To appear in proceedings of ICANN 2003

Multi-Scale Switching Linear Dynamical Systems

Onno Zoeter* and Tom Heskes

SNN, University of Nijmegen
Geert Grooteplein 21, 6525 EZ, Nijmegen, The Netherlands
{*orzoeter, tom*}*@snn.kun.nl*

Abstract. Switching linear dynamic systems can monitor systems that operate in different regimes. In this article we introduce a class of multi-scale switching linear dynamical systems that are particularly suited if such regimes form a hierarchy. The setup consists of a specific switching linear dynamical system for every level of coarseness. Jeffrey's rule of conditioning is used to coordinate the models at the different levels.
When the models are appropriately constrained, inference at finer levels can be performed independently for every subtree. This makes it possible to determine the required degree of detail on-line. The refinements of very improbable regimes need not be explored.
The computational complexity of exact inference in both the standard and the multi-class switching linear dynamical system is exponential in the number of observations. We describe an appropriate approximate inference algorithm based on expectation propagation and relate it to a variant of the Bethe free energy.

1 Introduction

In a *linear dynamical system* (LDS) a hidden state variable $\mathbf{x}_t$ is assumed to evolve with Markovian, linear Gaussian dynamics, of which only noisy measurements $\mathbf{z}_t$ are available. In a *switching linear dynamical system* (SLDS) this model is extended with discrete switch states s_t that denote the *regime* the system is in. Within every regime the state $\mathbf{x}_t$ evolves with different dynamics and also the observation model $p(\mathbf{z}_t|\mathbf{x}_t, s_t)$ might be different. The regime itself also follows a first-order Markov process. The model equations read

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t = j, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_t; A_j\mathbf{x}_{t-1}, Q_j) \quad p(\mathbf{z}_t|\mathbf{x}_t, s_t = j, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}_t; C_j\mathbf{x}_t, R_j)$$

$$p(s_t = j|s_{t-1} = i, \boldsymbol{\theta}) = \Pi_{i\to j} \ ,$$

with $\mathcal{N}(.;.,.)$ the Gaussian probability density function, and $\boldsymbol{\theta}$ the parameters in the model. The prior $p(s_1|\boldsymbol{\theta})$ is taken multinomial and $p(\mathbf{x}_1|s_1, \boldsymbol{\theta})$ Gaussian.

In this article we are concerned with models where the regimes s_t have natural refinements in sub-regimes as defined below. We will restrict ourselfs to models with two levels: a fine grained and a coarse grained level. Extensions to multiple levels are straightforward. At the coarse grained level, we refer to the discrete

* O. Zoeter is supported by the Dutch Competence Centre Paper and Board

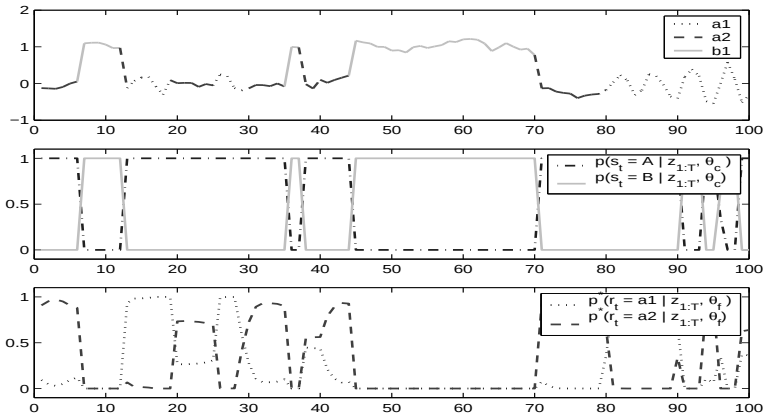

Fig. 1. A simple example: $M_c = \{A, B\}$, $\mathrm{ch}(A) = \{a_1, a_2\}$, and $\mathrm{ch}(B) = \{b_1\}$. Synthetic data was generated from $\boldsymbol{\theta}_f$ (top; colors indicate regimes as they were generated). First, posteriors for coarse regimes are inferred (middle). Then, with these results fixed, refinements of A and B are inferred independently (bottom; only results for submodel A are shown).

