Journal of Machine Learning Research 6 (2005) 1999–2026 Submitted 2/05; Published 12/05

Change Point Problems in Linear Dynamical Systems

Onno Zoeter O.ZOETER@SCIENCE.RU.NL
SNN, Biophysics
Radboud University Nijmegen
Geert Grooteplein 21
NL 6525 EZ, Nijmegen, The Netherlands

Tom Heskes T.HESKES@SCIENCE.RU.NL
Computer Science
Radboud University Nijmegen
Toernooiveld 1
NL 6525 ED Nijmegen, The Netherlands

Editor: Donald Geman

Abstract

We study the problem of learning two regimes (we have a normal and a prefault regime in mind) based on a train set of non-Markovian observation sequences. Key to the model is that we assume that once the system switches from the normal to the prefault regime it cannot restore and will eventually result in a fault. We refer to the particular setting as *semi-supervised* since we assume the only information given to the learner is whether a particular sequence ended with a stop (implying that the sequence was generated by the normal regime) or with a fault (implying that there was a switch from the normal to the fault regime). In the latter case the particular time point at which a switch occurred is not known.

The underlying model used is a *switching linear dynamical system (SLDS)*. The constraints in the regime transition probabilities result in an exact inference procedure that scales quadratically with the length of a sequence. Maximum aposteriori (MAP) parameter estimates can be found using an expectation maximization (EM) algorithm with this inference algorithm in the E-step. For long sequences this will not be practically feasible and an approximate inference and an approximate EM procedure is called for. We describe a flexible class of approximations corresponding to different choices of clusters in a Kikuchi free energy with weak consistency constraints.

Keywords: change point problems, switching linear dynamical systems, strong junction trees, approximate inference, expectation propagation, Kikuchi free energies

1. Introduction

In this article we investigate the problem of detecting a change in a dynamical system. An obvious practical application of such a model is the prediction of oncoming faults in an industrial process.

For simplicity the problem and algorithms are outlined for a model with four regimes, *normal*, *prefault*, *stop*, and *fault*, in Section 2 the extension to more regimes is discussed. The stop and fault regimes are special in the sense that they are *absorbing*. If the system reaches one of these states the process stops. A key assumption in the problem is that once the system reaches a prefault state, it can never recover.

The setup could be considered as a *change point problem*, although the name change point problem usually refers to a problem where the observations are independent if the underlying model parameters are known. In such settings the challenge is to determine if and where the parameters change their value. See Krishnaiah and Miao (1988) for a description of change point problems and references.

In this article we will be interested in slightly more complex problems where the observations are dependent, even if the parameters are known. The observations in the time-series are not assumed to be Markov. Instead, they are noisy observations of a latent first order Markov process.

The model discussed in this paper can be identified as a *switching linear dynamical system* (SLDS), with restricted dynamics in the regime indicators. The SLDS is a discrete time model and consists of T, d dimensional observations $\mathbf{y}_{1:T}$ and T, q dimensional latent states $\mathbf{x}_{1:T}$. The regime in every time-step is determined by (typically unobserved) discrete switches $s_{1:T}$. For $1 < t \leq T$, s_t is either normal or prefault. The last discrete indicator s_{T+1} is either a stop or a fault.

Within every regime the state transition and the observation model are linear Gaussian, and may differ per regime:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t, \theta) &= \mathcal{N}\left(\mathbf{x}_t; A_{s_t}\mathbf{x}_{t-1}, Q_{s_t}\right) , \\ p(\mathbf{y}_t|\mathbf{x}_t, s_t, \theta) &= \mathcal{N}\left(\mathbf{y}_t; C_{s_t}\mathbf{x}_t + \mu_{s_t}, R_{s_t}\right) . \end{aligned}$$

In the above $\mathcal{N}(.;.,.)$ denotes the Gaussian density function

$$\mathcal{N}(\mathbf{x}; \mathbf{m}, V) \equiv (2\pi)^{-(d/2)}|V|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top V^{-1}(\mathbf{x}-\mathbf{m})\right] .$$

The determinant of matrix V is denoted as $|V|$. The set of parameters in the model is denoted by θ. As mentioned the current regime is encoded by discrete random variables $s_{1:T}$ and are assumed to follow a first order transition model

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

The special characteristics of the regimes and their transitions are reflected by zeros in $\Pi_{s_{t-1}\to s_t}$, i.e. denoting the possible states (normal, fault, etc.) by their initial letter, we require $\Pi_{\mathrm{n}\to\mathrm{f}} = 0$, $\Pi_{\mathrm{p}\to\mathrm{n}} = 0$, $\Pi_{\mathrm{p}\to\mathrm{s}} = 0$, $\Pi_{\mathrm{s}\to j} = 0$, for all $j \neq \mathrm{s}$ and $\Pi_{\mathrm{f}\to j} = 0$, for all $j \neq \mathrm{f}$.

The first regime is always normal, i.e. $s_1 = \mathrm{n}$, and the first latent state is drawn from a Gaussian prior

$$p(\mathbf{x}_1|s_1 = \mathrm{n}, \theta) = \mathcal{N}(\mathbf{x}_1; \mathbf{m}_1, V_1) .$$

With these choices the entire model is *conditional Gaussian*; conditioned on the discrete variables $s_{1:T}$, the remaining variables are jointly Gaussian distributed. The conditional independencies implied by the model are depicted as a dynamical Bayesian network (Pearl, 1988) in Figure 1.

One of the properties of the conditional Gaussian distribution which leads often to computational problems is that it is not closed under marginalization. For instance, the state posterior over $\mathbf{x}_t$ given all observations is

$$\sum_{s_{1:T}} \int p(s_{1:T}, \mathbf{x}_{1:T}|\mathbf{y}_{1:T}, \theta)\, d\mathbf{x}_{1:t-1,t+1:T} = \sum_{s_{1:T}} p(\mathbf{x}_t|s_{1:T}, \mathbf{y}_{1:T}, \theta) p(s_{1:T}|\mathbf{y}_{1:T}, \theta) ,$$

which is not a conditional Gaussian, but a mixture of Gaussians with M^T components, with M the number of possible regimes in the system. However, as we will discuss in the next section

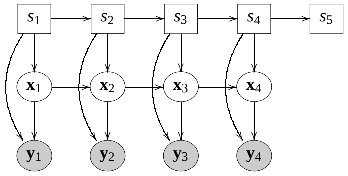

Figure 1: The dynamic Bayesian network for a switching linear dynamical system with four observations. Square nodes denote discrete, and ovals denote continuous random variables. Shading emphasizes that a particular variable is observed.

the assumption that a system cannot restore from a prefault state to a normal state results in a considerable simplification, since many of these components have zero weight.

In Section 3 we review how this sparsity can be exploited in an inference algorithm. As the basis for an EM algorithm it can then straightforwardly be used to compute MAP estimates for the model parameters. The exact inference algorithm has running time $\mathcal{O}(T^2)$. Hence for relatively short sequences the constrained transition model makes exact inference feasible. However for larger sequences the exact inference algorithm will be inappropriate. In Section 5 we introduce a flexible class of approximations which can be interpreted as a generalization of *expectation propagation* (Minka, 2001). It has running time $\mathcal{O}(T\kappa)$, with $0 \leq \kappa \leq \lceil \frac{T-2}{2} \rceil$, an integer parameter that can be set according to the available computational resources. With $\kappa = 0$ the approximation is equivalent to an iterated version (Heskes and Zoeter, 2002; Zoeter and Heskes, 2005) of *generalized pseudo Bayes 2* (Bar-Shalom and Li, 1993) with $\kappa = \lceil \frac{T-2}{2} \rceil$ exact inference is recovered. Section 6 discusses experiments with inference and MAP parameter estimation on synthetic data and a change point problem in EMG analysis.

2. Benefits of the Constrained Regime Transition Probabilities

An interesting aspect of the model introduced in Section 1 is that, by the restriction in the regime transitions, the number of possible regimes histories is considerably less than the 2^T possible histories which would be implied by a system with unconstrained transitions (see e.g. Cemgil et al., 2004; Fearnhead, 2003). If the absorbing state s_{T+1} is not observed, there are T possible regime histories in the current model. One normal sequence $s_{1:T} = \mathrm{n}$, and $T-1$ fault sequences: $s_{1:\tau} = \mathrm{n}, s_{\tau+1:T} = \mathrm{p}$, with $1 \leq \tau \leq T-1$. In the remainder of this paper we let τ denote the time-slice up to and including which the regime has been normal, i.e. with $\tau = T$ the entire sequence was normal. Under our assumptions, a fault has to be preceded by at least one prefault state, so if s_{T+1} is observed to be a fault the entirely normal sequence gets zero weight. So with s_{T+1} observed to be a fault the number of the possible histories is $T-1$. If s_{T+1} is observed to be a stop, then only the normal sequence has nonzero probability.

If the parameters in the model, θ, are known, the exact posterior over a continuous state

$$p(\mathbf{x}_t | \mathbf{y}_{1:T}, s_{T+1} = \mathrm{f}, \theta)$$

is a mixture of Gaussians with $T-1$ components, one for every regime history, and can be obtained by running the traditional Kalman filter and smoother $T-1$ times. In fact the posteriors can be computed in a slightly faster way by computing shared partial results only once. This algorithm is introduced in Section 3, and will form a suitable basis for the approximate algorithm from Section 5. Although we do not expect exact inference to be practical for large T, we can compare approximations with exact results for larger examples than in the regular SLDS case.

The restriction that there are only two non-absorbing regimes is only made for clarity of the exposition. In general the model has M non-absorbing regimes that form stages. No stage can be skipped, and once the system has advanced to the new stage it cannot recover to a previous one. The number of regime histories with non-zero probability in such a system is less than or equal to T^{M-1}. This can be seen by a simple inductive argument: if $M = 1$ there is only one possible history. If $M > 1$ there are $T-(M-1)+1 < T$ possible starting points for the M-th regime (including the starting point $T+1$, i.e. when regime M does not occur). The $M-1$ steps are deducted since the system needs at least $M-1$ steps to reach the M-th regime. Once the start of the M-th regime is fixed, we have a smaller problem with $M-1$ regimes of length at most T. So the number of distinct regime histories is bounded by $T \times T^{M-2}$. In principle this is still polynomial in T and for small M and limited T exact posteriors could be computed, but obviously the need for approximations is stronger with complexer models.

3. Inference

In this section we will introduce the exact recursive inference algorithm as a special case of the *sum-product algorithm* (Kschischang et al., 2001). At this point we assume θ known, leaving the MAP estimation problem to Section 4.

We are interested in one-slice and two-slice posteriors, $p(s_t, \mathbf{x}_t|\mathbf{y}_{1:T}, \theta)$ and $p(s_{t-1,t}, \mathbf{x}_{t-1,t}|\mathbf{y}_{1:T}, \theta)$ respectively.

By defining $\mathbf{u}_t \equiv \{s_t, \mathbf{x}_t\}$ we obtain a model that has the same conditional independence structure as the linear dynamical system and the HMM. From time to time we will use a sum notation to denote both the summation over the domain of the discrete variables, and the integration over the domain of the continuous variables in $\mathbf{u}_t$. The computational complexity in the current case is due to the parametric form of the (conditional) distributions over $\mathbf{u}_t$ as discussed in Section 1.

