Variational Learning for Switching State-Space Models

Zoubin Ghahramani
Geoffrey E. Hinton
Gatsby Computational Neuroscience Unit
University College London
17 Queen Square
London WC1N 3AR, UK

Email: zoubin@gatsby.ucl.ac.uk

Submitted to *Neural Computation*

Abstract

We introduce a new statistical model for time series which iteratively segments data into regimes with approximately linear dynamics and learns the parameters of each of these linear regimes. This model combines and generalizes two of the most widely used stochastic time series models—hidden Markov models and linear dynamical systems—and is closely related to models that are widely used in the control and econometrics literatures. It can also be derived by extending the mixture of experts neural network (Jacobs et al., 1991) to its fully dynamical version, in which both expert and gating networks are recurrent. Inferring the posterior probabilities of the hidden states of this model is computationally intractable, and therefore the exact Expectation Maximization (EM) algorithm cannot be applied. However, we present a variational approximation that maximizes a lower bound on the log likelihood and makes use of both the forward–backward recursions for hidden Markov models and the Kalman filter recursions for linear dynamical systems. We tested the algorithm both on artificial data sets and on a natural data set of respiration force from a patient with sleep apnea. The results suggest that variational approximations are a viable method for inference and learning in switching state-space models.

1 Introduction

Most commonly used probabilistic models of time series are descendants of either hidden Markov models (HMM) or stochastic linear dynamical systems, also known as state-space models (SSM). Hidden Markov models represent information about the past of a sequence through a single discrete random variable–the hidden state. The prior probability distribution of this state is derived from the previous hidden state using a stochastic transition matrix. Knowing the state at any time makes the past, present and future observations statistically independent. This is the *Markov* independence property that gives the model its name.

State-space models represent information about the past through a real-valued hidden state vector. Again, conditioned on this state vector, the past, present, and future observations are statistically independent. The dependency between the present state vector and the previous state vector is specified through the dynamic equations of the system and the noise model. When these equations are linear and the noise model is Gaussian, the state-space model is also known as a linear dynamical system or Kalman filter model.

Unfortunately, most real-world processes cannot be characterized by either purely discrete or purely linear–Gaussian dynamics. For example, an industrial plant may have multiple discrete modes of behavior, each of which has approximately linear dynamics. Similarly, the pixel intensities in an image of a translating object vary according to approximately linear dynamics for subpixel translations, but as the image moves over a larger range the dynamics change significantly and nonlinearly.

This paper addresses models of dynamical phenomena which are characterized by a combination of discrete and continuous dynamics. We introduce a probabilistic model called the switching state-space model inspired by the divide-and-conquer principle underlying the mixture of experts neural network (Jacobs et al., 1991). Switching state-space models are a natural generalization of hidden Markov models and state-space models in which the dynamics can transition in a discrete manner from one linear operating regime to another. There is a large literature on models of this kind in econometrics, signal processing, and other fields (Harrison and Stevens, 1976; Chang and Athans, 1978; Hamilton, 1989; Shumway and Stoffer, 1991;

Bar-Shalom and Li, 1993). Here we extend these models to allow for multiple real-valued state vectors, draw connections between these fields and the relevant literature on neural computation and probabilistic graphical models, and derive a learning algorithm for all the parameters of the model based on a structured variational approximation which rigorously maximizes a lower bound on the log likelihood.

The paper is organized as follows. In the following section we review the background material on state-space models, hidden Markov models, and hybrids of the two. In section 3, we describe the generative model—i.e. the probability distribution defined over the observation sequences—for switching state-space models. In section 4, we describe the learning algorithm for switching state-space models which is based on a structured variational approximation to the EM algorithm. In section 5 we present simulation results both in an artificial domain, to assess the quality of the approximate inference method, and in a natural domain. Finally, we conclude with section 6.

2 Background

2.1 State-space models

A state-space model defines a probability density over time series of real-valued observation vectors $\{Y_t\}$ by assuming that the observations were generated from a sequence of hidden state vectors $\{X_t\}$.[1] In particular, the state-space model specifies that given the hidden state vector at one time step the observation vector at that time step is statistically independent from all other observation vectors, and that the hidden state vectors obey the Markov independence property. The joint probability for the sequences of states X_t and observations Y_t can therefore be factored as:

$$P(\{X_t, Y_t\}) = P(X_1)P(Y_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(Y_t|X_t), \qquad (1)$$

The conditional independencies specified by equation (1) can be expressed graphically in the form of Figure 1. The simplest and most commonly used models of this kind assume that the transition and output functions are linear and time-invariant and the distributions of the state and observation variables are multivariate Gaussian. We will use the term state-space model to refer to this simple form of the model. For such models, the state transition function is

$$X_t = AX_{t-1} + w_t \qquad (2)$$

where A is the state transition matrix and w_t is zero-mean Gaussian noise in the dynamics, with covariance matrix $\mathcal{Q}$. $P(X_1)$ is assumed to be Gaussian. Equation (2) ensures that if $P(X_{t-1})$ is Gaussian, then so is $P(X_t)$. The output function is

$$Y_t = CX_t + v_t \qquad (3)$$

where C is the output matrix and v_t is zero-mean Gaussian output noise with covariance matrix R; $P(Y_t|X_t)$ is therefore also Gaussian:

$$P(Y_t|X_t) = (2\pi)^{-D/2}|R|^{-1/2}\exp\left\{-\frac{1}{2}(Y_t - CX_t)' R^{-1} (Y_t - CX_t)\right\}, \qquad (4)$$

where D is the dimensionality of the Y vectors.

Often, the observation vector can be divided into input (or predictor) variables and output (or response) variables. To model the input–output behavior of such a system—i.e. the conditional probability of output sequences given input sequences—the linear Gaussian SSM can be modified to have a state-transition function

$$X_t = AX_{t-1} + BU_t + w_t, \qquad (5)$$

where U_t is the input observation vector and B is the (fixed) input matrix.[2]

[1] A table describing the variables and the notation used throughout the paper is provided in Appendix A.

[2] One can also define the state such that $X_{t+1} = AX_t + BU_t + w_t$.

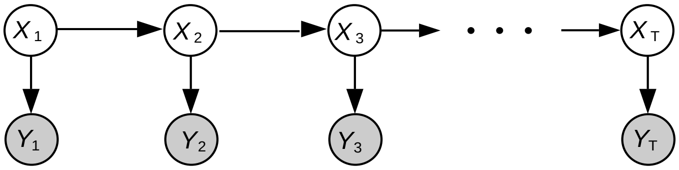

Figure 1: A directed acyclic graph (DAG) specifying conditional independence relations for a state-space model. Each node is conditionally independent from its non-descendents given its parents: The output Y_t is conditionally independent from all other variables given the state X_t; and X_t is conditionally independent from $X_1, \ldots, X_{t-2}$ given X_{t-1}. In this figure and the following figures, shaded nodes represent observed variables and unshaded nodes represent hidden variables.

The problem of *inference* or *state estimation* for a state-space model with known parameters consists of estimating the posterior probabilities of the hidden variables given a sequence of observed variables. Since the local likelihood functions for the observations are Gaussian and the priors for the hidden states are Gaussian, the resulting posterior is also Gaussian. Three special cases of the inference problem are often considered: filtering, smoothing, and prediction (Anderson and Moore, 1979; Goodwin and Sin, 1984). The goal of *filtering* is to compute the probability of the current hidden state X_t given the sequence of inputs and outputs up to time t—$P(X_t|\{Y\}_1^t, \{U\}_1^t)$.[3] The recursive algorithm used to perform this computation is known as the *Kalman filter* (Kalman and Bucy, 1961). The goal of *smoothing* is to compute the probability of X_t given the sequence of inputs and outputs up to time T, where $T > t$. The Kalman filter is used in the forward direction to compute the probability of X_t given $\{Y\}_1^t$ and $\{U\}_1^t$. A similar set of *backward* recursions from T to t complete the computation by accounting for the observations after time t (Rauch, 1963). We will refer to the combined forward and backward recursions for smoothing as the Kalman smoothing recursions (also known as the RTS or Rauch-Tung-Streibel smoother). Finally, the goal of *prediction* is to compute the probability of future states and observations given observations upto time t. Given $P(X_t|\{Y\}_1^t, \{U\}_1^t)$ computed as before, the model is simulated in the forward direction using equations (2) (or (5) if there are inputs) and (3) to compute the probability density of the state or output at future time $t + \tau$.

The problem of *learning* the parameters of a state-space model is known in engineering as the *system identification* problem, and in its most general form assumes access only to sequences of input and output observations. We focus on maximum likelihood learning, in which a single (locally optimal) value of the parameters is estimated, rather than Bayesian approaches which treat the parameters as random variables and compute or approximate the posterior distribution of the parameters given the data. One can also distinguish between on-line and off-line approaches to learning. On-line recursive algorithms, favored in real-time adaptive control applications, can be obtained by computing the gradient or the second derivatives of the log likelihood (Ljung and Söderström, 1983). Similar gradient-based methods can be obtained for off-line methods. An alternative method for off-line learning makes use of the Expectation Maximization (EM) algorithm (Dempster et al., 1977). This procedure iterates between an E-step that fixes the current parameters and computes posterior probabilities over the hidden states given the observations, and an M-step that maximizes the expected log likelihood of the parameters using the posterior distribution computed in the E-step. For linear Gaussian state-space models, the E-step is exactly the Kalman smoothing problem as defined above, and the M-step simplifies to a linear regression problem (Shumway and Stoffer, 1982; Digalakis et al., 1993). Details on the EM algorithm for state-space models can be found in Ghahramani and Hinton (1996b), as well as in the original Shumway and Stoffer (1982) paper.

2.2 Hidden Markov models

Hidden Markov models also define probability distributions over sequences of observations $\{Y_t\}$. The distribution over sequences is obtained by specifying a distribution over observations at each time step t given a *discrete* hidden state S_t, and the probability of transitioning from one hidden state to another. Using the Markov property, the joint probability for the sequences of states S_t and observations Y_t, can be factored in

[3] The notation $\{Y\}_1^t$ is short-hand for the sequence $Y_1, \ldots, Y_t$.

exactly the same manner as equation (1), with S_t taking the place of X_t:

$$P(\{S_t, Y_t\}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t). \tag{6}$$

Similarly, the conditional independencies in an HMM can be expressed graphically in the same form as Figure 1. The state is represented by a single multinomial variable that can take one of K discrete values, $S_t \in \{1, \ldots, K\}$. The state transition probabilities, $P(S_t|S_{t-1})$, are specified by a $K \times K$ transition matrix. If the observables are discrete symbols taking on one of L values, the observation probabilities $P(Y_t|S_t)$ can be fully specified as a $K \times L$ observation matrix. For a continuous observation vector, $P(Y_t|S_t)$ can be modeled in many different forms, such as a Gaussian, mixture of Gaussians, or a neural network. HMMs have been applied extensively to problems in speech recognition (Juang and Rabiner, 1991), computational biology (Baldi et al., 1994), and fault detection (Smyth, 1994).

Given an HMM with known parameters and a sequence of observations, two algorithms are commonly used to solve two different forms of the inference problem (Rabiner and Juang, 1986). The first computes the posterior probabilities of the hidden states using a recursive algorithm known as the *forward–backward* algorithm. The computations in the forward pass are exactly analogous to the Kalman filter for SSMs, while the computations in the backward pass are analogous to the backward pass of the Kalman smoothing equations. As noted by Bridle (personal communication, 1985) and Smyth, Heckerman and Jordan (1997), the forward–backward algorithm is a special case of exact inference algorithms for more general graphical probabilistic models (Lauritzen and Spiegelhalter, 1988; Pearl, 1988). The same observation holds true for the Kalman smoothing recursions. The other inference problem commonly posed for HMMs is to compute the single most likely sequence of hidden states. The solution to this problem is given by the *Viterbi* algorithm, which also consists of a forward and backward pass through the model.