state as s_t, and to the continuous state as $\mathbf{x}_t$. We denote the set of regimes that s_t can take on with M_c. At the fine level we use r_t, $\mathbf{y}_t$ and M_f. The hierarchy, or grouping, we have in mind is a parent-child relationship ch(.): for every $j \in M_f$ there is exactly one $m \in M_c$ for which $j \in \mathrm{ch}(m)$ holds.

In the multi-scale setup of this article there are two different models: one for the coarse and one for the fine level. First the state and regime are inferred in the coarse model. Then, given the posterior probabilities for the coarse parent regimes, a refinement is inferred in the second model. Jeffrey's rule of conditioning is used to ensure that the posterior weights of the children add up to that of the parent. A simple example with synthetic data is presented in Fig. 1 (Matlab code with the full model description is available at `www.snn.kun.nl/~orzoeter/multiscale.html`). The two model setup is discussed in Sect. 2. In Sect. 3 restrictions for the fine level model are introduced so that refinements of the coarse regimes can be inferred independently. This way, in an on-line application, only probable regimes need to be explored in greater detail. In Sect. 4 we show how a deterministic approximation overcomes the computational complexity implied by the SLDS.

2 The Fine Level Model

The model for the first level is a basic SLDS. For the fine level model we want to ensure that the posterior probabilities of being in a child of m sum to the posterior probability of being in m in the coarse model:

$$\sum_{r_1 \in \mathrm{ch}(s_1)} \cdots \sum_{r_T \in \mathrm{ch}(s_T)} p(r_{1:T}|\mathbf{z}_{1:T}, \boldsymbol{\theta}_f) = p(s_{1:T}|\mathbf{z}_{1:T}, \boldsymbol{\theta}_c) \ .$$

To enforce these constraints we introduce extra random variables $\tilde{s}_{1:T}$ that have a link satisfying

$$p(r_t = j | r_{t-1} = i, \tilde{s}_t = m, \boldsymbol{\theta}_{\mathrm{f}}) = 0 \ , \quad \text{if } j \neq \mathrm{ch}(m) \ , \tag{1}$$

i.e. the link rules out the combination of a parent m with a "cousin" j. The motivation behind the introduction of $\tilde{s}_t$ is that we can now put constraints on the *sum* over possible values of r_t: $\sum_{j \in \mathrm{ch}(m)} p(r_t = j | \mathbf{z}_{1:T}, \boldsymbol{\theta}_{\mathrm{f}})$, instead of only on individual values. If $p(s_{1:T} | \mathbf{z}_{1:T}, \boldsymbol{\theta}_{\mathrm{c}})$ is crisp, making the fine model agree with the coarse can be done by treating $\tilde{s}_{1:T}$ as observations ("hard clamping").

If, however, $p(s_{1:T} | \mathbf{z}_{1:T}, \boldsymbol{\theta}_{\mathrm{c}})$ is not crisp, for example if it is the result of inference at the coarse level, we have to ensure that the marginal over $\tilde{s}_{1:T}$ is kept fixed to $p^*(\tilde{s}_{1:T} = \mathrm{path}_i) \equiv p(s_{1:T} = \mathrm{path}_i | \mathbf{z}_{1:T}, \boldsymbol{\theta}_{\mathrm{c}})$ ("soft clamping"). This is done by Jeffrey's rule of conditioning:

$$p^*(\tilde{s}_{1:T}, r_{1:T}, \mathbf{y}_{1:T} | \mathbf{z}_{1:T}, \boldsymbol{\theta}_{\mathrm{f}}) \equiv p(r_{1:T}, \mathbf{z}_{1:T} | \tilde{s}_{1:T}, \mathbf{z}_{1:T}, \boldsymbol{\theta}_{\mathrm{f}}) p^*(\tilde{s}_{1:T}) \ , \tag{2}$$

We denote probabilities that have the constraints on $\tilde{s}_{1:T}$ enforced with $p^*(.)$.