Assuming θ and $\mathbf{y}_{1:T}$ fixed and given, the joint probability distribution over all the variables in the model can be written as a product of *factors*

$$p(s_{1:T+1}, \mathbf{x}_{1:T}, \mathbf{y}_{1:T}|\theta) = \prod_{t=1}^{T+1} \psi_t(\mathbf{u}_{t-1,t}) \ ,$$

with

$$\begin{aligned}
\psi_1(\mathbf{u}_1) &\equiv p(s_1|\theta)p(\mathbf{x}_1|s_1,\theta)p(\mathbf{y}_1|\mathbf{x}_1,s_1,\theta) \ , \\
\psi_t(\mathbf{u}_{t-1,t}) &\equiv p(s_t|s_{t-1},\theta)p(\mathbf{x}_t|\mathbf{x}_{t-1},s_t,\theta)p(\mathbf{y}_t|\mathbf{x}_t,s_t,\theta) \quad \text{for } t=2,\ldots,T, \qquad (1) \\
\psi_{T+1}(s_{T,T+1}) &\equiv p(s_{T+1}|s_T,\theta) \ ,
\end{aligned}$$

and $\mathbf{u}_0 \equiv \emptyset$ and $\mathbf{u}_{T+1} \equiv s_{T+1}$. The *factor graph* (Kschischang et al., 2001) implied by this choice of factors is shown in Figure 2. Note that we have simplified the figure by not showing the observations $\mathbf{y}_{1:T}$. These are always observed and are incorporated in the factors.

The sum-product algorithm implied by the factor graph from Figure 2 is presented in Algorithm 1. It is analogous to the forward-backward algorithm in the HMM. The computational com-

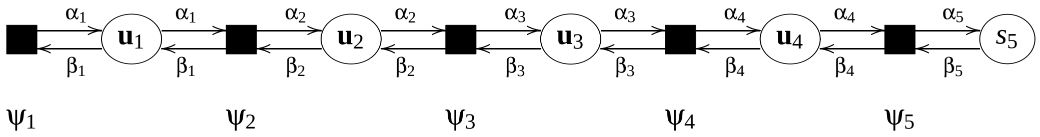

Figure 2: The factor graph corresponding to the change point model and message passing scheme for a model with four observations.

plexity of this algorithm is due to the conditional Gaussian factors and the implied increase in the complexity of the messages.

Algorithm 1 The sum-product algorithm for the SLDS

Forward pass Start the recursion with

$$p(\mathbf{u}_1|\mathbf{y}_1,\theta) \equiv \alpha_1(\mathbf{u}_1) = \frac{\psi_1(\mathbf{u}_1)}{Z_1}, \quad Z_1 = \sum_{\mathbf{u}_1} \psi_1(\mathbf{u}_1) .$$

For $t = 1, \dots T$

$$p(\mathbf{u}_t|\mathbf{y}_{1:t},\theta) \equiv \alpha_t(\mathbf{u}_t) = \frac{\sum_{\mathbf{u}_{t-1}} \alpha_{t-1}(\mathbf{u}_{t-1})\psi_t(\mathbf{u}_{t-1,t})}{Z_{t|t-1}} ,$$

with $p(\mathbf{y}_t|\mathbf{y}_{1:t-1},\theta) \equiv Z_{t|t-1} = \sum_{\mathbf{u}_{t-1,t}} \alpha_{t-1}(\mathbf{u}_{t-1})\psi_t(\mathbf{u}_{t-1,t})$.

Backward pass If s_{T+1} is not observed, start the recursion with

$$\beta_T(\mathbf{u}_T) = 1 .$$

If s_{T+1} is observed the definition of $\beta_T(\mathbf{u}_T)$ is changed accordingly: if $s_{T+1} = \mathrm{n}$, then $\beta_T(s_T = \mathrm{p}) = 0$. Similarly if $s_{T+1} = \mathrm{p}$ then $\beta_T(s_T = \mathrm{n}) = 0$.

For $t = T-1, T-2, \dots 1$

$$\frac{p(\mathbf{y}_{t+1:T}|\mathbf{u}_t,\theta)}{p(\mathbf{y}_{t+1:T}|\mathbf{y}_{1:t},\theta)} \equiv \beta_t(\mathbf{u}_t) = \frac{\sum_{u_{t+1}} \psi_{t+1}(\mathbf{u}_{t,t+1})\beta_{t+1}(\mathbf{u}_{t+1})}{\prod_{v=t+1}^{T} Z_{v|v-1}} .$$

After a forward-backward pass, single-slice and two-slice posteriors are given by

$$\begin{aligned} p(\mathbf{u}_t|\mathbf{y}_{1:T},\theta) &= \alpha_t(\mathbf{u}_t)\beta_t(\mathbf{u}_t) \\ p(\mathbf{u}_{t-1,t}|\mathbf{y}_{1:T},\theta) &= \frac{1}{Z_{t|t-1}}\alpha_{t-1}(\mathbf{u}_{t-1})\psi_t(\mathbf{u}_{t-1,t})\beta_t(\mathbf{u}_t) . \end{aligned}$$

In the forward pass the message $\alpha_t(\mathbf{u}_t) \equiv p(\mathbf{x}_t, s_t|\mathbf{y}_{1:t},\theta)$ is *not* conditional Gaussian, but a mixture of Gaussians conditioned on the regime indicator s_t. It has t components in total: conditioned on $s_t = \mathrm{n}$ the posterior contributes a single Gaussian component $p(\mathbf{x}_t|s_t = \mathrm{n}, \mathbf{y}_{1:t},\theta)$ conditioned on $s_t = \mathrm{p}$ the posterior $p(\mathbf{x}_t|s_t = \mathrm{p}, \mathbf{y}_{1:t}, \theta)$ is a mixture of Gaussians with $t-1$ components: each component corresponds to a possible starting point of the prefault regime.

In the smoothing pass an analogous growth of the number of components in the backward messages $\beta_t(\mathbf{u}_t)$ occurs, but now growing backwards in time. The single-slice posterior, which is obtained from the product of the forward and backward messages, has T components for all t.

Note that the linear complexity in T is special for the change point model with the restricted regime transitions. In general the number of components in the posterior would grow exponentially.

4. MAP Parameter Estimation

In Section 3 we have assumed that the model parameters θ were known. If they are not known, the expectation-maximization (EM) algorithm (Dempster et al., 1977) with Algorithm 1 in the expectation step, can be used to find maximum likelihood (ML) or maximum a posteriori (MAP) parameter settings. Appendix B lists the M-step updates for the change point model. Appendix C discusses sensible priors on the transition probabilities in Π.

The learning setting is *semi supervised*. We assume we are given a set of V training sequences $\left\{\mathbf{y}^{(v)}_{1:T_v}\right\}_{v=1}^{V}$ and that for some, possibly all, we observe $s^{(v)}_{T+1}$. All sequences v for which $s^{(v)}_{T+1} = \mathrm{s}$ can be used to estimate the parameters of the normal regime. If $s^{(v)}_{T+1} = \mathrm{f}$ or not observed, the change point from the normal to the prefault regime is inferred in the E-step. The updates from Appendix B then boil down to weighted variants of the linear dynamical system M-step updates, where the weights correspond to the posterior probabilities of being in a particular regime.

The EM algorithm is guaranteed to converge to a local maximum of the likelihood/parame-ter posterior. Different initial parameter estimates $\theta^{(0)}$ may lead the algorithm to converge to different local maxima. This is a known property of the EM algorithm for fitting a mixture of Gaussians. In the current model it can be hoped that the dependence on initialization is less than in the general mixtures of Gaussian case. If there are sequences that are known to be entirely normal (when $s_{T+1} = \mathrm{s}$) these sequences are only used to determine the characteristics of the normal regime. Also, due to the change point restriction, some ambiguity is resolved since it is known that the normal precedes the prefault regime.

5. Approximate Inference: Kikuchi Free Energies with Weak Consistency Constraints

The exact inference algorithm presented in Section 3 has the same form as the HMM and Kalman filter algorithms. The messages that are sent, $\alpha_t(\mathbf{u}_t)$ and $\beta_t(\mathbf{u}_t)$, are not in the conditional Gaussian family, but are conditional mixtures. As was discussed in Section 3, the number of components in the mixtures grows linear with t and $T-t$ respectively.

A straightforward approximation is to approximate these messages by a conditional Gaussian in every step. This implies that every message stores only two components, regardless of t. In the forward pass the best approximating conditional Gaussian can be defined in Kullback-Leibler (KL) sense. The approximating conditional Gaussian is then found by *moment matching* or a *collapse* (see Appendix A). The oldest use of this approach we are aware of is in Harrison and Stevens (1976).

A symmetric approximation for the backward pass working directly on the $\beta_t(\mathbf{u}_t) = \frac{p(\mathbf{y}_{t+1:T}|\mathbf{u}_t,\theta)}{p(\mathbf{y}_{t+1:T}|\mathbf{y}_{1:t},\theta)}$ messages cannot be formulated, since in contrast to the α messages, the β messages in general will not be proper distributions and hence a KL divergence is not defined.

This has led to other approaches that introduced additional approximations beyond the projection onto the conditional Gaussian family, (e.g. Shumway and Stoffer, 1991; Kim, 1994). The expectation propagation (EP) framework of Minka (2001) is very suited for this particular model and essentially formulates a backward pass symmetric to the approach outlined above (Zoeter and Heskes, 2005). There are at least two ways of looking at EP. In the first, EP is seen as an iteration scheme where at every step an exact model potential is added to the approximation followed by a projection onto a chosen approximating family. In the second, EP is derived from a particular variational problem. The EP algorithm is introduced in Section 5.2 using the second point of view, which facilitates the description of our generalization in Section 5.3. For a presentation of EP as an iteration of projections the reader is referred to Minka (2001).

The approximate filter and the EP algorithm share that they are greedy: the approximations are made locally. In the EP algorithm the local approximations are made as consistent as possible by iteration. There is no guarantee that the resulting means and covariances in the conditional Gaussian families equal the means and covariances of the exact posteriors. The strong junction tree framework of Lauritzen (1992) operates on trees with larger cliques and approximates messages on a global level. Thereby it *does* guarantee exactness of means and covariances. For the SLDS a strong junction tree has at least one cluster that effectively contains all discrete variables.

Section 5.3 introduces a generalization of the EP algorithm from Section 5.2. In the generalization, an extra integer parameter κ is introduced that allows a trade-off between computation time and accuracy. The EP algorithm from Zoeter and Heskes (2005) and the strong junction tree from Lauritzen (1992) are then on both extremes.

5.1 Exact Inference as an Energy Minimization Procedure

To facilitate the introduction of the expectation and the generalized expectation propagation algorithms, exact inference is introduced in this Section as a minimization procedure. Expectation propagation will follow from an approximation of the objective.