To learn maximum likelihood parameters for an HMM given sequences of observations, one can use the well-known *Baum-Welch* algorithm (Baum et al., 1970). This algorithm is a special case of EM that uses the forward–backward algorithm to infer the posterior probabilities of the hidden states in the E-step. The M-step uses expected counts of transitions and observations to re-estimate the transition and output matrices (or linear regression equations in the case where the observations are Gaussian distributed). Like state-space models, HMMs can be augmented to allow for input variables, such that they model the conditional distribution of sequences of output observations given sequences of inputs (Cacciatore and Nowlan, 1994; Bengio and Frasconi, 1995; Meila and Jordan, 1996).

2.3 Hybrids

A burgeoning literature on models which combine the discrete transition structure of HMMs with the linear dynamics of SSMs has developed in fields ranging from econometrics to control engineering, (Harrison and Stevens, 1976; Chang and Athans, 1978; Hamilton, 1989; Shumway and Stoffer, 1991; Bar-Shalom and Li, 1993; Deng, 1993; Kadirkamanathan and Kadirkamanathan, 1996; Chaer et al., 1997). These models are known alternately as hybrid models, state-space models with switching, and jump-linear systems. We briefly review some of this literature, including some related neural network models.[4]

Shortly after Kalman and Bucy solved the problem of state estimation for linear Gaussian state-space models attention turned to the analogous problem for switching models (Ackerson and Fu, 1970). Chang and Athans (1978) derive the equations for computing the conditional mean and variance of the state when the parameters of a linear state-space model switch according to arbitrary and Markovian dynamics. The prior and transition probabilities of the switching process are assumed to be known. They note that for M models (sets of parameters) and an observation length T, the exact conditional distribution of the state is a Gaussian mixture with M^T components. The conditional mean and variance, which require far less computation, are therefore only summary statistics.

Shumway and Stoffer (1991) consider the problem of learning the parameters of state-space models with a single real-valued hidden state vector and switching output matrices. The probability of choosing a

[4] A review of how state-space models and HMMs are related to simpler statistical models such as PCA, factor analysis, mixture of Gaussians, vector quantization and independent components analysis (ICA) can be found in Roweis and Ghahramani (1999).

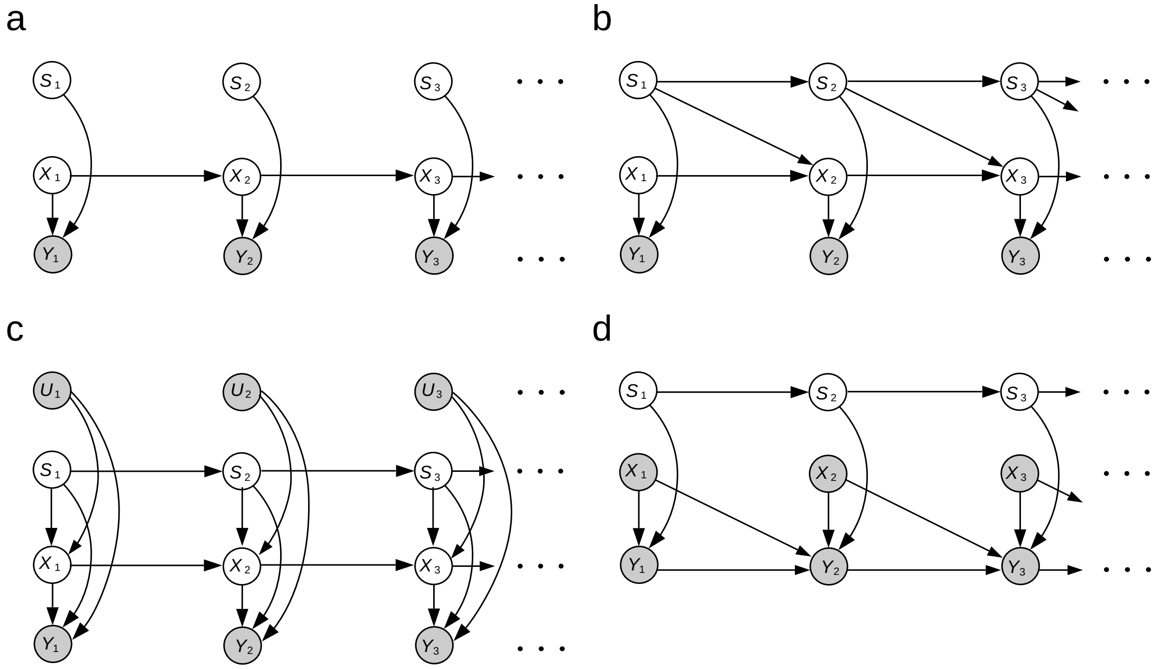

Figure 2: Directed acyclic graphs specifying conditional independence relations for various switching state-space models. (a) Shumway and Stoffer (1991): the output matrix (C in equation (3)) switches independently between a fixed number of choices at each time step. Its setting is represented by the discrete hidden variable S_t; (b) Bar-Shalom and Li (1993): both the output equation and the dynamic equation can switch and the switches are Markov; (c) Kim (1994); (d) Fraser and Dimitriadis (1993): outputs and states are observed. Here we have shown a simple case where the output depends directly on the current state, previous state and previous output.

particular output matrix is a pre-specified time-varying function, independent of previous choices (Figure 2a). A pseudo-EM algorithm is derived in which the E-step, which in its exact form would require computing a Gaussian mixture with M^T components, is approximated by a single Gaussian at each time step.

Bar-Shalom and Li (1993; sections 11.6) review models in which both the state dynamics and the output matrices switch, and where the switching follows Markovian dynamics (Figure 2b). They present several different methods for approximately solving the state-estimation problem in switching models (they do not discuss parameter estimation for such models). These methods, which are referred to as generalized pseudo-Bayesian (GPB) and interacting multiple models (IMM), are all based on the idea of collapsing into one Gaussian the mixture of M Gaussians which results from considering all the settings of the switch state at a given time step. This avoids the exponential growth of mixture components at the cost of providing an approximate solution. More sophisticated but computationally expensive methods that collapse M^2 Gaussians into M Gaussians are also derived. Kim (1994) derives a similar approximation for a closely related model which also includes observed input variables (Figure 2c). Furthermore, Kim discusses parameter estimation for this model, although without making reference to the EM algorithm. Other authors have used Markov chain Monte Carlo methods for state and parameter estimation in switching models (Carter and Kohn, 1994; Athaide, 1995) and in other related dynamic probabilistic networks (Dean and Kanazawa, 1989; Kanazawa et al., 1995).

Hamilton (1989; 1994, section 22.4) describes a class of switching models in which the real-valued observation at time t, Y_t, depends both on the observations at times $t-1$ to $t-r$ and on the discrete states at time t to $t-r$. More precisely, Y_t is Gaussian with mean that is a linear function of $Y_{t-1}, \ldots, Y_{t-r}$ and of binary indicator variables for the discrete states, $S_t, \ldots, S_{t-r}$. The system can therefore be seen as an $(r+1)^{\text{th}}$ order hidden Markov model driving an r^{th} order auto-regressive process, and are tractable for small r and number of discrete states in S.

Hamilton's models are closely related to Hidden Filter HMM (HFHMM; Fraser and Dimitriadis 1993). HFHMMs have both discrete and real-valued states. However, the real-valued states are assumed to be either observed or a known, deterministic function of the past observations (i.e. an embedding). The outputs depend on the states and previous outputs, and the form of this dependence can switch randomly (Figure 2d). Because at any time step the only hidden variable is the switch state, S_t, exact inference in this model can be carried out tractably. The resulting algorithm is a variant of the forward–backward procedure for HMMs. Kehagias and Petridis (1997) and Pawelzik et al. (1996) present other variants of this model.

Elliott et al. (1995; section 12.5) present an inference algorithm for hybrid (Markov switching) systems for which there is a separate observable from which the switch state can be estimated. The true switch states, S_t, are represented as unit vectors in $\Re^M$ and the estimated switch state is a vector in the unit square with elements corresponding to the estimated probability of being in each switch state. The real-valued state, X_t, is approximated as a Gaussian given the estimated switch state by forming a linear combination of the transition and observation matrices for the different SSMs weighted by the estimated switch state. Eliott et al. also derive control equations for such hybrid systems and discuss applications of the change-of-measures whitening procedure to a large family of models.

With regard to the literature on neural computation, the model presented in this paper is a generalization both of the mixture of experts neural network (Jacobs et al., 1991; Jordan and Jacobs, 1994) and the related mixture of factor analyzers (Hinton et al., 1996; Ghahramani and Hinton, 1996b). Previous dynamical generalizations of the mixture of experts architecture consider the case in which the gating network has Markovian dynamics (Cacciatore and Nowlan, 1994; Kadirkamanathan and Kadirkamanathan, 1996; Meila and Jordan, 1996). One limitation of this generalization is that the entire past sequence is summarized in the value of a single discrete variable (the gating activation), which for a system with M experts can convey on average at most $\log M$ bits of information about the past. In the models we consider in this paper both the experts and the gating network have Markovian dynamics. The past is therefore summarized by a state composed of the cross-product of the discrete variable and the combined real-valued state-space of all the experts. This provides a much wider information channel from the past. One advantage of this representation is that the real-valued state can contain componential structure. Thus, attributes such as the position, orientation, and scale of an object in an image, which are most naturally encoded as independent real-valued variables, can be accommodated in the state without the exponential growth required of a discretized HMM-like representation.

It is important to place the work in this paper in the context of the literature we have just reviewed. The hybrid models, state-space with switching and jump-linear systems we have described all assume that there is a single real-valued state vector. The model considered in this paper generalizes this to multiple real-valued state vectors.[5] Unlike the models described in Hamilton (1994), Fraser and Dimitradis (1993) and the current dynamical extensions of mixtures of experts, in the model we present the real-valued state vectors are hidden. The inference algorithm we derive, which is based on making a structured variational approximation, is entirely novel in the context of switching state-space models. Specifically, our method is unlike all the approximate methods we have reviewed in that it is not based on fitting a single Gaussian to a mixture of Gaussians by computing the mean and covariance of the mixture.[6] We derive a learning algorithm for all of the parameters of the model, including the Markov switching parameters. This algorithm maximizes a lower bound on the log likelihood of the data, rather than a heuristically motivated approximation to the likelihood. The algorithm has a simple and intuitive flavor: It decouples into forward-backward recursions on a hidden Markov model, and Kalman smoothing recursions on each state-space model. The states of the HMM determine the soft assignment of each observation to a state-space model; the prediction errors of the state-space models determine the observation probabilities for the HMM.

[5] Note that the state vectors could be concatenated into one large state vector with factorized (block-diagonal) transition matrices (cf. factorial hidden Markov model; Ghahramani and Jordan, 1997). However, this obscures the decoupled structure of the model.

[6] Both classes of methods can be seen as minimizing Kullback-Liebler (KL) divergences. However, the KL divergence is asymmetrical, and whereas the variational methods minimize it in one direction the methods that merge Gaussians minimize it in the other direction. We will return to this point in section 4.2.

