A Dynamic Bayesian Network Approach to Tracking Using Learned Switching Dynamic Models

Vladimir Pavlović[1], James M. Rehg, and Tat-Jen Cham

Compaq Computer Corporation
Cambridge Research Lab
Cambridge, MA 02139
{vladimir,rehg,tjc}@crl.dec.com

Abstract. Switching linear dynamic systems (SLDS) attempt to describe a complex nonlinear dynamic system with a succession of linear models indexed by a switching variable. Unfortunately, despite SLDS's simplicity exact state and parameter estimation are still intractable. Recently, a broad class of learning and inference algorithms for time-series models have been successfully cast in the framework of dynamic Bayesian networks (DBNs). This paper describes a novel DBN-based SLDS model. A key feature of our approach are two approximate inference techniques for overcoming the intractability of exact inference in SLDS. As an example, we apply our model to the human figure motion analysis. We present experimental results for learning figure dynamics from video data and show promising results for tracking, interpolation, synthesis, and classification using learned models.

1 Introduction

Many natural processes have complex, highly nonlinear and time-varying dynamics. For instance, economic trends, maneuvering targets, and the human figure all exhibit complex and rich dynamic behavior. Dynamics are essential to the analysis of these processes as well as to their realistic prediction (forecasting) and synthesis (simulation). Dynamic models can provide a powerful cue in the presence of missing/multiple measurements and measurement noise. A dynamic model imposes additional structure on the state space by specifying which state trajectories are possible (or probable) and by specifying the speed at which a trajectory evolves.

Unfortunately, state and parameter estimation problems in complex dynamic models can be a daunting task. State estimation in non-linear models is usually cast in frameworks whose origins lay in the theory of extended Kalman filters (c.f. [1]). Parameter estimation of such highly nonlinear models is often a result of tedious measurements and expert knowledge about the problem. For instance, consider the human figure modeling in the field of biomechanics. The dynamics of the figure are the result of its mass distribution, joint torques produced by the motor control system, and reaction forces

[1] Contact author: Vladimir Pavlović, Compaq Computer Corp., Cambridge Research Lab, 1 Kendall Sq., Cambridge, MA 02139, phone (617) 551-7699, fax (617) 551-7650, e-mail vladimir@crl.dec.com

resulting from contact with the environment (e.g. the floor). Research efforts in biomechanics, rehabilitation, and sports medicine have resulted in complex, specialized models of human motion (c.f. [11].) Such complex models have been used successfully to simulate [10] and to track human body motion [27].

This paper explores the alternative method of learning dynamic models from a training corpus of observed state space trajectories. In cases where sufficient training data is available, the learning approach promises flexibility and generality. A wide range of learning algorithms can be cast in the framework of dynamic Bayesian networks (DBNs) [7], a subclass of now famous Bayesian network models (c.f. [23, 13]). DBNs generalize two well-known signal modeling tools: Kalman filters [1] for continuous state linear dynamic systems (LDS) and Hidden Markov Models (HMMs) [24] for classification of discrete state sequences.

The DBN framework provides two distinct benefits: First, a broad variety of modeling schemes can be conceptualized in a single framework with an intuitively-appealing graphical notation (see Figure 1 for an example). Second, a broad corpus of exact and approximate statistical inference and learning techniques from the Bayesian network literature can be applied to dynamical systems. In particular, it has been shown that estimation in LDSs and inference in HMMs are special cases of inference in DBNs.

The focus of this paper is on a subclass of DBN models called Switching Linear Dynamic Systems [2, 26, 17, 9, 22]. Intuitively, these models attempt to describe a complex nonlinear dynamic system with a succession of linear models that are indexed by a switching variable. While other approaches such as learning weighted combinations of linear models are possible, the switching approach has an appealing simplicity and is naturally suited to the case where the dynamics are time-varying.

This paper makes two contributions. First, we derive two efficient algorithms for approximate state estimation in SLDSs. An approximate Viterbi inference algorithm and a structured variational inference algorithm are cast as in the framework of DBN inference. Second, we demonstrate the application of the SLDS framework to modeling the human figure dynamics. In particular, we demonstrate the learning of switching models of fronto-parallel walking and jogging motion from video data. We demonstrate the application of these learned models to segmentation, synthesis, and tracking tasks.

2 Switching Linear Dynamic System Model

Consider an SLDS described using the following set of continuous and discrete state-space equations:

$$x_{t+1} = A(s_{t+1})x_t + v_{t+1}(s_{t+1}), \quad y_t = Cx_t + w_t$$

for the continuous-valued linear dynamic system (LDS), and

$$Pr(s_{t+1}|s_t) = s'_{t+1} \Pi s_t$$

for the discrete switching model. We assumed that the LDS models a Gauss-Markov process with state noise $v_t(s_t) \sim \mathcal{N}(0, Q(s_t))$, measurement noise $w_t \sim \mathcal{N}(0, R)$, and initial state $x_0 \sim \mathcal{N}(x_0(s_0), Q_0(s_0))$. The switching model is assumed to be a discrete

first order Markov process. State variables of this model are written as s_t. They belong to the set of S discrete symbols $\{e_0, \ldots, e_{S-1}\}$, where e_i is the unit vector of dimension S with a non-zero element in the i-th position. The switching model is defined with the state transition matrix Π whose elements are $\Pi(i,j) = Pr(s_{t+1} = e_i | s_t = e_j)$, and an initial state distribution π_0.

