+1},\mathbf{y}_{1:t})}{p(\mathbf{x}_{t+1}|s_t,s_{t+1},\mathbf{y}_{1:t})}\,q(\mathbf{x}_{t+1}|s_{t+1},\mathbf{y}_{1:T})\\&\quad= \frac{q(\mathbf{x}_t|s_t,\mathbf{y}_{1:t})\psi(\mathbf{x}_{t+1}|\mathbf{x}_t,s_t,s_{t+1})}{p(\mathbf{x}_{t+1}|s_t,s_{t+1},\mathbf{y}_{1:t})}\,q(\mathbf{x}_{t+1}|s_{t+1},\mathbf{y}_{1:T}),\end{aligned}$$

where the approximation in (22) is due to the fact that if we condition on $\mathbf{y}_\tau$, with $\tau\le t$, $\mathbf{x}_{t+1}$ is not independent of s_t.

The required posterior over the discrete latent state, $p(s_t = j, s_{t+1}=k|\mathbf{y}_{1:T})$, is computed as

$$\begin{aligned}p(s_t,s_{t+1}|\mathbf{y}_{1:T}) &= q(s_{t+1}|\mathbf{y}_{1:T})p(s_t|s_{t+1},\mathbf{y}_{1:T})\\&\approx q(s_{t+1}|\mathbf{y}_{1:T})p(s_t|s_{t+1},\mathbf{y}_{1:t})\end{aligned}$$

$$= \frac{q(s_{t+1}|\mathbf{y}_{1:T})p(s_t, s_{t+1}|\mathbf{y}_{1:t})}{q(s_{t+1}|\mathbf{y}_{1:t})}$$
$$= \frac{q(s_{t+1}|\mathbf{y}_{1:T})\psi(s_{t+1}|s_t)q(s_t|\mathbf{y}_{1:t})}{q(s_{t+1}|\mathbf{y}_{1:t})}. \quad (23)$$

Note that the need for the extra approximations in (23) and in (22) comes from the fact that the posteriors for discrete and continuous latent states are treated separately. The posteriors are computed by conditioning on either the discrete or the continuous latent state, i.e. only half of $\mathbf{z}_t$. A property of a Markov chain is that conditioning on the *entire* latent state at t renders past, future and observation independent. This property is exploited in the regular Kalman smoother and EP, but is not used in the ABP. Summarising, the ABP smoother from Kim and Nelson (1999) requires two additional approximations beyond the projection onto the CG family. In contrast, EP requires no additional approximations.

6.2. Partial smoothing

The filter in Shumway and Stoffer (1991) is related to the filter in Section 3. However, instead of approximating filtered and smoothed estimates with a CG distribution it approximates them with a single mode. The scheme computes

$$p(\mathbf{x}_{t-1,t}, s_t|\mathbf{y}_{1:t}) \propto \psi(\mathbf{y}_t, \mathbf{x}_t, s_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{y}_{1:t}) \quad (24)$$

$$q(\mathbf{x}_t|\mathbf{y}_{1:t}) = \text{Collapse}\left(\sum_{s_t, \mathbf{x}_{t-1}} p(\mathbf{x}_{t-1,t}, s_t|\mathbf{y}_{1:t})\right), \quad (25)$$

with $q(\mathbf{x}_t|\mathbf{y}_{1:t})$ a Gaussian distribution (compared to a mixture with M modes in the ABP and in Algorithm 2). This forward pass is also known as *generalized pseudo Bayes 1 (GPB 1)* (Bar-Shalom and Fortmann, 1988). The filtering recursion is treated in Shumway and Stoffer (1991) as if it were *exact* if the switches only govern the observation model (no links from s_t to $\mathbf{y}_t$ in Fig. 1). However, even with such restrictions the greedy local projections result in an approximation of the exact one and two slice marginals. Following the argumentations in Lauritzen (1992) it can be seen that the combinatoric explosion described in Section 2 is a property of all conditionally Gaussian dynamic models. It is essentially caused by the unobserved continuous chain $\mathbf{x}_{1:T}$ which couples all discrete states $s_{1:T}$.

In the smoothing pass of Shumway and Stoffer (1991) the probabilities over discrete states s_t are *not* smoothed (only approximations of filtered regime posteriors are available by integrating $\mathbf{x}_{t-1,t}$ in (24)). The posterior over continuous states $\mathbf{x}_t$ are computed using the standard Kalman smoother recursion as if there are no switches.

7. Related deterministic approximations

The emphasis in this paper is on the symmetric backward pass for the generalized pseudo Bayes algorithm. This approximation is based on the assumption that a one-slice posterior can be well approximated by a conditional Gaussian. This assumption is reasonable if there is relatively little ambiguity about the current regime and the posterior is peaked in a single mode, or if the ambiguity about the discrete state is high yielding a very broad posterior over the continuous state. In systems where several regime histories, i.e. succesive settings for the discrete states, have non-negligible posterior weight, *and* their corresponding components in $\mathbf{x}$ space are peaked and far apart, an approximation with too few components will lead to a poor approximation.