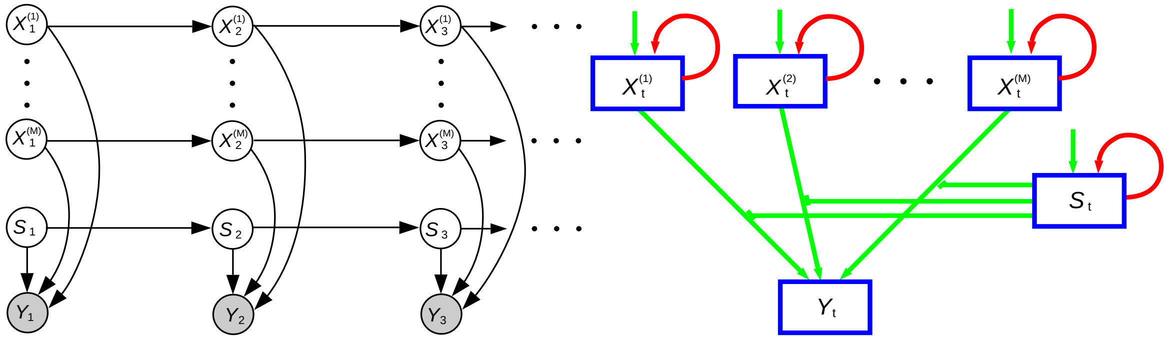

Figure 3: a) Graphical model representation for switching state-space models. S_t is the discrete switch variable and $X_t^{(m)}$ are the real-valued state vectors. b) Switching state-space model depicted as a generalization of the mixture of experts. The light arrows correspond to the connections in a mixture of experts. In a switching state-space model, the states of the experts and of the gating network also depend on their previous states (dark arrows).

3 The Generative Model

In switching state-space models, the sequence of observations $\{Y_t\}$ is modeled by specifying a probabilistic relation between the observations and a hidden state space comprising M real-valued state vectors, $X_t^{(m)}$, and one discrete state vector S_t. The discrete state, S_t, is modeled as a multinomial variable that can take on M values: $S_t \in \{1, \ldots, M\}$; for reasons that will become obvious we refer to it as the *switch* variable. The joint probability of observations and hidden states can be factored as

$$\begin{aligned} P(\{S_t, X_t^{(1)}, \ldots, X_t^{(M)}, Y_t\}) &= P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) \cdot \prod_{m=1}^{M} P(X_1^{(m)}) \prod_{t=2}^{T} P(X_t^{(m)}|X_{t-1}^{(m)}) \\ &\quad \cdot \prod_{t=1}^{T} P(Y_t|X_t^{(1)}, \ldots, X_t^{(M)}, S_t), \end{aligned} \tag{7}$$

which corresponds graphically to the conditional independencies represented by Figure 3. Conditioned on a setting of the switch state, $S_t = m$, the observable is multivariate Gaussian with output equation given by state-space model m. Notice that m is used as both an index for the real-valued state variables, and as a value for the switch state. The probability of the observation vector Y_t is therefore

$$P(Y_t|X_t^{(1)}, \ldots, X_t^{(M)}, S_t = m) = (2\pi)^{-\frac{D}{2}} |R|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(Y_t - C^{(m)} X_t^{(m)}\right)' R^{-1} \left(Y_t - C^{(m)} X_t^{(m)}\right)\right\} \tag{8}$$

where D is the dimension of the observation vector, R is the observation noise covariance matrix, and $C^{(m)}$ is the output matrix for state-space model m (cf. equation (4) for a single linear-Gaussian state-space model). Each real-valued state vector evolves according to the linear Gaussian dynamics of a state-space model with differing initial state, transition matrix, and state noise (equation (2)). For simplicity we will assume that all state vectors have identical dimensionality; the generalization of the algorithms we present to models with different size state-spaces is immediate. The switch state itself evolves according to the discrete Markov transition structure specified by the initial state probabilities $P(S_1)$ and the $M \times M$ state transition matrix $P(S_t|S_{t-1})$.

An exact analogy can be made to the "mixture of experts" architecture for modular learning in neural networks (figure 3b; Jacobs et al, 1991). Each state space model is a linear expert with Gaussian output noise model and linear-Gaussian dynamics. The switch state "gates" the outputs of the M state-space models, and therefore plays the role of a gating network with Markovian dynamics.

There are many possible extensions of the model above and we shall consider three obvious and straightforward ones:

(Ex1) Differing output covariances, $R^{(m)}$, for each state-space model;

(Ex2) Differing output means, $\mu_Y^{(m)}$, for each state-space model, such that each model is allowed to capture observations in a different operating range;

(Ex3) Conditioning on a sequence of observed input vectors, $\{U_t\}$.

4 Learning

An efficient learning algorithm for the parameters of a switching state-space model can be derived by generalizing the Expectation Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977). EM alternates between optimizing a distribution over the hidden states (the E-step) and optimizing the parameters given the distribution over hidden states (the M-step). Any distribution over the hidden states, $Q(\{S_t, X_t\})$, where $X_t = [X_t^{(1)}, \ldots X_t^{(M)}]$ is the combined state of the state-space models, can be used to define a lower bound, $\mathcal{B}$, on the log probability of the observed data:

$$\begin{aligned}
\log P(\{Y_t\}|\theta) &= \log \sum_{\{S_t\}} \int P(\{S_t, X_t, Y_t\}|\theta)\, d\{X_t\} && (9)\\
&= \log \sum_{\{S_t\}} \int Q(\{S_t, X_t\}) \left[\frac{P(\{S_t, X_t, Y_t\}|\theta)}{Q(\{S_t, X_t\})}\right] d\{X_t\} && (10)\\
&\geq \sum_{\{S_t\}} \int Q(\{S_t, X_t\}) \log \left[\frac{P(\{S_t, X_t, Y_t\}|\theta)}{Q(\{S_t, X_t\})}\right] d\{X_t\} = \mathcal{B}(Q, \theta), && (11)
\end{aligned}$$

where θ denotes the parameters of the model and we have made use of Jensen's inequality (Cover and Thomas, 1991) to establish (11). Both steps of EM increase the lower bound on the log probability of the observed data. The E-step holds the parameters fixed and sets Q to be the posterior distribution over the hidden states given the parameters,

$$Q(\{S_t, X_t\}) = P(\{S_t, X_t\}|\{Y_t\}, \theta). \qquad (12)$$

This maximizes $\mathcal{B}$ with respect to the distribution, turning the lower bound into an equality, which can be easily seen by substitution. The M-step holds the distribution fixed and computes the parameters that maximize $\mathcal{B}$ for that distribution. Since $\mathcal{B} = \log P(\{Y_t\}|\theta)$ at the start of the M-step, and since the E-step does not affect $\log P$, the two steps combined can never decrease $\log P$. Given the change in the parameters produced by the M-step, the distribution produced by the previous E-step is typically no longer optimal, so the whole procedure must be iterated.

Unfortunately, the exact E-step for switching state-space models is intractable. Like the related hybrid models described in section 2.3, the posterior probability of the real-valued states is a Gaussian mixture with M^T terms. This can be seen by using the semantics of directed graphs, in particular the d-separation criterion (Pearl, 1988), which implies that the hidden state variables in Figure 3, while marginally independent, become conditionally dependent given the observation sequence. This induced dependency effectively couples all of the real-valued hidden state variables to the discrete switch variable, as a consequence of which the exact posteriors become Gaussian mixtures with an exponential number of terms.[7]

In order to derive an efficient learning algorithm for this system, we relax the EM algorithm by approximating the posterior probability of the hidden states. The basic idea is that, since expectations with respect to P are intractable, rather than setting $Q(\{S_t, X_t\}) = P(\{S_t, X_t\}|\{Y_t\})$ in the E-step, a tractable distribution Q is used to *approximate* P. This results in an EM learning algorithm which maximizes a lower bound on the log likelihood. The difference between the bound $\mathcal{B}$ and the log likelihood is given by the Kullback-Liebler (KL) divergence between Q and P (Cover and Thomas, 1991):

$$\mathrm{KL}(Q\|P) = \sum_{\{S_t\}} \int Q(\{S_t, X_t\}) \log \left[\frac{Q(\{S_t, X_t\})}{P(\{S_t, X_t\}|\{Y_t\})}\right] d\{X_t\}. \qquad (13)$$

[7] The intractability of the E-step or smoothing problem in the simpler single-state switching model has been noted by Ackerson and Fu (1970), Chang and Athans (1978), Bar-Shalom and Li (1993), and others .

Since the complexity of exact inference in the approximation given by Q is determined by its conditional independence relations, not by its parameters, we can choose Q to have a tractable structure—a graphical representation which eliminates some of the dependencies in P. Given this structure, the parameters of Q are varied to obtain the tightest possible bound by minimizing (13). Therefore, the algorithm alternates between optimizing the parameters of the distribution Q to minimize (13) (the E-step) and optimizing the parameters of P given the distribution over the hidden states (the M-step). Like in exact EM, both steps increase the lower bound $\mathcal{B}$ on the log likelihood, however equality is not reached in the E-step.

We will refer to the general strategy of using a parameterized approximating distribution as a *variational approximation* and refer to the free parameters of the distribution as *variational parameters*. A completely factorized approximation is often used in statistical physics, where it provides the basis for simple yet powerful *mean field approximations* to statistical mechanical systems (Parisi, 1988). Theoretical arguments motivating approximate E-steps are presented in Neal and Hinton (1998; originally in a technical report in 1993). Saul and Jordan (1996) showed that approximate E-steps could be used to maximize a lower bound on the log likelihood, and proposed the powerful technique of *structured* variational approximations to intractable probabilistic networks. The key insight of Saul and Jordan's work, which the present paper makes use of, is that by judicious use of an approximation Q, exact inference algorithms can be used on the tractable substructures in an intractable network. A general tutorial on variational approximations can be found in Jordan et al. (1998).

The parameters of the switching state-space model are $\theta = \{A^{(m)}, C^{(m)}, \mathcal{Q}^{(m)}, \mu_{X_1}^{(m)}, \mathcal{Q}_1^{(m)}, R, \boldsymbol{\pi}, \Phi\}$, where $A^{(m)}$ is the state dynamics matrix for model m, $C^{(m)}$ is its output matrix, $\mathcal{Q}^{(m)}$ is its state noise covariance, $\mu_{X_1}^{(m)}$ is the mean of the initial state, $\mathcal{Q}_1^{(m)}$ is the covariance of the initial state, R is the (tied) output noise covariance, $\boldsymbol{\pi} = P(S_1)$ is the prior for the discrete Markov process, and $\Phi = P(S_t|S_{t-1})$ is the discrete transition matrix. Extensions (Ex1)–(Ex3) can be readily implemented by substituting $R^{(m)}$ for R, adding means $\mu_Y^{(m)}$ and input matrices $B^{(m)}$.