3 Independent Submodels

The fine level model described in Sect. 2 ensures consistency with the already inferred coarse model. It is however necessary to treat the refinements of all coarse regimes together in one large model for the second level. In this section we will alleviate this requirement by making appropriate choices for the state and regime transtition probabilities such that the sub-models become independent.

The idea is that the model is restricted such that whenever there is a switch in coarse regimes ($\tilde{s}_{t-1} \neq \tilde{s}_t$) the discrete and continuous latent states at the fine level uncouple (see Fig. 2). I.e. when $\tilde{s}_{t-1} \neq \tilde{s}_t$, $\{r_t, \mathbf{y}_t\}$ does not depend on $\{r_{t-1}, \mathbf{y}_{t-1}\}$. For the continous state this is accomplished by introducing a "reset" after a switch in $\tilde{s}$: the new state is drawn from a Gaussian prior:

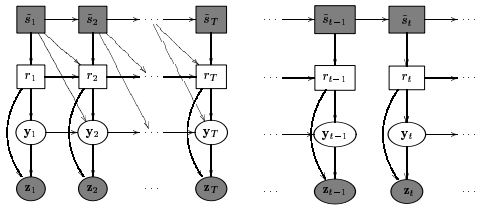

Fig. 2. Graphical structure corresponding to the fine level model. Left: when $\tilde{s}_{t-1} \neq \tilde{s}_t$, right: when $\tilde{s}_{t-1} \neq \tilde{s}_t$.

$$p(\mathbf{y}_t | \mathbf{y}_{t-1}, r_t = j, \tilde{s}_{t-1} = l, \tilde{s}_t = m, \boldsymbol{\theta}_{\mathrm{f}}) = \begin{cases} \mathcal{N}(\mathbf{y}_t; A_j \mathbf{y}_{t-1}, Q_j) & : \ l = m \\ \mathcal{N}(\mathbf{y}_t; \boldsymbol{\nu}_j, \Sigma_j) & : \ l \neq m \ . \end{cases}$$

The regime transition probability is constrained similarly:

$$p(r_t = j | r_{t-1} = i, \tilde{s}_{t-1} = l, \tilde{s}_t = m, \boldsymbol{\theta}_{\mathrm{f}}) = \begin{cases} 0 : j \notin \mathrm{ch}(m) & (3-a) \\ \Pi_{i \to j | m} : l = m, j \in \mathrm{ch}(m) & (3-b) \\ \pi_{j|m} : l \neq m, j \in \mathrm{ch}(m) \ . & (3-c) \end{cases}$$

Case (3-a) encodes the already discussed constraint that every child has only one parent (1). Case (3-b) describes the probability of jumping between regimes within one subtree ("brothers"). These probabilities are fully modeled. Case (3-c) states that if a jump at the coarse level occurs, the fine level regime is drawn from a prior. Note that, to make notation not too complex, we have left out a possible dependence of $\mathbf{y}_t$ on r_{t-1}, and a dependence of r_t on $\tilde{s}_{t-1}$. Crucial for the sub-models to become independent is only the conditional independence depicted in Fig. 2 for the case $\tilde{s}_{t-1} \neq \tilde{s}_t$.