Coupling between the LDS and the switching process stems from the dependency of the LDS parameters A and Q on the switching process state s_t. Namely,

$$A(s_t = e_i) = A_i, \;\; Q(s_t = e_i) = Q_i$$

In other words, switching state s_t determines which of S possible plant models is used at time t.

The complex state space representation is equivalently depicted by the DBN dependency graph in Figure 1. The dependency graph implies that the *joint distribution* P

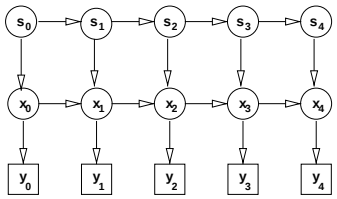


Fig. 1. Bayesian network representation (dependency graph) of an SLDS of duration five. s denote instances of the discrete valued switching states. x and y are continuous valued LDS states and measurements. Arcs in the graph show dependencies among variables.

over the variables of the SLDS can be written as

$$P(\mathcal{Y}_T, \mathcal{X}_T, \mathcal{S}_T) = \\ Pr(s_0) \prod_{t=1}^{T-1} Pr(s_t|s_{t-1}) Pr(x_0|s_0) \prod_{t=1}^{T-1} Pr(x_t|x_{t-1}, s_t) \prod_{t=0}^{T-1} Pr(y_t|x_t),$$

where $\mathcal{Y}_T, \mathcal{X}_T$, and $\mathcal{S}_T$ denote the sequences (of length T) of observations and hidden state variables. For instance, $\mathcal{Y}_T = \{y_0, \ldots, y_{T-1}\}$. From the Gauss-Markov assumption on the LDS (e.g. $x_{t+1}|_{x_t, s_{t+1}=e_i} \sim \mathcal{N}(A_i x_t, Q_i)$) and recalling the Markov switching model assumption, the joint pdf of the SLDS of duration T can be easily defined.

2.1 Hidden State Inference and Estimation

The goal of inference in complex DBNs is to estimate the posterior probability of the hidden states of the system (s_t and x_t) given some known sequence of observations $\mathcal{Y}_T$ and the known model parameters. Namely, we need to find the posterior

$$P(\mathcal{X}_T, \mathcal{S}_T | \mathcal{Y}_T) = Pr(\mathcal{X}_T, \mathcal{S}_T | \mathcal{Y}_T),$$

or , equivalently, its *sufficient statistics*. Given the form of P it is easy to show that these are the first and the second order statistics: mean and covariance among hidden states $x_t, x_{t-1}, s_t, s_{t-1}$.

If there were no switching dynamics, the inference would be straightforward – we could infer $\mathcal{X}_T$ from $\mathcal{Y}_T$ using LDS inference (RTS smoothing [25]). However, the presence of switching dynamics embedded in matrix Π makes exact inference more complicated. To see that, assume that the initial distribution of x_0 at $t = 0$ is Gaussian, at $t = 1$ the pdf of the physical system state x_1 becomes a mixture of S Gaussian pdfs since we need to marginalize over S possible but unknown plant models. At time t we will have a mixture of S^t Gaussians, which is clearly intractable for even moderate sequence lengths. It is therefore necessary to explore approximate inference techniques that will result in a tractable learning method.

2.2 Approximate Inference Using Viterbi Approximation

The task of Viterbi approximation approach is to find the most likely sequence of switching states s_t for a given observation sequence $\mathcal{Y}_T$. If the best sequence of switching states is denoted $\mathcal{S}_T^*$ we can then approximate the desired posterior $P(\mathcal{X}_T, \mathcal{S}_T|\mathcal{Y}_T)$ as[1]

$$P(\mathcal{X}_T, \mathcal{S}_T|\mathcal{Y}_T) = P(\mathcal{X}_T|\mathcal{S}_T, \mathcal{Y}_T)P(\mathcal{S}_T|\mathcal{Y}_T) \approx P(\mathcal{X}_T|\mathcal{S}_T, \mathcal{Y}_T)\ \delta(\mathcal{S}_T - \mathcal{S}_T^*),$$

i.e. the switching sequence posterior $P(\mathcal{S}_T|\mathcal{Y}_T)$ was approximated by its mode. It is well known how to apply Viterbi inference to discrete state hidden Markov models [24] and continuous state Gauss-Markov models [16]. Here we develop an algorithm for approximate Viterbi inference in SLDSs.