In a filtering pass it is relatively straightforward to keep more than M components in the approximation, either by selecting components with high weight, greedily merging components, or using other projection criteria. In general it is difficult to define a matching backward pass for such a general filter. In Kitagawa (1994) several proposals are made to start a backward filter that make only slight additional approximations beyond the ones from the forward pass. Messages in the forward and backward pass are approximated by greedily combining two components until a specified number of components is reached. With M components, this algorithm is computationally more expensive than the algorithm proposed here. Since the projection of components is not directly linked to values of s_t the method is more suited to infer the continuous part of the state space than the discrete part, but it gives the opportunity to use additional computer cycles to track more than M components.

In Lerner *et al.* (2000) a flexible filter is proposed where a fixed-lag smoother on only the discrete variables guides the projection onto an arbitrary number of K components. This will be particularly suited for on-line applications since essentially only a fixed number of time-slices need to be kept in memory.

In Zoeter and Heskes (2005) a generalization of the current framework is proposed that is again symmetric in forward-backward passes, i.e. does not introduce more approximations on the backward pass. However this algorithm can only work with a number of components M^κ, where κ is free to choose.

8. Experiments

In this section we compare the behavior of EP, the ABP and state-of-the-art sampling techniques. In the first part we will do this by generating artificial models and short data sequences. For these sequences all approximations can be compared with the exact results. In the second part we study the quality of the proposed approximation on longer sequences. Since for these longer sequences exact results are unattainable we compare the proposed method with Gibbs sampling.

8.1. Comparisons with exact posteriors

We generate models by drawing parameters from conjugate priors. The regime prior at $t = 1$ is multinomial. Its parameters are drawn from a uniform distribution and subsequently normalized. The parameters in the rows of $\Pi_{s_{t-1}\rightarrow s_t}$, the regime transition probabilities, are treated similarly. The elements of the initial state mean and the observation matrices C are drawn from a standard normal distribution. The state transition matrices A, are also constructed based on draws from the standard normal distribution. The covariances for the white Gaussian noise in the transition and observation models and for the initial state, are drawn from an inverse Wishart distribution with 10 degrees of freedom and scale matrix $0.01I$.

We drew 100 models with $\mathbf{x}_t \in R^3$, $\mathbf{y}_t \in R^2$ and two regimes, and generated a sequence of length eight for each of these models. Using (4) we computed the exact posteriors. For each task we computed approximate one-slice smoothed posteriors using the following methods.

EP1 Algorithm 2 with one forward-backward pass. The computational complexity[1] of this algorithm is $\mathcal{O}(M^2T)$, with M the number of regimes, and T the number of time-slices.

EP Algorithm 2 until convergence or at most 20 iterations. The computational complexity is $\mathcal{O}(M^2TI)$, with I the number of iterations.

ABP Using the approach from Kim and Nelson (1999) described in Section 6, the associated complexity is $\mathcal{O}(M^2T)$.

Gibbs 1000 samples generated using the efficient Gibbs sampler from Carter and Kohn (1996). In this sampler the continuous latent states $\mathbf{x}_{1:t}$ are intergrated out analytically. The computational cost is $\mathcal{O}(KMT)$, with K the number of samples. The first 20 samples of the MCMC chain are discarded. Note that with $K = 1020$ the computational complexity of this approach for the current setting with $T = 8$ and $M = 2$ is higher than that of exact computation which has associated complexity $\mathcal{O}(M^T)$.

RBPS-M Using Rao-Blackwellized particle smoothing (Doucet *et al.*, 2001; Chen and Liu, 2000). As in the Gibbs sampler, the performance of the particle smoother is significantly improved by analytically integrating out $\mathbf{x}_{1:T}$. The number of particles is taken identical to the number of regimes. The computational complexity is $\mathcal{O}(KMT)$, so choosing the number of particles, K, equal to M results in computational costs identical to EP1. Since the sequences are short, relatively few resampling steps are performed, and as a result the diversity at the start of the drawn sequences $s_{1:T}$ is still acceptable. So we did not implement extra "rejuvenation" methods (Doucet *et al.*, 2001) to increase variance on the smoothing pass.

RBPS-10M Using the Rao-Blackwellized particle smoother with the number of particles 10 times the number of regimes.

Ideally we would perhaps compute statistics such as $\mathrm{KL}(p_{\mathrm{exact}}(\mathbf{z}_t)||p_{\mathrm{approx}}(\mathbf{z}_t))$ for every time-slice and every approximation method. However this leads often to an infinite KL divergence for sampling approaches, since it is not uncommon that for a particular time-slice one of the regimes is not represented by at least one sample. Instead we compute the mean squared error for the posterior state mean, and the mean KL after collapsing both the exact and approximated posteriors onto a single Gaussian per state $\mathbf{x}_t$

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{KL}(\mathrm{Collapse}(p(\mathbf{x}_t|\mathbf{y}_{1:T},\boldsymbol{\theta}))||\mathrm{Collapse}(p_{\mathrm{approx}}(\mathbf{x}_t))),$$

i.e. a KL divergence that gives an indication of how well the mean and covariances of the exact and approximated posteriors match. Figure 3 gives the MSE based ranks of the algorithms we compared, an analogous figure with KL based ranks looks indistinguishable for this set of experiments and is not shown. Figure 4 gives a typical result.