We start by following the variational approach (e.g. Jaakkola, 2001) and turn the computation of $-\log Z \equiv -\log p(\mathbf{y}_{1:T}|\theta)$ into an optimization problem:

$$\begin{aligned}
-\log Z &= \min_{\tilde{p}} \left[-\log Z + \mathrm{KL}\left(\tilde{p}(\mathbf{u}_{1:T})||p(\mathbf{u}_{1:T}|\mathbf{y}_{1:T},\theta)\right)\right] && (2)\\
&= \min_{\tilde{p}} \left[-\log Z + \sum_{\mathbf{u}_{1:T}} \tilde{p}(\mathbf{u}_{1:T}) \log \frac{\tilde{p}(\mathbf{u}_{1:T})}{Z^{-1}\prod_{t=1}^{T}\psi_t(\mathbf{u}_{t-1,t})}\right] && (3)\\
&= \min_{\tilde{p}} \left[-\sum_{t=1}^{T}\sum_{\mathbf{u}_{t-1,t}} \tilde{p}(\mathbf{u}_{t-1,t}) \log \psi_t(\mathbf{u}_{t-1,t}) + \sum_{\mathbf{u}_{1:T}} \tilde{p}(\mathbf{u}_{1:T}) \log \tilde{p}(\mathbf{u}_{1:T})\right]. && (4)
\end{aligned}$$

In (2)–(4) the minimization is over all valid distributions $\tilde{p}(\mathbf{u}_{1:T})$ on the domain $\mathbf{u}_{1:T}$. The KL term in (2) is guaranteed to be positive and equals zero if and only if $\tilde{p}(\mathbf{u}_{1:T}) = p(\mathbf{u}_{1:T}|\mathbf{y}_{1:T},\theta)$ (Gibbs inequality). This guarantees the equality in (2).

In terms of $\mathbf{u}_t$ the exact posterior factors as

$$p(\mathbf{u}_{1:T}|\mathbf{y}_{1:T},\theta) = \frac{\prod_{t=2}^{T} p(\mathbf{u}_{t-1,t}|\mathbf{y}_{1:T},\theta)}{\prod_{t=2}^{T-1} p(\mathbf{u}_t|\mathbf{y}_{1:T},\theta)} \ , \tag{5}$$

so we can restrict the minimization in (4) to be over all valid distributions of the form (5):

$$\begin{aligned} -\log Z \quad = \quad & \min_{\{\tilde{p}_t,\tilde{q}_t\}} \Bigg[-\sum_{t=2}^{T} \sum_{\mathbf{u}_{t-1,t}} \tilde{p}_t(\mathbf{u}_{t-1,t}) \log \psi_t(\mathbf{u}_{t-1,t}) \\ & + \sum_{t=2}^{T} \sum_{\mathbf{u}_{t-1,t}} \tilde{p}_t(\mathbf{u}_{t-1,t}) \log \tilde{p}_t(\mathbf{u}_{t-1,t}) - \sum_{t=2}^{T-1} \sum_{\mathbf{u}_t} \tilde{q}_t(\mathbf{u}_t) \log \tilde{q}_t(\mathbf{u}_t) \Bigg] \ . \end{aligned} \tag{6}$$

The minimization is now with respect to one-slice beliefs $\tilde{q}_t(\mathbf{u}_t)$ and two-slice beliefs $\tilde{p}_t(\mathbf{u}_{t-1,t})$ under the constraints that these beliefs are properly normalized and consistent:

$$\tilde{p}_t(\mathbf{u}_t) = \tilde{q}_t(\mathbf{u}_t) = \tilde{p}_{t+1}(\mathbf{u}_t) \ . \tag{7}$$

To emphasize that the above constraints are exact, and to distinguish them from the *weak consistency constraints* that will be introduced below, we will refer to (7) as *strong consistency constraints*.

Minimizing the objective in (6) under normalization and strong consistency constraints (7) gives exact one- and two-slice posteriors. Since they are exact, the one-slice beliefs $\tilde{q}_t(\mathbf{u}_t)$ will have T components in our change point model and M^T components in a general SLDS.

5.2 Expectation Propagation

As we have seen in the previous section, exact inference inference can be interpreted as a minimization procedure under constraints. At the minimum, the variational parameters $\tilde{q}_t(\mathbf{u}_t)$ are equal to the exact single node marginals. Since these marginals have many components (T in our changepoint model, M^T in a general SLDS) even storing the results is computationally demanding.

To obtain an approximation the variational parameters $\tilde{q}_t(\mathbf{u}_t)$ are restricted to be conditional Gaussian. Recall that $\mathbf{u}_t \equiv \{s_t, \mathbf{x}_t\}$, so that the conditional Gaussian restriction implies that for every possible value for s_t, $\mathbf{x}_t$ follows a Gaussian distribution, instead of a mixture of Gaussians with a mixture component for every possible regime history for $s_{1:t-1,t+1:T}$. This restriction is analogous to the approximation in the generalized pseudo Bayes 2 (GPB 2) filter (Bar-Shalom and Li, 1993) where in every time update step mixtures of Gaussians are collapsed onto single Gaussians. In fact, as we will see shortly, GPB2 can be seen as a first forward pass in the algorithm that follows from our current approach.

The conditional Gaussian form of $\Psi_t(\mathbf{u}_{t-1,t})$ and the conditional Gaussian choice for $\tilde{q}_t(\mathbf{u}_t)$ imply that at the minimum in (6) $\tilde{p}_t(\mathbf{u}_{t-1,t})$ is conditionally Gaussian as well (see Appendix D).

If we restrict the form of $\tilde{q}_t(\mathbf{u}_t)$, but leave the consistency constraints exact as in (7), a minimum of the free energy has a very restricted form. The strong consistency constraints would imply that the two exact marginals $\sum_{\mathbf{u}_{t-1}} \tilde{p}_t(\mathbf{u}_{t-1,t}) = \tilde{p}_t(\mathbf{u}_t)$ and $\sum_{\mathbf{u}_t} \tilde{p}_t(\mathbf{u}_{t-1,t}) = \tilde{p}_{t-1}(\mathbf{u}_{t-1})$ are conditional Gaussians instead of conditional mixtures. This holds only if the continuous variables $x_{t-1,t}$ are independent of the discrete states $s_{t-1,t}$ in $\tilde{p}_t(\mathbf{u}_{t-1,t})$.

To obtain non-trivial approximations, the single-slice beliefs $\tilde{q}_t(\mathbf{u}_t)$ are restricted to be conditional Gaussian as outlined above, and in addition the consistency constraints are weakened. Instead of having equal marginals we only require overlapping beliefs to be consistent on their overlapping expectations

$$\langle f(\mathbf{u}_t)\rangle_{\tilde{p}_t} = \langle f(\mathbf{u}_t)\rangle_{\tilde{q}_t} = \langle f(\mathbf{u}_t)\rangle_{\tilde{p}_{t+1}} \ , \tag{8}$$

where $f(\mathbf{u}_t)$ is the vector of sufficient statistics of the conditional Gaussian family over $\mathbf{u}_t$ as defined in Appendix A.

With these restrictions $\tilde{q}_t(\mathbf{u}_t)$ is in general not the marginal of $\tilde{p}_t(\mathbf{u}_{t-1,t})$, so one-slice and two-slice beliefs satisfying (8) do not lead to a proper distribution of the form (5). As a result, although we started the derivation with the variational (mean-field) bound (2), the objective we aim to minimize is not guaranteed to be a bound on $-\log Z$.

The EP algorithm can be seen as fixed point iteration in the *dual space* of the constrained minimization problem (Zoeter and Heskes, 2005, Appendix D). This is in direct analogy to the interpretation of loopy belief propagation as fixed point iteration in the dual space of the Bethe free energy (Yedidia et al., 2005).

Algorithm 2 presents the generalization that will be derived next, but with $\kappa = 0$ it gives the basic update equations of this section. In a first forward pass, with all backward messages initialized as $\beta_t(\mathbf{u}_t) = 1$ (i.e. effectively with no backward messages), the updates are equivalent to the greedy projection filter GPB2.

As a final note we remark that this approximation, and even the update scheme, can also be derived from the iterative projection point of view of EP. To obtain Algorithm 2 with $\kappa = 0$, the approximating family should be chosen to be a product of independent conditional Gaussians (Zoeter and Heskes, 2005).

5.3 Generalized Expectation Propagation

Since we have associated the EP approach to an approximation of the Bethe free energy (6), we can extend the approximation analogously to Kikuchi's extension of the Bethe free energy (Yedidia et al., 2005).

In the EP free energy (6) the minimization is w.r.t. beliefs over *outer clusters*, $\tilde{p}_t(\mathbf{u}_{t-1,t})$, and their *overlaps*, $\tilde{q}_t(\mathbf{u}_{t-1,t})$. In the so-called *negative entropy*,

$$\sum_{t=2}^{T}\sum_{\mathbf{u}_{t-1,t}} \tilde{p}_t(\mathbf{u}_{t-1,t}) \log \tilde{p}_t(\mathbf{u}_{t-1,t}) - \sum_{t=2}^{T-1}\sum_{\mathbf{u}_t} \tilde{q}_t(\mathbf{u}_t) \log \tilde{q}_t(\mathbf{u}_t) \ ,$$

from (6), the outer clusters enter with a plus, the overlaps with a minus sign. These 1 and -1 factors can be interpreted as *counting numbers* that ensure that every variable effectively is counted once in the (approximate) entropy in (6). If the free energy is exact (i.e. no parametric choice for the beliefs, and strong consistency constraints), the local beliefs are exact marginals, and as in (5), the counting numbers can be interpreted as powers that dictate how to construct a global distribution from the marginals.

In Kikuchi's extension the outer clusters are taken larger. The minimization is then w.r.t. beliefs over outer clusters, their direct overlaps, the overlaps of the overlaps, etc. With each belief again proper counting numbers are associated.

One way to construct a valid Kikuchi based approximation is as follows (Yedidia et al., 2005). Choose outer clusters $\mathbf{u}_{outer(i)}$ and associate with them the counting number $c_{outer(i)} = 1$. The outer clusters should be such that all domains $\mathbf{u}_{t-1,t}$ of the model potentials $\Psi_t(\mathbf{u}_{t-1,t})$ are fully contained

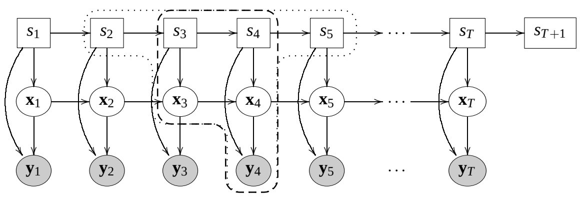

Figure 3: Cluster definitions for $\kappa = 0$ (dashed) and $\kappa = 1$ (dotted).

in at least one outer cluster. Then recursively define the overlaps of the outer clusters $\mathbf{u}_{over(i)}$, the overlaps of the overlaps, etc. The counting number associated with cluster γ is given by the Möbius recursion

$$c_\gamma = 1 - \sum_{\mathbf{u}_{\gamma'} \supset \mathbf{u}_\gamma} c_{\gamma'} \ . \tag{9}$$

A crucial observation for the SLDS is that it makes sense to take outer clusters larger than the cliques of a (weak) junction tree. If we do not restrict the parametric form of $\tilde{q}_t(\mathbf{u}_t)$ and keep exact constraints, the cluster choice in (5) gives exact results. However, the restriction that $\tilde{q}_t(\mathbf{u}_t)$ must be conditional Gaussian, and the weak consistency constraints imply an approximation: only part of the information from the past can be passed on to the future and vice versa. With weak constraints it is beneficial to take larger outer clusters and larger overlaps, since the weak consistency constraints are then over a larger set of sufficient statistics and hence "stronger".