More formally, we are looking for the switching sequence $\mathcal{S}_T^*$ such that

$$\mathcal{S}_T^* = \arg\max_{\mathcal{S}_T} P(\mathcal{S}_T|\mathcal{Y}_T).$$

It is easily to shown that a (suboptimal) solution to this problem can be obtain by recursive optimization of the probability of the best sequence at time t

$$\begin{aligned} J_{t,i} &= \max_{\mathcal{S}_{t-1}} P\left(\mathcal{S}_{t-1}, s_t = e_i, \mathcal{Y}_t\right) \\ &\approx \max_j \Big\{ P\left(y_t|s_t = e_i, s_{t-1} = e_j, S_{t-2}^*(j), \mathcal{Y}_{t-1}\right) P\left(s_t = e_i|s_{t-1} = e_j\right) \\ &\quad \max_{\mathcal{S}_{t-2}} P\left(\mathcal{S}_{t-2}, s_{t-1} = e_j, \mathcal{Y}_{t-1}\right) \Big\} \end{aligned} \tag{1}$$

Here $S_{t-2}^*(i)$ is the "best" switching sequence up to time $t-1$ when SLDS is in state i at time $t-1$, $S_{t-2}^*(i) = \arg\max_{S_{t-2}} J_{t-1,i}$.

To find the likelihood term $P(y_t)$ in Equation 1 note that concurrently with the recursion for each pair of consecutive switching state i, j at times $t, t-1$ one can

[1] $\delta(x) = 1$ for $x = \emptyset$ and zero otherwise.

update the statistics [2] of continuous states of LDS based on current "best" switching sequences using the Kalman filter inference (c.f. [1]). For instance,

$$\begin{aligned}
\hat{x}_{t|t,i} &\stackrel{\Delta}{=} \left\langle x_t | \mathcal{Y}_t, s_t = e_i, S^*_{t-1}(i) \right\rangle \\
\hat{x}_{t|t-1,i,j} &\stackrel{\Delta}{=} \left\langle x_t | \mathcal{Y}_{t-1}, s_t = e_i, s_{t-1} = e_j, S^*_{t-2}(j) \right\rangle \\
\hat{x}_{t|t,i,j} &\stackrel{\Delta}{=} \left\langle x_t | \mathcal{Y}_t, s_t = e_i, s_{t-1} = e_j, S^*_{t-2}(j) \right\rangle .
\end{aligned}$$

$\Sigma_{t|t,i}\Sigma_{t|t-1,i,j}\Sigma_{t|t,i,j}$ denote corresponding second order statistics. The likelihood term can then be easily computed as the probability of innovation $y_t - C\hat{x}_{t|t-1,i,j}$ of $j \to i$ transition, which has normal distribution with mean $C\hat{x}_{t|t-1,i,j}$ and variance $C\Sigma_{t|t-1,i,j}C' + R$.

The index of the most likely state j^* at $t-1$ corresponding to the maximum in Equation 1 is kept for every state i at time t in the state transition record $\psi_{t-1,i} = j*$. LDS statistics corresponding to $j*$ are updated accordingly, $\hat{x}_{t|t,i} = \hat{x}_{t|t,i,\psi_{t-1,i}}\Sigma_{t|t,i} = \Sigma_{t|t,i,\psi_{t-1,i}}$. Once all T observations have been fused to the "best" switching state sequence is the one that ends in $i^*_{T-1} = \arg\min_i J_{T-1,i}$. States of this sequence can be traced back through the state transition record $\psi_{t-1,i}$, $i^*_t = \psi_{t,i^*_{t+1}}$.

Once the "best" switching sequence is known, the switching model's sufficient statistics are simply $\langle s_t \rangle = e_{i^*_t}$ and $\langle s_t s'_{t-1} \rangle = e_{i^*_t} e'_{i^*_{t-1}}$. Sufficient LDS statistics for this switching sequence can be easily obtained using Rauch-Tung-Streiber (RTS) fixed interval smoothing [1].

The Viterbi inference algorithm for SLDSs can now be summarized as

Find most likely state sequence $\mathcal{S}^*_T$ using recursion in 1; Find DBN's sufficient statistics for $P(\mathcal{S}_T \mid \mathcal{Y}_T) = \delta(\mathcal{S}_T - \mathcal{S}^*_T)$;

Approximate Inference Using Structured Variational Inference General structured variational inference technique for Bayesian networks is described in [14]. The idea behind this method is to find distribution Q which is in some sense close to the desired conditional distribution P, but is easier to compute. One can then employ Q as an approximation of P, $P(\mathcal{X}_T, \mathcal{S}_T | \mathcal{Y}_T) \approx Q(\mathcal{X}_T, \mathcal{S}_T | \mathcal{Y}_T)$.

Namely, for a given set of observations $\mathcal{Y}_T$, a distribution $Q(\mathcal{X}_T, \mathcal{S}_T | \eta, \mathcal{Y}_T)$ with an additional set of *variational parameters* η is defined such that Kullback–Leibler divergence between $Q(\mathcal{X}_T, \mathcal{S}_T | \eta, \mathcal{Y}_T)$ and $P(\mathcal{X}_T, \mathcal{S}_T | \mathcal{Y}_T)$ is minimized with respect to η:

$$\eta^* = \arg\min_\eta \sum_{\mathcal{S}_T} \int_{\mathcal{X}_T} Q(\mathcal{X}_T, \mathcal{S}_T | \eta, \mathcal{Y}_T) \log \frac{P(\mathcal{X}_T, \mathcal{S}_T | \mathcal{Y}_T)}{Q(\mathcal{X}_T, \mathcal{S}_T | \eta, \mathcal{Y}_T)}.$$