While there are many possible approximations to the posterior distribution of the hidden variables that one could use for learning and inference in switching state-space models, we focus on the following:

$$Q(\{S_t, X_t\}) = \frac{1}{Z_Q}\left[\psi(S_1)\prod_{t=2}^{T}\psi(S_{t-1}, S_t)\right]\prod_{m=1}^{M}\psi(X_1^{(m)})\prod_{t=2}^{T}\psi(X_{t-1}^{(m)}, X_t^{(m)}), \tag{14}$$

where the ψ are unnormalized probabilities, which we will call *potential* functions and define soon, and Z_Q is a normalization constant ensuring that Q integrates to one. Although Q has been written in terms of potential functions rather than conditional probabilities, it corresponds to the simple graphical model shown in Figure 4. The terms involving the switch variables S_t define a discrete Markov chain and the terms involving the state vectors $X_t^{(m)}$ define M *uncoupled* state-space models. Like in mean field approximations we have approximated the stochastically coupled system by removing some of the couplings of the original system. Specifically, we have removed the stochastic coupling between the chains that results from the fact that the observation at time t depends on all the hidden variables at time t. However, we retain the coupling between the hidden variables at successive time steps since these couplings can be handled exactly using the forward–backward and Kalman smoothing recursions. This approximation is therefore structured, in the sense that not all variables are uncoupled.

The discrete switching process is defined by

$$\psi(S_1 = m) = P(S_1 = m)\, q_1^{(m)} \tag{15}$$

$$\psi(S_{t-1}, S_t = m) = P(S_t = m|S_{t-1})\, q_t^{(m)}, \tag{16}$$

where the $q_t^{(m)}$ are variational parameters of the Q distribution. These parameters scale the probabilities of each of the states of the switch variable at each time step, so that $q_t^{(m)}$ plays exactly the same role as the observation probability $P(Y_t|S_t = m)$ would play in a regular hidden Markov model. We will soon see that minimizing $\mathrm{KL}(Q\|P)$ results in an equation for $q_t^{(m)}$ which supports this intuition.

The uncoupled state-space models in the approximation Q are also defined by potential functions which are related to probabilities in the original system. These potentials are the prior and transition probabilities

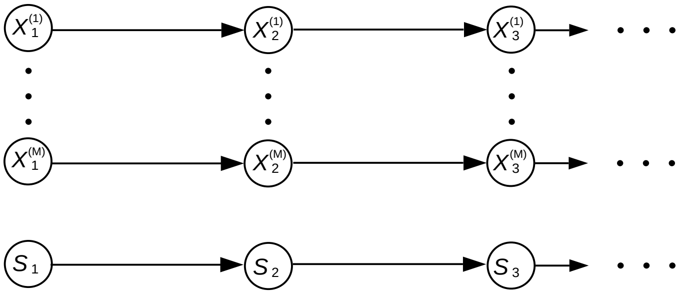

Figure 4: Graphical model representation for the structured variational approximation to the posterior distribution of the hidden states of a switching state-space model.

for $X^{(m)}$ multiplied by a factor that changes these potentials to try to account for the data:

$$\psi(X_1^{(m)}) = P(X_1^{(m)}) \left[P(Y_1|X_1^{(m)}, S_1 = m)\right]^{h_1^{(m)}} \tag{17}$$

$$\psi(X_{t-1}^{(m)}, X_t^{(m)}) = P(X_t^{(m)}|X_{t-1}^{(m)}) \left[P(Y_t|X_t^{(m)}, S_t = m)\right]^{h_t^{(m)}} \tag{18}$$

where the $h_t^{(m)}$ are variational parameters of Q. The vector h_t plays a role very similar to the switch variable S_t. Each component $h_t^{(m)}$ can range between 0 and 1. When $h_t^{(m)} = 0$ the posterior probability of $X_t^{(m)}$ under Q does not depend on the observation at time Y_t. When $h_t^{(m)} = 1$, the posterior probability of $X_t^{(m)}$ under Q includes a term which assumes that state-space model m generated Y_t. We call $h_t^{(m)}$ the ***responsibility*** assigned to state-space model m for the observation vector Y_t. The difference between $h_t^{(m)}$ and $S_t^{(m)}$ is that $h_t^{(m)}$ is a deterministic parameter, while $S_t^{(m)}$ is a stochastic random variable.

To maximize the lower bound on the log likelihood, $\mathrm{KL}(Q\|P)$ is minimized with respect to the variational parameters $h_t^{(m)}$ and $q_t^{(m)}$ separately for each sequence of observations. Using the definition of P for the switching state-space model (equation (7) and (8)) and the approximating distribution Q, the minimum of KL satisfies the following fixed point equations for the variational parameters (see Appendix B):

$$h_t^{(m)} = Q(S_t = m) \tag{19}$$

$$q_t^{(m)} = \exp\left\{-\frac{1}{2}\left\langle \left(Y_t - C^{(m)}X_t^{(m)}\right)' R^{-1} \left(Y_t - C^{(m)}X_t^{(m)}\right)\right\rangle\right\} \tag{20}$$

where $\langle\cdot\rangle$ denotes expectation over the Q distribution. Intuitively, the responsibility, $h_t^{(m)}$ is equal to the probability under Q that state-space model m generated observation vector Y_t, and $q_t^{(m)}$ is an unnormalized Gaussian function of the expected squared error if state-space model m generated Y_t.

To compute $h_t^{(m)}$ it is necessary to sum Q over all the S_τ variables not including S_t. This can be done efficiently using the forward–backward algorithm on the switch state variables, with $q_t^{(m)}$ playing exactly the same role as an observation probability associated with each setting of the switch variable. Since $q_t^{(m)}$ is related to the prediction error of model m on data Y_t, this has the intuitive interpretation that the switch state associated with models with smaller expected prediction error on a particular observation will be favored at that time step. However, the forward–backward algorithm ensures that the final responsibilities for the models are obtained after considering the entire sequence of observations.

To compute $q_t^{(m)}$ it is necessary to calculate the expectations of $X_t^{(m)}$ and $X_t^{(m)}X_t^{(m)'}$ under Q. We see this by expanding equation (20):

$$q_t^{(m)} = \exp\left\{-\frac{1}{2}Y_t'R^{-1}Y_t + Y_t'R^{-1}C^{(m)}\,\langle X_t^{(m)}\rangle - \frac{1}{2}tr\left[C^{(m)'}R^{-1}C^{(m)}\,\langle X_t^{(m)}X_t^{(m)'}\rangle\right]\right\}, \tag{21}$$

```
Initialize parameters of the model.

Repeat until bound on log likelihood has converged:

    E step Repeat until convergence of KL($Q \| P$):

        E.1 Compute $q_t^{(m)}$ from the prediction error of state-space model $m$ on
        observation $Y_t$
        E.2 Compute $h_t^{(m)}$ using the forward-backward algorithm on the HMM, with
        observation probabilities $q_t^{(m)}$
        E.3 For $m = 1$ to $M$
            Run Kalman smoothing recursions, using the data weighted by $h_t^{(m)}$

    M step

        M.1 Re-estimate parameters for each state-space model using the data
        weighted by $h_t^{(m)}$
        M.2 Re-estimate parameters for the switching process using Baum-Welch
        update equations.
```

Figure 5: Learning algorithm for switching state-space models.

where *tr* is the matrix trace operator and we have used $tr(AB) = tr(BA)$. The expectations of $X_t^{(m)}$ and $X_t^{(m)} X_t^{(m)'}$ can be computed efficiently using the Kalman smoothing algorithm on each state-space model, where for model m at time t, the data is weighted by the responsibilities $h_t^{(m)}$.[8] Since the h parameters depend on the q parameters, and vice-versa, the whole process has to be iterated, where each iteration involves calls to the forward–backward and Kalman smoothing algorithms. Once the iterations have converged, the E-step outputs the expected values of the hidden variables under the final Q.

The M-step computes the model parameters that optimize the expectation of the log likelihood (equation (34) in Appendix B), which is a function of the expectations of the hidden variables. For switching SSMs, all the parameter re-estimates can be computed analytically. For example, taking derivatives of the expectation of (34) with respect to $C^{(m)}$ and setting to zero we get

$$\hat{C}^{(m)} = \left(\sum_{t=1}^{T} \langle S_t^{(m)} \rangle \; Y_t \langle X_t^{(m)'} \rangle \right) \left(\sum_{t=1}^{T} \langle S_t^{(m)} \rangle \; \langle X_t^{(m)} X_t^{(m)'} \rangle \right)^{-1} \tag{22}$$

which is a weighted version of the re-estimation equations for SSMs. Similarly, the re-estimation equations for the switch process are analogous to the Baum-Welch update rules for HMMs. The learning algorithm for switching state-space models using the above structured variational approximation is summarized in Figure 5.

4.1 Deterministic Annealing

The KL divergence minimized in the E step of the variational EM algorithm can have multiple minima in general. One way to visualize these minima is to consider the space of all possible *segmentations* of an observation sequence of length T, where by segmentation we mean a discrete partition of the sequence between the state space models. If there are M SSMs, then there are M^T possible segmentations of the

[8] Weighting the data by $h_t^{(m)}$ is equivalent to running the Kalman smoother on the unweighted data using a time-varying observation noise covariance matrix $R_t^{(m)} = R/h_t^{(m)}$.

sequence. Given one such segmentation, inferring the optimal distribution for the real-valued states of the SSMs is a convex optimization problem, since these real-valued states are conditionally Gaussian. So the difficulty in the KL minimization lies in trying to find the best (soft) partition of the data.

Like in other combinatorial optimization problems, the possibility of getting trapped in local minima can be reduced by gradually annealing the cost function. We can employ a deterministic variant of the annealing idea by making the following simple modifications to the variational fixed point equations (19) and (20):

$$h_t^{(m)} = \frac{1}{\mathcal{T}} Q(S_t = m) \tag{23}$$

$$q_t^{(m)} = \exp\left\{ -\frac{1}{2\mathcal{T}} \left\langle \left(Y_t - C^{(m)} X_t^{(m)}\right)' R^{-1} \left(Y_t - C^{(m)} X_t^{(m)}\right) \right\rangle \right\}. \tag{24}$$

Here $\mathcal{T}$ is a *temperature* parameter, which is initialized to a large value and gradually reduced to 1. The above equations maximize a modified form of the bound $\mathcal{B}$ in (11), where the entropy of Q has been multiplied by $\mathcal{T}$ (Ueda and Nakano, 1995).

4.2 Merging Gaussians

Almost all the approximate inference methods that are described in the literature for switching state-space models are based on the idea of merging, at each time step, a mixture of M Gaussians into one Gaussian. The merged Gaussian is obtained simply by setting its mean and covariance equal to the mean and covariance of the mixture. Here we briefly describe, as an alternative to the variational approximation methods we have derived, how this more traditional Gaussian merging procedure can be applied to the model we have defined.

In the switching state-space models described in section 3 there are M different SSMs, with possibly different state-space dimensionalities, so it would be inappropriate to merge their states into one Gaussian. However, it is still possibly to apply a Gaussian merging technique by considering each SSM separately. In each SSM, m, the hidden state density produces at each time step a mixture of *two* Gaussians—one for the case $S_t = m$ and one for $S_t \neq m$. We merge these two Gaussians, weighted the current estimates of $P(S_t = m|Y_1, \ldots Y_t)$ and $1 - P(S_t = m|Y_1, \ldots Y_t)$, respectively. This merged Gaussian is used to obtain the Gaussian prior for $X_{t+1}^{(m)}$ for the next time step. We implemented a forward-pass version of this approximate inference scheme, which is analogous to the IMM procedure described in Bar-Shalom and Li (1993).

This procedure finds at each time step the "best" Gaussian fit to the current mixture of Gaussians for each SSM. If we denote the approximating Gaussian by Q and the mixture being approximated by P, "best" is defined here as minimizing KL$(P\|Q)$. Furthermore, Gaussian merging techniques are *greedy* in that the "best" Gaussian is computed at every time step and used immediately for the next time step. For a Gaussian Q, KL$(P\|Q)$ has no local minima, and it is very easy to find the optimal Q by computing the first two moments of P. Inaccuracies in this greedy procedure arise because the estimates of $P(S_t|Y_1, \ldots, Y_t)$ are based on this single merged Gaussian and not on the real mixture.

