Learning and Inferring Motion Patterns using Parametric Segmental Switching Linear Dynamic Systems

Sang Min Oh James M. Rehg Tucker Balch Frank Dellaert
GVU Center, College of Computing
Georgia Institute of Technology, Atlanta, GA, U.S.A.
{sangmin, rehg, tucker, dellaert}@cc.gatech.edu

Abstract

Switching Linear Dynamic System (SLDS) models are a popular technique for modeling complex nonlinear dynamic systems. An SLDS can describe complex temporal patterns more concisely and accurately than an HMM by using continuous hidden states. However, the use of SLDS models in practical applications is challenging for three reasons. First, exact inference in SLDS models is computationally intractable. Second, the geometric duration model induced in standard SLDSs limits their representational power. Third, standard SLDSs do not provide a principled way to interpret systematic variations governed by higher order parameters.

The contributions in this paper address all of these three challenges. First, we present a data-driven MCMC (DD-MCMC) sampling method for approximate inference in SLDSs. We show DD-MCMC provides an efficient method for estimation and learning in SLDS models. Second, we present segmental SLDSs (S-SLDS), where the geometric distributions of the switching state durations are replaced with arbitrary duration models. Third, we extend the standard SLDS model with additional global parameters that can capture systematic temporal and spatial variations. The resulting parametric SLDS model (P-SLDS) uses EM to robustly interpret parametrized motions by incorporating additional global parameters that underly systematic variations of the overall motion.

The overall development of the extensions for SLDSs provide a principled framework to interpret complex motions. The framework is applied to the honey bee dance interpretation task in the context of the on-going BioTracking project at the Georgia Institute of Technology. The experimental results suggest that the enhanced models provide an effective framework for a wide range of motion analysis applications.

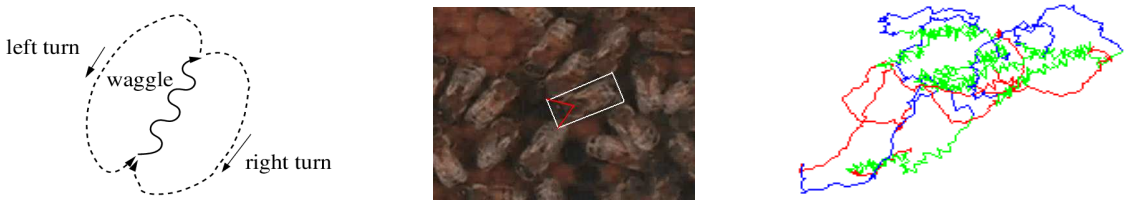
1 Introduction

A challenging problem in computer vision is to infer the behavioral patterns of a target in a segment of video. Even if we assume that targets can be reliably tracked, we still face the difficult problem of interpreting behavior. Manual interpretation of video by skilled field workers is common in domains such as biology. However, this is a time-consuming process that does not support large scale video analysis. Thus, it is desirable to develop methods that automatically infer the behavioral patterns of the targets. Moreover, in applications where there are a large range of behaviors which are difficult to specify manually, we need the ability to *learn* these behaviors from examples.

These requirements translate into two inference tasks that are of central importance. The first, 'labeling', is to automatically segment motion sequences according to different behavioral modes. The second task is what we call 'quantification', by which we mean the identification of global parameters that underly a given motion, e.g., the direction of a pointing gesture. These two inference tasks are not independent: a better understanding of the systematic variations in the data can improve the labeling results, and vice versa.

1.1 Biotracking

The application domain which motivates this work is a new research area which enlists visual tracking and AI modeling techniques in the service of biology [2, 3, 9, 51]. The current state of biological field work is still dominated by manual data



(a) (b) (c)

Figure 1: (a) A bee dance consists of three patterns : waggle, left turn, and right turn. (b) The box in the middle is a tracked bee. (c) An example honey bee dance trajectory. The track is automatically obtained using a vision-based tracker and manually labeled afterward. Key : waggle , right-turn , left-turn

interpretation, a time-consuming process. Automatic interpretation methods can provide field biologists with new tools for the quantitative study of animal behavior. A classical example of animal behavior and communication is the honey bee dance [20], depicted in a stylized form in Fig. 1(a). Honey bees communicate the location and distance to a food source through a dance that takes place within the hive. The dance is decomposed into three different regimes: "left turn", "right turn" and "waggle". The length (duration) and orientation of the waggle phase correspond to the distance and the orientation to the food source. Figure 1(b) shows a dancer bee that was tracked by a previously developed vision-based tracker described in [26]. After tracking, the obtained trajectory of the dancing bee is manually labeled as "left turn", "right turn" or "waggle" and is shown in Figure 1(c).

Our work on SLDS models is in support of three goals for automatic bee dance analysis. First, we aim to learn the motion patterns of honey bee dances from the labeled training sequences. Second, we should be able to automatically segment new sequences into the three dance modes reliably, i.e., the labeling problem. Finally, we face a quantification problem where the aim is to automatically deduce the message communicated, in this case: the distance and orientation to the food source. Note that both the labels and the global parameters are unknown, hence the problem is one of simultaneously inferring these hidden variables.

1.2 A Model-Based Approach

We adopt a model-based approach to the representation of behaviors and analysis of motion data. We assume that behaviors are made of sub-behaviors. In this case, we require a model which is expressive enough to capture individual sub-behaviors as well as the inter-relationships between them.

Hence, the basic generative model we adopt is the Switching Linear Dynamic System (SLDS) model [43, 44, 45]. In an SLDS model, there are multiple linear dynamic systems (LDS) that underly the motion, one for each behavioral mode that we assume. We can then model the complex behavior of the target by switching within this set of LDSs. In contrast to an HMM, which models a continuous trajectory using a set of piecewise constant functions, an SLDS provides the possibility to describe complex temporal patterns concisely and accurately. SLDS models have become increasingly popular in the vision and graphics communities as they provide an intuitive framework for describing the continuous but non-linear dynamics of real-world motion. For example, it has been used for human motion analysis [43, 44, 45, 47] and motion synthesis [56].

1.3 Overview

In this paper, we extend the scope and modeling power of standard SLDS models. We present a method for learning behavioral patterns from data. We describe inference methods for labeling the motion sequences while simultaneously quantifying the global parameters. When applying the standard SLDS model to the complex task of interpreting honey bee behavior, it quickly becomes clear that there are severe limitations in the original SLDS model that limit its applicability on real tasks. In this paper we discuss these three major problems and address them by extending the model in two novel ways. For each extension, we provide robust inference methods.

We discuss the three main limitations of the original SLDS model in Section 2, previewing each of the three main contributions along with related work in those areas. In Section 4, we introduce a data-driven MCMC-based inference method to address the intractability of exact inference in SLDSs. In Section 5, we present the segmental extension of a standard SLDS model, the "segmental SLDS" model (S-SLDS) with enhanced duration modeling capabilities. Then, in Section 6, we advance a parametric extension of SLDS (P-SLDS) which is able to infer systematic variations in the data. We combine S-SLDSs and P-SLDSs in Section 7 and show how we can learn and perform inference in the resulting parametric segmental SLDS (PS-SLDS). Finally, in Section 8, we describe the experimental data and demonstrate the improved labeling and quantification capabilities of the enhanced SLDS model through the experimental results on the honey bee dance decoding tasks.

2 Contributions and Related Work

In this paper, we address three limitations of the standard SLDS model: (1) the intractability of exact inference in SLDSs, (2) limitations in duration modeling, and (3) the absence of a principled way to quantify global parameters. We propose novel solutions to address these problems. First, we introduce a Data-Driven MCMC (DD-MCMC) inference method to identify the exact posterior of SLDSs in the presence of intractability. Secondly, a segmental SLDS model is proposed to improve the duration-modeling capabilities of standard SLDSs. Finally, we introduce a parametric extension of SLDSs that provide a principled means to quantify the embedded global parameters. In the following sections we discuss each of these contributions along with the related work that provided the inspiration for them.

The BioTracking project [2, 3] is an interdisciplinary research initiative between biology and multi-robot systems. This work builds on our previous work on automatic labeling of honey bee dances using HMMs [17]. However, in [17], the honey bees were tracked via a color segmentation tracker and HMMs were learned from two dimensional observations, i.e. locations of the bees. In contrast, the real-world dancer bee tracks are automatically obtained from a set of noisy video data by using a previously-developed appearance tracker described in [26]. In addition, SLDSs are used to learn and infer the motion patterns of bees and a new DD-MCMC method as well as novel SLDS extensions are presented in this work.

In comparison to our previous conference publications on this topic [40, 41], the current paper extends the SLDS model to include duration modeling, and presents the detailed learning and inference mechanisms for parametric segmental SLDS (PS-SLDS).

2.1 Robust inference via Data-Driven Markov Chain Monte Carlo Sampling

Inference in an SLDS model involves computing the posterior distribution on the hidden states, which consists of the (discrete) switching state and the (continuous) dynamic state. In the Biotracking application which motivates this work, the discrete state represents distinct honey bee behaviors while the dynamic state represents the bee's true motion. Given video-based measurements of the position and orientation of the bee over time, SLDS inference can be used to obtain a MAP estimate of the behavior and motion of the bee. In addition to its central role in applications such as MAP estimation, inference is also the crucial step in parameter learning via the EM algorithm [45].

It is known that exact inference in SLDS is intractable, as the size of Gaussian mixtures increases exponentially with time [31]. Thus, there have been research efforts to derive efficient approximate inference methods. The early examples include GPB2 [5], and Kalman filtering [10], and the pseudo-EM algorithm [52]. More recent examples include variational approximation [22, 43, 45, 39, 25], an approximate Viterbi method [44, 43, 45], expectation propagation [57], iterative Monte Carlo methods [15], sequential Monte Carlo methods [16], and Gibbs sampling [11, 48]. Approximate inference in SLDS models has focused primarily on two classes of techniques: stage-wise methods such as approximate Viterbi [45] or GPB2 [5] which maintain a constant representational size for each time step as data is processed sequentially, and structured variational methods which approximate the intractable exact model with a tractable, decoupled model [22, 39, 45, 25].

While these approaches are successful in some application domains, such as vision and graphics, they do not provide any mechanism for fine-grained control over the accuracy of the approximation. In fields such as biology where learned models can be used to answer scientific questions about animal behavior, scientists would like to characterize the accuracy of an approximation and they may be willing to pay an additional computational price for getting as close as possible to the true posterior. In our initial stage of experiments, we observed that the existing approximation methods, e.g., an approximate

Viterbi method demonstrated poor labeling performance. In such cases, it is necessary to validate the capacity of the model to verify whether such a poor labeling result is due to the approximation method itself or an inherent limitation in the ability of the model not being able to represent the temporal phenomenon adequately.

We describe a novel proposal distribution for Data-driven MCMC inference in Section 4, originally presented at AAAI [40]. In situations where a controllable degree of accuracy is required, Markov-Chain Monte-Carlo (MCMC) methods are attractive. Standard MCMC techniques, however, are often plagued by slow convergence rates. We therefore explore the use of Rao-Blackwellisation [12] and the Data-Driven MCMC paradigm [53] to improve convergence. The Data-Driven MCMC approach has been successfully applied in computer vision [53, 7, 27, 30] and robotic mapping [46].

2.2 Improved Duration modeling

The duration modeling capabilities of a standard SLDS are limited by the assumption which is imposed upon the transitions between the discrete switching states. As a consequence of Markov assumption, the probability of remaining in a given switching state follows a geometric distribution :

$$P(d) \quad = \quad a^{d-1}(1-a) \tag{1}$$

Above, d denotes the duration of a given switching state and a denotes the probability of a self-transition. One consequence of this model is that a duration of one time-step possesses the largest probability mass. This can be seen in Fig. 2 where the red curve depicts the geometric distribution.

In contrast, many natural temporal phenomena exhibit patterns of regularity in the duration over which a given model or regime is active. In such cases the standard SLDS model would not effectively encode the regularity of the data. A honey bee dance is a good example: a dancer bee will attempt to stay in the waggle regime for a certain duration to effectively communicate a message. In such cases, it is clear that the actual duration diverges from a geometric distribution.