The dependency structure of Q is chosen such that it closely resembles the dependency structure of the original distribution P. However, unlike P the dependency structure

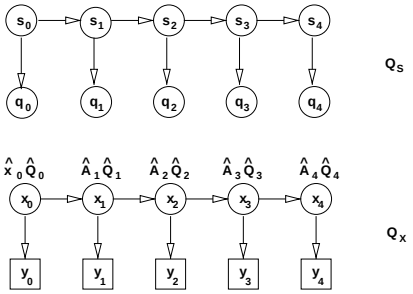


Fig. 2. Factorization of the original SLDS. Factorization reduces the fully coupled model into a seemingly decoupled pair of a HMM (Q_s) and a LDS (Q_x).

of Q *must* allow a computationally efficient inference. In our case we define Q by decoupling the switching and LDS portions of SLDS as shown in Figure 2. The two subgraphs of the original network are a hidden Markov model (HMM) Q_S with variational parameters $\{q_0, \ldots, q_{T-1}\}$ and a time-varying LDS Q_X with variational parameters $\{\hat{x}_0, \hat{A}_0, \ldots, \hat{A}_{T-1}, \hat{Q}_0, \ldots, \hat{Q}_{T-1}\}$. Because the subgraphs are *decoupled*, inference can be performed for each submodel separately, $Q(\mathcal{X}_T, \mathcal{S}_T|\eta, \mathcal{Y}_T) = Q_X(\mathcal{X}_T|\eta, \mathcal{Y}_T)$ $Q_S(\mathcal{S}_T|\eta)$. This is also reflected in the sufficient statistics of the posterior defined by the approximating network, e.g. $\langle x_t x_t{}' s_t \rangle = \langle x_t x_t{}' \rangle \langle s_t \rangle$.

The optimal values of the variational parameters η are obtained by minimizing the KL-divergence w.r.t. η. This leads to, for instance, the following recursive expression for the LDS Q_X's state transition matrix

$$\hat{A}_t = \hat{Q}_t \sum_{i=0}^{S-1} Q_i^{-1} A_i \langle s_t(i) \rangle . \tag{2}$$

Similarly, an expression can be found for optimal HMM variational parameters

$$\log q_t(i) = -\frac{1}{2} \left\langle (x_t - A_i x_{t-1})' \hat{Q}_i^{-1} (x_t - A_i x_{t-1}) \right\rangle - \frac{1}{2} \log |\hat{Q}_{t,i}|. \tag{3}$$

To obtain the expectation terms $\langle s_t \rangle = Pr(s_t|q_0, \ldots, q_{T-1})$ we use the inference in the HMM with output "probabilities" q_t [24]. Similarly, to obtain $\langle x_t \rangle = E[x_t|\mathcal{Y}_T]$ we perform LDS inference in the decoupled time-varying LDS via RTS smoothing. Since $\hat{A}_t, \hat{Q}_t$ in the decoupled LDS Q_X depends on $\langle s_t \rangle$ from the decoupled HMM Q_S and q_t depends on $\langle x_t \rangle, \langle x_t x_t{}' \rangle, \langle x_t x_{t-1}{}' \rangle$ from the decoupled LDS, the optimal parameter equations (e.g. 2 and 3) together with the inference solutions in the decoupled models form a set of fixed-point equations. Solution of this fixed-point set yields a tractable approximation to the intractable inference of the original fully coupled SLDS.

The variational inference algorithm for fully coupled SLDSs can now be summarized as:

[2] LDS statistics are state means and covariances. $\langle . \rangle$ denotes the expectation operator.

error = ∞;
Initialize $\langle s_t \rangle$;
while (error $>$ maxError) {
    Find $\hat{Q}_t, \hat{A}_t, \hat{x}_0$ from $\langle s_t \rangle$;
    Estimate $\langle x_t \rangle, \langle x_t {x_t}' \rangle$ and $\langle x_t {x_{t-1}}' \rangle$ from y_t
      using time-varying LDS inference;
    Find q_t from $\langle x_t \rangle, \langle x_t {x_t}' \rangle$ and $\langle x_t {x_{t-1}}' \rangle$;
    Estimate $\langle s_t \rangle$ from q_t using HMM inference.
    Update approximation error (KL divergence);
}

2.3 Maximum Likelihood Learning of Complex DBNs

Learning in complex DBNs can be formulated as the problem of ML learning in general Bayesian networks. Hence, a generalized EM algorithm [21] can be used to find optimal values of DBN parameters $\{A, C, Q, R, \Pi, \pi_0\}$. The expectation (E) step of EM is the inference itself. We outlined two approximate inference algorithms in the previous section. Note that the variational inference algorithm does not necessarily lead to non-decreasing likelihood of data in the EM (even though it usually does so in practice.) On the other hand, (a bound on) likelihood is guaranteed not to decrease when structural variational inference is used. See [14] for more details.