In contrast, variational methods seek to minimize KL$(Q\|P)$, which can have many local minima. Moreover, these methods are not greedy in the same sense: they iterate forward and backward in time until obtaining a locally optimal Q.

5 Simulations

5.1 Experiment 1: Variational Segmentation and Deterministic Annealing

The goal of this experiment was to assess the quality of solutions found by the variational inference algorithm, and the effect of using deterministic annealing on these solutions. We generated 200 sequences of length 200 from a simple model which switched between two SSMs. These SSMs and the switching process were defined by:

$$X_t^{(1)} = 0.99\, X_{t-1}^{(1)} + w_t^{(1)} \qquad w_t^{(1)} \sim \mathcal{N}(0,1) \tag{25}$$

$$X_t^{(2)} = 0.9\, X_{t-1}^{(2)} + w_t^{(2)} \qquad w_t^{(2)} \sim \mathcal{N}(0,10) \tag{26}$$

$$Y_t = X_t^{(m)} + v_t \qquad v_t \sim \mathcal{N}(0, 0.1) \tag{27}$$

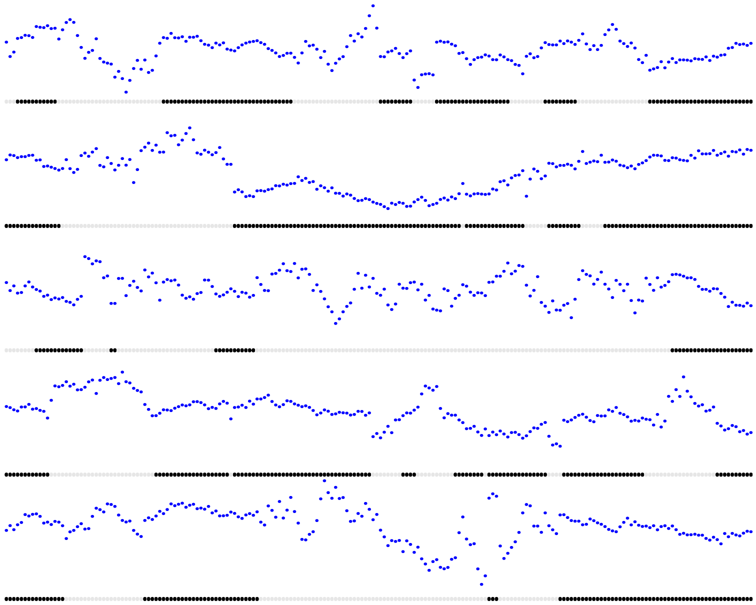

Figure 6: Five data sequences of length 200, with their true segmentations below them. In the segmentations, switch states 1 and 2 are represented with dark and light dots, respectively. Notice that it is difficult to correctly segment the sequences based only on knowing the dynamics of the two processes.

where the switch state m was chosen using priors $\boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}^{(2)} = 1/2$ and transition probabilities $\Phi_{11} = \Phi_{22} = 0.95$; $\Phi_{12} = \Phi_{21} = 0.05$. Five sequences from this data set are shown in in Figure 6, along with the true state of the switch variable.

We compared three different inference algorithms: variational inference, variational inference with deterministic annealing (section 4.1), and inference by Gaussian merging (section 4.2). For each sequence, we initialized the variational inference algorithms with equal responsibilities for the two SSMs and ran them for 12 iterations. The non-annealed inference algorithm ran at a fixed temperature of $\mathcal{T} = 1$; while the annealed algorithm was initialized to a temperature of $\mathcal{T} = 100$ which decayed down to 1 over the 12 iterations, using the decay function $\mathcal{T}_{i+1} = \frac{1}{2}\mathcal{T}_i + \frac{1}{2}$. To eliminate the effect of model inaccuracies we gave all three inference algorithms the true parameters of the generative model.

The segmentations found by the non-annealed variational inference algorithm showed little similarity to the true segmentations of the data (Figure 7). Furthermore, the non-annealed algorithm generally underestimated the number of switches, often converging on solutions with no switches at all. Both the annealed variational algorithm and the Gaussian merging method found segmentations that were more similar to the true segmentations of the data. Comparing percent correct segmentations, we see that annealing substantially improves the variational inference method, and that the Gaussian merging and annealed variational methods perform comparably (Figure 8). The average performance of the annealed variational method is only about 1.3% better than Gaussian merging.

5.2 Experiment 2: Modelling respiration in a patient with sleep apnea

Switching state-space models should prove useful in modelling time series which have dynamics characterized by several different regimes. To illustrate this point we examined a physiological data set from a patient tentatively diagnosed with sleep apnea, which is a medical condition in which patients intermittently stop breathing during sleep. The data was obtained from the repository of time series data sets associated with Santa Fe Time Series Analysis and Prediction Competition (Weigend and Gershenfeld, 1993) and is described

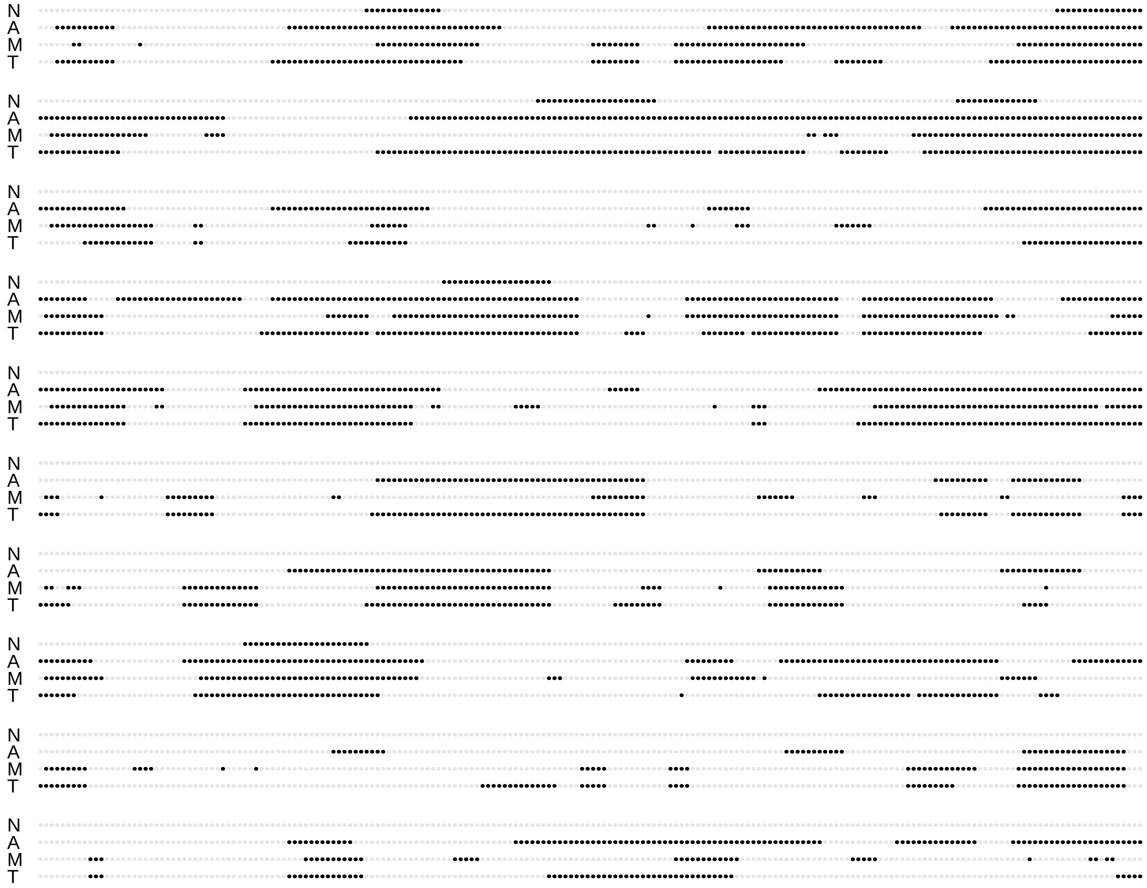

Figure 7: For ten different sequences of length 200, segmentations are shown as sequences of light and dark dots corresponding to the two SSMs generating this data. The rows are the segmentations found using the variational method with no annealing (N), the variational method with deterministic annealing (A), the Gaussian merging method (M), and the true segmentation (T). All three inference algorithms give real-valued $h_t^{(m)}$; hard segmentations were obtained by thresholding the final $h_t^{(m)}$ values at 0.5. The first five sequences are the ones shown in Figure 6.

in detail in Rigney et al. (1993).[9] The respiration pattern in sleep apnea is characterized by at least two regimes—no breathing and gasping breathing induced by a reflex arousal. Furthermore, in this patient there also seem to be periods of normal rhythmic breathing (Figure 9).

We trained switching state-space models, varying the random seed, the number of components in the mixture ($M = 2$ to 5), and the dimensionality of the state space in each component ($K = 1$ to 10), on a data set consisting of 1000 consecutive measurements of the chest volume. As controls we also trained simple state-space models (i.e. $M = 1$) varying the dimension of the state-space from $K = 1$ to 10, and simple hidden Markov models (i.e. $K = 0$) varying the number of discrete hidden states from $M = 2$ to $M = 50$. Simulations were run until convergence or for 200 iterations, whichever came first; convergence was assessed by measuring the change in likelihood (or bound on the likelihood) over consecutive steps of EM.

The likelihood of the simple SSMs and the HMMs was calculated on a test set consisting of 1000 consecu-

[9] The data is available on the web at `http://www.cs.colorado.edu/ ~andreas/ Time-Series/ SantaFe.html#setB`. We used samples 6201–7200 for training and 5201-6200 for testing.

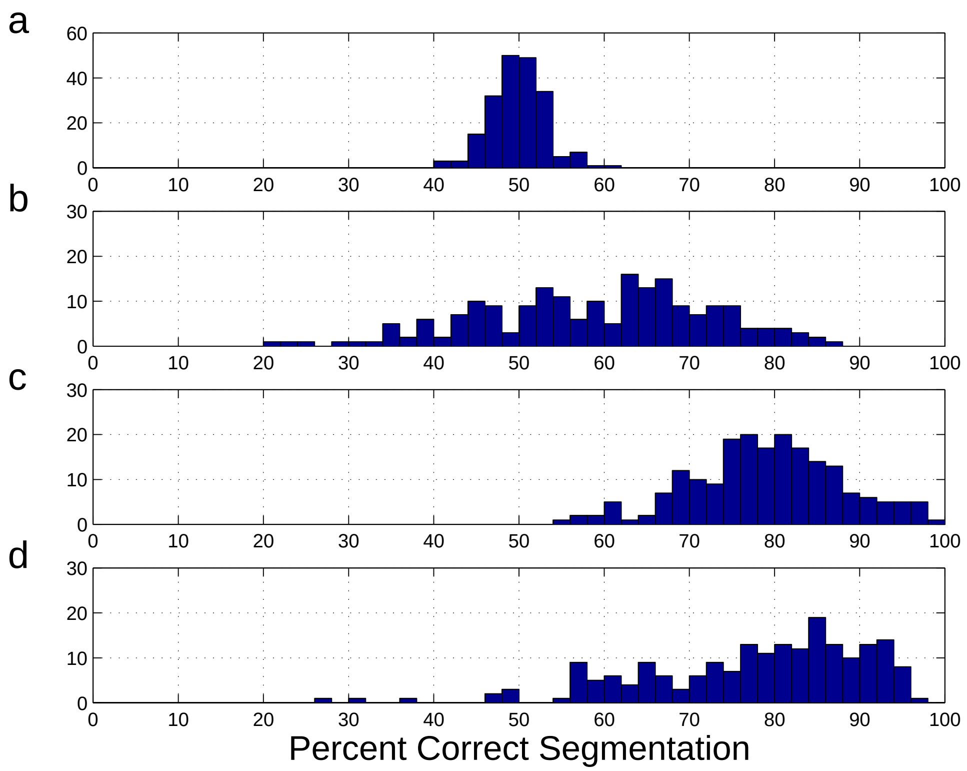

Figure 8: Histograms of percent correct segmentations: (a) control, using randon segmentation; (b) variational inference without annealing; (c) variational inference with annealing; (d) Gaussian merging. Percent correct segmentation was computed by counting the number of time steps for which the true and estimated segmentations agree.