The conditional independencies in the model allow us to write

$$p(r_{1:T}, \mathbf{y}_{1:T} | \mathbf{z}_{1:T}, \tilde{s}_{1:T}, \boldsymbol{\theta}_\mathrm{f}) \propto \prod_{\substack{t=2 \\ \tilde{s}_{t-1}=\tilde{s}_t}}^{T} p(\mathbf{z}_t, \mathbf{y}_t, r_t | \mathbf{y}_{t-1}, r_{t-1}, \tilde{s}_t, \boldsymbol{\theta}_\mathrm{f}) \prod_{\substack{t=1 \\ \tilde{s}_{t-1}\neq\tilde{s}_t}}^{T} p(\mathbf{z}_t, \mathbf{y}_t, r_t | \tilde{s}_t, \boldsymbol{\theta}_\mathrm{f})$$

$$= \prod_{m \in M_\mathrm{c}} \left\{ \prod_{\substack{t=2 \\ \tilde{s}_{t-1}=\tilde{s}_t=m}}^{T} p(\mathbf{z}_t, \mathbf{y}_t, r_t^{(m)} | \mathbf{y}_{t-1}, r_{t-1}^{(m)}, \tilde{s}_t = m, \boldsymbol{\theta}^{(m)}) \times \right.$$

$$\left. \prod_{\substack{t=1 \\ \tilde{s}_{t-1}\neq\tilde{s}_t=m}}^{T} p(\mathbf{z}_t, \mathbf{y}_t, r_t^{(m)} | \tilde{s}_t = m, \boldsymbol{\theta}^{(m)}) \right\}, \qquad (4)$$

where $\boldsymbol{\theta}^{(m)}$ are the disjoint parameter sets that together form $\boldsymbol{\theta}_\mathrm{f}$, and $r_t^{(m)}$ are variables ranging over $\mathrm{ch}(m)$. The boundary $t = 1$ can be taken into account by setting $\tilde{s}_0 \neq m$ for all m.

The marginal $p^*(\tilde{s}_{1:T})$ is fixed, so by (4) the posterior $p^*(\tilde{s}_{1:T}, r_{1:T}, \mathbf{y}_{1:T} | \mathbf{z}_{1:T}, \boldsymbol{\theta}_\mathrm{f})$ in (2) factors into independent subtree terms. Therefore these terms, and hence filtered or smoothed one-slice posteriors, can be computed independently.

4 Approximate Inference

4.1 The Coarse Level Model

The computational complexity of exact inference in an SLDS is exponential in the number of observations. One way to see this is to look at the posterior

$$p(\mathbf{x}_t | \mathbf{z}_{1:T}) = \sum_{s_{1:T}} p(\mathbf{x}_t | s_{1:T}, \mathbf{z}_{1:T}) p(s_{1:T} | \mathbf{z}_{1:T}) \ ,$$

where for notational convenience we drop the dependence on $\boldsymbol{\theta}_\mathrm{c}$. Every regime history $s_{1:T}$ gives rise to a different Gaussian $p(\mathbf{x}_t | s_{1:T}, \mathbf{z}_{1:T})$. Since $s_{1:T}$ is not observed, we have to take every possible history into account and integrate these out. So the exact posterior is a mixture with $|M_\mathrm{c}|^T$ components. In this section we therefore describe a greedy approximate inference strategy for the SLDS. An adaptation for the fine level SLDS from Sect. 2 is presented in Sect. 4.2.

The approximation is a particular form of *expectation propagation* [4]. We present it here in the spirit of the sum-product algorithm [3]. For ease of notation and interpretability we treat $\mathbf{u}_t = \{s_t, \mathbf{x}_t\}$ together as one *conditionally*

Gaussian distributed random variable. Slightly abusing notation, we will use the sum sign for the combined operation of summing out s_t and integrating out $\mathbf{x}_t$. The posterior distribution can be written as a product of *local potentials*, ψ_t:

$$p(\mathbf{u}_{1:T}|\mathbf{z}_{1:T}) \propto \prod_{t=1}^{T} \psi_t(\mathbf{u}_{t-1}, \mathbf{u}_t), \text{ with } \psi_t(\mathbf{u}_{t-1}, \mathbf{u}_t) \equiv p(\mathbf{z}_t|\mathbf{u}_t)p(\mathbf{u}_t|\mathbf{u}_{t-1}) ,$$