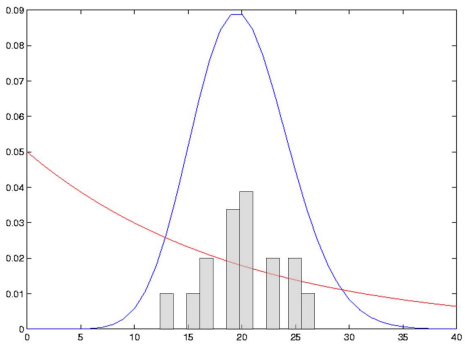


Figure 2: A Gaussian model is closer to the duration distribution of training data (shown as the overlayed histogram) than a geometric duration model.

To illustrate this point, we learned a duration model for the waggle phase using a Gaussian density and a conventional geometric distribution, using one of the manually labeled dance sequences depicted in Figure 15. Figure 2 shows the learned geometric and Gaussian distributions for comparison. It can be observed that the Gaussian model is much closer to the training data than the conventional geometric model.

The limitation of a geometric distribution has been previously addressed by the HMM community, and HMM models with enhanced duration capabilities have been developed [18, 33, 50, 42]. The variable duration HMM (VD-HMM) was introduced in [18]: state durations are modeled explicitly in a variety of PDF forms. Later, a different parameterization of the state durations was introduced where the state transition probabilities are modeled as functions of time, which are referred to as non-stationary HMMs (NS-HMM) [33]. It has since been shown that the VD-HMM and the NS-HMM are duals [14]. In addition, segmental HMM with random effects was developed in the data mining community [21, 29]. Ostendorf et.al. provide an excellent discussion of segmental HMMs in [42].

We adopt similar ideas to arrive at SLDS models with enhanced duration modeling.

2.3 Inference of Global Parameters

In many applications we are more interested in the global parameters that underly the behavior rather than the exact categorization of the sub-motions. Unfortunately, the standard SLDS does not provide a principled way to quantify temporal and spatial variations with respect to a fixed (canonical) underlying behavioral template. E.g., the dynamics and observations of a pointing gesture would vary based on the speed of the motion and the direction being indicated.

Previously, Wilson & Bobick addressed this problem by presenting a parametric HMMs (P-HMM) [55]. In a P-HMM, the parametric observation models learned are conditioned on global observation parameters, such that globally parameterized gestures can be recognized. P-HMMs have been used to interpret human gestures, showing superior recognition performance in comparison to standard HMMs. A similar approach was taken in the style-machines work by Brand and Hertzmann [8]. A related transformation-invariant learning approach for video analysis is described in [19].

Inspired by P-HMM, we extend the standard SLDS model, resulting in a parametric SLDS (P-SLDS). As in a P-HMM, the P-SLDS model incorporates global parameters that underly systematic spatial variations of the overall target motion. In addition, while P-HMM only introduced global observation parameters which describe spatial variations in the outputs, we additionally introduce dynamic parameters which capture temporal variations. As mentioned earlier, the problems of global parameter quantification and labeling can be solved simultaneously. Hence, we formulate expectation-maximization (EM) methods for learning and inference in P-SLDS and present it in Section 6. A preliminary version of P-SLDS work appeared in [41].

3 Background

3.1 Linear Dynamic Systems

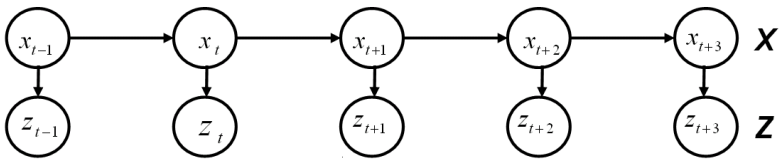


Figure 3: A linear dynamic system (LDS)

A Linear Dynamic System (LDS) is a time-series state-space model consisting of a linear Gaussian dynamics model and a linear Gaussian observation model. The graphical representation of an LDS is shown in Fig. 3. The Markov chain at the top represents the state evolution of the continuous hidden states x_t. The prior density p_1 on the initial state x_1 is assumed to be normal with mean μ_1 and covariance Σ_1, i.e., $x_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$.

The state x_t is obtained by the product of state transition matrix F and the previous state x_{t-1}, corrupted by zero-mean white noise w_t with covariance matrix Q:

$$x_t = Fx_{t-1} + w_t \text{ where } w_t \sim \mathcal{N}(0, Q) \tag{2}$$

In addition, the measurement z_t is generated from the current state x_t through the observation matrix H, and corrupted by zero-mean observation noise v_t:

$$z_t = Hx_t + v_t \text{ where } v_t \sim \mathcal{N}(0, V) \tag{3}$$

Thus, an LDS model M is defined by the tuple $M \triangleq \{(\mu_1, \Sigma_1), (F, Q), (H, V)\}$. Exact inference in an LDS can be performed using the RTS smoother [5], an efficient variant of belief propagation for linear Gaussian models. Further details on LDSs can be found in [5, 34, 49].

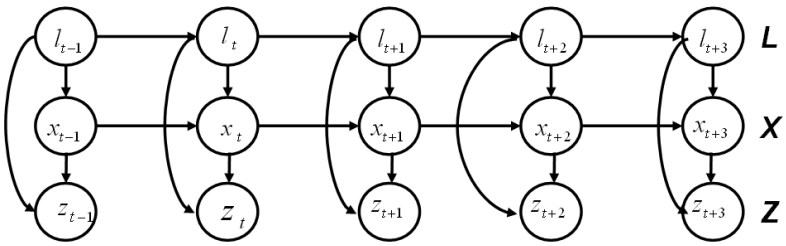


Figure 4: Switching linear dynamic systems (SLDS)

3.2 Switching Linear Dynamic Systems

In a switching LDS (SLDS), we assume the existence of n distinct LDS models $M \triangleq \{M_l | 1 \leq l \leq n\}$. The graphical model corresponding to an SLDS is shown in Fig. 4. The middle chain, representing the hidden state sequence $X \triangleq \{x_t | 1 \leq t \leq T\}$, together with the observations $Z \triangleq \{z_t | 1 \leq t \leq T\}$ at the bottom, is identical to an LDS in Fig. 3. However, we now have an additional discrete Markov chain $L \triangleq \{l_t | 1 \leq t \leq T\}$ that determines which of the n models M_l is used at every time-step. We call $l_t \in M$ the label at time t and L a label sequence.

In addition to a set of LDS models M, we specify two additional parameters: a multinomial distribution $\pi(l_1)$ over the initial label l_1 and an $n \times n$ transition matrix B that defines the switching behavior between the n distinct LDS models. In summary, a standard SLDS model is defined by the tuple $\Theta \triangleq \Big\{\pi, B, M \triangleq \{M_l | 1 \leq l \leq n\}\Big\}$.

Switching linear dynamic system (SLDS) models have been studied in a variety of research communities ranging from computer vision [44, 43, 45, 38, 10], computer graphics [56, 47], tracking [6], signal processing [15, 16] and speech recognition [48], to econometrics [28], visualization [57], machine learning [32, 22, 40, 41, 39, 25], control systems [54] and statistics [52]. While one can find several versions of SLDS in the literature, our work is most closely related to the model structure and extensions described in [44, 43, 45, 40, 41, 39].

3.3 Learning and Inference in SLDS

The Expectation Maximization (EM) algorithm [13] can be used to obtain the maximum-likelihood parameters $\hat{\Theta}$. The hidden variables in EM are the label sequence L and the state sequence X. Given the observation data Z, EM iterates between the two steps:

- E-step : Inference to obtain the posterior distribution

$$f^i(L, X) \triangleq P(L, X | Z, \Theta^i) \tag{4}$$

  over the hidden variables L and X, using a current guess for the SLDS parameters Θ^i.

- M-step : maximize the expected log-likelihoods with respect to Θ:

$$\Theta^{i+1} \leftarrow \underset{\Theta}{\text{argmax}} \ \langle \log P(L, X, Z | \Theta) \rangle_{f^i(L,X)} \tag{5}$$

Above, $\langle \cdot \rangle_W$ denotes the expectation of a function $(\cdot)$ under a distribution W. The intractability of the exact E-step in Eq. 4 motivates the development of sampling-based approximate inference techniques discussed in Section 4.

4 Inference via Data-Driven MCMC

All Markov Chain Monte Carlo (MCMC) methods work similarly [23]: they generate a sequence of *samples* with the property that the collection of samples approximates the desired target distribution. To accomplish this, a *Markov chain* is defined

over the space of interest. The transition probabilities are set up in a very specific way such that the *stationary distribution* of the Markov chain is exactly the target distribution. This guarantees that if we run the chain for a sufficiently long time, the sample distribution converges to the target distribution.

In this section, we introduce a novel sampling-based inference method with the property that the estimated posterior distribution will theoretically converge to the correct posterior distribution, e.g., $P(L|Z)$, in an efficient manner. Fast convergence is achieved by incorporating a data-driven approach where we introduce proposal priors and label-cue models.

4.1 Rao-Blackwellised MCMC

In our solution, we propose to use the Rao-Blackwellised posterior $P(L|Z)$, rather than the joint posterior $P(L, X|Z)$. The result is the dramatic reduction in the size of the sampling space from (L, X) to L. This results in an improved approximation of the label distribution $P(L|Z)$, which is exactly the quantity of interest in many application. This reduction is justified by the Rao-Blackwell theorem [12]. Rao-Blackwellisation is achieved through the analytic integration of the continuous states X given a sample label sequence $L^{(r)}$. In this scheme, we can compute the probability of the r th sample labels $P(L^{(r)}|Z)$ up to a normalizing constant via the marginalization of the joint PDF :

$$P(L^{(r)}|Z) \quad \propto \quad \int_X P(L^{(r)}, X, Z) \tag{6}$$

Note that we omit the implicit dependence on the model parameters Θ for brevity. The joint PDF $P(L^{(r)}, X, Z)$ in the r.h.s. of Eq. 6 can be evaluated via inference in the time-varying LDS given the sample label sequence $L^{(r)}$. Specifically, the inference over the continuous hidden states X in the middle chain of Fig. 4 can be performed by RTS smoothing [5]. The resulting posterior is a time-series of Gaussians on X and can be effectively integrated out.

We use the Metropolis-Hastings (MH) algorithm [24, 36] to generate a sequence of samples $L^{(r)}$. The pseudo-code for the algorithm is shown in Algorithm 3 in Appendix A.

4.2 Learning and Inference

The key to the Data-Driven paradigm [53] is the use of cues which are present in the data to provide an efficient MCMC proposal distribution Q. An efficient proposal Q can result in substantially faster convergence [1]. Even though MCMC is guaranteed to converge, a naive exploration of the high dimensional state space L is prohibitive. Thus, the design of a proposal distribution which enhances the ability of the sampler to efficiently explore the space with high probability mass is crucial. Our data-driven approach consists of two phases : *learning* and *inference*.

In the learning phase, we collect temporal cues from the training data. Then, a set of models of cues which we call 'label-cue models', i.e. $\{P(c|l_i)|1 \leq i \leq n\}$, are learned in a supervised manner based on the collection of cues. For example, the change of heading angles is used as a cue in the honey bee dance application. From the stylized dance in Fig. 1(a), we observe that the heading angles will jitter but stay constant on average during the waggling, but generally increase or decrease during the right turn or left turn phases. Thus, a cue c_t for a frame is set to be the change of heading angles within the corresponding window. In this case, the label-cue model predicts the label in each time-step based on the cue. Note that the heading angles are measured clockwise.

For every time sample, we examine its local neighborhood within the boundary of a window with a fixed size to collect cues (the heading angle change). The cues are shown at the top of Fig. 5. Then, the collected cues are classified according to the training labels. The label-cue models are learned by fitting a conditional Gaussian to the cues for each of the possible labels, as shown at the bottom of Fig. 5. The estimated means and the standard deviations show that the average change of heading angles are -5.77, -0.10 and 5.79 radians, corresponding approximately to left-turn, waggle, and right-turn.

In the inference phase, we first collect the temporal cues from the test data without access to the labels as shown at the top of Fig. 6. Then, the proposal prior is evaluated based on the collected cues and the learned label-cue models. By a proposal prior $P(\tilde{l}_t|c_t)$, we denote the distribution on the labels which is a rough approximation to the true posterior $P(l_t|Z)$. However, the raw proposal prior often over-fits test data as shown in Fig. 6. Thus, we use the smoothed estimates as the final proposal prior, shown in Fig. 7. At the bottom of Fig. 7, the ground truth labels are shown below the final proposal prior for comparison. It can be observed that the obtained prior provide an excellent guide to the labels of the dance segments.