Given the sufficient statistics from the inference phase, *parameter update equations* in maximization (M) step are obtained by maximizing $\langle \log P(\mathcal{X}_T, \mathcal{S}_T, \mathcal{Y}_T) \rangle$ with respect to the parameter of interest. For instance, updated values of the state transition parameters are easily shown to be

$$\hat{A}_i = \left(\sum_{t=1}^{T-1} \left\langle x_t x'_{t-1} s_t(i) \right\rangle \right) \left(\sum_{t=1}^{T-1} \left\langle x_{t-1} x'_{t-1} s_t(i) \right\rangle \right)^{-1}$$

$$\hat{\Pi} = \left(\sum_{t=1}^{T-1} \left\langle s_t s'_{t-1} \right\rangle \right) \mathrm{diag} \left(\sum_{t=1}^{T-1} \langle s_t \rangle \right)^{-1}.$$

All the variable statistics are evaluated before updating any parameters. Notice that the above equations represent a generalization of the parameter update equations of classical (non-switching) LDS models [8].

3 Previous Work

SLDS models and their equivalents have been studied in statistics, time-series modeling, and target tracking since early 1970's. Bar-Shalom [2] and Kim [17] have developed a number of approximate pseudo-Bayesian inference techniques based on mixture component truncation or collapsing is SLDSs. They did not address the issue of learning system parameters. Shumway and Stoffer [26] presented a systematic view of inference

and learning in SLDS while assuming known prior switching state distributions at each time instance, $Pr(s_t) = \pi_t(i)$ and no temporal dependency between switching states. Krishnamurthy and Evans [18] imposed Markov dynamics on the switching model. However, they assumed that noisy measurements of the switching states are available.

Ghahramani [9] introduced a DBN-framework for learning and approximate inference in one class of SLDS models. His underlying model differs from ours in assuming the presence of S independent, white noise-driven LDSs whose measurements are selected by the Markov switching process. An alternative input-switching LDS model was proposed by Pavlovic et al. [22] and utilized for mouse motion classification. A switching model framework for particle filters is described in [12] and applied to dynamics learning in [3]. Manifold learning [5] is another approach to constraining the set of allowable trajectories within a high dimensional state space. An HMM-based approach is described in [4].

4 Experiments

We applied our DBN-based SLDS framework to the modeling of motion of the human figure. Most current models of the human figure dynamics belong to one of two model groups. One assumes highly complex, hand-crafted biomechanical models. This approach has been used successfully to produce computer graphics animations of human motion [10] and to track upper body motion in a user-interface setting [27]. On the other end of the spectrum are simple LDS models. Most previous figure trackers which have used a dynamic model employed a simple smoothness prior such as a constant velocity Kalman filter [15].

Two categories of fronto-parallel motion were present in our data: walking and jogging. Fronto-parallel motions exhibit interesting dynamics and are free from the difficulties of 3-D reconstruction. Experiments can be conducted easily using a single video source, while self-occlusions and cluttered backgrounds make the tracking problem non-trivial.

We adopted the 2-D Scaled Prismatic Model proposed by Morris and Rehg [19] to describe the kinematics of the figure. The kinematic model lies in the image plane, with each link having one degree of freedom (DOF) in rotation and another DOF in length. A chain of SPM transforms can model the image displacement and foreshortening effects produced by 3-D rigid links. The appearance of each link in the image is described by a template of pixels which is manually initialized and deformed by the link's DOF's.

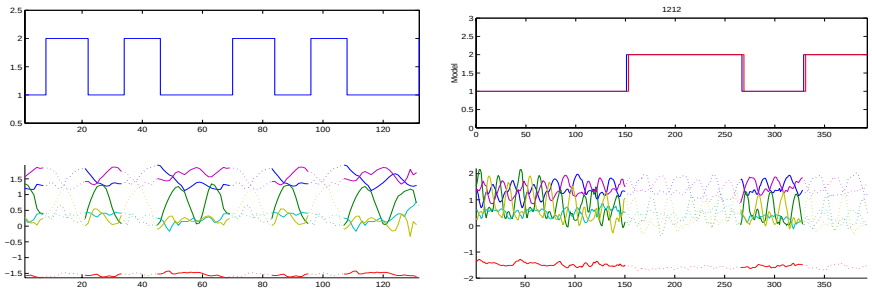
In our figure tracking experiments we analyzed the motion of the legs, torso, and head, and ignoring the arms. Our kinematic model had eight DOF's, corresponding to rotations at the knees, hip, and neck. A sample configuration of our figure model is shown in Figure 4.2.

4.1 Classification

The first task we addressed was learning an SLDS model for walking and running. Our training set consisted of 18 sequences of six individuals jogging (two examples of three people) and walking at a moderate pace (two examples of six people.) Each sequence

was approximately 50 frames duration. The training data consisted of the joint angle states of the SPM in each image frame, which was obtained manually.

Each of the two motion types were each modeled as multi–state[3] SLDSs and then combined into a single complex SLDS. Measurement matrix in all cases was assumed to be identity, $C = I$. Initial state segmentation within each motion type was obtained using unsupervised clustering in a state space of some simple dynamics model (e.g. constant velocity model.) Parameters of the model (A, Q, R, x_0, Π, π_0) were then reestimated using the EM-learning framework with approximate Viterbi inference. This yielded refined segmentation of switching states within each of the models. An example of the learned switching state sequence within a single "jog" training example is shown in Figure 3(a).