and $\psi_1(\mathbf{u}_0, \mathbf{u}_1) \equiv p(\mathbf{z}_1|\mathbf{u}_1)$. We are interested in one and two-slice marginals of the joint posterior. To avoid having to construct the entire posterior, these marginals are computed by local operations where *messages* are sent between nodes in a graph. We distinguish *variable nodes* that are associated with $\mathbf{u}_t$'s and *function nodes* that are associated with ψ_t's. The message from ψ_t forward to $\mathbf{u}_t$ is called $\alpha_t(\mathbf{u}_t)$ and the message from ψ_t back to $\mathbf{u}_{t-1}$ is referred to as $\beta_{t-1}(\mathbf{u}_{t-1})$. In a chain, variable nodes simply pass on the messages they receive. The message passing scheme is depicted in Fig. 3.

We denote the approximation of $p(\mathbf{u}_t|\mathbf{y}_{1:T})$ by $q_t(\mathbf{u}_t)$. It is computed by multiplying all incoming messages from neighboring function nodes: $q_t(\mathbf{u}_t) \propto \alpha_t(\mathbf{u}_t)\beta_t(\mathbf{u}_t)$. Associated with every function node is an approximate two-slice belief that we denote by $p(\mathbf{u}_{t-1}, \mathbf{u}_t|\mathbf{y}_{1:T}) \approx \hat{p}_t(\mathbf{u}_{t-1}, \mathbf{u}_t)$. Throughout the procedure we will ensure, using greedy approximations, that both $\hat{p}_t$ and q_t are conditionally Gaussian (CG) distributed. New messages from function node ψ_t to variable node $\mathbf{u}_{t'}$, where t' can be $t-1$ or t are computed as follows.

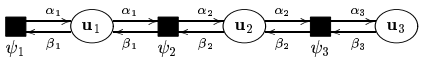

Fig. 3. Message propagation.

1. Construct a two-slice belief by multiplying the potential corresponding to the local function node ψ_t with *all* messages from neighboring variable nodes to ψ_t, yielding

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

2. The one-slice marginal $\hat{p}_t(\mathbf{u}_{t'}) = \sum_{\mathbf{u}_{t''}} \hat{p}_t(\mathbf{u}_{t''}, \mathbf{u}_t)$, with $t'' = \{t-1, t\}\backslash t'$, is not CG, but more complex: for every value of $s_{t'}$, $\mathbf{x}_{t'}$ follows a *mixture* of Gaussians. Find $q_{t'}(\mathbf{u}_{t'})$ that approximates $\hat{p}_t(\mathbf{u}_{t'})$ best in Kullback-Leibler (KL) sense:

$$q_{t'}(\mathbf{u}_{t'}) = \operatorname*{argmin}_{q \in \mathrm{CG}} \sum_{\mathbf{u}_{t'}} \hat{p}_t(\mathbf{u}_{t'}) \log \frac{\hat{p}_t(\mathbf{u}_{t'})}{q_{t'}(\mathbf{u}_{t'})} .$$

It can be shown that $q_{t'}$ follows by "collapsing" the mixture

$$q_{t'}(\mathbf{u}_{t'}) \propto \mathrm{Collapse}\Big(\sum_{\mathbf{u}_{t''}} \hat{p}_t(\mathbf{u}_{t''}, \mathbf{u}_t)\Big) .$$

Where $\mathrm{Collapse}\big(p_{ij}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ij}, \Sigma_{ij})\big) \equiv p_j\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_j, \Sigma_j)$, with $p_j = \sum_i p_{ij}$, $\boldsymbol{\mu}_j = \sum_i p_{i|j}\boldsymbol{\mu}_{ij}$, $\Sigma_j = \sum_i p_{i|j}(\Sigma_{ij} + (\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_j)^T)$, and $p_{i|j} = p_{ij}/p_j$.