From Fig. 3 we see that Algorithm 2 nearly always outperforms traditional methods. The results from Fig. 4 show that all methods give reasonable results (although Fig. 4 only shows results for a single task, the results for the others are comparable). Although both the particle smoother and the Gibbs sampler would give exact results in the limit of infinite K, the rate at which this is attained is slow. Comparing the results between the first iteration for EP with the forward-backward pass in ABP from Kim and Nelson (1999), we see that the extra approximations in the backward pass in Kim and Nelson (1999) indeed have a negative effect.

8.2. Comparisons with Gibbs sampling

In this section we study the quality of the approximations on sequences with 100 observations. In our experiments we

[1] The "big-O" notation gives the order of the computation time disregarding constant factors which depend on a.o. the particular implementation. The complexities of operations on Gaussian potentials depend on the dimensions of the observations and the states, but are equal for all methods. We therefore treat these operations as constants in the complexity description.

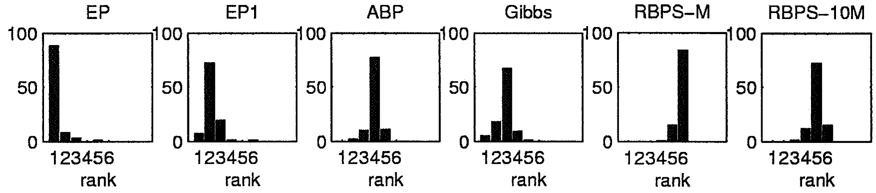

Fig. 3 Histogram of ranks of the tested algorithms on 100 artificially generated problems. Rank 1 indicates smallest distance with exact state posterior in MSE sense, rank 6 indicates largest distance

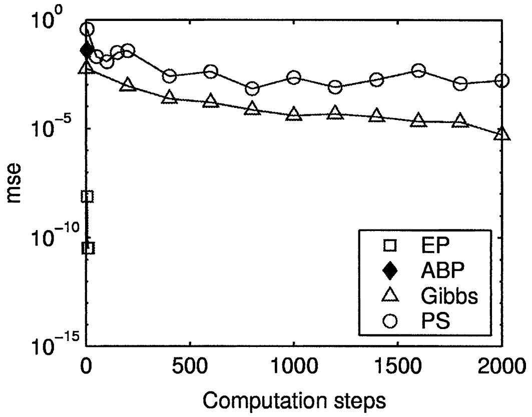

Fig. 4 Results from a "typical" task from the experiments represented in Fig. 3. Mean squared error is shown versus "computation steps". One iteration of EP is equivalent to M^2 such computation steps, drawing K samples for the Gibbs sampler or working with K particles in the particle smoother is equivalent to KM steps. This makes the x-axis roughly proportional to CPU time. For the particle smoother, 14 different runs with different values for K were used, a line connects these for clarity

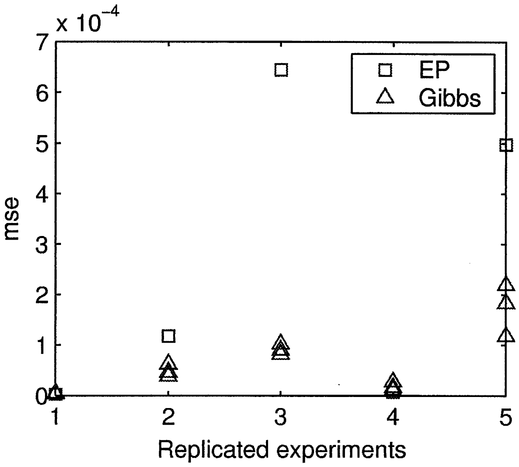

Fig. 6 Results from five experiments similar to the experiment from Fig. 5 based on three instead of five individual chains. For every replication the mean squared distance in the posterior state means between EP and the ground truth, and three individual Gibbs chains and the ground truth are plotted. The ground truth is taken to be the state means based on the combined three Gibbs chains

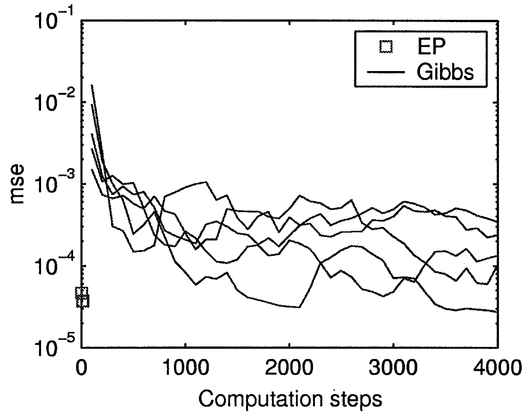

Fig. 5 Results from an experiment with a long observation sequence. Mean squared errors in the posterior state means for five independent Gibbs runs and EP are shown. The estimate based on samples from the combined five Gibbs runs is taken as ground truth. Observe that the distance of the deterministic approximation to the ground truth is of the same order as the individual Gibbs runs

compare the deterministic approximation with five independent runs of the Gibbs sampler, drawing 2000 samples each. This corresponds roughly to a minute computation time for the deterministic approach and a weekend for the five runs.