(a) (b)

Fig. 3. (a) Segmentation of two-state SLDS model states within single "jog" motion sequence. (b) Segmentation of mixed walking/running sequence. Top graph shows correct segmentation (dotted red line) and estimated segmentation (solid blue line). Bottom graph depicts the segmentation of the estimated LDS states.

To test the classification ability of our learned model we next considered segmentation of sequences of *complex* motion, i.e., motion consisting of alternations of "jog" and "walk."[4] Identification of different motion "regimes" was conducted using the approximate Viterbi inference. Estimates of "best" switching states $\langle s_t \rangle$ indicated which of the two models can be considered to be driving the corresponding motion segment. One example of this segmentation is depicted in Figure 3(b).

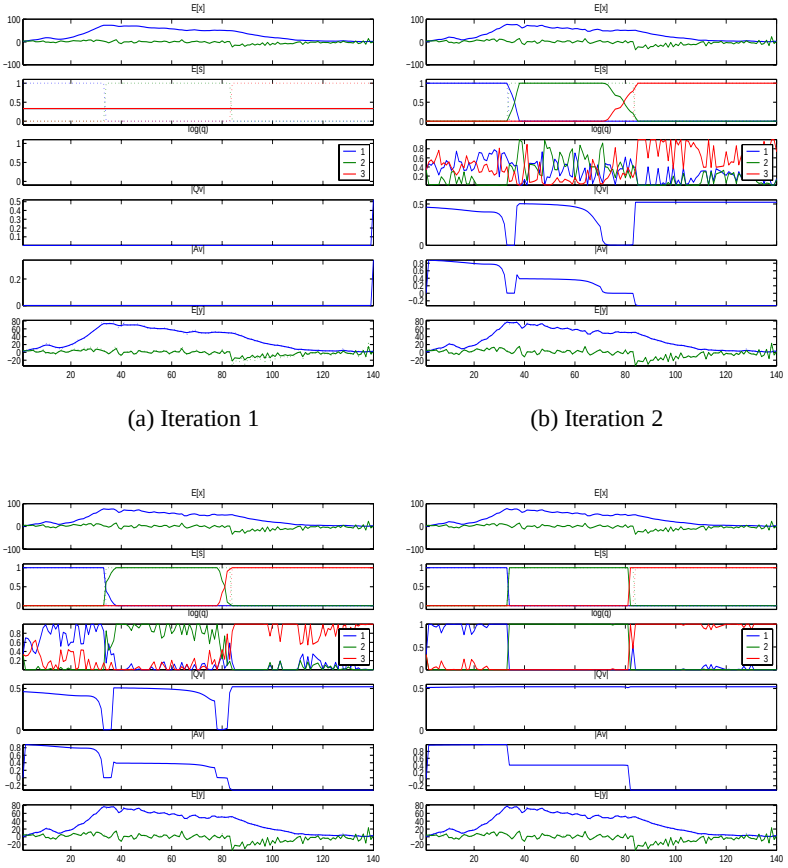
Classification experiments on a set of 20 test sequences gave an error rate[5] of 2.9% over a total of 8312 classified data points.

[3] We explored SLDS models with two to six states.

[4] Test sequences were constructed by concatenating in random order randomly selected and noise corrupted training sequences. Transitions between sequences were smoothed using B-spline smoothing.

[5] Classification error was defined as the difference between inferred segmentation and true segmentation accumulated over all sequences, $e = \sum_{t=0}^{T-1} |\langle s_t \rangle - s_{\text{true},t}|$.

Additional classification experiments were performed using the structured variational inference technique. Figure 4 depicts state estimates and variational parameters in iterations 1 through 4 of variational inference. Initial uncertain switching state distribu-



(a) Iteration 1 (b) Iteration 2

(c) Iteration 3 (d) Iteration 4

Fig. 4. Iterations of variational inference. The graphs depict: continuous state estimates $\langle x \rangle$, switching state estimates $\langle s \rangle$, HMM variational parameter $\log q$, determinants of LDS variational parameters $|Q_v|$ and $|A_v|$, and the true (dotted) and estimated measurements $E[y]$.

tion $\langle s \rangle$ leads to low variational state noise variance $\hat{Q}$ (whose determinant is indicated by $|Q_v|$ in Figure 4) and low variational state transition matrix (whose determinant is indicated by $|A_v|$ in Figure 4). Through further iterations the variational inference algorithm converges to the true switching state sequence.

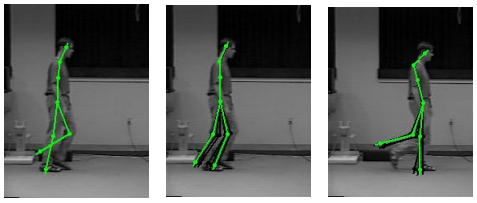
4.2 Tracking

A second experiment explored the utility of the SLDS model in improving tracking of the human figure from video. The difficulty in this case is that feature (joint angle) measurements are not readily available from a sequence of image intensities. Hence, we use the SLDS as a multi-hypothesis predictor that initializes multiple local template searches in the image space. Instead of choosing S^2 multiple hypotheses $\hat{x}_{t|t-1,i,j}$ at each time step we pick the best S hypothesis with the highest switching probability, i.e., $\hat{x}_{t|t-1,i,i_t^*}$ where $i_t^* = \arg\max_j \{\Pi(i,j) J_{t-1,j}\}$.