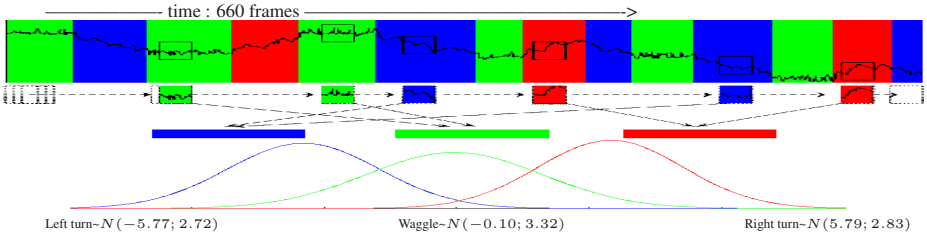


Figure 5: Learning phase. Three label-cue models are learned from the training data. See text for detailed descriptions.

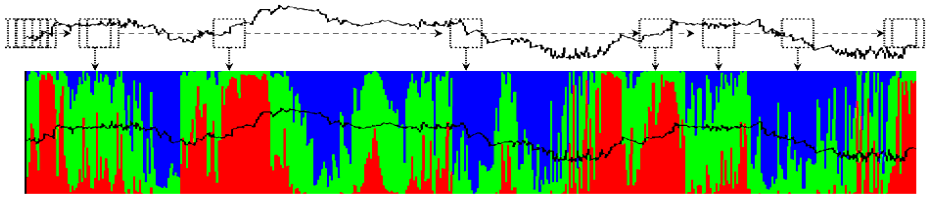


Figure 6: Inference phase. Raw proposal prior is evaluated based on the collected temporal cues.

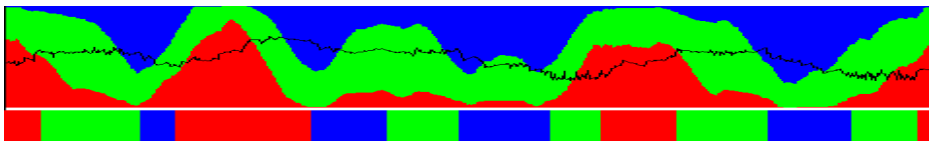
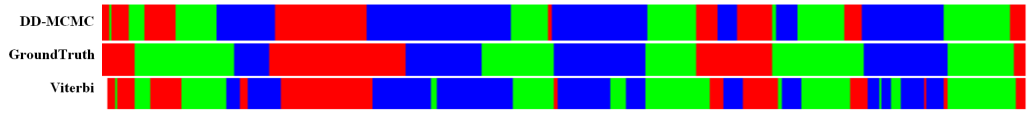


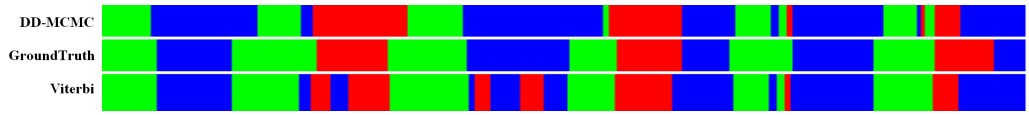
Figure 7: Final proposal prior and the ground truth labels. Key : waggle , right-turn , left-turn .

Afterwards, the obtained proposal prior $P(\tilde{L})$ is used to construct the data-driven proposal Q. Then, MH algorithm balances the whole MCMC procedure in such a way that the MCMC inference on labels converges to the true posterior $P(L|Z)$. The details of learning and inference in DD-MCMC method are described in Appendix A.

4.3 Experimental Results



(a) Sequence 1



(b) Sequence 2

Figure 8: DD-MCMC MAP, ground truth, Viterbi labels. Key : waggle , right-turn , left-turn .

DD-MCMC is a Bayesian inference algorithm. It can be used as a labeling method by computing the MAP label sequence from the estimated posterior distribution $P(L|Z)$, where the label in the MAP sequence at each time step is the individually most-likely label in $P(L|Z)$. The resulting MCMC MAP labels, the ground-truth, and the approximate Viterbi labels for

two data sequences in the database are shown from the top to bottom in Fig. 8. It can be observed that DD-MCMC delivers solutions that agree very well with the ground truth. On the other hand, the approximate Viterbi labels at the bottom over-segment the data (insertion errors). The insertion errors of approximate Viterbi highlight one of the limitations of the class of stagewise greedy algorithms for SLDS where they fail to consider the long-term correlations in the data.

Some errors between the MAP labels and the ground truth still occur (see Figure 8), due to the systematic irregular motions of the tracked bees. In these cases, even an expert biologist will have difficulty figuring out all the correct dance labels solely based on the observation data, without access to the video.

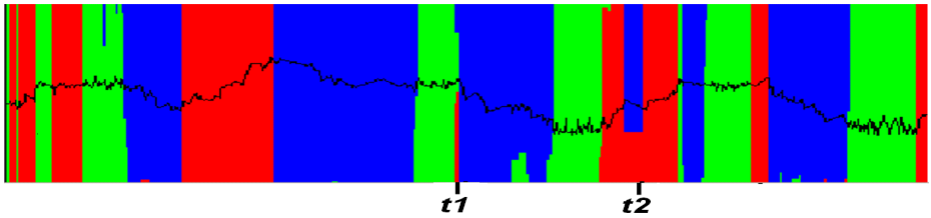


Figure 9: Posterior distribution $P(L|Z)$ is discovered from sequence 1. The heading angle of a bee is superimposed on the figure as an indicator of a dancer's dance mode. The label inference results around t_1 introduce an insertion error due to strong noise. In addition, the systematic irregular pattern of a tracked bee around t_2 results in another error. Key : waggle , right-turn , left-turn

To further analyze the inference capabilities of a standard SLDS within our application, we investigated the posterior distribution $P(L|Z)$ which is estimated from the first dataset using DD-MCMC inference, see Fig. 9. Careful examination of the posterior in the over-segmented intervals reveals that the tracking data is very noisy and most of the over-segmentations are induced due to this factor, e.g., see t_1 marked in Fig. 9. On the other hand, as an example of a *systematic* noise, examination of the original video shows that the tracked bee systematically walks to the left around the time-step marked as t_2 in Fig. 9 due to the collision with other bees around it for about 20 frames while it was turning right. Consequently, the MCMC posterior shows the two eminent hypotheses for those frames, e.g., t_2.

4.4 Discussion

While DD-MCMC inference improves upon the Viterbi method, the results are still not completely satisfactory for the bee dance application. The MAP label results contain several over-segmentations. In addition, it can be observed that the average waggle duration based on these labels diverges significantly from the ground truth from the resulting short waggles from the insertion errors.

We introduce segmental SLDS and parametric SLDS as extensions of our modeling framework which are designed to resolve the problems mentioned above. From the visualized posterior in Fig. 9, we notice two limitations of the standard SLDS model in our bee application. First, we observe that the limited duration modeling power of SLDS weakens its labeling capabilities on bee data. It can be observed that slight noise introduces an over-segmentation errors even though such noise appears only for a few frames. Second, the absence of systematic means to quantify global parameters should be addressed. The estimation of global dance angle and average waggle duration solely dependent on labeling estimates can deviate substantially from the ground truth. Moreover, the global information can provide a better cue for the overall labeling process.

It should be noted that DD-MCMC is computationally demanding, although it is efficient in comparison to some of the standard MCMC methods such as Gibbs sampling. For example, it proposed 4,500 samples on average to converge in the experiments. As each proposed label sequence requires a temporal smoothing step for Rao-Blackwellised inference, the computation required for each sample in DD-MCMC is approximately equal to the entire cost of the approximate Viterbi (VI) method. As a consequence, the DD-MCMC method used approximately 4,500 times more computation than VI method in our application. In the sections that follow, we will describe two extensions of the SLDS model that improve its ability to handle complex motion data. In each case we compare the performance of the extended SLDS models with the standard

approach. Since the choice of approximate inference method is potentially a confounding factor in model comparison, we restrict our attention to standard, deterministic approximation techniques, e.g., approximate Viterbi and variational method.

5 Segmental SLDS

We introduce the segmental SLDS (S-SLDS) model, which improves upon the standard SLDS model by relaxing the Markov assumption at a time-step level to a coarser *segment level*. The development of the S-SLDS model is motivated by the regularity in durations that is exhibited by honey bee dances. As discussed in Section 2.2, a dancer bee will attempt to stay in the waggle regime for a certain duration to effectively communicate the distance to the food source. In such a case, the geometric distribution induced in a standard SLDS is not an appropriate choice. Fig. 2 shows that a geometric distribution assigns the highest probability to the duration of a single time step. As a result, the label inference in standard SLDSs is susceptible to over-segmentation.

In an S-SLDS, the durations are first modeled explicitly and then non-stationary duration functions are derived from them. Both of them are learned from data. As a consequence, the S-SLDS model has more descriptive power and can yield more accurate results than the standard SLDSs. Nonetheless, we show that one can always convert a learned S-SLDS model into an equivalent standard SLDS, operating in a different label space. The approach has the significant advantage of allowing us to reuse the large array of approximate inference and learning techniques developed for SLDSs.

5.1 Conceptual View on the Generative Process of S-SLDS

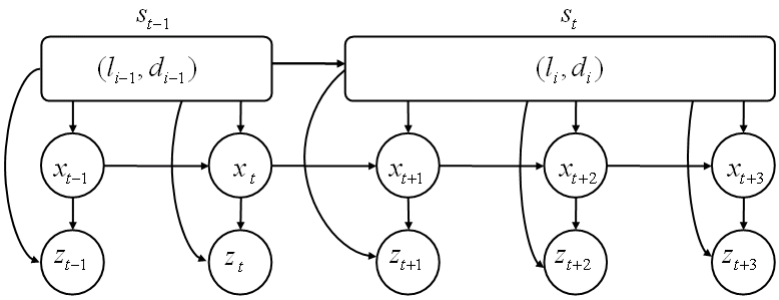


Figure 10: A schematic view of an S-SLDS with explicit duration models.

In an S-SLDS, we deal with segments of finite duration, i.e. each segment $s_i \triangleq (l_i, d_i)$ is described by a tuple consisting of a label l_i and a duration d_i. Within each segment, a fixed LDS model M_l is used to generate the continuous state sequence for the duration d_i. Similar to SLDSs, we take an S-SLDS to have an initial distribution $\pi(l_1)$ over the initial label l_1 of the first segment s_1, and an $n \times n$ semi Markov label transition matrix $\tilde{B}$ that defines the switching behavior between the segment labels. The tilde denotes that the matrix is a semi-Markov transition matrix. Additionally, we associate each label l with a fixed *duration model* D_l, represented as a multinomial. We denote the set of n duration models as $D \triangleq \{D_l(d) | 1 \leq l \leq n\}$, and refer to them in what follows as *explicit duration models*. In summary, an S-SLDS is defined by a tuple $\Theta \triangleq \left\{\pi, \tilde{B}, D \triangleq \{D_l | l = 1..n\}, M \triangleq \{M_l | l = 1..n\}\right\}$.

A schematic depiction of an S-SLDS is illustrated in Fig. 10. The top chain in the figure is a series of segments where each segment is depicted as a rounded box. In the model, the current segment $s_i \triangleq (l_i, d_i)$ generates a next segment s_{i+1} in the following manner: first, the current label l_i generates the next label l_{i+1} based on the label transition matrix $\tilde{B}$; then, the next duration d_{i+1} is generated from the duration model for the label l_{i+1}, i.e. $d_{i+1} \sim D_{l_{i+1}}(d)$. The dynamics for the continuous hidden states and observations are identical to a standard SLDS : a segment s_i evolves the set of continuous hidden states X with a corresponding LDS model M_{l_i} for the duration d_i, then the observations Z are generated given the labels L and the set of continuous states X.

5.2 Graphical Representation of S-SLDS

In this section we present a graphical representation of an S-SLDS which transforms the conceptual generative model described in Section 5.1 into an equivalent model that uses a conventional Markov switching process at every time-step. To maintain the same duration semantics, we introduce *counter variables* $C \triangleq \{c_t | 1 \leq t \leq T\}$. The resulting graphical model of S-SLDS is illustrated in Fig. 11, and is identical to the graphical model of an SLDS in Fig. 4, but with additional top chain representing a series of counter variables C.