Figure 5 shows a comparison of the posterior means based on the individual Gibbs runs and the deterministic approximation and the posterior means computed from the combined samples of all the five runs. As can be seen all approximations lie relatively close, which probably indicates that all aproximations give reasonably correct results.

Since for any three vectors a, b, and c

$$\begin{aligned}(a-c)^{\top}(a-c) &= 2((a-b)^{\top}(a-b)+(b-c)^{\top}(b-c))\\ &\quad -(a+c-2b)^{\top}(a+c-2b)\\ &\leq 2((a-b)^{\top}(a-b)+(b-c)^{\top}(b-c)),\end{aligned}$$

we have an effective bound on the mean squared error of EP. Taking a, b, and c to be the EP, Gibbs, and exact posterior state means respectively, we get that the EP MSE is bounded by two times the sum of the Gibbs MSE and the Gibbs-EP mean squared distance:

$$E_{\mathrm{EP}} \leq 2\left(E_{\mathrm{Gibbs}} + D_{\mathrm{Gibbs,EP}}\right). \tag{26}$$

So from (26), and the analogous bound for the Gibbs error, we see that the difference in error between the Gibbs and EP approximation is of the order $D_{\mathrm{Gibbs,EP}}$. This Gibbs-EP mean squared distance can be read of from Fig. 5 and is relatively small.

Figure 5 shows results from a single experiment. In Fig. 6 the results of five replications of this experiment (based on three instead of five individual Gibbs chains) are shown. In the five replications the mean squared distance between EP and Gibbs is consistently small.

The experiments in this section give empirical evidence that our proposed approximation does not break down on longer sequences. For the analogous approximate filter in a fully discrete network (where the projection is onto a factorized distribution) Boyen and Koller (1998) show that the errors incurred by the approximation disappear at a geometric rate due to the stochastic nature of the transition model. Intuitively, as exact and approximate estimates are propagated through the transition model some of the "information" in both is forgotten, resulting in a smearing effect which makes both predicted distributions closer. Although the experiments in this section support a conjecture that such a proof can be extended to the conditional Gaussian case, the required technical conditions on the model and the proof itself remain work for future research.

In Zoeter and Heskes (2003) the approximation method is used to visualize high-dimensional time-series using projections onto a piece-wise linear manifold.

9. Discussion

We have introduced a novel deterministic approximation scheme for the well known inference problem in conditionally Gaussian state space models. Whereas the complexity of exact inference scales exponentially with the number of observations, the new approximate method requires computation time linear in the number of observations.

The approach can be seen as a symmetric backward pass to previously proposed assumed density filtering methods. In the literature several alternative backward passes have been introduced. An important benefit of the method described in this article is that the underlying philospphy for the forward and backward passes are the same and that, unlike the previously known deterministic methods, no additional approximations need to be made in the backward pass. Also no specific assumptions such as invertibility of transition or observation models are needed. Since both the forward and the backward passes perform greedy and local approximations it makes sense to iterate passes to find a "best" approximation. Section 5 describes a variant of the so-called Bethe free energy that is related to such a scheme. Fixed points of iterations of forward-backward passes correspond to extrema of this energy. The fixed points have a natural interpretation closely related to properties of the exact posterior. Given this theoretical support and the fact that the new method empirically seems to perform better than previously suggested backward passes we tend to conclude that these variants have now been superseded.

In Section 8 we have also presented comparisons with state-of-the-art sampling approaches: Gibbs sampling and particle smoothing. The experiments showed that all methods gave reasonably accurate results. Experiments with small problems where the approximations could be compared with exact results indicate that the rate of convergence of Gibbs sampling can be slow compared to the deterministic method. Even given a thousand fold extra computation time the estimation errors of the Gibbs sampler were typically larger than those for the deterministic approximation. Experiments with larger problems can no longer be compared with exact results. However the solutions found with Gibbs sampling and the deterministic approach are very close. This suggests that the quality of the deterministic approach does not break down in larger problems. The proposed method will be in particular suitable for problems that include both off-line phases (e.g. for parameter learning) and on-line phases (e.g. for monitoring tasks).

Several aspects of the proposed method require further research. Firstly, the essential dimensions in the SLDS model space are not known. In contrast for e.g. the class of Ising models (pairwise binary models) particular instances can be characterized based on the type of interactions (attractive, repulsive, or mixed) and their respective strengths. The experimental results reported here for switching linear dynamical systems are based on a particular mechanism for drawing models. The results are expected to be indicative for a wide range of models, but surely there will be models for which the approximations in the proposed algorithm is too severe. The model used in Lerner and Parr (2001) to obtain a reduction from the subset sum problem is an example. But this is an extreme example where few approximations are expected to give reasonable results. For models of intermediate difficulty, more flexible approximation schemes such as the Gibbs sampler might outperform the deterministic approach described here. A precise characterisation of such models is work for future research.

Another issue is numerical stability. With a straightforward implementation of the operations on conditional Gaussian potentials, problems with ill-conditioned covariance matrices and poor scaling of probabilities might occur. Just as with the canonical Kalman filter and the hidden Markov model, careful representations are required to obtain robust code.