Given the predicted means for the figure locations, state-space observations are obtained by local image registration, or hill-climbing. This identifies the state-space modes in the likelihood function given by the template model. A larger set of measurements could be explored through sampling, as described in [6]. Given these observations of figure state, the regular SLDS filtering yields SLDS state priors.

Figure 5 shows stills from a representative example of SLDS tracking of walking motion. In this experiment, simple template features were used to model the appearance of the figure. Each link in the model has an associated template, which is initialized manually in the first frame and applied throughout the sequence. Template features are not robust to appearance changes such as lighting effects or the wrinkling of cloth. As a result, a template-based tracker can benefit substantially from an accurate dynamical model.

A constant velocity predictor does poorly in this case, leading to tracking failure by frame seven (shown in Figure 4.b). The learned SLDS model gives improved predictions leading to more robust tracking.



(a) (b) (c)

Fig. 5. (a) Tracker (in white) using constant velocity predictor drifts off track by frame 7. (b) SLDS-based tracker is on track at frame 7. Model (switching state) 3 has the highest likelihood. Black lines show prior mean and observation. (c) SLDS tracker at frame 20.

4.3 Synthesis and Interpolation

In Section 2 we introduced SLDS as a *generative* model. Nonetheless, SLDS is most commonly employed as a classifier (e.g. Section 4.1.) To test the power of the learned

SLDS framework we examined its use in synthesizing realistic–looking motion sequences and interpolating motion between missing frames.

In the first set of experiments the learned walk/jog SLDS was used to generate a "synthetic walk." Two stick figure motion sequences of the noise driven model are shown in Figure 6. Depending on the amount of noise used to drive the model the stick figure exhibits more or less "natural"–looking walk. Departure from the realistic walk becomes more evident as the simulation time progresses. This behavior is not unexpected as the SLDS in fact learns locally consistent motion patterns.

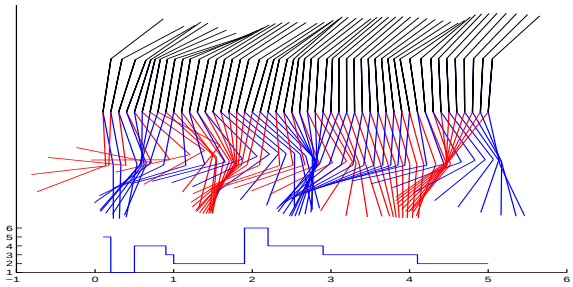


Fig. 6. Synthesized walk motion over 50 frames using SLDS as a generative model. States of the synthesized motion are shown on the bottom. Right leg is in blue color.

Another realistic situation may call for filling-in a small number of missing frames from a large motion sequence. SLDS can then be utilized as an interpolation function. In a set of experiments we employed the learned walk/jog model to interpolate a walk motion over two sequences with missing frames (see Figure 7.) The visual quality of the interpolation and the motion synthesized from it was high (left column in Figure 7.) As expected, the sparseness of the measurement set had definite bearing on this quality.

5 Conclusions

We have introduced a new approach to dynamics learning based on switching linear models. We have proposed two approximation techniques, Viterbi and structural variational inference, which overcomes the exponential complexity of exact inference. Simplicity of approximate Viterbi inference is contrasted by the lack of an exact bound on the approximation error. This is a problem in general with greedy Viterbi-style approximations, as well as with Markov chain Monte Carlo methods [20]. On the other hand, more complex structured variational inference guarantees minimization of this error by considering global approximation of intractable SLDS distribution.

Our preliminary experiments have demonstrated promising results in classification of human motion, improved visual tracking performance, and motion synthesis and interpolation using our SLDS framework. We demonstrated accurate discrimination between walking and jogging motions. We showed that SLDS models provide more robust