We define symmetric extensions of the outer clusters as depicted in Figure 3. The size of the clusters is indicated by $0 \leq \kappa \leq \lceil \frac{T-2}{2} \rceil$:

$$\begin{aligned} \mathbf{u}_{outer(i)} &= \left\{ s_{i:i+2(\kappa+1)-1}, x_{i+\kappa,i+\kappa+1} \right\}, \quad \text{for } i > 1 \wedge i < T - 2\tau + 2 & (10) \\ \mathbf{u}_{over(i)} &= \mathbf{u}_{outer(i)} \cap \mathbf{u}_{outer(i+1)} \ . & (11) \end{aligned}$$

In the outer clusters only the discrete space is extended because the continuous part can be integrated out analytically and the result stays in the conditional Gaussian family. The first and the last outer cluster have a slightly larger set. In addition to the set (10) the first cluster also contains $\mathbf{x}_{1:i+\kappa-1}$ and the last also $\mathbf{x}_{i+\kappa+2:T}$. This implies a choice where the number of outer clusters is as small as possible at the cost of a larger continuous part in the first and the last cluster. A slightly different choice would have more clusters, but only two continuous variables in every outer cluster.

To demonstrate the construction of clusters and the computation of their associated counting numbers we will look at the case of $\kappa = 1$. Below the clusters are shown schematically, with outer clusters on the top row, and recursively the overlaps of overlaps, etc.

$$\begin{array}{ccccccc} s_{1,2,3,4} & & s_{2,3,4,5} & & s_{3,4,5,6} & & s_{4,5,6,7} \\ x_{1,2,3} & & x_{3,4} & & x_{4,5} & & x_{5,6,7} \\ & s_{2,3,4} & & s_{3,4,5} & & s_{4,5,6} & \\ & x_3 & & x_4 & & x_5 & \\ & & s_{3,4} & & s_{4,5} & & \\ & & & s_4 & & & \end{array}$$

The outer clusters all have counting number 1. The direct overlaps each have two larger clusters in which they are contained. Their associated counting numbers follow from (9) as $1 - 2 = -1$.

The overlaps of overlaps have five clusters in which they are contained, their counting numbers are $1-(3-2)=0$. The clusters on the lowest level have nine parents, which results in a counting number $1-(4-3+0)=0$. It is easily verified that with $\kappa=0$ we obtain the cluster and counting number choice of Section 5.2.

A second crucial observation for the SLDS is that the choice of outer clusters (10) implies that we only have to consider outer clusters and direct overlaps, i.e. the phenomenon that all clusters beyond the direct overlaps get an associated counting number of 0 in the example above extends to all κ. This is a direct result of the fact that the clusters from (10) form the cliques and separators in a (weak) junction tree. I.e. another way to motivate a generalization with the cluster choice (10) is to replace (5) with

$$p(\mathbf{u}_{1:T}|\mathbf{y}_{1:T},\theta)=\frac{\prod_{i=1}^{N}p(\mathbf{u}_{outer(i)}|\mathbf{y}_{1:T},\theta)}{\prod_{j=1}^{N-1}p(\mathbf{u}_{over(j)}|\mathbf{y}_{1:T},\theta)}\,, \tag{12}$$

and use this choice in (4) to obtain an extension of (6). In (12), $N=T-2\kappa-1$ denotes the number of outer clusters in the approximation.

The aim then becomes to minimize

$$\begin{aligned}\mathcal{F}_{\text{GEP}} \quad = \quad & -\sum_{i=1}^{N}\sum_{\mathbf{u}_{outer(i)}}\tilde{p}_i(\mathbf{u}_{outer(i)})\log\Psi^{(i)}(\mathbf{u}_{outer(i)})\\ & +\sum_{i=1}^{N}\sum_{\mathbf{u}_{outer(i)}}\tilde{p}_i(\mathbf{u}_{outer(i)})\log\tilde{p}_i(\mathbf{u}_{outer(i)})\\ & -\sum_{i=1}^{N-1}\sum_{\mathbf{u}_{over(i)}}\tilde{q}_i(\mathbf{u}_{over(i)})\log\tilde{q}_i(\mathbf{u}_{over(i)})\,,\end{aligned} \tag{13}$$

w.r.t. the potentials $\tilde{p}_i(\mathbf{u}_{outer(i)})$, and $\tilde{q}_i(\mathbf{u}_{over(i)})$. For $i=2,3,\ldots N-1$, the potentials $\Psi^{(i)}(\mathbf{u}_{over(i)})$ are identical to the potentials $\psi_{i+\kappa+1}(\mathbf{u}_{i+\kappa,i+\kappa+1})$ from (1). At the boundaries they are a product of potentials that are "left over":

$$\begin{aligned}\Psi^{(1)} \quad &= \quad \prod_{j=1}^{\kappa+2}\psi_j(\mathbf{u}_{j-1,j})\\ \Psi^{(N)} \quad &= \quad \prod_{j=T-\kappa}^{T}\psi_j(\mathbf{u}_{j-1,j})\,,\end{aligned}$$

with $\Psi^{(1)}=\prod_{j=1}^{T}\psi_j(\mathbf{u}_{j-1,j})$ if $N=1$.

The approximation in the *generalized EP free energy*, $\mathcal{F}_{GEP}$, arises from the restriction that $\tilde{q}_i(\mathbf{u}_{over(i)})$ is conditional Gaussian and from the fact that overlapping potentials are only required to be weakly consistent

$$\left\langle f(\mathbf{u}_{over(i)})\right\rangle_{\tilde{p}_i}=\left\langle f(\mathbf{u}_{over(i)})\right\rangle_{\tilde{q}_i}=\left\langle f(\mathbf{u}_{over(i)})\right\rangle_{\tilde{p}_{i+1}}\,.$$

The benefit of the (weak) junction tree choice of outer clusters and overlaps is that we can employ the same algorithm for the $\kappa=0$ as for the $\kappa>0$ case. Algorithm 2 can be seen as a single-loop minimization heuristic. As mentioned above, and as shown in Appendix D, the algorithm can be interpreted as fixed point iteration in the space of Lagrange multipliers that are added to (13) to enforce the weak consistency constraints. Just as for EP itself, convergence of Algorithm 2 is not guaranteed.

In Algorithm 2 the messages are initialized as conditional Gaussian potentials, such that

$$\tilde{q}(\mathbf{u}_{over(i)}) = \alpha_i(\mathbf{u}_{over(i)})\beta_i(\mathbf{u}_{over(i)})$$

are normalized. A straightforward initialization would be to initialize all messages with 1. If at the start all products of matching messages are normalized, we can interpret the product of local normalizations $\prod_{i=1}^{N} Z_i$ as an approximation of the normalization constant Z.

Algorithm 2 Generalized EP for an SLDS

Compute a forward pass by performing the following steps for $i = 1, 2, \ldots, N-1$, with $i' \equiv i$, and a backward pass by performing the same steps for $i = N, N-1, \ldots, 2$, with $i' \equiv i-1$. Iterate forward-backward passes until convergence. At the boundaries keep $\alpha_0 = \beta_N = 1$.

1. Construct an outer-cluster belief,

$$\tilde{p}_i(\mathbf{u}_{outer(i)}) = \frac{\alpha_{i-1}(\mathbf{u}_{over(i-1)})\Psi^{(i)}(\mathbf{u}_{outer(i)})\beta_i(\mathbf{u}_{over(i)})}{Z_i} \ ,$$

   with $Z_i = \sum_{\mathbf{u}_{outer(i)}} \alpha_{i-1}(\mathbf{u}_{over(i-1)})\Psi^{(i)}(\mathbf{u}_{outer(i)})\beta_i(\mathbf{u}_{over(i)})$.

2. Marginalize to obtain a one-slice marginal

$$\tilde{p}_i(\mathbf{u}_{over(i')}) = \sum_{\mathbf{u}_{outer(i)} \backslash \mathbf{u}_{over(i')}} \tilde{p}_i(\mathbf{u}_{outer(i)}) \ .$$

3. Find $\tilde{q}_{i'}(\mathbf{u}_{over(i')})$ that approximates $\tilde{p}_i(\mathbf{u}_{over(i')})$ best in Kullback-Leibler (KL) sense:

$$\tilde{q}_{i'}(\mathbf{u}_{over(i')}) = \text{Collapse}\left(\tilde{p}_i(\mathbf{u}_{over(i')})\right) \ .$$

4. Infer the new message by division.

$$\alpha_i(\mathbf{u}_{over(i)}) = \frac{\tilde{q}_i(\mathbf{u}_{over(i)})}{\beta_i(\mathbf{u}_{over(i)})} \ , \quad \beta_{i-1}(\mathbf{u}_{over(i-1)}) = \frac{\tilde{q}_{t-1}(\mathbf{u}_{over(i-1)})}{\alpha_{i-1}(\mathbf{u}_{over(i-1)})} \ .$$

Figure 4 gives a graphical representation of Algorithm 2 for $\kappa = 0$. Figure 5 gives a similar schema for $\kappa = 1$. The two figures show what information is lost when the one-slice beliefs are collapsed.

The choice of $0 \leq \kappa \leq \left\lceil \frac{T-2}{2} \right\rceil$ now allows a trade off between computational complexity and degrees of freedom in the approximation. With $\kappa = 0$, we obtain the EP/Bethe free energy equivalent to Zoeter and Heskes (2005). With $\kappa = \left\lceil \frac{T-2}{2} \right\rceil$ there is only one cluster and we obtain a *strong* junction tree, and the found posteriors are exact. Just as with the Kikuchi extension of belief propagation, there is no guaranteed monotonic improvement for intermediate κ's (Kappen and Wiegerinck, 2002). However, in the change point model, where there are no loops and larger clusters only imply more statistics being propagated between time-slices, we expect improvements

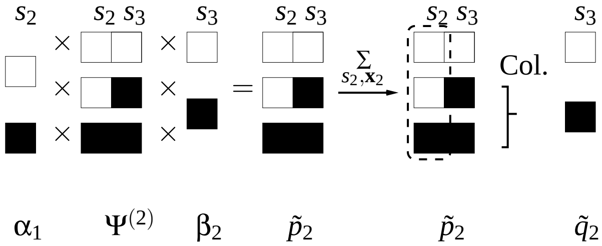

Figure 4: A schematic representation of steps 1, 2 and 3 from Algorithm 2 with $\kappa = 0$, for a sequence with more than 3 observations. The potential $\Psi^{(2)}(\mathbf{u}_{2,3})$ contains three Gaussian components: $p(\mathbf{y}_3, \mathbf{x}_3|\mathbf{x}_2, s_2 = \mathrm{n}, s_3 = \mathrm{n})$, $p(\mathbf{y}_3, \mathbf{x}_3|\mathbf{x}_2, s_2 = \mathrm{n}, s_3 = \mathrm{p})$, and $p(\mathbf{y}_3, \mathbf{x}_3|\mathbf{x}_2, s_2 = \mathrm{p}, s_3 = \mathrm{p})$. The (p, n) assignment gets zero weight by the non-recovery assumption and is therefore not shown. Combinations with absorbing states are excluded since the sequence does not stop at 3. Every component is encoded by a row, with white squares denoting normal, and black squares prefault regimes. The messages $\alpha_1(\mathbf{x}_2, s_2)$ and $\beta_2(\mathbf{x}_3, s_3)$ are conditional Gaussian by construction and hence each have two components: one corresponding to normal and one to prefault. Exact marginalization gives $\tilde{p}_2(\mathbf{x}_3, s_3)$, which still consists of three components. To emphasize that s_2 is not part of the domain, it is enclosed by a dashed rectangle. Conditioned on $s_3 = \mathrm{p}$, $\tilde{p}_2(\mathbf{x}_3|s_3 = \mathrm{p})$ is a mixture. This mixture is collapsed to obtain a conditional Gaussian approximation $\tilde{q}_2(\mathbf{u}_3)$.