3. Infer the new message by division. All messages not sent from ψ_t remain fixed, in particular β_t and α_{t-1}, so new messages are computed as

$$\alpha_t(\mathbf{u}_t) = \frac{q_t(\mathbf{u}_t)}{\beta_t(\mathbf{u}_t)} \ , \quad \beta_{t-1}(\mathbf{u}_{t-1}) = \frac{q_{t-1}(\mathbf{u}_{t-1})}{\alpha_{t-1}(\mathbf{u}_{t-1})} \ .$$

For filtering the messages are initialized with $\mathbf{1}$, and steps 1. to 3. are performed for $t = 1 : T$ sequentially. For the best smoothed posterior the above steps are iterated (e.g. using forward-backward passes) until convergence.

Fixed points of steps 1. to 3. correspond to stationary points of a "Bethe free energy"

$$\mathcal{F}_{\mathrm{EP}}(\hat{p}, q) = \sum_t \sum_{\mathbf{u}_{t-1,t}} \hat{p}_t(\mathbf{u}_{t-1,t}) \log \frac{\hat{p}_t(\mathbf{u}_{t-1,t})}{\psi_t(\mathbf{u}_{t-1,t})} - \sum_t \sum_{\mathbf{u}_t} q_t(\mathbf{u}_t) \log q_t(\mathbf{u}_t) \ , \quad (5)$$

subject to the constraints that all $\hat{p}_t$'s and q_t's sum to 1, and "weak" consistency constraints:

$$\mathrm{Collapse}\Big(\sum_{\mathbf{u}_{t-1}} \hat{p}_t(\mathbf{u}_{t-1}, \mathbf{u}_t) \Big) = q_t(\mathbf{u}_t) = \mathrm{Collapse}\Big(\sum_{\mathbf{u}_{t+1}} \hat{p}_{t+1}(\mathbf{u}_t, \mathbf{u}_{t+1}) \Big) \ .$$

This relationship is analogous to the one between the Bethe free energy and loopy belief propagation [6]. The only difference is that the strong consistency constraints are replaced by the weak ones: i.e. here overlapping beliefs only have to agree on their *expectations*. The proof of this claim is similar to the one in [6] and follows by constructing the Lagrangian and setting its derivatives to 0. In the resulting stationary conditions the Lagrange multipliers added for the weak consistency constraints have a one-to-one correspondence with messages α_t and β_t: the multipliers form the canonical parameters of the messages. Given this relationship the mapping between fixed points of the message passing scheme and stationary of points of $\mathcal{F}_{\mathrm{EP}}$ follows easily.

Iterating above steps can be seen as a procedure that greedily tries to find one and two-slice marginals that approximate the exact beliefs as good as possible and are pairwise consistent *after a collapse*. The details of the approximation are beyond the scope of this text. We refer the interested reader to [2].

4.2 The Fine Level Model

At the fine level we treat the marginal $p^*(\tilde{s}_{1:T}) \equiv p(s_{1:T}|\mathbf{z}_{1:T}, \boldsymbol{\theta}_{\mathrm{c}})$, as fixed and use Jeffrey's rule of conditioning and the extra $\tilde{s}_t$ nodes to ensure that the fine level model is consistent with the coarse level model. In the approximate inference procedure described in the previous section $p(s_{1:T}|\mathbf{z}_{1:T}, \boldsymbol{\theta}_{\mathrm{c}})$ is approximated by overlapping two-slice marginals. We use these to enforce consistency: $p(\tilde{s}_{t-1}, \tilde{s}_t|\mathbf{z}_{1:T}, \boldsymbol{\theta}_{\mathrm{f}}) = \hat{p}_t(s_{t-1}, s_t)$. So, effectively, distant interactions are disregarded.

