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 provides the possibility to 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 several 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 systematic way to robustly interpret systematic variations governed by higher order parameters.

The contributions in this paper address all three challenges above. First, we present a data-driven MCMC sampling method for SLDSs as a robust and efficient approximate inference method. Second, we present segmental switching linear dynamic systems (S-SLDS), where the geometric distributions are replaced with arbitrary duration models. Third, we extend the standard model with a parametric model 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 proposed inference methods and extensions for SLDSs provide a robust 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 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 that are being exhibited by 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 by skilled operators, as is common in domains such as biology, is a time-consuming and error-prone process. Thus, it is desirable to develop methods that automatically infer the behavioral patterns of the targets. In addition, in applications where there is large variability in the behaviors, we need a framework in which we can *learn* these behaviors from examples.

In particular, we are interested in two inference tasks that are of central importance. The first, ’labeling’, is to automatically segment the 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 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]. The current state of biological field work is still dominated by manual data

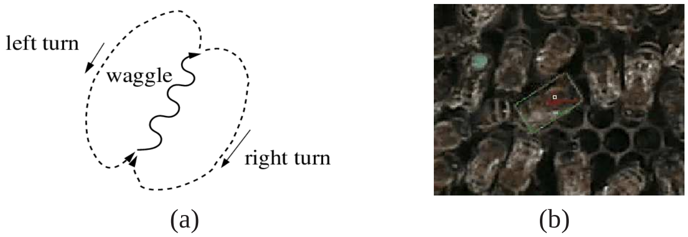


Figure 1: (a) A bee dance is in three patterns : waggle, left turn, and right turn. (b) The box in the middle is a tracked bee.

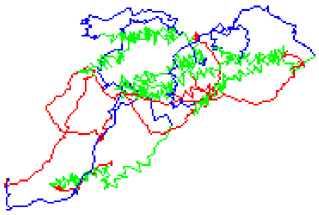


Figure 2: An example honey bee dance trajectory. The track is automatically obtained using a vision-based tracker and manually labeled afterward. Key : waggle , right-turn , left-turn

interpretation, a time-consuming and error-prone 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, 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: “turn left”, “turn right” and “waggle”. The length (duration) and orientation of the waggle phase corresponds 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 [21]. After tracking, the obtained trajectory of the dancing bee is manually labeled as “turn left”, “turn right” or ”waggle” and is shown in Figure 2.

The research goals in this application domain are three-fold. First, we aim to learn the motion patterns of honey bee dances from the obtained training dance sequences. Second, we should be able to automatically segment new dance sequences into 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 take a model-based approach, in which we employ a computational model of behavior in order to interpret the data. In our case the motions are complex, i.e. they are comprised of sub-behaviors. The model we use should be expressive enough to accurately model the individual sub-behaviors, while at the same time able to capture the inter-relationships between them.

Hence, the basic generative model we adopt is the Switching Linear Dynamic System (SLDS) model [35, 36, 37]. 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, 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 classification [35, 36, 37, 39] and motion synthesis [47].

1.3 Contributions

In this paper, we present a framework that learns behavioral patterns from data and provides robust inference methods that label the motion sequences while simultaneously quantifying the global parameters, significantly extending the scope and modeling power of standard SLDS models. 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, as well as providing robust inference methods for each of these.

We discuss the three main limitations of the original SLDS model in Section 3, 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 Background

2.1 Linear Dynamic Systems

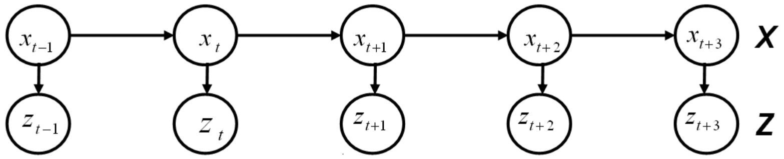


Figure 3: A linear dynamic system (LDS)

An 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:

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

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:

$$z_t = Hx_t + v_t \text{ where } v_t \sim \mathcal{N}(0, V) \tag{2}$$

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 done exactly using the RTS smoother [5], an efficient belief propagation implementation. For further details on LDSs, the reader is referred to [5, 27, 41].