A last topic is indicators of quality. Just as for the sampling approaches it is important that for the deterministic approach there is (at least a coarse) indication whether the approximation is acceptable or not. Empirically, convergence problems, or very slow convergence is a sign that the approximation may be off. But robust and theoretically motivated indicators require more research.

Appendix A. Operations on conditional Gaussian potentials

To allow for simple notation in the main text this appendix introduces the conditional Gaussian (CG) distribution. A discrete variable s and a continuous variable $\mathbf{x}$ are jointly CG distributed if the marginal of s is multinomial distributed and, conditioned on s, $\mathbf{x}$ is Gaussian distributed. Let $\mathbf{x}$ be d-dimensional and let S be the set of values s can take. In moment form the joint distribution reads

$$
\begin{aligned}
p(s, \mathbf{x}) = \; & \pi_s (2\pi)^{-d/2} |\Sigma_s|^{-1/2} \\
& \times \exp\left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_s)^\top \Sigma_s^{-1} (\mathbf{x} - \boldsymbol{\mu}_s) \right],
\end{aligned}
$$

with moment parameters $\{\pi_s, \boldsymbol{\mu}_s, \Sigma_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top\}$, where π_s is positive for all s and satisfies $\sum_s \pi_s = 1$ and Σ_s is a positive definite matrix. The definition of $\Sigma_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top$ instead of Σ_s is motivated by (29) below. For compact notation sets with elements dependent on s will implicitly ranges over $s \in S$. In canonical form the CG distribution is given by

$$
p(s, \mathbf{x}) = \exp\left[g_s + \mathbf{x}^\top \mathbf{h}_s - \frac{1}{2} \mathbf{x}^\top K_s \mathbf{x} \right], \tag{27}
$$

with canonical parameters $\{g_s, \mathbf{h}_s, K_s\}$.

The so-called *link function* $g(\cdot)$ maps canonical parameters to moment parameters:

$$
\begin{aligned}
g(\{g_s, \mathbf{h}_s, K_s\}) &= \left\{ \pi_s, \boldsymbol{\mu}_s, \Sigma_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top \right\} \\
\pi_s &= \exp(g_s - \bar{g}_s) \\
\boldsymbol{\mu}_s &= K_s^{-1} \mathbf{h}_s \\
\Sigma_s &= K_s^{-1} \; ,
\end{aligned}
$$

with $\bar{g}_s \equiv \frac{1}{2} \log |\frac{K_s}{2\pi}| - \frac{1}{2} \mathbf{h}_s^\top K_s \mathbf{h}_s$, the part of g_s that depends on $\mathbf{h}_s$ and K_s. The link function is unique and invertible:

$$
\begin{aligned}
g^{-1}\left(\left\{ \pi_s, \boldsymbol{\mu}_s, \Sigma_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top \right\} \right) &= \{g_s, \mathbf{h}_s, K_s\} \\
g_s &= \log \pi_s - \frac{1}{2} \log |2\pi \Sigma_s| \\
& \quad - \frac{1}{2} \boldsymbol{\mu}_s^\top \Sigma_s^{-1} \boldsymbol{\mu}_s \\
\mathbf{h}_s &= \Sigma_s^{-1} \boldsymbol{\mu}_s \\
K_s &= \Sigma_s^{-1} \; .
\end{aligned}
$$

A conditional Gaussian *potential* is a generalization of the above distribution in the sense that it has the same form as in (27) but need not integrate to 1. K_s is restricted to be symmetric, but need not be positive definite. If K_s is positive definite the moment parameters are determined by $g(\cdot)$. In this section we will use $\phi(s, \mathbf{x}; \{g_s, \mathbf{h}_s, K_s\})$ to denote a CG potential over s and $\mathbf{x}$ with canonical parameters $\{g_s, \mathbf{h}_s, K_s\}$.

Multiplication and division of CG potentials are the straightforward extensions of the analogous operations for multinomial and Gaussian potentials. In canonical form:

$$
\begin{aligned}
& \phi(s, \mathbf{x}; \{g_s, \mathbf{h}_s, K_s\}) \phi(s, \mathbf{x}; \{g'_s, \mathbf{h}'_s, K'_s\}) \\
& \quad = \phi(s, \mathbf{x}; \{g_s + g'_s, \mathbf{h}_s + \mathbf{h}'_s, K_s + K'_s\}) \\
& \phi(s, \mathbf{x}; \{g_s, \mathbf{h}_s, K_s\}) / \phi(s, \mathbf{x}; \{g'_s, \mathbf{h}'_s, K'_s\}) \\
& \quad = \phi(s, \mathbf{x}; \{g_s - g'_s, \mathbf{h}_s - \mathbf{h}'_s, K_s - K'_s\}).
\end{aligned}
$$

With the above definition of multiplication we can define a unit potential

$$
1(s, x) \equiv \phi(s, \mathbf{x}; \{0, \mathbf{0}, 0\}),
$$

which satisfies $1(s, x) p(s, \mathbf{x}) = p(s, x)$ for all CG potentials $p(s, \mathbf{x})$. We will sometimes use the shorthand 1 for the unit potential when its domain is clear from the text.