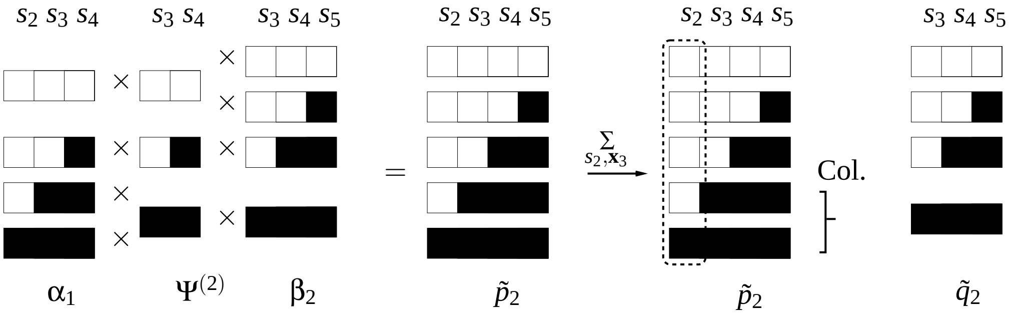

Figure 5: The steps in Algorithm 2 with $\kappa = 1$ are analogous to the steps with $\kappa = 0$ as depicted in Figure 4. With $\kappa = 1$ the two components that are approximated are expected to be very similar: they have been updated with the same transition and observation models in the last three time-slices.

to be extremely likely. In fact, we have not seen the performance degrade with larger κ in any of our experiments.

6. Experiments

In Section 6.1 we explore the properties of the learning algorithm and approximate inference in a controled setting with artificial data. Section 6.2 presents experiments with EMG data.

6.1 Synthetic Data

As discussed in Section 4 the constraints in the regime transitions aid in learning. When a stop is observed in the trainset, the entire sequence is guaranteed to be normal. Also, the fact that the normal regime precedes the prefault regime resolves the invariance under relabeling that would be present in a general switching linear dynamical systems setting. Experiments with artificially generated data shows that even with a relatively small trainset the two regimes can be learned fairly reliably.

We ran experiments where 15 train and 5 test sequences were generated from randomly drawn change point models. The classes of the train sequences (stop or fault) were observed, the classes of the test sequences were unknown. Figure 6 is not an a-typical result. In many experiments we find that both the classification (determining whether the sequence ended in a stop or in a fault) and the determination of the change point were often (near) perfect.

The MAP

$$\tau_{\text{map}} = \underset{\tau}{\operatorname{argmax}}\, p(s_{1:\tau} = \mathrm{n}, s_{\tau+1:T} = \mathrm{p}|\mathbf{y}_{1:T}, \theta_{\text{ML}})\ ,$$

is taken as the predicted change point. In 10 replications the mean squared error between the actual and the inferred change point was 6.6 (standard deviation 11.75, median 0).

These results are encouraging, but may also be largely due to the fact that arbitrarily drawn models may not pose a serious challenge. Qualitatively the replicated experiments show that for most replications the errors are close to 0 (as in Figure 6). This explains the low median. In a few replications the model has learned normal and prefault classes that are different from the true generating model and hence result in large errors. In these replications we still see the “arbitrariness” of the fitted clusters that is common to the mixtures of Gaussians learning. We do not investigate a proper characterization of “difficult” and “easy” models here, but discuss some of the possible pitfalls with the approach in Section 6.2.

To explore the properties of the approximations developed in Section 5, we ran 10 experiments where a single sequence of length 10 was generated from a randomly drawn model. For every sequence, approximate single node posteriors $\tilde{q}(\mathbf{x}_t|\mathbf{y}_{1:T}, \theta)$ were computed using Algorithm 2 with $\kappa = 0, 1, 2, 3, 4$. Figure 7 shows the maximum absolute error in the single node posterior means as a function of κ. The lines show the average over the 10 experiments, the maximum encountered, and the minimum. For sequences with length 10, $\kappa = 4$ is guaranteed to give exact results. So in theory, the lines in Figure 7 should meet at $\kappa = 4$. The discrepancies in Figure 7 are explained by different round off errors in our implementations of Algorithm 2 and the strong junction tree.

As expected the approximations are very good and improve with the size of κ. It must be emphasized however, that the improvement with larger κ can be expected based on intuition, but is not guaranteed.

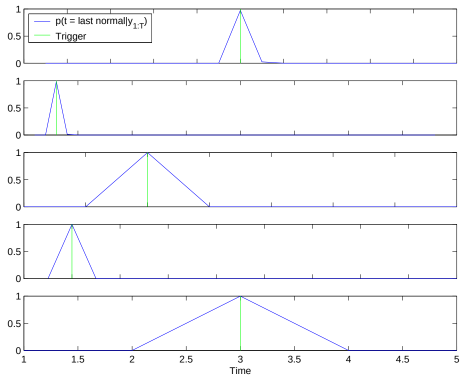

Figure 6: Shown are the inferred and true change points on 5 test sequences. The EM algorithm from Section 4 was presented with 15 artificially generated train sequences, all of which resulted in an observed fault.

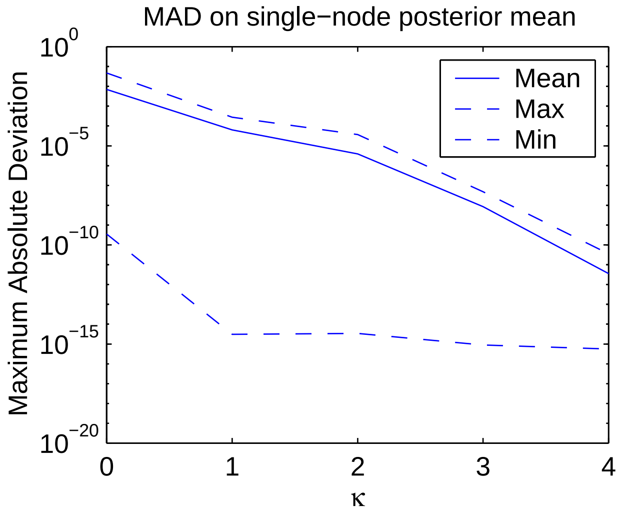

Figure 7: Maximum absolute deviation between exact and approximate single-slice posterior state mean as a function of κ. Shown are the mean, maximum and minimum over ten replications. In all replications T=10, so $\kappa = 4$ gives exact results. The small differences between the mean, maximum and minimum deviations that are observed in the plot for $\kappa = 4$ are caused by different round off errors in the generalized EP and the original strong junction tree implementations.

6.2 Detecting Changes in EMG Signals

The algorithm from Section 4 was used to detect changes in EMG patterns in the stumbling experiments from Schillings et al. (1996).

In Schillings et al. (1996) bipolar electromyography (EMG) activity in human subjects were recorded. The subjects were walking on a treadmill at 4 km/h. By releasing an object suspended from an electromagnet the subjects could be tripped at a specified phase in the walking pattern. Partially obscured glasses and earplugs ensured that the tripping was unexpected.

In the experiments video recordings and a pressure-sensitive strip attached to the obstacle signaled the tripping onset. We extracted an interesting change point problem from this experiment by only looking at the EMG signals measured at the biceps femoris at the contra lateral side, i.e. by only looking at a signal which is indicative of the activity of the large muscle in the upper leg at the non-obstructed side.

In our experiments we used data for a single subject. The dataset consisted of 15 control trials where no object was released and 8 stumbling experiments. All sequences were of equal length, and started roughly at the same phase in the walking pattern. The control trials were treated as normal sequences and the 8 others as fault sequences. In the first experiments the class of the sequences were assumed to be known and the aim for the algorithm was to determine the change points.

The original series were raised to a power of -.2 to obtain signals that seemed in agreement with the additive noise assumptions. The initial parameter settings for the normal regime was in *Fourier form* (West and Harrison, 1997). The chosen harmonic components were obtained from a discrete Fourier transform. Based on the residuals of the 15 normal sequences a model with 4 harmonics was selected.

There were three different phases of training. In the first, only the normal sequences were considered and the transition matrix A_{n} was kept fixed. In the second phase, again only the normal regime was considered, but A_{n} was also fitted. In all phases of learning κ was set to 50, i.e. inference results were indistinguishable from exact. The result of the first two phases is characterized by the left plot in Figure 8. In the third phase all parameters were fitted. The prefault model was initialized as an outlier model, i.e. the parameters for the prefault regime were copies of the normal regime, but the noise covariances were larger. The characteristics of the entire model are depicted in the right plot of Figure 8.

After convergence, the mean absolute distance between the MAP change point and the triggers in the 8 fault sequences is 4.25, with standard deviation 1.49. Figure 9 shows the posteriors $p(s_{1:t} = \mathrm{n}, s_{t+1:T} = \mathrm{p}|\mathbf{y}_{1:T}, \theta)$ and the trigger signals for the 8 fault sequences. There are two typical errors: the inferred change point for a few sequences is several steps too early, for a few it comes too late. Figure 10 gives the characteristics for the second and the third fault sequences. The MAP of the second sequence falls a few time steps after the trigger. From the left plot in Figure 10, we might judge that the actual response in the biceps femoris actually starts close to the inferred point. These characteristics are also visible in the other sequences with ‘late’ inferred change points. On the other hand, the sequences with too early inferred change points (e.g. the right plot in Figure 10), *do* show a weakness of the current setup. The degrees of freedom that are available in the prefault submodel are used to also explain outliers at the end of the normal regime. This is likely to be a problem in the model specification; there is nothing to prevent a discontinuity in the expected muscle activity at a change (as can be seen in the right plot in Figure 8 and in the plots in Figure 10). Adapting

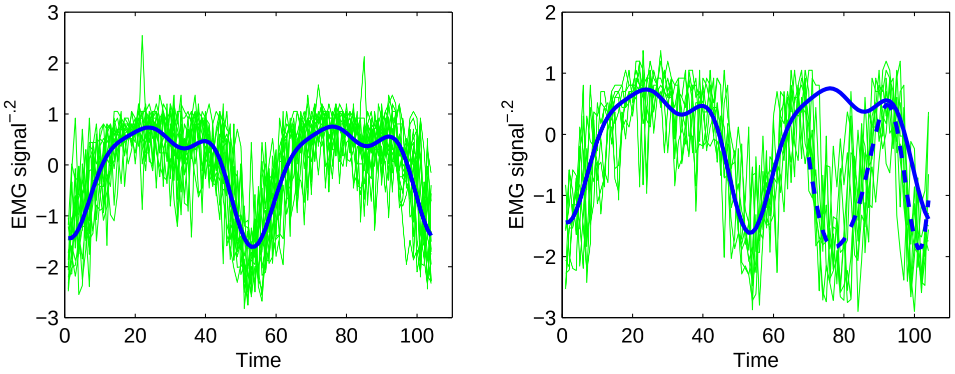

Figure 8: Characteristics of the learned model. The left plot shows the transformed EMG signals for all normal sequences in thin solid lines. The thick solid line shows the model prediction with the regime indicators clamped to normal. The right plot shows all EMG signals from stumbling trials. The light thick line shows model predictions with all regime indicators clamped to normal just as in the left plot. The dark thick line shows the model predictions with the regime indicators clamped to normal from 1 to 70 and to prefault from 71 to 104. This change point was hand picked and roughly coincides with the average trigger time.