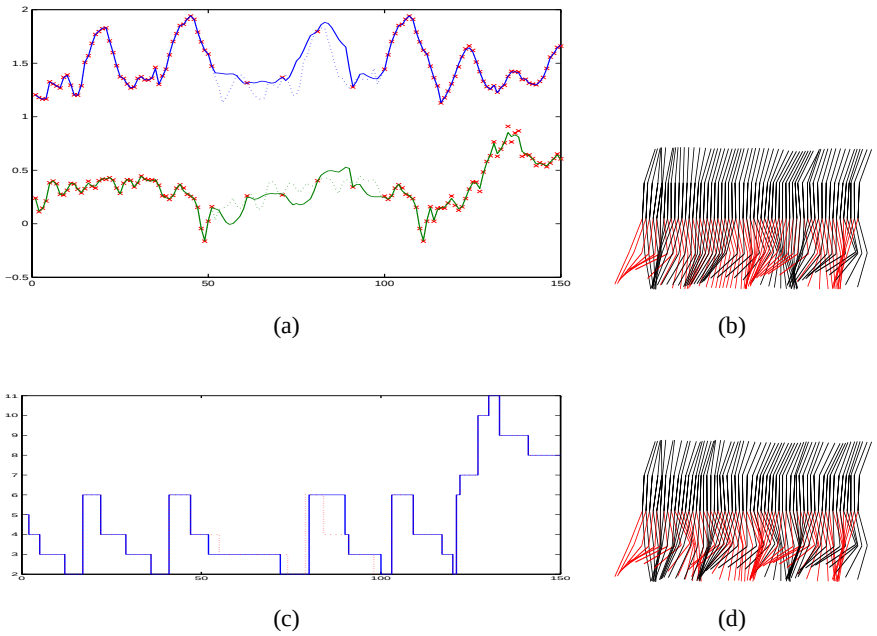


Fig. 7. SLDS as an interpolation function. A motion sequences with missing measurements between frames 50 and 100 was interpolated using an SLDS model. Symbols red 'x' in figure (a) indicate known measurement points. Solid lines show interpolated joint angle values. Dotted lines indicate ground truth (smoothing with no missing measurements.) Figure (c) depicts corresponding SLDS states. Stick figure motion generated from interpolated data is shown in figure (d). Figure (b) shows true stick figure motion.

tracking performance than simple constant velocity predictors. The fact that these models can be learned from data may be an important advantage in figure tracking, where accurate physics-based dynamical models may be prohibitively complex.

We are currently building a more comprehensive collection of frontoparallel human motion. We plan to build SLDS models for wide variety of motions and performers and evaluate their performance.

References

[1] B. D. O. Anderson and J. B. Moore. *Optimal filtering*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1979.

[2] Y. Bar-Shalom and X.-R. Li. *Estimation and tracking: principles, techniques, and software*. YBS, Storrs, CT, 1998.

[3] A. Blake, B. North, and M. Isard. Learning multi-class dynamics. In *NIPS '98*, 1998.

[4] M. Brand. Pattern discovery via entropy minimization. Technical Report TR98-21, Mitsubishi Electric Research Lab, 1998. Available at *http://www.merl.com*.

[5] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proceedings of International Conference on Computer Vision*, pages 494–499, Cambridge, MA, June 1995.
[6] T.-J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *Computer Vision and Pattern Recognition*, pages 239–245, 1999.
[7] T. Dean and K. Kanazawa. A model for reasoning about persistance and causation. *Computational Intelligence*, 5(3), 1989.
[8] Z. Ghahramani. Learning dynamic Bayesian networks. In C. L. Giles and M. Gori, editors, *Adaptive processing of temporal information*, Lecture notes in artificial intelligence. Springer-Verlag, 1997.
[9] Z. Ghahramani and G. E. Hinton. Switching state-space models. submitted for publication, 1998.
[10] J. K. Hodgins, W. L. Wooten, D. C. Brogan, and J. F. O'Brien. Animating human athletics. In *Computer Graphics (Proc. SIGGRAPH '95)*, pages 71–78, 1995.
[11] V. T. Inman, H. J. Ralston, and F. Todd. *Human Walking*. Williams and Wilkins, 1981.
[12] M. Isard and A. Blake. A mixed-state CONDENSATION tracker with automatic model-switching. In *Proceedings of International Conference on Computer Vision*, pages 107–112, Bombay, India, 1998.
[13] F. V. Jensen. *An introduction to Bayesian Networks*. Springer-Verlag, 1995.
[14] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in graphical models*. Kluwer Academic Publishers, 1998.
[15] I. A. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *Computer Vision and Pattern Recognition*, pages 81–87, San Fransisco, CA, June 18-20 1996.
[16] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction. *Journal of Basic Engineering (ASME)*, D(83):95–108, 1961.
[17] C.-J. Kim. Dynamic linear models with markov-switching. *Journal of Econometrics*, 60:1–22, 1994.
[18] V. Krishnamurthy and J. Evans. Finite-dimensional filters for passive tracking of markov jump linear systems. *Automatica*, 34(6):765–770, 1998.
[19] D. D. Morris and J. M. Rehg. Singularity analysis for articulated object tracking. In *Computer Vision and Pattern Recognition*, pages 289–296, Santa Barbara, CA, June 23–25 1998.
[20] R. M. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, pages 71–113, 1992.
[21] R. M. Neal and G. E. Hinton. A new view of the EM algorithm that justifies incremental and other variants. In M. Jordan, editor, *Learning in graphical models*, pages 355–368. Kluwer Academic Publishers, 1998.
[22] V. Pavlovic, B. Frey, and T. S. Huang. Time-series classification using mixed-state dynamic Bayesian networks. In *Computer Vision and Pattern Recognition*, pages 609–615, June 1999.
[23] J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Mateo, CA, 1998.
[24] L. R. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1993.
[25] H. E. Rauch. Solutions to the linear smoothing problem. *IEEE Trans. Automatic Control*, AC-8(4):371–372, October 1963.
[26] R. H. Shumway and D. S. Stoffer. Dynamic linear models with switching. *Journal of the American Statistical Association*, 86(415):763–769, September 1991.
[27] C. R. Wren and A. P. Pentland. Dynamic models of human motion. In *Proceeding of the Third International Conference on Automatic Face and Gesture Recognition*, pages 22–27, Nara, Japan, 1998.