In a similar spirit we can define multiplication and division of potentials with different domains. If the domain of one of the potentials (the denominator in case of division) forms a subset of the domain of the other we can *extend* the smaller to match the larger and perform a regular multiplication or division as defined above. The continuous domain of the small potential is extended by adding zeros in $\mathbf{h}_s$ and K_s at the corresponding positions. The discrete domain is extended by replicating parameters, e.g. extending s to $[s \; t]^\top$ we use parameters $g_{st} = g_s$, $\mathbf{h}_{st} = \mathbf{h}_s$, and $K_{st} = K_s$.

Marginalization is less straightforward for CG potentials. Integrating out continuous dimensions is analogous to marginalization in Gaussian potentials and is only defined if the corresponding moment parameters are defined. Marginalization is then defined as converting to moment form, 'selecting' the appropriate rows and columns from $\boldsymbol{\mu}_s$ and Σ_s, and converting back to canonical form. More problematic is the marginalization over discrete dimensions of the CG potential. Summing out s results in a distribution $p(\mathbf{x})$ which is a mixture of Gaussians with mixing weights $p(s)$, i.e. the CG family *is not closed under summation*. In the text we will sometimes use, somewhat sloppily, the $\sum$ notation for both summing out discrete and integrating out continuous dimensions.

We define *weak marginalization* (Lauritzen, 1992), as exact marginalization followed by a *collapse*: a projection of the exact marginal onto the CG family. The projection minimizes the Kullback-Leibler divergence $KL(p||q)$ between

p, the exact (strong) marginal and q, the weak marginal:

$$\begin{aligned} q(s,\mathbf{x}) &= \underset{q\in CG}{\operatorname{argmin}}\, \mathrm{KL}(p||q) \\ &\equiv \underset{q\in CG}{\operatorname{argmin}} \sum_{s,\mathbf{x}} p(s,\mathbf{x}) \log \frac{p(s,\mathbf{x})}{q(s,\mathbf{x})}. \end{aligned}$$

This projection has the property that, conditioned on s the weak marginal has the same mean and covariance as the exact marginal. The weak marginal can be computed by *moment matching* (Whittaker, 1989). If $p(\mathbf{x}|s)$ is a mixture of Gaussians for every s with mixture weights $\pi_{r|s}$, means $\boldsymbol{\mu}_{sr}$, and covariances Σ_{sr} (e.g. the exact marginal $\sum_r p(s,r,\mathbf{x})$ of CG distribution $p(s,r,\mathbf{x})$), the moment matching procedure is defined as

$$\begin{aligned} \text{Collapse}\,(p(s,\mathbf{x})) &\equiv p(s)\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_s,\Sigma_s) \\ \boldsymbol{\mu}_s &\equiv \sum_r \pi_{r|s}\boldsymbol{\mu}_{sr} \\ \Sigma_s &\equiv \sum_r \pi_{r|s}\left(\Sigma_{sr} + (\boldsymbol{\mu}_{sr}-\boldsymbol{\mu}_s)\right. \\ &\qquad \left.\times\, (\boldsymbol{\mu}_{sr}-\boldsymbol{\mu}_s)^\top\right). \end{aligned}$$

Note that this projection, contrary to exact marginalization, is not linear, and hence in general:

$$\text{Collapse}\,(p(s,\mathbf{x})q(\mathbf{x})) \neq \text{Collapse}\,(p(s,\mathbf{x}))\,q(\mathbf{x}).$$

In even more compact notation, with $\delta_{s,m}$ the Kronecker delta function, we can write a CG potential as

$$\begin{aligned} p(s,\mathbf{x}) &= \exp[\nu^\top f(s,\mathbf{x})], \text{ with} \\ f(s,\mathbf{x}) &\equiv [\delta_{s,m},\ \delta_{s,m}\mathbf{x}^\top,\ \delta_{s,m}\mathrm{vec}(\mathbf{x}\mathbf{x}^\top)^\top | m\in S]^\top \qquad (28) \\ \nu &\equiv \left[g_s,\ \mathbf{h}_s^\top, -\frac{1}{2}\mathrm{vec}(K_s)^\top \middle| s\in S\right]^\top \end{aligned}$$

the sufficient statistics, and the canonical parameters respectively. In this notation the moment parameters follow from the canonical parameters as

$$\begin{aligned} g(\nu) &= \langle f(s,\mathbf{x})\rangle_{\exp[\nu^\top f(s,\mathbf{x})]} \\ &\equiv \sum_s \int d\mathbf{x} f(s,\mathbf{x}) \exp[\nu^\top f(s,\mathbf{x})]. \qquad (29) \end{aligned}$$

Appendix B. Proof of Theorem 1

In this section we present the proof of Theorem 1. The proof and intuition are analogous to the result that fixed points of loopy belief propagation can be mapped to extrema of the Bethe free energy (Yedidia *et al.*, 2001).