2.2 Switching Linear Dynamic Systems

In an 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

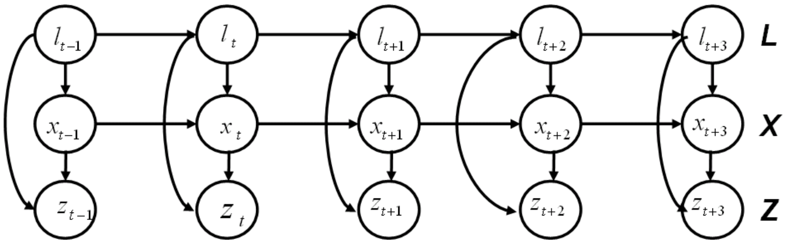


Figure 4: Switching linear dynamic systems (SLDS)

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

Switching linear dynamic system (SLDS) models have been studied in a variety of research communities ranging from computer vision [36, 35, 37, 30, 8, 43], computer graphics [43, 47, 39], tracking [6], signal processing [12, 13] and speech recognition [40], to econometrics [22], visualization [48], machine learning [25, 17, 31, 32, 33, 20], control systems [45] and statistics [42]. 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 [36, 35, 37, 31, 32, 33].

2.3 Learning and Inference in SLDS

The EM algorithm [10] 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

$$f^i(L, X) \triangleq P(L, X | Z, \Theta^i) \tag{3}$$

  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:

$$\Theta^{i+1} \leftarrow \underset{\Theta}{\operatorname{argmax}} \ \langle \log P(L, X, Z | \Theta) \rangle_{f^i(L,X)} \tag{4}$$

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.3 motivates the development of approximate inference techniques discussed in more detail below.

3 Contributions and Related Work

In this paper, we address three limitations of the standard SLDS model: (1) intractability of exact inference in SLDSs, (2) limitations in duration modeling, and (3) absence of a systematic 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 investigate the exact posterior of SLDSs in the presence of intractability. Secondly, a segmental SLDS model is proposed to improve the limited duration modeling power of standard SLDSs. Finally, we introduce a parametric extension of SLDSs that provide a systematic means to quantify the embedded global parameters. In the sections below 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. One of the authors' previous work on automatic labeling of honey bee dances using HMMs [14] is most closely related to the work in this paper. However, in [14], 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 [21]. In addition, SLDSs are used to learn and infer the motion patterns of bees and new DD-MCMC method and novel SLDS extensions are presented in this work.

In comparison to our previous conference publications on this topic [31, 32], the current paper extends the SLDS model to include duration modeling, and presents the detailed learning and inference mechanisms for parametric segmental SLDS which combines the advantages of two extended models, i.e., S-SLDS and P-SLDS.

3.1 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 [37].

It is known that the exact inference in SLDS is intractable as the size of Gaussian mixtures increases exponentially with time [24]. Thus, there have been research efforts to derive efficient approximate inference methods. The early examples include GPB2 [5], and Kalman filtering [8], and the pseudo-EM algorithm [42]. More recent examples include a variational approximation [17, 35, 37, 33], an approximate Viterbi method [36, 35, 37], expectation propagation [48], iterative Monte Carlo methods [12], sequential Monte Carlo methods [13], and Gibbs sampling [40]. Approximate inference in SLDS models has focused primarily on two classes of techniques: stage-wise methods such as approximate Viterbi [37] 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 [17, 33, 37].

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 and etc., demonstrated poor labeling performance. In such cases, it is necessary to validate the capacity of the model to verify whether such a poor labling result is due to the approximation method itself or the inherent limitation of the model not being able to represent the temporal phemomenon adequately.

We describe a novel proposal distribution for Data-driven MCMC inference in Section 4, originally presented at AAAI [31]. 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-Blackwellization [9] and the Data-Driven MCMC paradigm [44] to improve convergence. The Data-Driven MCMC approach has been successfully applied in computer vision [23, 44] and robotic mapping [38].