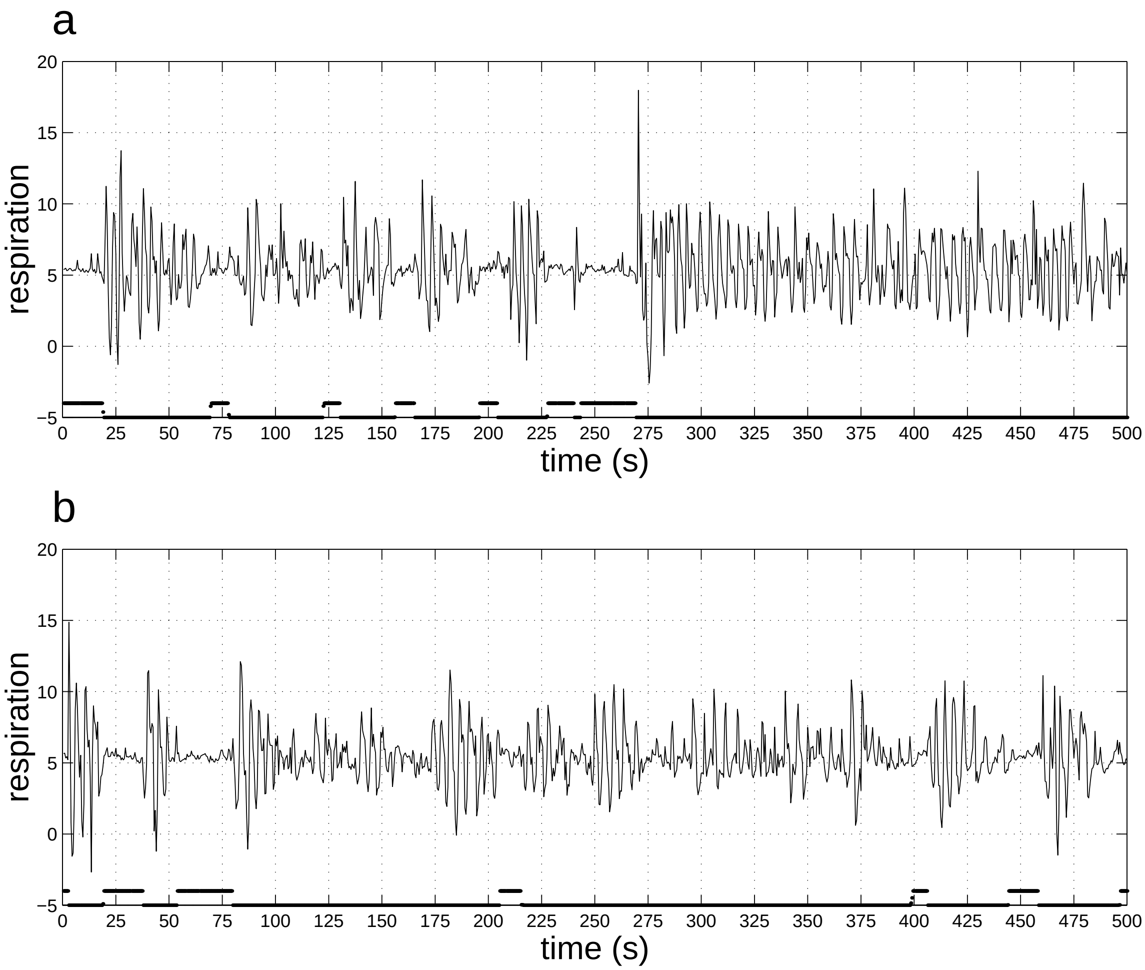

Figure 9: Chest volume (respiration force) of a patient with sleep apnea during two non-continuous time segments of the same night (measurements sampled at 2 Hz). (a) Training data. Apnea is characterized by extended periods of small variability in chest volume, followed by bursts (gasping). Here we see such behaviour around $t = 250$, followed by normal rhythmic breathing. (b) Test data. In this segment we find several instances of apnea and an approximately rhythmic region. (The thick lines at the bottom of each plot are explained in the main text.)

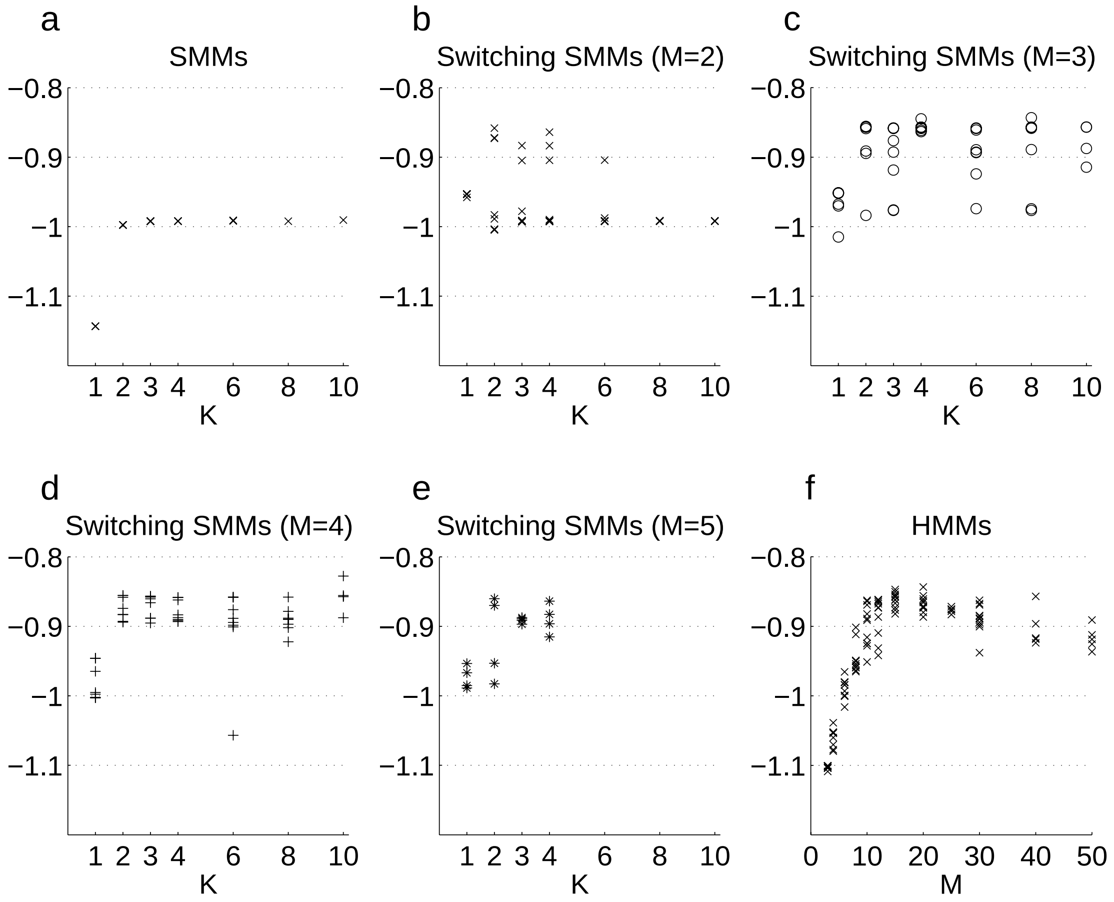

Figure 10: Log likelihood (nats per observation) on the test data from a total of almost 400 runs of simple state-space models, switching state-space models with differing numbers of components, and hidden Markov models.

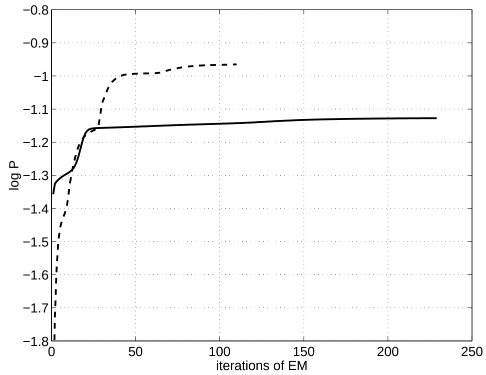

Figure 11: Learning curves for a state space model ($K = 4$) and a switching state-space model ($M = 2, K = 2$).

tive measurements of the chest volume. For the switching SSMs the likelihood is intractable so we calculated the lower bound on the likelihood, $\mathcal{B}$. The simple SSMs modeled the data very poorly for $K = 1$, and the performance was flat for values of $K = 2$ to 10 (Figure 10a). The large majority of runs of the switching state-space model resulted in models with higher likelihood than those of the simple SMMs (Figure 10b-e). One consistent exception should be noted: for values of $M = 2$ and $K = 6$ to 10, the switching SSM performed almost identically to the simple SSM. Exploratory experiments suggest that in these cases a single component takes responsibility for all the data, so the model has $M = 1$ effectively. This may be a local minimum problem or a result of poor initialization heuristics. Looking at the learning curves for simple and switching state space models it is easy to see that there are plateaus at the solutions found by the simple one-component SSMs which the switching SSM can get caught in (Figure 11).

The likelihoods for hidden Markov models with around $M = 15$ were comparable to those of the best switching state-space models (Figure 10f). So purely in terms of coding efficiency, switching SSMs have little advantage over HMMs on this data.

However, it is useful to contrast the solutions learned by HMMs with the solutions learned by the switching SSMs. The thick dots at the bottom of the Figures 9a and b show the responsibility assigned to one of two components in a fairly typical switching SSM with $M = 2$ components of state size $K = 2$. This component has clearly specialized to modeling the data during periods of apnea, while the other component models the gasps and periods of rhythmic breathing. These two switching components provide a much more intuitive model of the data than the 10-20 discrete components needed in an HMM with comparable coding efficiency.[10]

6 Discussion

The main conclusion we can draw from the first series of experiments is that even when given the correct model parameters, the problem of segmenting a switching time series into its components is difficult. There are combinatorially many alternatives to be considered, and the energy surface suffers from many local minima, so local optimization approaches like the variational method we used are limited by the quality of the initial conditions. Deterministic annealing can be thought of as a sophisticated initialization procedure for the hidden states: the final solution at each temperature provides the initial conditions at the next. We found that annealing substatially improved the quality of the segmentations found.