We will first describe an approximation scheme for the general fine level model and deal with independent sub-models later. The approximation is based

on a free energy identical to (5) but now with definition $\mathbf{u}_t \equiv \{\tilde{s}_t, r_t, \mathbf{y}_t\}$ and the constraints that all $\hat{p}_t$'s and q_t's sum to one replaced by

$$\sum_{r_{t-1,t}, \mathbf{y}_{t-1,t}} \hat{p}_t(\mathbf{u}_{t-1}, \mathbf{u}_t) = p^*(\tilde{s}_{t-1}, \tilde{s}_t)$$

(since $\sum_{\tilde{s}_{t-1,t}} p^*(\tilde{s}_{t-1}, \tilde{s}_t) = 1$ proper normalization is automatically enforced). In the way new messages are computed only the first step needs to be changed:

1′. Construct a two-slice belief that has the correct marginal over $\tilde{s}_{t-1,t}$:

$$\hat{p}_t(\mathbf{u}_{t-1}, \mathbf{u}_t) \propto \frac{\alpha_{t-1}(\mathbf{u}_{t-1})\psi_t(\mathbf{u}_{t-1}, \mathbf{u}_t)\beta_t(\mathbf{u}_t)}{\sum_{\substack{r_{t-1,t} \\ \mathbf{y}_{t-1,t}}} \alpha_{t-1}(\mathbf{u}_{t-1})\psi_t(\mathbf{u}_{t-1}, \mathbf{u}_t)\beta_t(\mathbf{u}_t)} p^*(\tilde{s}_{t-1,t}) \; .$$

Fixed points of this new message passing scheme correspond to stationary points of $\mathcal{F}_{\mathrm{EP}}$ with the changed normalization constraints. The proof is analogous to the proof for the standard case presented in Sect. 4.1. (see also [5] for a related algorithm). The intuition behind the free energy is similar: the adapted message passing scheme tries to find one and two-slice marginals that approximate the exact beliefs as good as possible, are pairwise consistent after a collapse, *and* are consistent with the soft-assignments to regimes in the coarse level model.

Having established a way to infer posteriors for a general fine level model, we now adapt the above message passing scheme such that independent sub-models as described in Sect. 3 can be handled independently.

One possible way to adapt the scheme is to work with discrete variables $r^{(m+)}$ that range over $\mathrm{ch}(m) \cup \bar{m}$, where $\bar{m}$ is a special state that encodes "not in subtree m". This would however imply some inefficiency, since when we are refining regime m we are not interested in a continuous state associated with $\bar{m}$, nor in the mode $\hat{p}_t(\bar{m}, \bar{m})$. Instead, the inference algorithm for sub-model m only computes the required parts of the two-slice joint $\hat{p}_t$. In the remainder define $\mathbf{u}_t \equiv \{r_t^{(m)}, \mathbf{y}_t\}$ and

$$\begin{aligned} \psi_t^{(mm)}(\mathbf{u}_{t-1,t}) &\equiv p(\mathbf{z}_t, \mathbf{u}_t | \mathbf{u}_{t-1}, \tilde{s}_{t-1} = \tilde{s}_t = m, \boldsymbol{\theta}^{(m)}) \\ \psi_t^{(\bar{m}m)}(\mathbf{u}_t) &\equiv p(\mathbf{z}_t, \mathbf{u}_t | \tilde{s}_{t-1} \neq \tilde{s}_t = m, \boldsymbol{\theta}^{(m)}) \; , \end{aligned}$$

with $\psi_1^{(mm)}(\mathbf{u}_{0,1}) \equiv 0$ and $\psi_1^{(\bar{m}m)}(\mathbf{u}_0) \equiv p(\mathbf{z}_1, \mathbf{u}_1 | \tilde{s}_1, \boldsymbol{\theta}^{(m)})$.

To infer the refinements of the coarse regime m we use the message passing scheme of the general fine level, but with steps 1. and 2. adapted as follows.