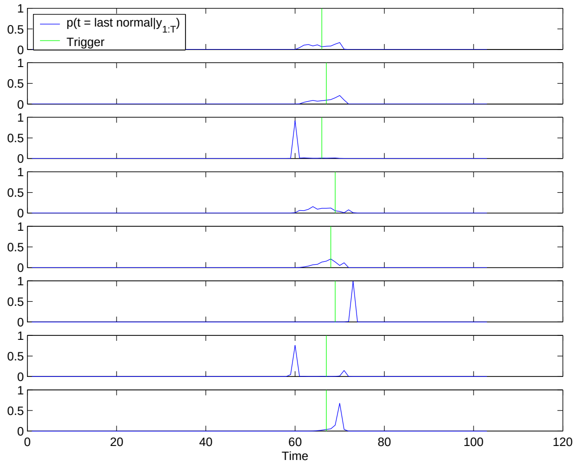

Figure 9: Stumble detection based on a single EMG signal. Vertical bars show the true stumbling trigger. The solid curves show the posterior probability that t was the last normal time point.

the model such that the expected muscle activity is continuous even during a change point might resolve some of the observed overfitting.

To test the classification performance we ran 23 leave-one-out experiments. In every experiment 22 sequences were presented in the training phase. The sequence that was left out was presented after training with s_{T+1} not observed. A simple classification scheme was used; every sequence for

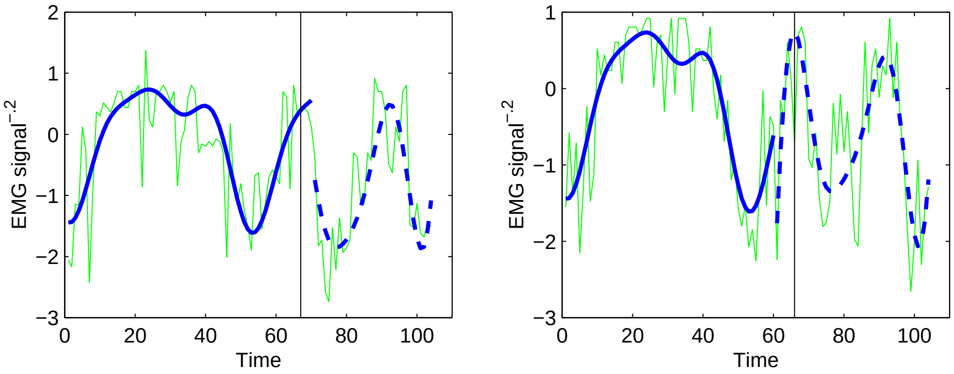

Figure 10: Thin lines show the EMG recordings of a single sequence. The vertical bars show the moment at which the stumble was triggered. The thick lines give an indication of the learned models. They were constructed by clamping the discrete states of the model to the MAP change point value and computing the predicted mean EMG signal (light lines represent the normal, dark lines the prefault regime). The left plot shows the second (from the top) sequence from Figure 9 and gives an acceptable detection of the prefault regime. The right plot shows the third sequence and represents a typical overfit: the model uses its degrees of freedom to fit outliers preceding the change point in some of the sequences. This explains the too early warnings in Figure 9.

which $p(s_{1:T} = \mathrm{n}|\mathbf{y}_{1:T}, \theta) \geq .5$ was classified as normal. With this scheme all abnormal sequences, and 13 out of 15 normal sequences were correctly classified.

7. Discussion

Motivated by fault and change detection problems in dynamical systems we have introduced a switching linear dynamical system with constrained regime transition probabilities. The system is assumed to start in a normal regime and to either result in an absorbing stop state or change to a prefault regime. Once the system reaches a prefault regime, it cannot recover and eventually has to result in a fault.

These model assumptions have several advantages. As discussed in Sections 2 and 3, the assumption that the system cannot recover can be exploited to yield an algorithm that computes exact state and regime posteriors in time polynomial in the number of observations.

Another advantage is with learning. An observed stop implies that the system did not change, and an observed fault implies that it did. So if a set of training sequences exists for which the exact change points are unknown, but for which the resulting absorbing states are observed, these model assumptions provide an interesting semi-supervised learning setting. The experiments from Section 6 indicate that these extra assumptions help to solve some of the problems with local minima that occur in general mixtures of Gaussians and SLDS learning. Although overfitting may still occur, careful initialization may be necessary, and violations of the linear Gaussian assumptions may pose problems.

Since the number of observations, T, may grow very large we have introduced an approximate inference algorithm in Section 5.

The algorithm, generalized expectation propagation (GEP), can be derived as a fixed point iteration that aims to minimize a variant of a Kikuchi free energy. One way of interpreting the algorithm is that it sends messages along a weak junction tree as if it was a strong junction tree. This is analogous to the interpretation of loopy belief propagation as an algorithm that sends messages on a loopy graph as if it was operating on a tree.

The change point model has two pleasant properties that makes the application of GEP particularly elegant. The first is the fact that the conditional independencies in the underlying model form a chain. Therefore we can straightforwardly choose outer clusters in the Kikuchi approximation such that they form a (weak) junction tree. We have shown that the resulting GEP updates then simplify since only outer clusters and direct overlaps need to be considered, i.e. from an implementation point of view the algorithm is not more complicated than the ordinary EP algorithm. Also, since there are no loops disregarded, increasing the cluster size leads to relatively "well behaved" approximations; they satisfy the perfect correlation and non-singularity conditions from Welling et al. (2005). Increasing the size of the clusters in our approximation implies that more statistics are passed from past to future and vice versa. This makes an improvement in the approximation very likely (although an improvement is only guaranteed for $\kappa \propto T$ at which point it becomes exact). This is in contrast to the generalization of belief propagation on e.g. complete graphs, which is notorious for the fact that with unfortunate choices of clusters the quality degrades with larger clusters (Kappen and Wiegerinck, 2002). In our experiments with the change point model, we have never observed a degradation of the quality with an increase of κ. This suggests that κ should be set as large as computing power permits.

The first pleasant property of the change point model leads to the observation that in approximations with weak consistency constraints it makes sense to take clusters larger than is necessary to form a (weak) junction tree. This property is shared with all models that have (weak) junction trees with reasonable cluster sizes, in particular chains and trees.

The second property is due to the no-recovery assumption property in the change point model. This implies that exact inference is polynomial in T, and also that approximate inference is polynomial in κ, which makes a wide range of κ's feasible. In a general SLDS exact inference scales exponential in T and approximate inference exponential in κ.

Although we did not discuss this in Section 5, the GEP algorithm is not restricted to trees or chains. In models with cycles and complicated parametric families, an algorithm can send messages as if it is sending messages on a strong junction tree, whereas the underlying cluster choices do not form a tree, neither a weak nor a strong one. See Heskes and Zoeter (2003) for a discussion.

Algorithm 2 is conjectured to be a proper generalization of the EP framework. Although tree EP (Minka and Qi, 2004) results in approximations that are related to (variants of) Kikuchi free energies it is unlikely that a tree or another clever choice of the approximating family would result in Algorithm 2. Since the overlapping $\tilde{q}(\mathbf{u}_{over(i)})$ are not strongly consistent they cannot easily be interpreted as marginals of a proper approximating family on which the EP algorithm would project.

Acknowledgments

We would like to thank I. Schillings for providing the EMG data that was used in the experiments of Section 6.2.

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

$$p(s,\mathbf{x}) = \pi_s (2\pi)^{-d/2} |\Sigma_s|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_s)^\top \Sigma_s^{-1} (\mathbf{x}-\mu_s) \right] ,$$

with moment parameters $\{\pi_s, \mu_s, \Sigma_s + \mu_s\mu_s^\top\}$, 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 $\Sigma_s + \mu_s\mu_s^\top$ instead of Σ_s is motivated by (16) 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

$$p(s,\mathbf{x}) = \exp\left[g_s + \mathbf{x}^\top \mathbf{h}_s - \frac{1}{2}\mathbf{x}^\top K_s \mathbf{x} \right] , \tag{14}$$

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

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

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

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

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

A conditional Gaussian *potential* is a generalization of the above distribution in the sense that it has the same form as in (14) 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 $g(.)$. 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:

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

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

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

which satisfies $1(s,\mathbf{x})p(s,\mathbf{x}) = p(s,\mathbf{x})$ 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 μ_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:

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

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 μ_{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

$$\begin{aligned}\mathrm{Collapse}\left(p(s,\mathbf{x})\right) &\equiv p(s)\mathcal{N}\left(\mathbf{x};\mu_s,\Sigma_s\right) \\ \mu_s &\equiv \sum_r \pi_{r|s}\mu_{sr} \\ \Sigma_s &\equiv \sum_r \pi_{r|s}\left(\Sigma_{sr} + (\mu_{sr}-\mu_s)(\mu_{sr}-\mu_s)^\top\right).\end{aligned}$$

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

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

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

$$\begin{aligned} p(s,\mathbf{x}) &= \exp[\nu^\top f(s,\mathbf{x})], \text{ with} \\ f(s,\mathbf{x}) &\equiv [\delta_{s,m}\; \delta_{s,m}\mathbf{x}^\top\; \delta_{s,m}\text{vec}(\mathbf{x}\mathbf{x}^\top)^\top | m \in S]^\top \\ \nu &\equiv [g_s\; \mathbf{h}_s^\top\; -\frac{1}{2}\text{vec}(K_s)^\top | s \in S]^\top \end{aligned} \tag{15}$$

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

$$g(\nu) = \langle f(s,\mathbf{x})\rangle_{\exp[\nu^\top f(s,\mathbf{x})]} \equiv \sum_s \int d\mathbf{x} f(s,\mathbf{x}) \exp[\nu^\top f(s,\mathbf{x})] \; . \tag{16}$$

Appendix B. The M-step Updates

We define θ as the set of all parameters

$$\theta \equiv \left\{\Pi_{i\to j}, \mathbf{m}_1, V_1, A_j, C_j, \mu_j, r_j^2 | (i,j) \in G\right\} \; ,$$

and G as the set of allowed regime transitions

$$G \equiv \{(\mathrm{n},\mathrm{n}),\; (\mathrm{n},\mathrm{p}),\; (\mathrm{n},\mathrm{s}),\; (\mathrm{p},\mathrm{p}),\; (\mathrm{p},\mathrm{f})\} \; ,$$

with the shorthands $\mathrm{n}, \mathrm{p}, \mathrm{s}, \mathrm{f}$, for normal, prefault, stop, and fault regimes respectively.

For now we assume a flat prior on θ, i.e. we compute ML instead of MAP estimates.