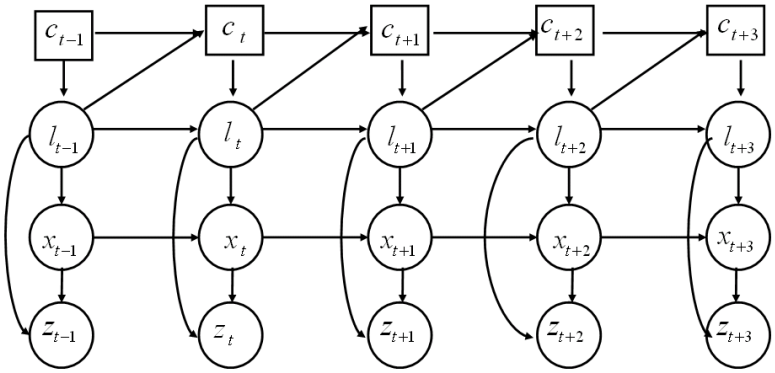


Figure 11: Graphical representation of an S-SLDS

The variables C maintain an incremental counter which evolves on the basis of *non-stationary transition functions* (NSTFs) $U \triangleq \{U_l(c) | 1 \leq l \leq n\}$. An NSTF U_l for the current label l_t defines the conditional dependency of the next counter variable c_{t+1} given the current counter variable c_t and the label l_t :

$$U_l(c_t) \quad = \quad P(c_{t+1} | c_t, l)$$

The system can either increment the counter, i.e. $c_{t+1} \leftarrow c_t + 1$, or reset it to one, i.e. $c_{t+1} \leftarrow 1$. If the counter variable c_{t+1} is reset, then a label transition occurs, i.e. a new segment is initialized. A new label l_{t+1} is chosen based on the label transition matrix $\tilde{B}$. If the counter simply increments, then the new label is set to be the current label l_t, i.e. $l_{t+1} \leftarrow l_t$.

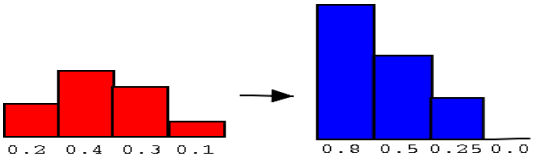


Figure 12: Conversion from explicit duration model D (left) to an equivalent NSTF U (right). As an example, $U(2) = D(2)/\{D(2) + D(3) + D(4)\} = 0.4/0.8 = 0.5$.

We first describe how to convert explicit duration models to equivalent NSTFs. Then, we discuss how the computed NSTFs are used for inference in SLDSs in Sec. 5.4. Given a time series data set, it is straightforward to estimate the parameters of explicit duration models D, as discussed in Sec. 5.1. However, in order to incorporate these durations into the SLDS framework, it is necessary to transform the explicit duration models D into equivalent NSTFs U. To do this, we can observe that the explicit duration models D and the NSTFs U are analogous to the duration models of VD-HMMs [18] and NS-HMMs [33] respectively. Hence, we can exploit the duality between VD-HMMs and NS-HMMs, which is described in

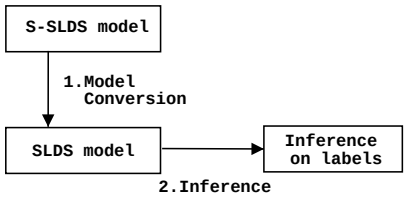


Figure 13: Inference in S-SLDS.

[14]. The equivalent NSTFs U are evaluated from the explicit duration models D as follows :

$$U_l(c_t) \quad = \quad 1 - \underbrace{\left(D_l(c_t) / \sum_{d=c_t}^{D_l^{max}} D_l(d) \right)}_{\bar{U}_l(c_t)} \tag{7}$$

Above, D_l^{max} denotes the maximum duration allowed for the l th model. Intuitively, the second term $\bar{U}_l(c_t)$ on the r.h.s. in Eq. 7 denotes the probability for a segment with a label l to reset the counter variable $c_{t+1} \leftarrow 1$. It represents the ratio of the probability of current duration c_t over the sum of durations equal or greater than c_t in the corresponding duration model D_l. An example is illustrated in Fig. 12 to show the evaluation of an NSTF from an explicit duration model. In summary, an S-SLDS model is completely defined by a tuple $\Theta \triangleq \left\{ \pi, \tilde{B}, U \triangleq \{U_l | 1 \leq l \leq n\}, M \triangleq \{M_l | 1 \leq l \leq n\} \right\}$ where the NSTFs U are obtained from the explicit duration models D.

5.3 Learning in a Segmental SLDS

Learning in S-SLDS is analogous to learning in SLDS, using EM. The initial distribution π, and LDS model parameters M are learned in exactly the same manner as in SLDS. However, it is necessary to learn the additional duration models D and the semi-Markov transition matrix $\tilde{B}$. These two additional model parameters only influence the label sequence L, and hence the ML estimates of these two parameters can be evaluated from a segmental representation of the label sequence L, i.e., $L = \cup_{j=1}^{|s|} s_j$. The specific functional forms of ML estimation depend upon the choice of duration models. An example is given in Section 7, where we learn the duration models in Gaussian forms from honey bee dance sequences. However, Gaussian models encode probabilities for non-existing negative durations as well. Hence, the only positive part of the learned Gaussian models were used in our work. Note that the choice on the form of probability distributions depend on the duration characteristics of data. For example, Gamma or log-normal distributions which only encode probability regions on positive durations can also be adopted.

5.4 Inference for Segmental SLDSs

We describe a convenient inference procedure for S-SLDS which simply reuses the existing SLDS inference modules by re-parameterizing itself into an equivalent SLDS model. This is an important advantage as it allows us to readily reuse the large array of approximate inference algorithms discussed in Section 3.3. In other words, the inference in S-SLDS is identical to that of the standard SLDS, simply with additional conversion from an S-SLDS to its corresponding SLDS. Note that the conversion algorithm described in this section is an independent convenience procedure which differs from the conversion for NSTFs described in Eq. 7.

The overall idea of inference is depicted in Figure 13. In step 1, we convert an S-SLDS model into an equivalent SLDS model. Then, we perform step 2 (inference) using any of the approximate inference algorithms for the standard SLDSs. Once the inference results are obtained via available standard SLDS inference methods, obtained SLDS results are converted back to S-SLDS form and the inference in S-SLDS concludes.

The model conversion from an S-SLDSs to an equivalent SLDS is possible by applying the standard technique of merging multiple discrete variables into meta variables. Specifically, all possible pairs of a label l_t and a counter value c_t are merged

and form a set of "lc" variables where $\mathcal{LC} \triangleq \{(l, c_i)|1 \leq l \leq n, 1 \leq c_i \leq D_l^{max}\}$. To obtain a complete SLDS model, an equivalent $n' \times n'$ transition matrix B', where $n' \triangleq \sum_{l=1}^{n} D_l^{max}$, is constructed from the semi-Markov transition matrix $\tilde{B}$ and the NSTFs U, as follows :

$$B'_{(l_i,c_i),(l_j,c_j)} = \begin{cases} U_{l_i}(c_i) & \text{increment} \\ \tilde{B}_{l_i,l_j}(1 - U_{l_i}(c_i)) & \text{reset} \\ 0 & \text{not allowed} \end{cases} \quad (8)$$

In Eq. 8, the three cases for the counter variable differ as follows : (increment) $l_i = l_j$ and $c_j = c_i + 1$, (reset) $c_j = 1$, and (not allowed) for all other cases. In addition, the initial label distribution π' for the equivalent SLDS can similarly be constructed from the S-SLDS initial distribution π :

$$\pi'(l_i, c_i) = \begin{cases} \pi(l_i) & \text{if } c_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

5.5 Computational Considerations

We have established that an equivalent SLDS can always be constructed from an arbitrary S-SLDS. However, if we reuse the original learning and inference algorithms for SLDSs in a naive manner, the cost of inference will be on the order of $O(TD_{max}^2|L|^2)$ for S-SLDSs, while it takes $O(T|L|^2)$ for SLDSs without duration models, where $D_{max} \triangleq \max_{1 \leq l \leq n}\{D_l^{max}\}$ denotes the maximum duration among all labels. Thus, there is a considerable computational overhead, by a factor of $O(D_{max}^2)$. This increased asymptotic running time overhead applies to all the approximate inference algorithms[1] with pairwise computations in general.

Nonetheless, we can still maintain linear efficiency w.r.t. the maximum duration D_{max} by exploiting the sparseness of the constructed SLDS matrix B'. It can be observed from Eq. 8 that the SLDS matrix B' is mostly sparse, i.e. only a few transitions are allowed between the states in $\mathcal{LC}$. In fact, only $|L| + 1$ transitions allowed for every lc state. The allowable transitions include the resets to $|L|$ labels and one increment transition. Hence, we can achieve an overall performance of $O(TD_{max}|L|^2)$ via exploiting this fact, which results in reduced overhead by a factor of $O(D_{max})$. The number is derived from the fact that there are total $O(D_{max}|L|)$ states at time $t-1$, and the number of transitions allowed for each state to time t reduces to $O(|L|)$ from $O(D_{max}|L|)$. This reduction in complexity allows us to incorporate a duration model with a large D_{max} and maintain computational efficiency. As a consequence, we can adopt the more powerful duration modeling capabilities of an S-SLDS at the cost of a modest complexity increase over the standard SLDS model.

Note that the Markov Chain properties in S-SLDS have a very regular structure which could be potentially exploited in the proposal distribution of a DD-MCMC sampler. Efficient application of DD-MCMC to the S-SLDS model structure is an interesting topic for future research.

6 The Parametric SLDS Model

As discussed in Section 2.3, the standard SLDS does not provide a principled means to quantify global variations in the motion patterns. For example, honey bees communicate the orientation and distance to food sources through the (spatial) dance angles and (temporal) waggle durations of their stylized dances which take place in the hive. As a result, these global motion parameters which encode the message of the bee dance are the variables that we are most interested in estimating.

Moreover, it should be noted that the labeling and quantification problems are not independent. For example, it can be observed in Fig. 1(a) that an angle estimate which is very close to the ground truth would provide a strong cue for the labeling of the overall motions.

In this section, we present a parametric SLDS (P-SLDS) model which makes it possible to quantify the global variables and solve both labeling and quantification problems in an iterative manner. The resulting P-SLDS learns canonical behavioral templates from data with additional information on the associated global parameters. P-SLDS effectively decodes the global

[1]Examples include approximate Viterbi methods [45] and variational methods [45, 22, 39] which require the computations between all possible state pairs from the previous time-step to the next time-step.

parameters while it simultaneously labels the sequences. This is done using an expectation-maximization (EM) algorithm [13, 35], presented in detail below.

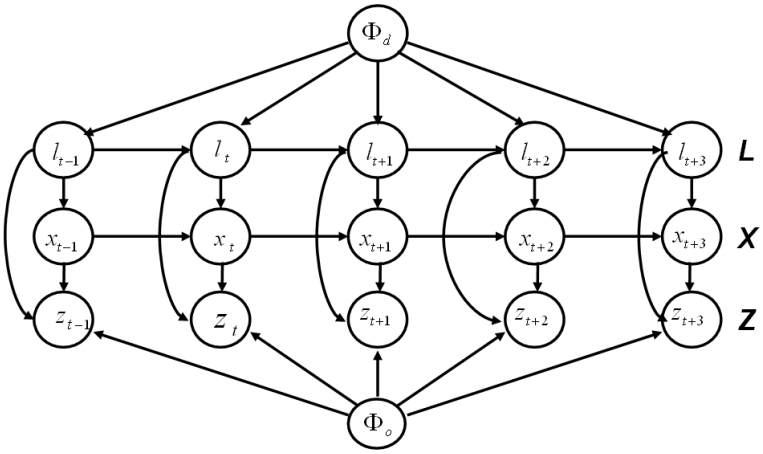


Figure 14: Parametric SLDS (P-SLDS)

6.1 Graphical representation of P-SLDS

In P-SLDSs, the discrete state transition probabilities and output probabilities are parameterized by a set of global parameters $\Phi = \{\Phi_d, \Phi_o\}$. The parameters Φ are global in that they systematically affect the entire sequence. The graphical model of P-SLDS is shown in Fig. 14. Note that there are two classes of global parameters : the dynamics parameters Φ_d and the observation parameters Φ_o.