1″. Construct the required parts of the two-slice marginal as

$$\begin{aligned} \hat{p}_t^{(mm)}(\mathbf{u}_{t-1,t}) &= p^*(\tilde{s}_{t-1} = \tilde{s}_t = m) \Big(Z_t^{(mm)} \Big)^{-1} \alpha_{t-1}(\mathbf{u}_{t-1}) \psi_t^{(mm)}(\mathbf{u}_{t-1,t}) \beta_t(\mathbf{u}_t) \\ \hat{p}_t^{(\bar{m}m)}(\mathbf{u}_t) &= p^*(\tilde{s}_{t-1} \neq \tilde{s}_t = m) \Big(Z_t^{(\bar{m}m)} \Big)^{-1} \psi_t^{(\bar{m}m)}(\mathbf{u}_t) \beta_t(\mathbf{u}_t) \\ \hat{p}_t^{(m\bar{m})}(\mathbf{u}_{t-1}) &= p^*(\tilde{s}_{t-1} = m \neq \tilde{s}_t) \big(Z_t^{(m\bar{m})} \Big)^{-1} \alpha_{t-1}(\mathbf{u}_{t-1}) \; , \end{aligned}$$

with $Z_t^{(mm)}$, $Z_t^{(\bar{m}m)}$, and $Z_t^{(m\bar{m})}$ the proper normalization constants of the r.h.s. *before* weighting with p^*.

$2''$. In a forward pass compute $q_t(\mathbf{u}_t) = \text{Collapse}\left(\hat{p}_t^{(mm)}(\mathbf{u}_t) + \hat{p}_t^{(\bar{m}m)}(\mathbf{u}_t)\right)$. In a backward pass compute $q_{t-1}(\mathbf{u}_{t-1}) = \text{Collapse}\left(\hat{p}_t^{(mm)}(\mathbf{u}_{t-1}) + \hat{p}_t^{(m\bar{m})}(\mathbf{u}_{t-1})\right)$.

The exposition has been restricted to two levels, but we can extend the approach to any number of scales. The approximations of $p^*(r_{t-1} = r_t = j|\mathbf{z}_{1:T})$, $p^*(r_{t-1} = j \neq r_t|\mathbf{z}_{1:T})$, and $p^*(r_{t-1} \neq j = r_t|\mathbf{z}_{1:T})$ form the constraints for the refinements of regime j and can be computed from the $\hat{p}_t$'s.

5 Discussion

We have introduced a class of switching linear dynamical system models that allows iterative refinement of regimes. If properly restricted, the models at levels of finer detail can be inferred independently. One of the advantages of this is that relatively complex models can be tracked at reasonable computational costs since only those regimes that have reasonable probability need to be refined. For instance to refine coarse regime A in Fig. 1, $t = 45 : 70$ can be disregarded.

The hierarchy and independence between sub-models allows recursive maximum likelihood fitting of parameters using an EM algorithm. In [7] this approach is used to interactively fit a hierarchical SLDS. An appealing line of future research is to use the multi-scale setup as a basis for greedy model learning.

The notion of multiple scales in statistical models is not new. There are many uses of multi-scale models in various disciplines. Our method shares with [1] the "top-down" construction of the hierarchy and the use of Jeffrey's rule to synchronize models at different scales. To our knowledge the work presented here is the first to enforce constraints in hybrid models for which exact inference is intractable. The approximate inference method from Sect. 4.2 is very general. Extensions to trees, or even structures containing cycles, are possible. It therefore paves the way for interesting combinations of previously proposed multi-scale models.

References

1. M. Ferreira, M. West, H. Lee, and D. Higdon. A class of multi-scale time series models. Unpublished.
2. T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *UAI-2002*, pages 216–223, 2002.
3. F. Kschischang, B. Frey, and H. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.
4. T. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of UAI-2001*, pages 362–369, 2001.
5. Y. Teh and M. Welling. The unified propagation and scaling algorithm. In *Advances in Neural Information Processing Systems 14*, page (in press). MIT Press, 2002.
6. J. Yedidia, W. Freeman, and Y. Weiss. Generalized belief propagation. In *NIPS 13*, pages 689–695, 2001.
7. O. Zoeter and T. Heskes. Hierarchical visualization of time-series data using switching linear dynamical systems. Submitted.