In the M-step we maximize the expected complete data log-likelihood $\hat{\mathcal{L}}$ with respect to θ. The expected complete data log-likelihood is defined as:

$$\hat{\mathcal{L}}(\mathbf{y}_{1:T}, s_{T+1}|\theta) \;\; \equiv \;\; E_{p(s_{1:T}, \mathbf{x}_{1:T}|\mathbf{y}_{1:T}, s_{T+1}, \theta_{\text{old}})}\left[\log p(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, s_{1:T+1}|\theta)\right] \; .$$

Using the conditional independencies implied by the model and the constraints in the regime prior and transitions we can rewrite it as:

$$\begin{aligned} \hat{\mathcal{L}}(\mathbf{y}_{1:T}, s_{T+1}|\theta) \;\; = \;\; & p(s_1 = n|\mathbf{y}_{1:T}, s_{T+1}, \theta_{\text{old}}) E_{p(\mathbf{x}_1|s_1=n, \mathbf{y}_{1:T}, s_{T+1}, \theta_{\text{old}})} \log \mathcal{N}(\mathbf{x}_1; \mathbf{m}_1, V_1) \\ & + \sum_{(i,j)\in G} \sum_{t=2}^{T+1} p(s_t = j, s_{t-1} = i|\mathbf{y}_{1:T}, s_{T+1}, \theta_{\text{old}}) \log \Pi_{i\to j} \\ & + \sum_{j\in\{n,p\}} \sum_{t=2}^{T} p(s_t = j|\mathbf{y}_{1:T}, s_{T+1}, \theta_{\text{old}}) \\ & \quad E_{p(\mathbf{x}_{t-1,t}|s_t=j, \mathbf{y}_{1:T}, s_{T+1}, \theta_{\text{old}})} \log \mathcal{N}(\mathbf{x}_t; A_j \mathbf{x}_{t-1}, Q_j) \\ & + \sum_{j\in\{\mathrm{n},\mathrm{p}\}} \sum_{t=1}^{T} p(s_t = j|\mathbf{y}_{1:T}, s_{T+1}, \theta_{\text{old}}) \\ & \quad E_{p(\mathbf{x}_t|s_t=j, \mathbf{y}_{1:T}, s_{T+1}, \theta_{\text{old}})} \log \mathcal{N}(\mathbf{y}_t; C_j \mathbf{x}_t + \mu_j, r_j I) \; . \end{aligned}$$

Note that from the model assumptions $p(s_1 = \mathrm{n}|\mathbf{y}_{1:T}, s_{T+1}, \theta_{\mathrm{old}}) = 1$.

The *M*-step updates for the parameters follow by adding Lagrange multipliers for the normalization constraints and setting partial derivatives to 0.

We use $\langle\cdot\rangle_{\cdot}$ to denote *weighted* expectations, and $p_t\,(ij)$ as a shorthand for the relevant posterior e.g.

$$\begin{aligned}\langle f(\mathbf{x}_{t-1}, \mathbf{x}_t)\rangle_{p_t(ij)} &= p(s_{t-1} = i, s_t = j|\mathbf{y}_{1:T}, s_{T+1}, \theta_{\mathrm{old}}) \\ &\quad \times \int d\mathbf{x}_{t-1,t}\, f(\mathbf{x}_{t-1}, \mathbf{x}_t) p(\mathbf{x}_{t-1,t}|s_{t-1} = i, s_t = j, \mathbf{y}_{1:T}, s_{T+1}, \theta_{\mathrm{old}})\,.\end{aligned}$$

In this notation $\langle 1\rangle_{p_t(ij)}$ simply gives a weighting factor. In the statistics above, and hence in the update equations below, we recognize forms similar to a regular LDS but now with a weighting term that would not be present in the non-switching case.

The updates for $\Pi_{i\to j}$ are weighted versions of the standard HMM updates. The prior is deterministic (all sequences start in the normal regime) and fixed.

The updates read:

$$\begin{aligned}A_j^{\mathrm{new}} &= \left(\sum_{t=2}^{T}\left\langle \mathbf{x}_t\mathbf{x}_{t-1}^{\top}\right\rangle_{p_t(\cdot j)}\right)\left(\sum_{t=2}^{T}\left\langle \mathbf{x}_{t-1}\mathbf{x}_{t-1}^{\top}\right\rangle_{p_t(\cdot j)}\right)^{-1} \\ Q_j^{\mathrm{new}} &= \frac{\left(\sum_{t=2}^{T}\left\langle \mathbf{x}_t\mathbf{x}_t^{\top}\right\rangle_{p_t(\cdot j)} - A_j^{\mathrm{new}}\sum_{t=2}^{T}\left\langle \mathbf{x}_t\mathbf{x}_{t-1}^{\top}\right\rangle_{p_t(\cdot j)}^{\top}\right)}{\sum_{t=2}^{T}\langle 1\rangle_{p_t(\cdot j)}} \\ m_1^{\mathrm{new}} &= \langle \mathbf{x}_1\rangle_{p_1(\mathrm{n})} \\ V_1^{\mathrm{new}} &= \left\langle \mathbf{x}_1\mathbf{x}_1^{\top}\right\rangle_{p_1(\mathrm{n})} - m_1^{\mathrm{new}}\,(m_1^{\mathrm{new}})^{\top} \\ \Pi_{i\to j}^{\mathrm{new}} &\propto \sum_{t=2}^{T+1}\langle 1\rangle_{p_t(ij)} \quad \forall_{(i,j)\in G}\,.\end{aligned}$$

We compute the new output matrix C_j and the new mean μ_j jointly by adding μ_j as an extra column to C_j and adding an entry to the continuous state that is always 1. We define

$$\begin{aligned}P_{t,j} &\equiv \begin{bmatrix}\left\langle \mathbf{x}_t\mathbf{x}_t^{\top}\right\rangle & \langle \mathbf{x}_t\rangle \\ \langle \mathbf{x}_t\rangle^{\top} & \langle 1\rangle\end{bmatrix} \\ \mathbf{m}_{t,j} &\equiv \begin{bmatrix}\langle \mathbf{x}_t\rangle \\ \langle 1\rangle\end{bmatrix} \\ \tilde{C}_j^{\mathrm{new}} &\equiv \begin{bmatrix}C_j^{\mathrm{new}}\ \mu_j^{\mathrm{new}}\end{bmatrix}\,,\end{aligned}$$

with the weighted expectations $\langle\cdot\rangle$ over $p_t(j)$, to arrive at

$$\begin{aligned}\tilde{C}_j^{\mathrm{new}} &= \left(\sum_{t=1}^{T}\mathbf{y}_t\mathbf{m}_{t,j}^{\top}\right)\left(\sum_{t=1}^{T}P_{t,j}\right)^{-1} \\ r_j^{2\ \mathrm{new}} &= \frac{\left(\sum_{t=1}^{T}\mathbf{y}_t^{\top}\mathbf{y}_t\,\langle 1\rangle_{p_t(j)} - \mathrm{tr}\left[\left(\tilde{C}_j^{\mathrm{new}}\right)^{\top}\left(\sum_{t=1}^{T}\mathbf{y}_t\mathbf{m}_{t,j}^{\top}\right)\right]\right)}{d\sum_{t=1}^{T}\langle 1\rangle_{p_t(j)}}\,,\end{aligned}$$

where d is the dimensionality of the observations $\mathbf{y}_t$.

When $s_{T+1} = \mathrm{f}$, or if it is not observed, posterior distributions such as $p(\mathbf{x}_{t-1,t}|s_{t-1} = i, s_t = j, \mathbf{y}_{1:T}, s_{T+1} = \mathrm{f}, \theta_{\mathrm{old}})$ are mixture of Gaussians (the $s_{T+1} = \mathrm{s}$ case results in a straightforward LDS variant). For the updates described above first and second moments of these mixtures are required. They can be computed analytically and simply boil down to the weighted sum of the means and second moments of the individual components. For example, $\left\langle \mathbf{x}_t \mathbf{x}_{t-1}^\top \right\rangle_{p_t(\cdot \mathrm{n})}$, is based on a mixture with $T-t-1$ components (if s_{T+1} is observed to be a fault), each corresponding to a possible end of the normal regime.

$$\begin{aligned}\left\langle \mathbf{x}_t \mathbf{x}_{t-1}^\top \right\rangle_{p_t(\cdot \mathrm{n})} &= \sum_{\tau=t}^{T-1} \left\langle \mathbf{x}_t \mathbf{x}_{t-1}^\top \right\rangle_{p_t:\tau(\cdot \mathrm{n:n})} \\ &\equiv \sum_{\tau=t}^{T-1} p(s_{1:\tau} = \mathrm{n}|\mathbf{y}_{1:T}, s_{T+1} = \mathrm{f}, \theta_{\mathrm{old}}) \\ &\quad \times \int d\mathbf{x}_{t-1,t} \mathbf{x}_t \mathbf{x}_{t-1}^\top p(\mathbf{x}_{t-1,t}|s_{1:\tau} = \mathrm{n}, s_{T+1} = \mathrm{f}, \theta_{\mathrm{old}}) \; .\end{aligned}$$

If the trainset consists of V sequences instead of one, in the above update steps all sums $\sum_{t=a}^{b}$ are replaced by $\sum_{v=1}^{V} \sum_{t=a}^{b_v}$. Only the update for $\mathbf{m}_1$ and V_1 change. The posterior over $\mathbf{x}_1$ is a mixture of Gaussians with one mixture component for every sequence. The required sufficient statistics follow again by a collapse.

Appendix C. Prior Distributions

In practice, if the underlying models for normal and prefault regimes are relatively “far apart”, we expect that the model parameters can be inferred reliably. For example if the prefault regime has an entirely different offset in the observation model, the prefault subsequences lie in an entirely different region of sensor space, which makes it easy to distinguish between the two. However in many practical applications we expect the difference not to be so profound. In this Section we introduce sensible priors on the parameters such that a priori knowledge can be incorporated.

Our main concern is with priors on the regime transition probabilities. There are three free parameters in the transition probabilities model: $\Pi_{\mathrm{n}\to\mathrm{n}}$, $\Pi_{\mathrm{n}\to\mathrm{p}}$ and $\Pi_{\mathrm{p}\to\mathrm{p}}$ ($\Pi_{\mathrm{n}\to\mathrm{s}} \equiv 1 - (\Pi_{\mathrm{n}\to\mathrm{n}} + \Pi_{\mathrm{n}\to\mathrm{p}})$, and $\Pi_{\mathrm{p}\to\mathrm{f}} \equiv 1 - \Pi_{\mathrm{p}\to\mathrm{p}}$ by construction).

The conjugate prior for $\Pi_{\mathrm{p}\to\mathrm{p}}$ is

$$p(\Pi_{\mathrm{p}\to\mathrm{p}}|\nu_{\mathrm{p}}, \lambda_{\mathrm{p}}) \propto \left(\Pi_{\mathrm{p}\to\mathrm{p}}\right)^{\nu_{\mathrm{p}}\lambda_{\mathrm{p}}} \left(1 - \Pi_{\mathrm{p}\to\mathrm{p}}\right)^{\nu_{\mathrm{p}}} \; .$$

The parameters ν_{p} and λ_{p} have a natural interpretation as the number of sequences and the average number of p → p transitions in a hypothesized set of “pseudo observed” sequences.