The first experiment also indicates that the much simpler Gaussian merging method performs comparably to annealed variational inference. The Gaussian merging methods have the advantage that at each time step the cost function minimized has no local minima. This may account for how well they perform relative to the non-annealed variational method. On the other hand, the variational methods have the advantage

[10] By using further assumptions to constrain the model, such as continuity of the real-valued hidden state at switch times, it should be possible to obtain even better performance on this data.

that they iteratively improve their approximation to the posterior, and they define a lower bound on the likelihood. Our results suggest that it may be very fruitful to use the Gaussian merging method to initialize the variational inference procedure. Furthermore, it is possible to derive variational approximations for other switching models described in the literature, and a combination of Gaussian merging and variational approximation may provide a fast and robust method for learning and inference in those models.

The second series of experiments suggests that on a real data set believed to have switching dynamics, the switching state-space model can indeed uncover multiple regimes. When it captures these regimes, it generalizes to the test set much better than the simple linear dynamical model. Similar coding efficiency can be obtained by using hidden Markov models, which due to the discrete nature of the state space can model nonlinear dynamics. However, in doing so, the hidden Markov models had to use 10-20 discrete states, which makes their solutions less interpretable.

Variational approximations provide a very powerful tool for inference and learning in complex probabilistic models. We have seen that when applied to the switching state-space model they can incorporate within a single framework well-known exact inference methods like Kalman smoothing and the forward-backward algorithm. Variational methods can be applied to many of the other classes of intractable switching models described in section 2.3. However, training more complex models also makes apparent the importance of good methods for model selection and initialization.

To summarize, switching state-space models are a dynamical generalization of mixture of experts neural networks, are closely related to well-known models in econometrics and control, and combine the representations underlying hidden Markov models and linear dynamical systems. For domains in which we have some a priori belief that there are multiple, approximately linear dynamical regimes, switching state space models provide a natural modeling tool. Variational approximations provide a method to overcome the single most difficult problem in learning switching SSMs, which is that the inference step is intractable. Deterministic annealing further improves on the solutions found by the variational method.

A Notation

Symbol	Size	Description
variables		
Y_t	$D \times 1$	observation vector at time t
$\{Y_t\}$	$D \times T$	sequence of observation vectors $[Y_1, Y_2, \ldots Y_T]$
$X_t^{(m)}$	$K \times 1$	state vector of state-space model (SSM) m at time t
X_t	$KM \times 1$	entire real-valued hidden state at time t: $X_t = [X_t^{(1)}, \ldots, X_t^{(M)}]$
S_t	$M \times 1$	switch state variable (represented either as discrete variable taking on values in $\{1, \ldots M\}$, or as an $M \times 1$ vector $S_t = [S_t^{(1)}, \ldots S_t^{(M)}]'$ where $S_t^{(m)} \in \{0, 1\}$)
model parameters		
$A^{(m)}$	$K \times K$	state dynamics matrix for SSM m
$C^{(m)}$	$D \times K$	output matrix for SSM m
$\mathcal{Q}^{(m)}$	$K \times K$	state noise covariance matrix for SSM m
$\mu_{X_1}^{(m)}$	$K \times 1$	initial state mean for SSM m
$\mathcal{Q}_1^{(m)}$	$K \times K$	initial state noise covariance matrix for SSM m
R	$D \times D$	output noise covariance matrix
$\boldsymbol{\pi}$	$M \times 1$	initial state probabilities for switch state
Φ	$M \times M$	state transition matrix for switch state
variational parameters		
$h_t^{(m)}$	1×1	responsibility of SSM m for Y_t
$q_t^{(m)}$	1×1	related to expected squared error if SSM m generated Y_t
miscellaneous		
X'		matrix transpose of X
$\|X\|$		matrix determinant of X
$\langle X \rangle$		expected value of X under the Q distribution
dimensions		
D		size of observation vector
T		length of a sequence of observation vectors
M		number of state-space models
K		size of state vector in each state-space model

B Derivation of the variational fixed-point equations

In this appendix we derive the variational fixed-point equations used in the learning algorithm for switching state space models. The plan is the following. First we write out the probability density P defined by a switching state space model. For convenience we will express this probability density in the log domain, through its associated energy function or *hamiltonian*, H. The probability density is related to the hamiltonian through the usual Boltzmann distribution (at a temperature of 1),

$$P(\cdot) = \frac{1}{Z} \exp\{-H(\cdot)\},$$

where Z is a normalization constant required such that $P(\cdot)$ integrates to unity. Expressing the probabilities in the log domain does not affect the resulting algorithm. We then similarly express the approximating distribution Q through its hamiltonian H_Q. Finally, we obtain the variational fixed point equations by setting to zero the derivatives of the KL divergence between Q and P with respect to the variational parameters of Q.

The joint probability of observations and hidden states in a switching state-space model is (equation (7))

$$P(\{S_t, X_t, Y_t\}) = \left[P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) \right] \prod_{m=1}^{M} \left[P(X_1^{(m)}) \prod_{t=2}^{T} P(X_t^{(m)}|X_{t-1}^{(m)}) \right] \prod_{t=1}^{T} P(Y_t|X_t, S_t). \tag{28}$$

We proceed to dissect this expression into its constituent parts. The initial probability of the switch variable at time $t = 1$ is given by

$$P(S_1) = \prod_{m=1}^{M} (\boldsymbol{\pi}^{(m)})^{S_1^{(m)}}, \tag{29}$$

where S_1 is represented by an $M \times 1$ vector $[S_1^{(1)} \ldots S_1^{(M)}]$ where $S_1^{(m)} = 1$ if the switch state is in state m, and 0 otherwise. The probability of transitioning from a switch state at time $t-1$ to a switch state at time t is given by

$$P(S_t|S_{t-1}) = \prod_{m=1}^{M} \prod_{n=1}^{M} (\Phi^{(m,n)})^{S_t^{(m)} S_{t-1}^{(n)}}. \tag{30}$$

The initial distribution for the hidden state variable in state-space model m is Gaussian with mean $\mu_{X_1}^{(m)}$ and covariance matrix $\mathcal{Q}_1^{(m)}$:

$$P(X_1^{(m)}) = (2\pi)^{-K/2} |\mathcal{Q}_1^{(m)}|^{-1/2} \exp\left\{ -\frac{1}{2} \left(X_1 - \mu_{X_1}^{(m)} \right)' (\mathcal{Q}_1^{(m)})^{-1} \left(X_1 - \mu_{X_1}^{(m)} \right) \right\}. \tag{31}$$

The probability distribution of the state in state-space model m at time t given the state at time $t-1$ is Gaussian with mean $A^{(m)} X_{t-1}^{(m)}$ and covariance matrix $\mathcal{Q}^{(m)}$:

$$P(X_t^{(m)}|X_{t-1}^{(m)}) = (2\pi)^{-K/2} |\mathcal{Q}^{(m)}|^{-1/2} \exp\left\{ -\frac{1}{2} \left(X_t^{(m)} - A^{(m)} X_{t-1}^{(m)} \right)' (\mathcal{Q}^{(m)})^{-1} \left(X_t^{(m)} - A^{(m)} X_{t-1}^{(m)} \right) \right\}. \tag{32}$$

Finally, using (8) we can write:

$$P(Y_t|X_t, S_t) = \prod_{m=1}^{M} \left[(2\pi)^{-D/2} |R|^{-1/2} \exp\left\{ -\frac{1}{2} \left(Y_t - C^{(m)} X_t^{(m)} \right)' R^{-1} \left(Y_t - C^{(m)} X_t^{(m)} \right) \right\} \right]^{S_t^{(m)}} \tag{33}$$

since the terms with exponent equal to 0 vanish in the product.

Combining (28)-(33) and taking the negative of the logarithm, we obtain the hamiltonian of a switching state-space model (ignoring constants):

$$H = \frac{1}{2} \sum_{m=1}^{M} \log |\mathcal{Q}_1^{(m)}| + \frac{1}{2} \sum_{m=1}^{M} \left(X_1^{(m)} - \mu_{X_1}^{(m)} \right)' (\mathcal{Q}_1^{(m)})^{-1} \left(X_1^{(m)} - \mu_{X_1}^{(m)} \right)$$

$$
\begin{aligned}
&+\frac{(T-1)}{2}\sum_{m=1}^{M}\log|\mathcal{Q}^{(m)}| + \frac{1}{2}\sum_{m=1}^{M}\sum_{t=2}^{T}\left(X_t^{(m)} - A^{(m)}X_{t-1}^{(m)}\right)'(\mathcal{Q}^{(m)})^{-1}\left(X_t^{(m)} - A^{(m)}X_{t-1}^{(m)}\right) \\
&+\frac{T}{2}\log|R| + \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T} S_t^{(m)}\left(Y_t - C^{(m)}X_t^{(m)}\right)' R^{-1}\left(Y_t - C^{(m)}X_t^{(m)}\right) \\
&-\sum_{m=1}^{M} S_1^{(m)}\log\boldsymbol{\pi}^{(m)} - \sum_{t=2}^{T}\sum_{m=1}^{M}\sum_{n=1}^{M} S_t^{(m)}S_{t-1}^{(n)}\log\Phi^{(m,n)}. \qquad (34)
\end{aligned}
$$

The hamiltonian for the approximating distribution can be analogously derived from the definition of Q (equation (14)):

$$
Q(\{S_t, X_t\}) = \frac{1}{Z_Q}\left[\psi(S_1)\prod_{t=2}^{T}\psi(S_{t-1}, S_t)\right]\prod_{m=1}^{M}\psi(X_1^{(m)})\prod_{t=2}^{T}\psi(X_{t-1}^{(m)}, X_t^{(m)}). \qquad (35)
$$

The potentials for the initial switch state and switch state transitions are

$$
\begin{aligned}
\psi(S_1) &= \prod_{m=1}^{M}(\boldsymbol{\pi}^{(m)}q_1^{(m)})^{S_1^{(m)}} \qquad (36) \\
\psi(S_{t-1}, S_t) &= \prod_{m=1}^{M}\prod_{n=1}^{M}\left(\Phi^{(m,n)}q_t^{(m)}\right)^{S_t^{(m)}S_{t-1}^{(n)}} \qquad (37)
\end{aligned}
$$

The potential for the initial state of state-space model m is

$$
\psi(X_1^{(m)}) = P(X_1^{(m)})\ \left[P(Y_1|X_1^{(m)}, S_1 = m)\right]^{h_1^{(m)}} \qquad (38)
$$

and the potential for the state at time t given the state at time $t-1$ is

$$
\psi(X_{t-1}^{(m)}, X_t^{(m)}) = P(X_t^{(m)}|X_{t-1}^{(m)})\ \left[P(Y_t|X_t^{(m)}, S_t = m)\right]^{h_t^{(m)}}. \qquad (39)
$$

The hamiltonian for Q is obtained by combining these terms and taking the negative logarithm:

$$
\begin{aligned}
H_Q \quad = \quad & \frac{1}{2}\sum_{m=1}^{M}\log|\mathcal{Q}_1^{(m)}| + \frac{1}{2}\sum_{m=1}^{M}\left(X_1^{(m)} - \mu_{X_1}^{(m)}\right)'(\mathcal{Q}_1^{(m)})^{-1}\left(X_1^{(m)} - \mu_{X_1}^{(m)}\right) \\
&+\frac{(T-1)}{2}\sum_{m=1}^{M}\log|\mathcal{Q}^{(m)}| + \frac{1}{2}\sum_{m=1}^{M}\sum_{t=2}^{T}\left(X_t^{(m)} - A^{(m)}X_{t-1}^{(m)}\right)'(\mathcal{Q}^{(m)})^{-1}\left(X_t^{(m)} - A^{(m)}X_{t-1}^{(m)}\right) \\
&+\frac{T}{2}\sum_{m=1}^{M}\log|R| + \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T} h_t^{(m)}\left(Y_t - C^{(m)}X_t^{(m)}\right)' R^{-1}\left(Y_t - C^{(m)}X_t^{(m)}\right) \\
&-\sum_{m=1}^{M} S_1^{(m)}\log\boldsymbol{\pi}^{(m)} - \sum_{t=2}^{T}\sum_{m=1}^{M}\sum_{n=1}^{M} S_t^{(m)}S_{t-1}^{(n)}\log\Phi^{(m,n)} - \sum_{t=1}^{T}\sum_{m=1}^{M} S_t^{(m)}\log q_t^{(m)}. \qquad (40)
\end{aligned}
$$