Theorem 1. *The collection of beliefs $\hat{p}_t(\mathbf{z}_{t-1,t})$ and $\hat{q}_t(\mathbf{z}_t)$ form fixed points of Algorithm 2 if and only if they are zero gradient points of $\mathcal{F}_{\mathrm{EP}}$ under the appropriate constraints.*

Proof: The properties of the fixed points of message passing follow from the description of Algorithm 2. We get the CG form (28) of messages α_t and β_t and their relationship with one and two slice marginals

$$\hat{p}_t(\mathbf{z}_{t-1,t}) \propto \alpha_{t-1}(\mathbf{z}_{t-1})\psi_t(\mathbf{z}_{t-1,t})\beta_t(\mathbf{z}_t) \qquad (30)$$

$$\hat{q}_t(\mathbf{z}_t) \propto \alpha_t(\mathbf{z}_t)\beta_t(\mathbf{z}_t) \qquad (31)$$

by construction, and consistency after a collapse

$$\langle f(\mathbf{z}_t)\rangle_{\hat{p}_t} = \langle f(\mathbf{z}_t)\rangle_{\hat{q}_t} = \langle f(\mathbf{z}_t)\rangle_{\hat{p}_{t+1}}, \qquad (32)$$

as a property of a fixed point.

To identify the nature of stationary points of $\mathcal{F}_{\mathrm{EP}}$ we first construct the Lagrangian by adding Lagrange multipliers $\boldsymbol{\alpha}_t(\mathbf{z}_t)$ and $\boldsymbol{\beta}_t(\mathbf{z}_t)$ for the forward and backward consistency constraints and $\gamma_{t-1,t}$ and γ_t for the normalization constraints.

$$\begin{aligned} &\mathcal{L}_{\mathrm{EP}}(\hat{p},\hat{q},\boldsymbol{\alpha},\boldsymbol{\beta},\gamma) \\ &= \sum_{t=1}^{T}\sum_{\mathbf{z}_{t-1,t}} \hat{p}_t(\mathbf{z}_{t-1,t}) \log\frac{\hat{p}_t(\mathbf{z}_{t-1,t})}{\psi_t(\mathbf{z}_{t-1,t})} \\ &\quad - \sum_{t=2}^{T-1}\sum_{\mathbf{z}_t} \hat{q}_t(\mathbf{z}_t)\log\hat{q}_t(\mathbf{z}_t) - \sum_{t=2}^{T-1}\boldsymbol{\alpha}_{t-1}(\mathbf{z}_{t-1})^\top \\ &\quad \times \left[\sum_{\mathbf{z}_{t-1,t}} f_{t-1}(\mathbf{z}_{t-1})\hat{p}_t(\mathbf{z}_{t-1,t}) - \sum_{\mathbf{z}_{t-1}} f_{t-1}(\mathbf{z}_{t-1})\hat{q}_{t-1}(\mathbf{z}_{t-1})\right] \\ &\quad - \sum_{t=2}^{T-1}\boldsymbol{\beta}_t(\mathbf{z}_t)^\top\left[\sum_{\mathbf{z}_{t-1,t}} f_t(\mathbf{z}_t)\hat{p}_t(\mathbf{z}_{t-1,t}) - \sum_{\mathbf{z}_t} f_t(\mathbf{z}_t)\hat{q}_t(\mathbf{z}_t)\right] \\ &\quad - \sum_{t=1}^{T}\gamma_{t-1,t}\left[\sum_{\mathbf{z}_{t-1,t}}\hat{p}_t(\mathbf{z}_{t-1,t}) - 1\right] \\ &\quad - \sum_{t=2}^{T-1}\gamma_t\left[\sum_{\mathbf{z}_t}\hat{q}_t(\mathbf{z}_t) - 1\right]. \end{aligned}$$

Note that $\boldsymbol{\alpha}_t(\mathbf{z}_t)$ and $\boldsymbol{\beta}_t(\mathbf{z}_t)$ (in boldface to distinguish them from messages and to emphasize that they are vectors) are vectors of canonical parameters as defined in Appendix A.

The stationarity conditions follow by setting the partial derivatives to 0. Taking derivatives w.r.t. $\hat{p}_t(\mathbf{z}_{t-1,t})$ and

$\hat{q}_t(\mathbf{z}_t)$ gives

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mathrm{EP}}}{\partial \hat{p}_t(\mathbf{z}_{t-1,t})} &= \log \hat{p}_t(\mathbf{z}_{t-1,t}) + 1 - \log \psi_t(\mathbf{z}_{t-1,t}) \\ &\quad - \boldsymbol{\alpha}_{t-1}(\mathbf{z}_{t-1})^\top f_{t-1}(\mathbf{z}_{t-1}) \\ &\quad - \boldsymbol{\beta}_t(\mathbf{z}_t)^\top f_t(\mathbf{z}_t) - \gamma_{t-1,t} \\ \frac{\partial \mathcal{L}_{\mathrm{EP}}}{\partial \hat{q}_t(\mathbf{z}_t)} &= -\log \hat{q}_t(\mathbf{z}_t) - 1 + \boldsymbol{\alpha}_t(\mathbf{z}_t)^\top f_t(\mathbf{z}_t) \\ &\quad + \boldsymbol{\beta}_t(\mathbf{z}_t)^\top f_t(\mathbf{z}_t) - \gamma_t.\end{aligned}$$