The *dynamics parameters* Φ_d represent the factors that cause temporal variations. The different values of the dynamics parameters Φ_d result in different switching behavior between behavioral modes. In the case of the honey bee dance, a food source that is far away leads a dancer bee to stay in each dance regime longer, resulting in a dance with larger radius which will show less frequent transitions between dance regimes. In terms of S-SLDS model, the global dynamics parameters are associated with duration models. In contrast, the *observation parameters* Φ_o represent factors that cause spatial variations. A good example is a pointing gesture, where the indicating direction changes the overall arm motions. In the honey bee dance case, one can consider standard SLDS as a behavioral template that can be stretched in time by global dynamic parameters Φ_d and spatially rotated by global observation parameters Φ_o.

The common underlying behavioral template is defined by canonical parameters Θ. Note that the canonical parameters Θ are embedded in the conditional dependency arcs in Fig. 14. In the bee dancing example, the canonical parameters describe the prototyped stylized bee dance. However, the individual dynamics in the different bee dances systematically vary from the prototyped dance due to the changing food source locations which are represented by the global parameters Φ.

Notice that the discrete state transitions in the top chain of Fig. 14 are instantiated by Θ and Φ_d, and the observation model at the bottom is instantiated by Θ and Φ_d while the continuous state transitions in the middle chain are instantiated solely by the canonical parameters Θ. In other words, the dynamics parameters Φ_d, vary the prototyped switching behaviors, and the observation parameters Φ_o vary the prototyped observation model. The intuition behind the quantification of global parameters is that they can be effectively discovered by finding the global parameters that best describe the discrepancies between the new observations and the behavioral template. In other words, the global parameters are estimated by minimizing the residual error that remains between the template and the observation sequence.

The result of *parameterizing* the SLDS model is the incorporation of additional conditioning variables in the initial state distribution $P(l_1|\Theta, \Phi_d)$, the discrete state transition table $P(l_t|l_{t-1}, \Theta, \Phi_d)$, and the observation model $P(z_t|l_t, x_t, \Theta, \Phi_o)$. There are three possibilities for the nature of the parameterization: (a) the PDF is a linear function of the global parameters Φ, (b) the PDF is a non-linear function of Φ, and (c) no functional form for the PDF is available. In the latter case of (c),

Algorithm 1 EM1 for Learning in P-SLDS

- E-step 1: obtain the posterior distribution

$$f_L^i(X) \stackrel{\Delta}{=} P(X|\Theta^i, \bar{D}) \tag{9}$$

over the hidden state sequence X, based on a current guess of the canonical parameters Θ^i.

- M-step 1: maximize the expected log-likelihood :

$$\Theta^{i+1} \leftarrow \underset{\Theta}{\text{argmax}} \ \left\langle \log P(\bar{L}, X, \bar{Z}|\Theta, \bar{\Phi}) \right\rangle_{f_L^i(X)} \tag{10}$$

general function approximators such as neural network may be used, as suggested in [55]. In the Section 6.2 and 6.3, we discuss learning and inference in P-SLDS under the assumption that functional forms are available.

We assume the global parameters are available during the learning phase when the behavior model is considered. However, during the testing phase, we are given only the observation sequence and we estimate the global parameters Φ jointly with L.

6.2 Learning in P-SLDS

In the learning phase, P-SLDS learns a canonical behavior template from motion data where the individual dynamics may vary due to different underlying global parameters, but we assume that these parameters are known in our training data. Learning in P-SLDS entails estimating the P-SLDS canonical parameters Θ, given the data $\bar{D} \stackrel{\Delta}{=} \left\{\bar{\Phi} = \{\bar{\Phi}_d, \bar{\Phi}_o\}, \bar{L}, \bar{Z}\right\}$ where the data $\bar{D}$ comprises a set of global parameters $\bar{\Phi} = \{\bar{\Phi}_d, \bar{\Phi}_o\}$, a label sequence $\bar{L}$, and the observations $\bar{Z}$. The upper bars indicate that the values are known. We employ EM [13, 35] with the continuous state X as the only hidden variables to find an ML estimate of the canonical parameters $\hat{\Theta}$.

The E-step in Eq. 9 is equivalent to inference in an LDS model. In more detail, since the global parameters $\bar{\Phi}$, the current P-SLDS parameters Θ^i, the label sequence $\bar{L}$, and the observations $\bar{Z}$ are all known, inference over the continuous hidden states X in E-step can be performed through Kalman-smoothing [4]. Given the posterior distribution $f_L^i(X)$ in Eq. 9, we then update the parameters Θ^{i+1} as in Eq. 10.

In case the parameterized dependencies such as $P(l_t|l_{t-1}, \Theta, \Phi_d)$ are linear functions of the global parameters Φ, the M-step in Eq. 10 can be solved analytically. However, in the case where the parametric dependencies are non-linear, an exact M-step is infeasible and must be solved by alternative optimization methods, e.g., conjugate gradient or Levenberg-Marquardt methods.

6.3 Inference in P-SLDS

We use the learned P-SLDS canonical parameters Θ to perform the quantification of the global parameters Φ and inference on the label sequence L , given the observations $\bar{Z}$. Note that the canonical parameter set Θ is fixed once they are learned from the training dataset $\bar{D}$, and we now interpret a novel dataset via inference where neither the global parameters Φ nor the label sequence L are known.

We use EM to quantify the optimal global parameters Φ as shown in Algorithm 2. Note that we use Algorithm 1 to learn the canonical model parameters Θ, while Algorithm 2 is used to estimate the global parameters Φ with simultaneous inference of the labels L. More details on the EM algorithm in Algorithm 2 are described below. In the following sections, we use the abbreviation "$\mathcal{LLH}$" to denote log-likelihood.

6.3.1 E-step 2

The exact E-step in Eq. 12 is proved to be intractable [31]. Thus, we need to rely on the approximate inference methods. Here, we present a derivation of E-step based on approximate Viterbi (VI) method [43]. Note that our derivation can be

Algorithm 2 EM2 for Inference in P-SLDS

- E-step 2 : obtain the posterior distribution:

$$f_I^i(L,X) \stackrel{\Delta}{=} P(L,X|\bar{Z},\Theta,\Phi^i) \tag{12}$$

  over the hidden label sequence L and the state sequence X, using a current guess for the global parameters Φ^i.

- M-step 2 : maximize the expected log-likelihood:

$$\Phi^{i+1} \leftarrow \underset{\Phi}{\text{argmax}} \ \left\langle \log P(L,X,\bar{Z}|\Theta,\Phi) \right\rangle_{f_I^i(L,X)} \tag{13}$$

extended in a straight-forward way to other approximate inference methods. At the ith EM iteration, the joint posterior over the hidden variables L and X is approximated by a peaked posterior over X with the Viterbi decoded label sequence $\hat{L}^i$:

$$\begin{aligned} P(L,X|\bar{Z},\Phi^i) &= P(X|L,\bar{Z},\Phi^i)P(L|\bar{Z},\Phi^i) \\ &\approx P(X|\hat{L}^i,\bar{Z},\Phi^i)\delta(\hat{L}^i) \end{aligned} \tag{11}$$

$$f_I^i(X) \stackrel{\Delta}{=} P(X|\hat{L}^i,\bar{Z},\Phi^i)\delta(\hat{L}^i)$$

We omit the implicit conditional dependence on the fixed canonical parameters $\Theta$for clarity.

6.3.2 M-step 2

Using the approximate posterior $f_I^i(X)$ obtained in Eq. 11, the expected complete log-likelihood ($\mathcal{LLH}$) in Eq. 13 is approximated as:

$$\begin{aligned} \mathcal{L}^i(\Phi) &\stackrel{\Delta}{=} \sum_L \int_X \log P(L,X,\bar{Z}|\Phi) P(L,X|\bar{Z},\Phi^i) \\ &\approx \int_X \log P(\hat{L}^i,X,\bar{Z}|\Phi) f_I^i(X) \end{aligned} \tag{14}$$

Using the chain rule, this factors as:

$$P(\hat{L}^i,X,\bar{Z}|\Phi) = P(\hat{L}^i|\Phi_d)P(X,\bar{Z}|\hat{L}^i,\Phi_o) \tag{15}$$

Note that we now only condition on relevant global parameters, e.g. the label sequence $\hat{L}^i$ is only conditioned on Φ_d. Substituting (15) into the expected $\mathcal{LLH}$ $\mathcal{L}^i(\Phi)$ in (14), we obtain a more succinct form of $\mathcal{L}^i(\Phi)$ in which the term $\log P(\hat{L}^i|\Phi_d)$ is moved outside the integral:

$$\begin{aligned} \mathcal{L}^i(\Phi) &= \log P(\hat{L}^i|\Phi_d) + \int_X \log P(X,\bar{Z}|\hat{L}^i,\Phi_o) f_I^i(X) \\ &= \mathcal{L}^i(\Phi_d) + \mathcal{L}^i(\Phi_o) \end{aligned} \tag{16}$$

Here we introduced two convenience terms, the dynamic log-likelihood $\mathcal{L}(\Phi_d)$ and the observation log-likelihood $\mathcal{L}(\Phi_o)$:

$$\mathcal{L}^i(\Phi_d) \stackrel{\Delta}{=} \log P(\hat{L}^i|\Phi_d) \tag{17}$$

$$\mathcal{L}^i(\Phi_o) \stackrel{\Delta}{=} \int_X \log P(X,\bar{Z}|\hat{L}^i,\Phi_o) f_I^i(X) \tag{18}$$

In Eq. 16, we can observe that the total expected $\mathcal{LLH}$ $\mathcal{L}^i(\Phi)$ is maximized by independently updating the global observation parameters Φ_o and dynamic parameters Φ_d, i.e. we obtain the updated global parameters Φ_d^{i+1} and Φ_o^{i+1} by maximizing the dynamic $\mathcal{LLH}$ $\mathcal{L}^i(\Phi_d)$ and the the observation $\mathcal{LLH}$ $\mathcal{L}^i(\Phi_o)$ respectively.

Now we can further factorize the dynamic $\mathcal{LLH}$ $\mathcal{L}^i(\Phi_d)$ in Eq. 17 and the observation $\mathcal{LLH}$ $\mathcal{L}^i(\Phi_o)$ in Eq. 18. Then, we obtain the fully factorized $\mathcal{LLH}$ terms as :

$$\mathcal{L}^i(\Phi_d) = \log P(\hat{l_1}^i|\Phi_d) + \log \sum_{t=2}^{|Z|} P(\hat{l_t}^i|\hat{l}_{t-1}^i\Phi_d) \tag{19}$$

$$\begin{aligned}\mathcal{L}^i(\Phi_o) &= \int_X \log \left\{ P(\bar{Z}|X, \hat{L}^i, \Phi_o) P(X|\hat{L}^i) \right\} f_I^i(X) \\ &\equiv \int_X \log P(\bar{Z}|X, \hat{L}^i, \Phi_o) f_I^i(X) \\ &= \sum_{t=1}^{|Z|} \int_{x_t} \log P(\bar{z}_t|x_t, \hat{l_t}^i, \Phi_o) f_I^i(x_t)\end{aligned} \tag{20}$$

Above, the term $f_I^i(x_t)$ denotes the marginal on x_t from the full posterior $f_I^i(X)$, i.e. $f_I^i(x_t) \triangleq \int_{X/x_t} f_I^i(X)$. Note that the term $\int_X \log P(X|\hat{L}^i) f_I^i(X)$ disappears in the second line of Eq. 20 as it is not a function of the global observation parameter Φ_o and does not help to improve the likelihood $\mathcal{L}^i(\Phi_o)$. In the case where we are modeling data with parametric S-SLDS models, the global dynamic parameters Φ_d are associated with the duration models of the S-SLDS, and Eq. 19 is not directly applicable because label transitions occur between segments. Hence, once we obtain the Viterbi labels $\hat{L}^i$, the label sequence is converted into a list of segments, i.e., $\hat{L}^i = \cup_{j=1}^{|s|} s_j$ where $s_j \triangleq (l_j, d_j)$, as described in Section 5.1. Then, the dynamic $\mathcal{LLH}$ for parametric S-SLDSs can be evaluated as follows :

$$\begin{aligned}\mathcal{L}^i(\Phi_d) &= \sum_{j=1}^{|s|} \log P(s_j|\Phi_d) \\ &= \sum_{j=1}^{|s|} \log D_{l_j}(d_j)\end{aligned} \tag{21}$$

The observation $\mathcal{LLH}$ for parametric S-SLDS is evaluated as in Eq. 20. The details of the M-step will depend upon the application domain. In the case where the parametric forms are linear in the global parameters Φ, the M-step is obtained analytically. Otherwise, alternative optimization methods can be used to maximize the non-linear $\mathcal{LLH}$ function, as described in Section 6.2.