Comparing H_Q with H we see that the interaction between the $S_t^{(m)}$ and the $X_t^{(m)}$ variables has been eliminated, while introducing two sets of variational parameters: the responsibilities $h_t^{(m)}$ and the bias terms on the discrete Markov chain, $q_t^{(m)}$. In order to obtain the approximation Q which maximizes the lower bound on the log likelihood, we minimize the KL divergence $\mathrm{KL}(Q\|P)$ as a function of these variational parameters

$$
\begin{aligned}
\mathrm{KL}(Q\|P) &= \sum_{\{S_t\}}\int Q(\{S_t, X_t\})\log\frac{Q(\{S_t, X_t\})}{P(\{S_t, X_t\}|\{Y_t\})}d\{X_t\} \qquad (41) \\
&= \langle H - H_Q\rangle - \log Z_Q + \log Z, \qquad (42)
\end{aligned}
$$

where $\langle \cdot \rangle$ denotes expectation over the approximating distribution Q and Z_Q is the normalization constant for Q. Both Q and P define distributions in the exponential family. As a consequence, the zeros of the derivatives of KL with respect to the variational parameters can be obtained simply by equating derivatives of $\langle H \rangle$ and $\langle H_Q \rangle$ with respect to corresponding sufficient statistics (Ghahramani, 1997):

$$\frac{\partial \langle H_Q - H \rangle}{\partial \langle S_t^{(m)} \rangle} = 0 \tag{43}$$

$$\frac{\partial \langle H_Q - H \rangle}{\partial \langle X_t^{(m)} \rangle} = 0 \tag{44}$$

$$\frac{\partial \langle H_Q - H \rangle}{\partial \langle P_t^{(m)} \rangle} = 0 \tag{45}$$

where $P_t^{(m)} = \langle X_t^{(m)} X_t^{(m)\prime} \rangle - \langle X_t^{(m)} \rangle \langle X_t^{(m)} \rangle'$ is the covariance of $X_t^{(m)}$ under Q. Many terms cancel when we subtract the two hamiltonians

$$H_Q - H = \sum_{m=1}^{M} \sum_{t=1}^{T} \frac{1}{2} \left(h_t^{(m)} - S_t^{(m)} \right) \left(Y_t - C^{(m)} X_t^{(m)} \right)' R^{-1} \left(Y_t - C^{(m)} X_t^{(m)} \right) - S_t^{(m)} \log q_t^{(m)} \tag{46}$$

Taking derivatives we obtain

$$\frac{\partial \langle H_Q - H \rangle}{\partial \langle S_t^{(m)} \rangle} = -\log q_t^{(m)} - \frac{1}{2} \left\langle \left(Y_t - C^{(m)} X_t^{(m)} \right)' R^{-1} \left(Y_t - C^{(m)} X_t^{(m)} \right) \right\rangle \tag{47}$$

$$\frac{\partial \langle H_Q - H \rangle}{\partial \langle X_t^{(m)} \rangle} = -\left(h_t^{(m)} - \langle S_t^{(m)} \rangle \right) \left((Y_t - C^{(m)} \langle X_t^{(m)} \rangle)' R^{-1} C^{(m)} \right) \tag{48}$$

$$\frac{\partial \langle H_Q - H \rangle}{\partial P_t^{(m)}} = \frac{1}{2} \left(h_t^{(m)} - \langle S_t^{(m)} \rangle \right) \left(C^{(m)\prime} R^{-1} C^{(m)} \right) \tag{49}$$

From (47) we get the fixed-point equation (20) for $q_t^{(m)}$. Both (48) and (49) are satisfied when $h_t^{(m)} = \langle S_t^{(m)} \rangle$. Using the fact that $\langle S_t^{(m)} \rangle = Q(S_t = m)$ we get (19).

References

Ackerson, G. A. and Fu, K. S. (1970). On state estimation in switching environments. *IEEE Transactions on Automatic Control*, AC-15(1):10–17.

Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering.* Prentice-Hall, Englewood Cliffs, NJ.

Athaide, C. R. (1995). *Likelihood Evaluation and State Estimation for Nonlinear State Space Models.* Ph.D. Thesis, Graduate Group in Managerial Science and Applied Economics, University of Pennsylvania, Philadelphia, PA.

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. (1994). Hidden Markov models of biological primary sequence information. *Proc. Nat. Acad. Sci. (USA)*, 91(3):1059–1063.

Bar-Shalom, Y. and Li, X.-R. (1993). *Estimation and Tracking.* Artech House, Boston, MA.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41:164–171.

Bengio, Y. and Frasconi, P. (1995). An input–output HMM architecture. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, *Advances in Neural Information Processing Systems 7*, pages 427–434. MIT Press, Cambridge, MA.

Cacciatore, T. W. and Nowlan, S. J. (1994). Mixtures of controllers for jump linear and non-linear plants. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems 6*, pages 719–726. Morgan Kaufmann Publishers, San Francisco, CA.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81:541–553.

Chaer, W. S., Bishop, R. H., and Ghosh, J. (1997). A mixture-of-experts framework for adaptive kalman filtering. *IEEE Trans. on Systems, Man and Cybernetics.*

Chang, C. B. and Athans, M. (1978). State estimation for discrete systems with switching parameters. *IEEE Transactions on Aerospace and Electronic Systems*, AES-14(3):418–424.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* Wiley, New York.

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persitence and causation. *Computational Intelligence*, 5(3):142–150.

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

Deng, L. (1993). A stochastic model of speech incorporating hierarchical nonstationarity. *IEEE Trans. on Speech and Audio Processing*, 1(4):471–474.

Digalakis, V., Rohlicek, J. R., and Ostendorf, M. (1993). ML estimation of a Stochastic Linear System with the EM Algorithm and its Application to Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 1(4):431–442.

Elliott, R. J., Aggoun, L., and Moore, J. B. (1995). *Hidden Markov Models: Estimation and Control.* Springer-Verlag, New York.

Fraser, A. M. and Dimitriadis, A. (1993). Forecasting probability densities by using hidden Markov models with mixted states. In Wiegand, A. S. and Gershenfeld, N. A., editors, *Time series prediction: Forecasting the future and understanding the past*, SFI Studies in the Sciences of Complexity, Proc. Vol. XV, pages 265–282. Addison-Wesley, Reading, MA.

Ghahramani, Z. (1997). On structured variational approximations. Technical Report CRG-TR-97-1 [http://www.gatsby.ucl.ac.uk/∼zoubin/papers/struct.ps.gz], Department of Computer Science, University of Toronto.

Ghahramani, Z. and Hinton, G. E. (1996a). Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2 [http://www.gatsby.ucl.ac.uk/∼zoubin/papers/tr-96-2.ps.gz], Department of Computer Science, University of Toronto.

Ghahramani, Z. and Hinton, G. E. (1996b). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1 [http://www.gatsby.ucl.ac.uk/∼zoubin/papers/tr-96-1.ps.gz], Department of Computer Science, University of Toronto.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, 29:245–273.

Goodwin, G. and Sin, K. (1984). *Adaptive filtering prediction and control.* Prentice-Hall.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton University Press, Princeton, NJ.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting (with discussion). *J. Royal Statistical Society B*, 38:205–247.

Hinton, G. E., Dayan, P., and Revow, M. (1996). Modeling the manifolds of Images of handwritten digits. *Submitted for Publication.*

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixture of local experts. *Neural Computation*, 3:79–87.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1998). An Introduction to variational methods in graphical models. In Jordan, M. I., editor, *Learning in Graphical Models*. Kluwer Academic Publishers.

Jordan, M. I. and Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214.

Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33:251–272.

Kadirkamanathan, V. and Kadirkamanathan, M. (1996). Recursive estimation of dynamic modular RBF networks. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems 8*, pages 239–245. MIT Press.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction. *Journal of Basic Engineering (ASME)*, 83D:95–108.

Kanazawa, K., Koller, D., and Russell, S. J. (1995). Stochastic simulation algorithms for dynamic probabilistic networks. In Besnard, P. and Hanks, S., editors, *Uncertainty in Artificial Intelligence. Proceedings of the Eleventh Conference.*, pages 346–351. Morgan Kaufmann Publishers, San Francisco, CA.

Kehagias, A. and Petrides, V. (1997). Time series segmentation using predictive modular neural networks. *Neural Computation*, 9(8):1691–1710.

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

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistical Society B*, pages 157–224.

Ljung, L. and Söderström, T. (1983). *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, MA.

Meila, M. and Jordan, M. I. (1996). Learning fine motion by Markov mixtures of experts. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8*. MIT Press.

Neal, R. M. and Hinton, G. E. (1998). A new view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, *Learning in Graphical Models*. Kluwer Academic Press.

Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley, Redwood City, CA.

Pawelzik, K., Kohlmorgen, J., and Müller, K.-R. (1996). Annealed competition of experts for a segmentation and classification of switching dynamics. *Neural Computation*, 8(2):340–356.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Rabiner, L. R. and Juang, B. H. (1986). An Introduction to hidden Markov models. *IEEE Acoustics, Speech & Signal Processing Magazine*, 3:4–16.

Rauch, H. E. (1963). Solutions to the linear smoothing problem. *IEEE Transactions on Automatic Control*, 8:371–372.

Rigney, D., Goldberger, A., Ocasio, W., Ichimaru, Y., Moody, G., and Mark, R. (1993). Multi-channel physiological data: Description and analysis. In Weigend, A. and Gershenfeld, N., editors, *Time Series Prediction: Forecasting the future and understanding the past*, SFI Studies in the Sciences of Complexity, Proc. Vol. XV, pages 105–129. Addison-Wesley, Reading, MA.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345.

Saul, L. and Jordan, M. I. (1996). Exploiting tractable substructures in Intractable networks. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems 8*. MIT Press.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Analysis*, 3(4):253–264.

Shumway, R. H. and Stoffer, D. S. (1991). Dynamic linear models with switching. *J. Amer. Stat. Assoc.*, 86:763–769.

Smyth, P. (1994). Hidden Markov models for fault detection in dynamic systems. *Pattern Recognition*, 27(1):149–164.

Smyth, P., Heckerman, D., and Jordan, M. I. (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9:227–269.

Ueda, N. and Nakano, R. (1995). Deterministic annealing variant of the EM algorithm. In Tesauro, G., Touretzky, D., and Alspector, J., editors, *Advances in Neural Information Processing Systems 7*, pages 545–552. Morgan Kaufmann.

Weigend, A. and Gershenfeld, N. (1993). *Time Series Prediction: Forecasting the future and understanding the past.* SFI Studies in the Sciences of Complexity, Proc. Vol. XV. Addison-Wesley, Reading, MA.