Setting above derivitives to 0 and filling in the solutions for $\gamma_{t-1,t}$ and γ_t (which simply form the log of the normalisation constants) results in

$$\begin{aligned}\hat{p}_t(\mathbf{z}_{t-1,t}) &\propto \mathrm{e}^{\boldsymbol{\alpha}_{t-1}(\mathbf{z}_{t-1})^\top f_{t-1}(\mathbf{z}_{t-1})} \psi_t(\mathbf{z}_{t-1,t}) \mathrm{e}^{\boldsymbol{\beta}_t(\mathbf{z}_t)^\top f_t(\mathbf{z}_t)} \\ \hat{q}_t(\mathbf{z}_t) &\propto \mathrm{e}^{\boldsymbol{\alpha}_t(\mathbf{z}_t)^\top f_t(\mathbf{z}_t) + \boldsymbol{\beta}_t(\mathbf{z}_t)^\top f_t(\mathbf{z}_t)}\ .\end{aligned}$$

The conditions $\frac{\partial \mathcal{L}_{\mathrm{EP}}}{\partial \boldsymbol{\alpha}_t(\mathbf{z}_t)} = 0$ and $\frac{\partial \mathcal{L}_{\mathrm{EP}}}{\partial \boldsymbol{\beta}_t(\mathbf{z}_t)} = 0$ retrieve the forward-equals-backward constraints (32).

So if we identify $\boldsymbol{\alpha}_t$ as the vector of the canonical parameters of the message α_t and $\boldsymbol{\beta}_t$ as the vector of the canonical parameters of the message β_t, we see that the conditions for stationarity of $\mathcal{F}_{\mathrm{EP}}$ and fixed points of Algorithm 2 are the same. □

As can be seen from the above proof, iteration of the forward-backward passes can be interpreted as fixed point iteration in terms of Lagrange multipliers.

References

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

Bar-Shalom Y. and Li. X.-R. 1993. Estimation and Tracking: Principles, Techniques, and Software. Artech House.

Boyen X. and Koller D. 1998. Tractable inference for complex stochastic processes. In Proceedings of the 14th Annual Conference on Uncertanty in Artificial Intelligence.

Carter C. and Kohn R. 1996. Markov chain Monte Carlo in conditionally Gaussian state space models. Biometrika 83(3):589–601.

Chen R. and Liu J. S. 2000. Mixture Kalman filters. Journal of the Royal Statistical Society, Series B 62:493–508.

Cover T. M. and Thomas J. A. 1991. Elements of Information Theory. John Wiley & Sons.

Doucet A., de Freitas N., Murphy K., and Russel S. 2001. Rao-Blackwellized particle filtering for dynamic Bayesian networks. In Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2001), San Francisco, CA. Morgan Kaufmann Publishers.

Hamilton J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57.

Harrison P. J. and Stevens C. F. 1976. Bayesian forecasting. Journal of the Royal Statistical Society Society B 38:205–247.

Helmick R. E., Blair W. D., and Hoffman S. A. 1995. Fixed-interval smoothing for Markovian switching systems. IEEE Transactions on Information Theory 41:1845–1855.

Heskes T. and Zoeter O. 2002. Expectation propagation for approximate inference in dynamic Bayesian networks. In Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI), San Francisco, CA. Morgan Kaufmann Publishers.

Kim C. -J. and Nelson Ch. R. 1999. State-Space Models with Regime Switching. MIT Press.

Kitagawa G. 1994. The two-filter formula for smoothing and an implementation of the Gaussian-sum filter. Annals of the Institute of Statistical Mathematics 46(4):605–623.

Kitagawa G. 1996. Monte Carlo filter and smoother for non-gaussian nonlinear state space models. Journal of Computational and Graphical Statistics 5(1):1–25.

Kschischang F. R., Frey B. J. and Loeliger H.-A. 2001. Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory 47(2):498–519.

Lauritzen S. L. 1992. Propagation of probabilities, means, and variances in mixed graphical association models. Journal of the American Statistical Association 87:1098–1108.

Lerner U. and Parr R. 2001. Inference in hybrid networks: Theoretical limits and practical algorithms. In Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2001). Morgan Kaufmann Publishers.

Lerner U., Parr R., Koller D. and Biswas G. 2000. Bayesian fault detection and diagnosis in dynamic systems. In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI), pp. 531–537.

Minka T. 2001 Expectation propagation for approximate Bayesian inference. In Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2001). Morgan Kaufmann Publishers.

Shumway R. H. and Stoffer D. S. 1991. Dynamic linear models with switching. Journal of the American Statistical Association, 86:763–769.

Whittaker J. 1989. Graphical Models in Applied Multivariate Statistics. John Wiley & Sons.

Yedidia J., Freeman W., and Weiss Y. 2001. Generalized belief propagation. In NIPS 13, pp. 689–695.

Zoeter O. and Heskes T. 2005. Change point problems in linear dynamical systems. Journal of Machine Learning Research 6:1999–2026.

Zoeter O. and Heskes T. 2003. Hierarchical visualization of time-series data using switching linear dynamical systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(10):1202–1214.