A similar reasoning holds for the parameters $\Pi_{\mathrm{n}\to\mathrm{n}}$, $\Pi_{\mathrm{n}\to\mathrm{s}}$, and $\Pi_{\mathrm{n}\to\mathrm{p}}$. Suppose we observe $V_{\mathrm{ns}} + V_{\mathrm{np}}$ sequences with on average $\bar{l}_{\mathrm{n}}$ n → n transitions, and V_{ns} of these ended in a stop and V_{np} switched to prefault. The probability of observing this set S of sequences is

$$p(S|\Pi_{\mathrm{n}\to\mathrm{n}}, \Pi_{\mathrm{n}\to\mathrm{s}}, \Pi_{\mathrm{n}\to\mathrm{p}}) = (\Pi_{\mathrm{n}\to\mathrm{n}})^{(V_{\mathrm{ns}}+V_{\mathrm{np}})\bar{l}_{\mathrm{n}}} (\Pi_{\mathrm{n}\to\mathrm{s}})^{V_{\mathrm{ns}}} \left(\Pi_{\mathrm{n}\to\mathrm{p}}\right)^{V_{\mathrm{np}}} \; .$$

The conjugate prior is

$$p(\Pi_{\mathrm{n}\to\mathrm{n}}, \Pi_{\mathrm{n}\to\mathrm{s}}, \Pi_{\mathrm{n}\to\mathrm{p}}|\nu_{\mathrm{ns}}, \nu_{\mathrm{np}}, \lambda_n) \propto (\Pi_{\mathrm{n}\to\mathrm{n}})^{(\nu_{\mathrm{ns}}+\nu_{\mathrm{np}})\lambda_{\mathrm{n}}} (\Pi_{\mathrm{n}\to\mathrm{s}})^{\nu_{\mathrm{ns}}} \left(\Pi_{\mathrm{n}\to\mathrm{p}}\right)^{\nu_{\mathrm{np}}} \; .$$

MAP estimates can be computed by changing the M-step slightly. Instead of maximizing the likelihood, the EM algorithm now aims to maximize

$$p(\theta|\mathbf{y}_{1:T}, s_{T+1}) \propto p(\mathbf{y}_{1:T}, s_{T+1}|\theta) p(\theta|\nu_{\mathrm{np}}, \nu_{\mathrm{ns}}, \lambda_{\mathrm{n}}, \nu_{\mathrm{p}}, \lambda_{\mathrm{p}}) \ .$$

The E-step stays the same, but the M-step updates are now found by maximizing

$$\widehat{MAP}(\mathbf{y}_{1:T}, s_{T+1}, \theta) \equiv \hat{\mathcal{L}}(\mathbf{y}_{1:T}, s_{T+1}|\theta) p(\theta|\nu_{\mathrm{np}}, \nu_{\mathrm{ns}}, \lambda_{\mathrm{n}}, \nu_{\mathrm{p}}, \lambda_{\mathrm{p}}) \ .$$

The required changes in the M-step updates are minor and intuitive. Only the update step for transition probabilities changes and becomes

$$\Pi^{\mathrm{new}}_{i\to j} \propto \sum_{t=2}^{T+1} \langle 1 \rangle_{p_t(ij)} + \nu_{ij} \quad \forall_{(i,j)\in G} \ ,$$

where

$$\begin{aligned} \nu_{\mathrm{nn}} &\equiv (\nu_{\mathrm{np}} + \nu_{\mathrm{ns}})\lambda_{\mathrm{n}} \\ \nu_{\mathrm{pp}} &\equiv \nu_{\mathrm{p}}\lambda_{\mathrm{p}} \\ \nu_{\mathrm{ps}} &\equiv \nu_{\mathrm{p}} \ . \end{aligned}$$

Appendix D. The Fixed Point Interpretation of Algorithm 2

In this section we show that fixed points of Algorithm 2 are stationary points of the generalized EP free energy (13), and that the algorithm can be interpreted as fixed point iteration in dual space. 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., 2005).

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 $\mathcal{F}_{\mathrm{GEP}}$ 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 (15) of messages α_t and β_t and their relationship with one and two slice marginals

$$\begin{aligned} \tilde{p}_i(\mathbf{u}_{outer(i)}) &\propto \alpha_{i-1}(\mathbf{u}_{over(i-1)})\Psi^{(i)}(\mathbf{u}_{outer(i)})\beta_i(\mathbf{u}_{over(i)}) \\ \tilde{q}_i(\mathbf{u}_{over(i)}) &\propto \alpha_i(\mathbf{u}_{over(i)})\beta_i(\mathbf{u}_{over(i)}) \end{aligned}$$

by construction, and weak consistency

$$\left\langle f(\mathbf{u}_{over(i)}) \right\rangle_{\tilde{p}_i} = \left\langle f(\mathbf{u}_{over(i)}) \right\rangle_{\tilde{q}_i} = \left\langle f(\mathbf{u}_{over(i)}) \right\rangle_{\tilde{p}_{i+1}} \ , \tag{17}$$

as a property of a fixed point.

To identify the nature of stationary points of $\mathcal{F}_{\text{GEP}}$ we first construct the Lagrangian by adding Lagrange multipliers $\alpha_i(\mathbf{u}_{over(i)})$ and $\beta_i(\mathbf{u}_{over(i)})$ for the forward and backward consistency constraints and $\gamma_{outer(i)}$ and $\gamma_{over(i)}$ for the normalization constraints.

$$\begin{aligned}
&\mathcal{L}_{\text{GEP}}(\tilde{p},\tilde{q},\alpha,\beta,\gamma)\\
&= \sum_{i=1}^{N}\sum_{\mathbf{u}_{outer(i)}} \tilde{p}_i(\mathbf{u}_{outer(i)}) \log \frac{\tilde{p}_i(\mathbf{u}_{outer(i)})}{\Psi^{(i)}(\mathbf{u}_{outer(i)})}\\
&\quad -\sum_{i=1}^{N-1}\sum_{\mathbf{u}_{over(i)}} \tilde{q}_i(\mathbf{u}_{over(i)}) \log \tilde{q}_i(\mathbf{u}_{over(i)})\\
&\quad -\sum_{i=2}^{N} \alpha_{i-1}(\mathbf{u}_{over(i-1)})^\top \left[\sum_{\mathbf{u}_{outer(i)}} f(\mathbf{u}_{over(i-1)})\tilde{p}_i(\mathbf{u}_{outer(i)}) - \sum_{\mathbf{u}_{over(i-1)}} f(\mathbf{u}_{over(i-1)})\tilde{q}_{i-1}(\mathbf{u}_{over(i-1)})\right]\\
&\quad -\sum_{i=1}^{N-1} \beta_i(\mathbf{u}_{over(i)})^\top \left[\sum_{\mathbf{u}_{outer(i)}} f(\mathbf{u}_{over(i)})\tilde{p}_i(\mathbf{u}_{outer(i)}) - \sum_{\mathbf{u}_{over(i)}} f(\mathbf{u}_{over(i)})\tilde{q}_i(\mathbf{u}_{over(i)})\right]\\
&\quad -\sum_{i=1}^{N} \gamma_{outer(i)}\left[\sum_{\mathbf{u}_{outer(i)}} \tilde{p}_i(\mathbf{u}_{outer(i)}) - 1\right] - \sum_{i=1}^{N-1} \gamma_{over(i)}\left[\sum_{\mathbf{u}_{over(i)}} \tilde{q}_i(\mathbf{u}_{over(i)}) - 1\right] .
\end{aligned}$$

Note that $\alpha_i(\mathbf{u}_{over(i)})$ and $\beta_i(\mathbf{u}_{over(i)})$ (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. $\tilde{p}_i(\mathbf{u}_{outer(i)})$ and $\tilde{q}_i(\mathbf{u}_{over(i)})$ gives

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\text{GEP}}}{\partial \tilde{p}_i(\mathbf{u}_{outer(i)})} &= \log \tilde{p}_i(\mathbf{u}_{outer(i)}) + 1 - \log \Psi^{(i)}(\mathbf{u}_{outer(i)})\\
&\quad - \alpha_{i-1}(\mathbf{u}_{over(i-1)})^\top f(\mathbf{u}_{over(i-1)}) - \beta_i(\mathbf{u}_{over(i)})^\top f(\mathbf{u}_{over(i)}) - \gamma_{outer(i)}\\
\frac{\partial \mathcal{L}_{\text{GEP}}}{\partial \hat{q}_i(\mathbf{u}_{over(i)})} &= -\log \hat{q}_i(\mathbf{u}_{over(i)}) - 1 + \alpha_i(\mathbf{u}_{over(i)})^\top f(\mathbf{u}_{over(i)}) + \beta_i(\mathbf{u}_{over(i)})^\top f(\mathbf{u}_{over(i)}) - \gamma_{over(i)} .
\end{aligned}$$

Setting above derivatives to 0 and filling in the solutions for $\gamma_{outer(i)}$ and $\gamma_{over(i)}$ (which implies the normalization of the potentials) results in

$$\begin{aligned}
\tilde{p}_i(\mathbf{u}_{outer(i)}) &\propto e^{\alpha_{i-1}(\mathbf{u}_{over(i-1)})^\top f(\mathbf{u}_{over(i-1)})}\Psi^{(i)}(\mathbf{u}_{outer(i)})e^{\beta_i(\mathbf{u}_{over(i)})^\top f(\mathbf{u}_{over(i)})}\\
\tilde{q}_i(\mathbf{u}_{over(i)}) &\propto e^{\alpha_i(\mathbf{u}_{over(i)})^\top f(\mathbf{u}_{over(i)}) + \beta_i(\mathbf{u}_{over(i)})^\top f(\mathbf{u}_{over(i)})} .
\end{aligned}$$

The conditions $\frac{\partial \mathcal{L}_{\text{GEP}}}{\partial \alpha_i(\mathbf{u}_{over(i)})} = 0$ and $\frac{\partial \mathcal{L}_{\text{GEP}}}{\partial \beta_i(\mathbf{z}_{over(i)})} = 0$ retrieve the forward-equals-backward constraints (17).

So if we identify α_i as the vector of the canonical parameters of the message α_i and β_i as the vector of the canonical parameters of the message β_i, we see that the conditions for stationarity of $\mathcal{F}_{\text{GEP}}$ 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

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

A. T. Cemgil, H. J. Kappen, and D. Barber. A generative model for music transcription. *Accepted to IEEE Transactions on Speech and Audio Processing*, 2004.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.

P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. Technical report, Dept. of Math. and Stat., Lancaster University, 2003.

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

T. Heskes and O. Zoeter. Generalized belief propagation for approximate inference in hybrid Bayesian networks. In C. Bishop and B. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2003.

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

T. Jaakkola. Tutorial on variational approximation methods. In *Advanced Mean Field Methods, Theory and Practice*. MIT Press, 2001.

H. J. Kappen and W. Wiegerinck. Novel iteration schemes for the cluster variation method. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 415–422, Cambridge, MA, 2002. MIT Press.

C.-J. Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60:1–22, 1994.

P. R. Krishnaiah and B. Q. Miao. Review about estimation of change points. In P. R. Krishnaiah and C. R. Rao, editors, *Handbook of Statistics*, volume 7. Elsevier, 1988.

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

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

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

T. Minka and Y. Qi. Tree-structured approximations by expectation propagation. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufmann, 1988.

A. Schillings, B. van Wezel, T. Mulder, and J. Duysens. Mechanically induced stumbling during human treadmill walking. *Journal of Neuroscience Methods*, 67:11–17, 1996.

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

M. Welling, Y. W. Teh, and T. Minka. Structured regions graphs: morphing ep intro gbp. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI-2005)*, Corvallis, Oregon, 2005. AUAI Press.

M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 2nd edition, 1997.

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

J. Yedidia, W. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, July 2005.

O. Zoeter and T. Heskes. Deterministic approximate inference techniques for conditionally Gaussian state space models. Submitted, 2005.