7 Modeling the Honey Bee Dance

We describe a model of the honey bee dance based on our parameterized segmental SLDS (PS-SLDS) model. The bee dance is parameterized by both classes of global parameters. The global dynamics parameter set $\Phi_d \triangleq \{\Phi_{d,l} | 1 \leq l \leq n\}$ is chosen to be correlated with the average duration of each dance regime, where $n = 3$. The global observation parameter Φ_o is chosen to be the angle orientation of the bee dance.

7.1 Canonical parameters

The canonical parameters in honey bee dances are : a tuple of initial label distribution π, semi-Markov segmental Markov transition matrix $\tilde{B}$, LDS model parameters M and variances in durations in each behavioral modes Σ :

$$\Theta \triangleq \left\{\pi, \tilde{B}, M \triangleq \{M_l | 1 \leq l \leq n\}, \Sigma \triangleq \{\Sigma_l | 1 \leq l \leq n\}\right\}$$

Note that the canonical parameter tuple Θ is fixed once it is learned from data, as mentioned in Section 6. The choice of canonical parameters are based on the knowledge of honey bee dances [20]. For example, it is reasonable to assume that the initial label distribution π and the segment label transition matrix $\tilde{B}$ between different dance regimes do not vary across the dance sequences. In addition, the dancer bees try to regulate their waggle durations to convey the dance messages effectively. Hence, the amount of variation in the duration of each dance regime is assumed to be constant. Hence, they are learned and represented as the variances Σ.

7.2 Dynamics model

We set the global dynamic parameters of the l th model $\Phi_{d,l}$ to be the average duration μ_l of lth dance regime, i.e., $\Phi_d \triangleq \{\mu_l | 1 \leq l \leq n\}$. Accordingly, each parameterized duration model D_l of S-SLDSs is modeled as a Gaussian distribution as follows :

$$D_l(c_t) = \mathcal{N}(\mu_l; \Sigma_l) \tag{22}$$

Above, the duration mean μ_l is a global dynamic parameter which is re-estimated at every EM iteration in P-SLDS learning (described in Section 6.3) while the variance Σ_l is a fixed canonical parameter. Then, the explicit duration model in Eq. 22 is discretized into a histogram with maximum duration length $D_l^{max} = 100$. In the video database, a dance regime with extremely long duration lasted for about 75 frames. Thus, the choice of the maximum duration length D_l^{max} would be sufficient to represent the duration model. Once the histogram duration model D_l is learned, we convert the model into an NSTF U_l , as discussed in Section 5.2.

The M-step update for a dynamics parameter $\Phi_{d,l}$ can be obtained by differentiating the dynamic $\mathcal{LLH}$ in Eq. 23 :

$$\begin{aligned}
\mathcal{L}^i(\Phi_d) &= \sum_{j=1}^{|s|} \log D_{l_j}(d_j) \\
&= \sum_{l=1}^{N} \left(\sum_{\forall l_j = l} \log D_l(d_j) \right) \\
&= -\frac{1}{2} \sum_{l=1}^{N} \left(\sum_{\forall l_j = l} \log \Sigma_l + \frac{(d_j - \mu_l)^2}{\Sigma_l} \right)
\end{aligned} \tag{23}$$

$$\frac{\partial \log P(\hat{L}|\Phi_d)}{\partial \mu_l} = \frac{2 \sum_{\forall l_j = l} (d_j - \mu_l)}{\Sigma_l} = 0$$

$$\mu_l^{new} \leftarrow \frac{\sum_{\forall l_j = l} d_j}{|s_l|} \tag{24}$$

In fact, the M-step update in Eq. 24 for the global dynamic parameters $\mu^{new} \triangleq \{\mu_l^{new} | 1 \leq l \leq n\}$ turns out to be equivalent to re-estimating the mean durations of distinct dance phases from the estimated segmented label sequence $\hat{L}^i = \cup_{j=1}^{|s|} s_j$.

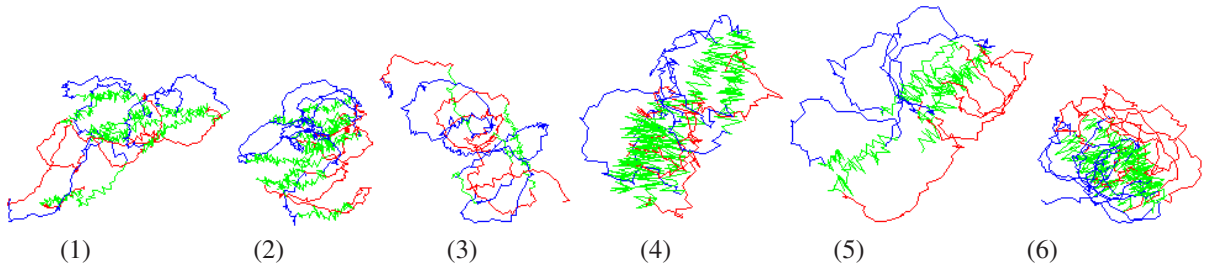


Figure 15: Bee dance sequences used in the experiments. Each dance trajectory is the output of a vision-based tracker. Tables 1 and 2 give the global motion parameters for each of the numbered sequences. Key : waggle, right-turn, left-turn.

7.3 Observation model

The observation data are time-series sequences of vectors $z_t = [x_t, y_t, \cos(\theta_t), \sin(\theta_t)]^T$ where (x_t, y_t) and θ_t denote, respectively, the 2D coordinates and the heading angle of the tracked dancer bee at time t. The angle of zero corresponds to the direction of the positive x-axis and increases in the clockwise direction. The trigonometric function elements in the observations were introduced to bound the effects of angular factors within $[-1, 1]$. Note that the observed temporary heading angle θ_t differs from the global dance angle Φ_o.

We use the following parameterized observation model $P(z_t|l_t, x_t, \Phi_o)$,

$$z_t \sim \mathcal{N}(R(\Phi_o)H_{\hat{l}_t}x_t, V_{\hat{l}_t}) \tag{25}$$

where $R(\Phi_o)$ is the rotation matrix, and $H_{\hat{l}_t}$ and $V_{\hat{l}_t}$ denote the observation parameters of the $\hat{l}_t$th component LDS, corresponding to label $\hat{l}_t$ of the Viterbi sequence $\hat{L}$. We also define $\alpha_t(\Phi_o)$ to denote the projected-then-rotated vector of the corresponding state x_t:

$$\alpha_t(\Phi_o) \triangleq R(\Phi_o)H_{l_t}x_t \tag{26}$$

Combining terms, we obtain the observation $\mathcal{LLH}\ \mathcal{L}^i(\Phi_o) \equiv$

$$-\sum_{t=1}^{|Z|} \left\langle [z_t - \alpha_t(\Phi_o)]^T V_{\hat{l}_t}^{-1} [z_t - \alpha_t(\Phi_o)] \right\rangle_{f_l^i(x_t)} \tag{27}$$

where we have omitted redundant constant terms. Intuitively, the optimization in (27) is to find an updated dance angle Φ_o^{i+1} which minimizes the sum of the expected Mahalanobis distances between the observations z_t's and the projected-then-rotated states $\alpha_t(\Phi_o)$'s. However, since the non-linearities are involved due to the rotation, there is no analytical solution to this maximization problem. Specifically, Eq. 27 involves quadratic trigonometric function terms such as $sin(\Phi_o)^2$. Thus, we perform 1D gradient ascent on the obtained functional and it is still guaranteed to increase the likelihood of the model in the spirit of Generalized EM [37].

8 Experimental Results

The experimental results show that PS-SLDS provides reliable global parameter quantification capabilities along with improved recognition abilities in comparison to the standard SLDS. The six dancer bee tracks obtained from the videos are shown in Fig. 15. Sample output from our vision-based tracker [26] is shown in Fig. 1(b), where the dancer bee is automatically tracked inside the white rectangle.

We performed experiments with 6 video sequences[2] with length 1058, 1125, 1054, 757, 609 and 814 frames, respectively. Once the sequence observations Z were obtained, the trajectories were preprocessed so that the mean of each track is located at (100,100). Note from Fig. 15 that the tracks are noisy and much more irregular than the idealized stylized dance prototype shown in Fig. 1(a). The red, green and blue colors represent right-turn, waggle and left-turn phases. The ground-truth labels are marked manually for comparison and learning purposes. The dimensionality of the continuous hidden states was set to four.

We adopted a leave-one-out (LOO) strategy for evaluation. The parameters are learned from five out of six datasets, and the learned model is applied to the left-out dataset to perform labeling and simultaneous quantification of angle/average waggle duration. Six experiments were performed using both PS-SLDS and the standard SLDS, so that we test on each sequence once. The PS-SLDS estimates of angle/average waggle durations (AWD) are directly obtained from the results of global parameter quantification. On the other hand, the SLDS estimates are heuristically obtained by averaging the transition numbers or averaging the heading angles from the inferred "waggle" segments.

8.1 Learning from Training Data

The parameters of both PS-SLDS and standard SLDS are learned from the data sequences depicted in Fig. 15. The standard SLDS model parameters were learned as described in Section 3.3. The canonical parameters tuple described in Section 7.1 are all learned solely based on the observations Z without any restriction on the parameter values. However, the prior distribution π on the first label was set to be a uniform distribution.

To learn the PS-SLDS model parameters, the ground truth waggle angles and AWDs were evaluated from the data. Then, each sequence was preprocessed (rotated) in such a way that the waggles head in the same direction based on the evaluated ground truth waggle angles. This preprocessing was performed to allow the PS-SLDS model to learn the canonical parameters which represent the behavioral template of the dance. Note that six sets of model parameters are learned through the LOO approach and the global angle of the test sequence is not known a priori during the learning phase. In addition to the model parameters learned by the standard SLDS, PS-SLDS learns additional duration models D, and semi-Markov transition matrix $\tilde{B}$, as described in Section 5.

8.2 Inference on Test Data

During the testing phase, the learned parameter set was used to infer the labels of the left-out test sequence. An approximate Viterbi (VI) method [43, 45] and variational approximation (VA) methods [22, 43, 45, 39] were used to infer the labels in standard SLDSs. The initial probability distributions for the VA method were initialized based on the VI labels. Our initialization scheme assigned VI labels a probability of 0.8 and the other two labels at every time-step were assigned probabilities of 0.1. We used the VI method due to its simplicity and speed. Our experiments compare the performance of SLDS and PS-SLDS models based on VI and VA methods.

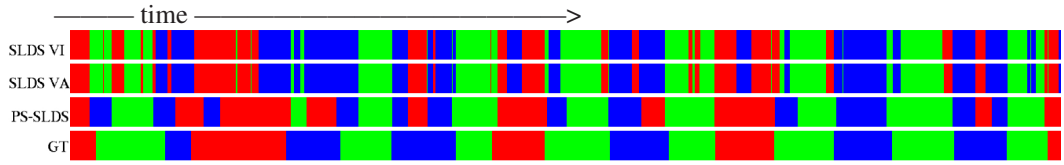
8.3 Qualitative Results

Our experimental results demonstrate the superior recognition capabilities of the proposed PS-SLDS model over the original SLDS model. The label inference results on all data sequences are shown in Fig. 16. The four color-coded strips in each figure represent SLDS VI, SLDS VA, PS-SLDS VI and the ground-truth labels from the top to the bottom. The x-axis represents time flow and the color is the label at that corresponding video frame.

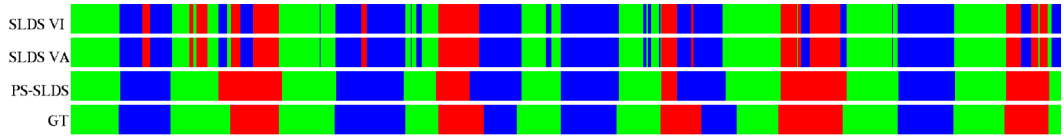
The superior recognition abilities of PS-SLDS can be observed from the presented results. The PS-SLDS results are closer to the ground truth or comparable to SLDS results in all sequences. In particular, the sequences 1, 2 and 3 are challenging because the tracking results obtained from the vision-based tracker are more noisy. In addition, the patterns of switching between dance modes and the durations of the dance regime are more irregular than the other sequences.