3.2 Improved Duration modeling

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

$$P(d) \quad = \quad a^{d-1}(1-a) \tag{5}$$

Above, d denotes the duration of a given switching state and a denotes Markov transition probability to make a self-transition which has a value between zero and one. As a consequence, a duration of one time-step come to possess the largest probability mass.

In contrast, many natural temporal phenomena exhibit patterns of regularity in the duration for which a given model or regime is active. In such cases the standard SLDS model would be inappropriate to effectively encode the regularity of durations in data. A honey bee dance is an 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.

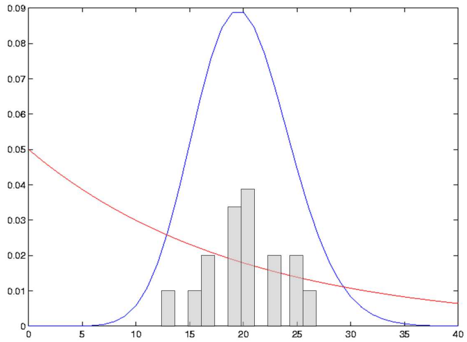


Figure 5: A realistic Gaussian and a limited geometric duration model. Models are learned from the data shown as the overlayed histogram.

For example, we learned a duration model for the waggle phase using a realistic Gaussian density and a conventional geometric distribution from one of the manually labeled dance sequences depicted in Figure 16. Figure 5 shows the learned geometric and Gaussian distributions for comparison. It can be observed that the learned geometric duration model does not exhibit any pattern of regularity in durations. Hence, standard SLDS models are inappropriate for data which exhibits temporal patterns that deviate from geometric distributions.

The limitation of a geometric distribution was also previously addressed by the HMM communities, and HMM models with enhanced duration capabilities were introduced [15, 26, 34]. HMMs has been widely studied by the speech recognition and the machine learning communities to enhance its duration modeling capabilities. The variable duration HMM (VD-HMM) was introduced in [15]: 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) [26]. It has since been shown that the VD-HMM and the NS-HMM are duals [11]. Ostendorf et.al. provides an excellent discussion on segmental HMMs [34].

We adopt similar ideas to arrive at SLDS models with enhanced duration modeling. The resulting segmental SLDS model is described in Section. 5.

3.3 Inference on global parameters

The standard SLDS does not provide a systematic 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 pointed at. In many applications we are more interested in these global underlying parameters rather than the exact categorization of the sub-motions.

Previously, Wilson & Bobick presented parametric HMMs [46]. In a PHMM, the parametric observation models learned are conditioned on global observation parameters, such that globally parameterized gestures can be recognized. PHMMs 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 [7]. A transformation-invariant learning approach for static images were addressed in [16].

We extend the standard SLDS model in a similar way, resulting in a parametric SLDS (P-SLDS) model. As in a PHMM, the P-SLDS model we propose incorporates global parameters that underly systematic spatial variations of the overall target motion. In addition, while PHMM only introduced global observation parameters which cause spatial variations, we additionally introduce dynamic parameters which capture temporal variations.

As mentioned earlier, the problem of global parameter quantification and labeling can be simultaneously solved. Hence, we formulate expectation-maximization (EM) methods for learning and inference in P-SLDS and present it in Section 6.

4 Inference via Data-Driven MCMC

In this section, we introduce a novel sampling-based method that theoretically converges to the correct posterior distribution on label sequences $P(L|Z)$. Faster convergence is achieved by incorporating a data-driven approach where we introduce proposal priors and label-cue models.

All MCMC methods work similarly [18]: 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.

4.1 Rao-Blackwellized MCMC

In our solution, we propose to pursue the Rao-Blackwellised posterior $P(L|Z)$, rather than the joint posterior $P(L, X|Z)$. The effect is the dramatic reduction of sampling space from L, X to L. This results in an improved approximation on the labels L, which are exactly the variables of interest in our application. This change is justified by the Rao-Blackwell theorem [9]. The Rao-Blackwellisation is achieved via the analytic integration on 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 the inference in the time-varying LDS with the varying but known parameters. 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 [19, 29] 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

We propose to use a Data-Driven paradigm [44] where the cues present in the data provide an efficient MCMC proposal distribution Q. It is crucial to provide an efficient proposal Q, which results in 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 motivated. 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 based on the collected cues in a supervised manner. For example, the change of heading angles is derived 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. Note that the heading angles are measured clockwise.