It can be observed that most of the over-segmentations that appear in the SLDS labeling results disappear in the PS-SLDS labeling results. PS-SLDS estimates still introduce some errors, especially in sequences 1 and 3. However, keeping in mind that even a human expert can introduce labeling noise, the labeling capabilities of PS-SLDS are fairly good.

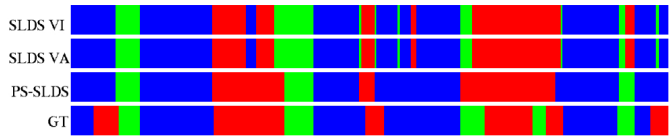
[2]The experimental data used in this work are available at : `http://borg.cc.gatech.edu/biotracking`



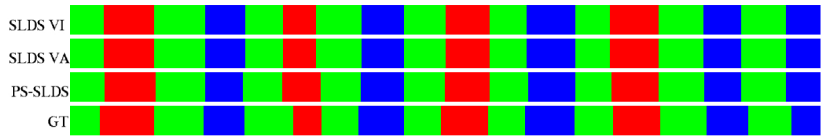
(a) Sequence 1



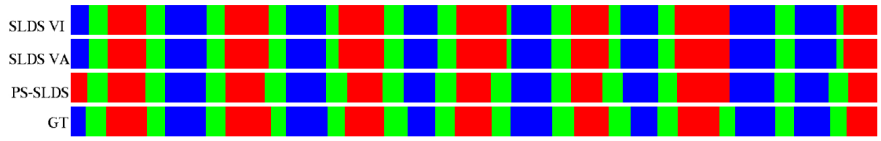
(b) Sequence 2



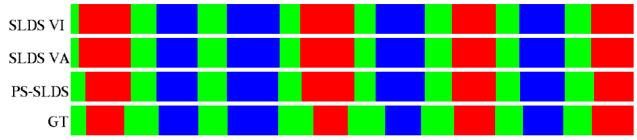
(c) Sequence 3



(d) Sequence 4



(e) Sequence 5



(f) Sequence 6

Figure 16: Label inference results. Estimates from SLDS and P-SLDS models are compared to manually-obtained ground truth (GT) labels. Key : waggle, right-turn, left-turn.

8.4 Quantitative Results

The quantitative results on the angle/average waggle duration quantification show the robust global parameter quantification capabilities of PS-SLDS. Table. 1 shows (from top to bottom) : the absolute errors of the PS-SLDS estimate along with the error rates (%) in parenthesis, SLDS estimates based on the VI and the VA methods, and the ground truth angle. The best estimates are accented in bold fonts. The SLDS estimates are obtained by the heuristic of averaging the heading angles in the sequences that were labeled as "waggle" in the inference step. All of the error values are the difference between estimated results and known ground truth values.

Based on the six tests, PS-SLDS and SLDS show comparable waggle angle estimation capabilities. There is no distinguishable gap in performance between VI and VA methods. Our hypothesis is that the over-segmentation errors do not effect the waggle angle estimate as much as it effects average waggle duration estimates. Note that the maximum error of PS-SLDS angle estimate was 0.11 radians for the fifth dataset, which is fairly good considering the noise in the tracking results.

The quantitative results on average waggle duration (AWD) quantification show the advantages of PS-SLDS in quantifying the global dynamics parameters of interest. AWD is an indicator of the distance to the food source from the hive and is valuable data for insect biologists. Table. 2 shows (from top to the bottom) : the absolute errors and error rates of the PS-SLDS estimates, the SLDS estimates of VI and VA methods and the ground truth AWDs. Again, the best estimates are marked in bold fonts where PS-SLDS estimates are consistently superior to the SLDS estimates. The SLDS estimates are obtained by evaluating means of the waggle durations in the inferred segments. The results again show that PS-SLDS estimates match the ground-truth closely. In particular, we want to highlight the quality of the PS-SLDS AWD estimates for sequences 2, 3, 4 and 5. In contrast, the SLDS estimates in these cases are inaccurate. More specifically, the SLDS estimates deviate far from the ground truth in most cases except for the sequence 4. The reliability of AWD estimates of PS-SLDS model show the benefit of the duration modeling and the canonical parameters supported by the enhanced models.

Finally, Table 3 shows the overall accuracy of the inferred labels for the PS-SLDS, SLDS DD-MCMC, SLDS VI, and SLDS VA results. It can be observed that PS-SLDS provides very accurate labeling results w.r.t. the ground truth. Moreover, PS-SLDS consistently improves upon the standard SLDSs across all six datasets. The overall experimental results show that PS-SLDS model is promising and provides a robust framework for the bee application. It should be noted that SLDS DD-MCMC is the most computationally intensive method, and PS-SLDS still improves on SLDS DD-MCMC in a consistent manner.

8.5 Discussion

It can be difficult to choose the right dimensionality for the hidden continuous states X. In our experiments, dimensions less than four resulted in poor classification. It is conjectured that such small dimensions do not provide hypothesis space rich enough to represent the motion patterns of dancer bees. On the other hand, some experiments with higher dimension (>10) suffered from over-segmentation when the model was trained using the limited available training data.

In addition, we investigated the synthesis capabilities of the model. The learned SLDSs and PS-SLDSs were used to generate honey bee dances. However, the synthesis results were not satisfactory. It is expected that more realistic trajectories can be generated with higher dimensional models (≥ 4) although the limited amount of training data is the major impediment to better generalization. The on-going work on the synthesis power of the model remains as a future work.

9 Conclusion

In this paper, we addressed the problem of learning and inferring behavioral patterns of a target based on tracked video data. In our approach, SLDSs are investigated as a promising framework to model complex motions. Accordingly, the labeling and quantification problems in computer vision are framed as learning and inference problems.

In our work, we encountered three challenges in developing an effective modeling system based on a standard SLDS model : (1) intractability of inference, (2) limited duration modeling, and (3) absence of principled means for quantification. We addressed these issues by introducing three extensions of the standard SLDS paradigm.

First, we addressed the intractability of inference in SLDSs by introducing a novel data-driven MCMC (DD-MCMC) method. The proposed method can effectively discover the true posterior over the hidden labels.

Sequence	1	2	3	4	5	6
PS-SLDS	0.09 (30)	**0.01** (4)	0.03 (3)	0.11 (8)	**0.11** (5)	**0.06** (8)
SLDS VI	**0.05** (16)	0.03 (12)	**0.02** (2)	**0.09** (7)	0.18 (9)	0.09 (11)
SLDS VA	**0.05** (16)	0.03 (12)	**0.02** (2)	**0.09** (7)	0.18 (9)	0.09 (11)
Ground Truth	-0.30	-0.25	1.13	-1.33	-2.08	-0.80

Table 1: Absolute errors in the global rotation angle estimates from PS-SLDS and SLDS in radians. The numbers in parenthesis are error rates (%). Last row contains the ground truth rotation angles. Sequence numbers refer to Fig. 15.

Sequence	1	2	3	4	5	6
PS-SLDS	**13.7** (27)	**0.91** (2)	**1.9** (9)	**0.22** (<1)	**0.4** (2)	**5.6** (17)
SLDS VI	40.8 (79)	28.9 (62)	11.1 (52)	0.44 (1)	3.6 (19)	8 (25)
SLDS VA	40.7 (79)	28.9 (62)	11.1 (52)	0.44 (1)	3.6 (19)	8 (25)
Ground Truth	51.6	46.6	21.4	41.1	19.4	32.6

Table 2: Absolute errors in the Average Waggle Duration (AWD) estimates for PS-SLDS and SLDS in frames. The numbers in parenthesis are error rates (%). Last row contains the ground truth AWD. Sequence numbers refer to Fig. 15.

Sequence	1	2	3	4	5	6
PS-SLDS	**75.9**	**92.4**	**83.1**	**93.4**	**90.4**	**91.0**
SLDS DD-MCMC	74.0	86.1	81.3	**93.4**	90.2	90.4
SLDS VI	71.6	82.9	78.9	92.9	89.7	89.2
SLDS VA	71.6	82.8	78.9	92.9	89.7	89.2

Table 3: Accuracy of label inference in percentage. Sequence numbers refer to Fig. 15.

Second, we presented a segmental SLDS model to enhance the duration modeling capabilities of SLDSs. The proposed S-SLDS can incorporate arbitrary duration models which are not supported by the standard SLDS model. Nonetheless, we also demonstrated that the proposed S-SLDS model can be converted into an equivalent standard SLDS model by introducing meta-variables. This conversion ensures that the large array of approximate inference algorithms developed for standard SLDSs can be applied in S-SLDSs.

Third, parametric SLDS (P-SLDS) is introduced as an extension to provide a systematic means to quantify the global parameters which induce systematic temporal and spatial variations in the motion. The proposed model can simultaneously infer the hidden labels and the global parameters in an iterative manner via the EM algorithm.

Finally, we presented experimental results on real-world honey bee dance sequences, where the honey bee dances were modeled using a parametric segmental SLDS (PS-SLDS) model, i.e. combination of P-SLDS and S-SLDS. Both the qualitative and quantitative results show that the enhanced SLDS model can robustly infer the labels and global parameters. A large number of over-segmentations in labeling which appeared in standard SLDSs are not present in the PS-SLDS results. In addition, the results on the quantification abilities of PS-SLDS show that PS-SLDS can provide estimates which are very close to the ground truth . It was also shown that PS-SLDS consistently improves on SLDSs in overall accuracy. The consistent results show that PS-SLDS improves upon SLDS for the honey bee dance data and suggest that they may be promising for other applications.

We have demonstrated that both DD-MCMC and PS-SLDS provide significant performance improvements over SLDS in analyzing bee dance data. It would be interesting to investigate the benefits of these methods in higher dimensional data spaces and additional problem domains. We plan to address this in our future work.

Acknowledgments

The authors would like to thank Zia Khan and Grant Schindler for providing the bee tracking data. Portions of this work were supported in part by Samsung Lee Kun Hee scholarship awarded to Sang Min Oh, NSF Award IIS-0433012 awarded to James M. Rehg, and NSF Award IIS-0219850 awarded to Tucker Balch and Frank Dellaert, and NSF CAREER Award IIS-0448111 awarded to Frank Dellaert. We thank the reviewers for their insightful comments which helped to improve the presentation of this work.

Appendix A. Data-Driven MCMC for SLDS

A. 1. Metropolis Hastings

Data-driven MCMC method adopts the Metropolis-Hastings (MH) framework [36, 24] to generate samples from arbitrary distributions. The pseudo-code for the MH algorithm is shown in Algorithm 3 (adapted from [23]).

Algorithm 3 Pseudo-code for Metropolis-Hastings (MH)

1. Start with a valid initial label sequence $L^{(1)}$.
2. Propose a new label sequence $L^{(r)'}$ from $L^{(r)}$ using a *proposal density* $Q(L^{(r)'}; L^{(r)})$.
3. Calculate the *acceptance ratio*
$$a = \frac{P(L^{(r)'}|Z)}{P(L^{(r)}|Z)} \frac{Q(L^{(r)}; L^{(r)'})}{Q(L^{(r)'}; L^{(r)})} \tag{28}$$
where $P(L|Z)$ is the *target distribution*.
4. If $a >= 1$ then accept $L^{(r)'}$, i.e., $L^{(r+1)} \leftarrow L^{(r)'}$.
Otherwise, accept $L^{(r)'}$ with probability $\min(1, a)$. If the proposal is rejected, then we keep the previous sample, i.e., $L^{(r+1)} \leftarrow L^{(r)}$.

Intuitively, step 2 proposes "moves" from the previous sample $L^{(r)}$ to the next sample $L^{(r)'}$ in the space of label sequences L, which is driven by a proposal distribution $Q(L^{(r)'}; L^{(r)})$. The evaluation of a and the acceptance mechanism in steps 3 and 4 have the effect of modifying the transition probabilities of the chain in such a way that its stationary distribution is exactly $P(L|Z)$.

A. 2. Learning

In the learning phase, we collect temporal cues from the training data. Then, a set of models of cues which we call 'label-cue models' are constructed based on the collected cues, i.e. $\{P(c|l_i)|1 \leq i \leq n\}$. By a temporal cue c_t, we mean a cue at time t that can provide a guess for the corresponding label l_t. A cue c_t is a certain statistic obtained by observing the data within the fixed time range of z_t. For example, the change of angles are collected as temporal cues in the bee application, as illustrated in Fig. 5.

Then, a set of n label-cue models $LC \triangleq \{P(c|l_i)|1 \leq i \leq n\}$ are learned from the classified cues where the cues are classified with respect to the training labels. Here, n corresponds to the number of existing patterns, the number of LDSs in our case. Each label-cue model $P(c|l_i)$ is an estimated generative model and describes the distribution of cue c given the label l_i. The learned label-cue models are used later in the inference phase to construct a proposal prior.

A. 3. Inference

In the inference phase, we first collect the temporal cues from the test data without access to the labels. Then, the learned label-cue models are applied to the cues and the proposal priors are constructed. A proposal prior $P(\tilde{l}_t|c_t)$ is a distribution

on the labels, which is a rough approximation to the true posterior $P(l_t|Z)$. When a cue c_t is obtained from a test data, we construct a corresponding proposal prior $P(\tilde{l_t}|c_t)$ as follows :

$$P(\tilde{l_t}|c_t) \quad \triangleq \quad \frac{P(c_t|l_i)}{\sum_{i=1}^{n} P(c_t|l_i)} \tag{29}$$

Above, a proposal prior $P(\tilde{l_t}|c_t)$ is obtained from the normalized likelihoods of all labels. The prior describes the likelihood that each label generates the cue. By evaluating all the proposal priors across the test sequence, we obtain a full set of proposal priors $P(\tilde{L}) \triangleq \{P(\tilde{l_t}|c_t)|1 \leq t \leq T\}$ over the entire label sequence. However, the resulting proposal priors were found to be sensitive to the noise in the data. Thus, we smooth the estimates and use the resulting distribution. The proposed approach is depicted graphically in Fig. 5,6,7 for the case of the bee dance domain.

The proposal priors $P(\tilde{L})$ and the SLDS discrete Markov transition PDF B constitute the data-driven proposal Q. While the proposal priors $P(\tilde{L})$ provide the data-driven characteristics, the Markov PDF B adds the model characteristics to a new sample. Consequently, the constructed proposal Q proposes samples that nicely embrace both the data and the intrinsic Markov properties. The proposal scheme comprises two sub-procedures. First, it selects a local region to update based on the proposal priors. Rather than updating the entire sequence of a previous sample $L^{(r)}$, it selects a local region in $L^{(r)}$ and then proposes a locally updated new sample $L^{(r')}$. The local update scheme improves the space exploration capabilities of MCMC and results in faster convergence. Secondly, the proposal priors $P(\tilde{L})$ and the discrete transition PDF B are used to assign the new labels within a selected region. The second step has the effect of proposing a sample which reflects both the data and Markov properties of SLDSs. The choice of the second step, product of two PDFs, proposes smoother and more plausible label sequences in general than other options, e.g., mixture of two PDFs. The two sub-steps are described in detail below.

In the first step, scoring schemes are used to select a local region within a sample. First, the previous sample labels $L^{(r)}$ are divided into a set of segments at a regular interval. Then, each segment is scored with respect to the proposal priors $P(\tilde{L})$, i.e. the affinities between the labels in each segment and the proposal priors are evaluated. Any reasonable affinity and scoring schemes are applicable. Finally, a segment is selected for an update via sampling based on the inverted scores.

In the second step, new labels l'_t's are sequentially assigned within a selected segment using the assignment function in Eq. 30 where $B_{l'_t|l'_{t-1}} \triangleq P(l'_t|l'_{t-1})$. The implicit dependence of $\tilde{l_t}$ on c_t in Eq. 29 is omitted for brevity.

$$P(l'_t) \quad = \quad \beta\delta(l_t) + \bar{\beta}\left\{\frac{B_{l'_t|l'_{t-1}} P(\tilde{l'_t})}{\sum_{l'_t=1}^{n} B_{l'_t|l'_{t-1}} P(\tilde{l'_t})}\right\} \tag{30}$$

Above, the first term with the sampling ratio β denotes the probability to keep the previous label l_t , i.e. $l'_t \leftarrow l_t$. The second term with coefficient $\bar{\beta} \triangleq 1 - \beta$ proposes a sampling of a new label l'_t.

References

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50:5–43, 2003.

[2] T. Balch, F. Dellaert, A. Feldman, A. Guillory, C. Isbell, Z. Khan, A. Stein, and H. Wilde. How A.I. and multi-robot systems research will accelerate our understanding of social animal behavior. *Proceedings of IEEE*, 94(7):1145–1463, July 2006.

[3] T. Balch, Z. Khan, and M. Veloso. Automatically tracking and analyzing the behavior of live insect colonies. In *Proc. Autonomous Agents 2001*, pages 521–528, Montreal, 2001.

[4] Y. Bar-Shalom and T.E. Fortmann. *Tracking and data association*. Academic Press, New York, 1988.

[5] Y. Bar-Shalom and X. Li. *Estimation and Tracking: principles, techniques and software*. Artech House, Boston, London, 1993.

[6] Y. Bar-Shalom and E. Tse. Tracking in a cluttered environment with probabilistic data-association. *Automatica*, 11:451–460, 1975.

[7] A. Barbu and S.-C. Zhu. Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 27(8):1239–1253, August 2005.

[8] Matthew Brand and Aaron Hertzmann. Style machines. In *SIGGRAPH : Proc. of Conference on Computer Graphics and Interactive Technologies*, pages 183–192, 2000.

[9] K. Branson and S. Belongie. Tracking multiple mouse contours (without too many samples). In *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1039–1046, 2005.

[10] C. Bregler. Learning and Recognizing Human Dynamics in Video Sequences. In *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 568–574, 1997.

[11] C. Carter and R. Kohn. Markov chain Monte Carlo in Conditionally Gaussian State Spaece Models. *Biometrika*, 83:589–601, 1996.

[12] G. Casella and C.P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, March 1996.

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

[14] P. M. Djuric and J-H. Chun. An MCMC Sampling Approach to Estimation of Nonstationary Hidden Markov Models. *IEEE Trans. Signal Processing*, 50(5):1113–1123, 2002.

[15] A. Doucet and C. Andrieu. Iterative Algorithms for State Estimation of Jump Markov Linear Systems. *IEEE Trans. Signal Processing*, 49(6):1216–1227, 2001.

[16] A. Doucet, N. J. Gordon, and V. Krishnamurthy. Particle filters for state estimation of jump Markov linear systems. *IEEE Trans. Signal Processing*, 49(3):613–624, 2001.

[17] A. Feldman and T Balch. Representing honey bee behavior for recognition using human trainable models. *Adaptive Behavior*, 12:241–250, 2004.

[18] J. Ferguson. Variable duration models for speech. In *Symposium on the Aplication of HMMs to Text and Speech*, pages 143–179, 1980.

[19] B. Frey and N. Jojic. Transformation-Invariant Clustering and Dimensionality Reduction Using EM. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 25(1):1–17, January 2003.

[20] K. Frisch. *The Dance Language and Orientation of Bees*. Harvard University Press, 1967.

[21] X. Ge and P. Smyth. Deformable Markov model templates for time-series pattern matching. In *Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, pages 81–90, 2000.

[22] Z. Ghahramani and G. E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):963–996, 1998.

[23] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov chain Monte Carlo in practice*. Chapman and Hall, 1996.

[24] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[25] A. Howard and T. Jebara. Dynamical systems trees. In *Proc.* 20^{th} *Conf. on Uncertainty in AI (UAI)*, pages 260–267, Banff, Canada, July 2004.

[26] Z. Khan, T. Balch, and F. Dellaert. A Rao-Blackwellized particle filter for EigenTracking. In *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 980–986, 2004.

[27] Z. Khan, T. Balch, and F. Dellaert. MCMC data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 28(12):1960–1972, 2006.

[28] C.-J. Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1-2):1–22, 1994.

[29] S. Kim and P. Smyth. Segmental Hidden Markov Models with Random Effects for Waveform Modeling. *J. of Machine Learning Research*, 7:945–969, October 2006.

[30] M. W. Lee and I. Cohen. A model-based approach for estimating human 3d poses in static images. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 28(6):905–916, 2006.

[31] U. Lerner and R. Parr. Inference in hybrid networks: Theoretical limits and practical algorithms. In *Proc.* 17^{th} *Conf. on Uncertainty in AI (UAI)*, pages 310–318, Seattle, WA, August 2001.

[32] U. Lerner, R. Parr, D. Koller, and G. Biswas. Bayesian fault detection and diagnosis in dynamic systems. In *Proc.* 17^{th} *AAAI National Conference on AI*, pages 531–537, Austin, TX, 2000.

[33] S. E. Levinson. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1(1):29–45, 1990.

[34] P. Maybeck. *Stochastic Models, Estimation and Control*, volume 1. Academic Press, New York, 1979.

[35] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. John Wiley & Sons, 1997.

[36] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[37] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. Kluwer Academic Press, 1998. Also published by MIT Press, 1999.

[38] B. North, A. Blake, M. Isard, and J. Rottscher. Learning and classification of complex dynamics. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 22(9):1016–1034, 2000.

[39] S. M. Oh, A. Ranganathan, J.M. Rehg, and F. Dellaert. A Variational Inference Method for Switching Linear Dynamic Systems. Technical Report GIT-GVU-05-16, GVU Center, College of Computing, 2005.

[40] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Data-Driven MCMC for Learning and Inference in Switching Linear Dynamic Systems. In *Proc.* 22^{nd} *AAAI National Conference on AI*, pages 944–949, Pittsburgh, PA, 2005.

[41] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Learning and Inference in Parametric Switching Linear Dynamic Systems. In *Proc. of Intl. Conf. on Computer Vision (ICCV)*, volume 2, pages 1161–1168, 2005.

[42] M. Ostendorf, V. V. Digalakis, and O. A. Kimball. From HMM's to Segment models : A Unified View of Stochastic Modeling for Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378, 1996.

[43] V. Pavlović and J.M. Rehg. Impact of Dynamic Model Learning on Classification of Human Motion. In *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 788–795, 2000.

[44] V. Pavlović, J.M. Rehg, T.-J. Cham, and K. Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proc. of Intl. Conf. on Computer Vision (ICCV)*, volume 1, pages 94–101, 1999.

[45] V. Pavlović, J.M. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *Advances in Neural Information Processing Systems (NIPS)*, pages 981–987, 2000.

[46] A. Ranganathan and F. Dellaert. Data driven MCMC for appearance-based topological mapping. In *Robotics: Science and Systems I*, pages 209–216, 2005.

[47] L. Ren, A. Patrick, A. Efros, J. Hodgins, and J. M. Rehg. A Data-Driven Approach to Quantifying Natural Human Motion. *ACM Trans. on Graphics, Special Issue: Proc. of 2005 SIGGRAPH Conf.*, 24(3):1090–1097, August 2005.

[48] A-V.I. Rosti and M.J.F. Gales. Rao-blackwellised Gibbs sampling for switching linear dynamical systems. In *Proceedings of Intl. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, volume 1, pages 809–812, 2004.

[49] S. Roweis and Z. Ghahramani. A Unifying Review of Linear Gaussian Models. *Neural Computation*, 11(2):305–345, 1999.

[50] M. Russel. A segmental HMM for speech pattern matching. In *Proceedings of Intl. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, pages 499–502, 1993.

[51] P. Dollár S. Belongie, K. Branson and V. Rabaud. Monitoring Animal Behavior in the Smart Vivarium. In *Measuring Behavior*, pages 70–72, 2005.

[52] R.H. Shumway and D.S. Stoffer. Dynamic linear models with switching. *Journal of the American Statistical Association*, 86:763–769, 1992.

[53] Z.W. Tu and S.C. Zhu. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 24(5):657–673, 2002.

[54] R. Vidal, A. Chiuso, and S. Soatto. Observability and identifiability of jump linear systems. In *Proceedings of IEEE Conference on Decision and Control*, volume 4, pages 3614–3619, 2002.

[55] A. D. Wilson and A. F. Bobick. Parametric Hidden Markov Models for Gesture Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 21(9):884–900, 1999.

[56] Y.Li, T.Wang, and H-Y. Shum. Motion texture : A two-level statistical model for character motion synthesis. In *SIGGRAPH : Proc. of Conference on Computer Graphics and Interactive Technologies*, 2002.

[57] O. Zoeter and T. Heskes. Hierarchical visualization of time-series data using switching linear dynamical systems. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 25(10):1202–1215, October 2003.