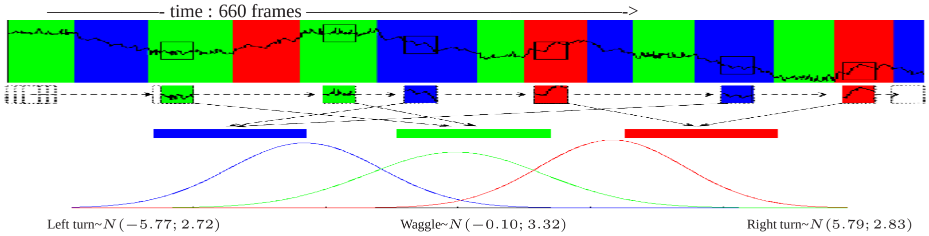


Figure 6: Learning phase. Three label-cue models are learned from the training data. See text for detailed descriptions.

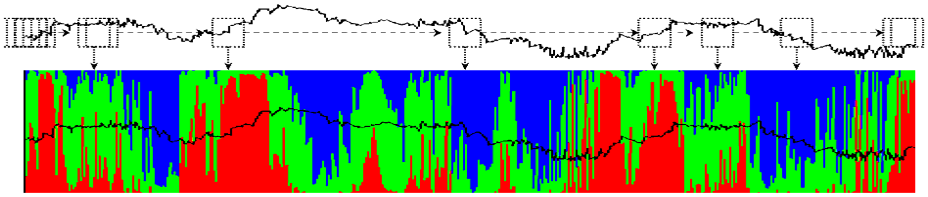


Figure 7: Inference phase. Raw proposal priors are evaluated based on the collected temporal cues.

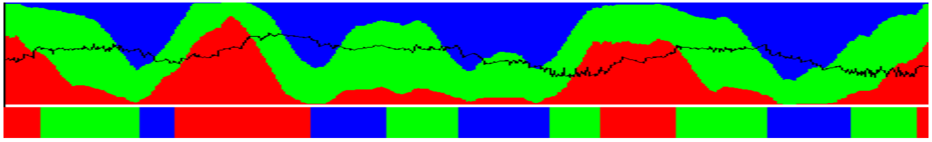


Figure 8: Final proposal priors and the ground truth labels. Key : waggle, right-turn, left-turn.

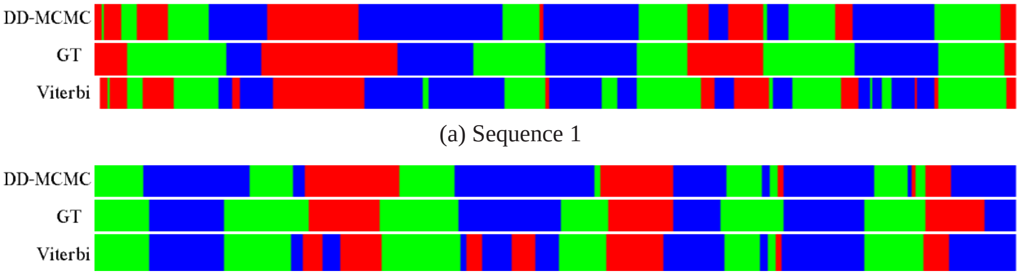
Specifically, a cue window slides over the entire angle data while it collects cues as shown at the top of Fig.6. Then, the collected cues are classified according to the training labels. Then, the label-cue (LC) models are learned in the form of three Gaussians in our example, as shown at the bottom of Fig.6. The estimated means and the standard deviations show that the average change of heading angles are -5.77, -0.10 and 5.79 radians, as expected.

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.7. Then, the proposal priors are 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.7. Thus, we use the smoothed estimates as the final proposal priors, shown in Fig.8. At the bottom of Fig.8, the ground truth labels are shown below the final proposal priors for comparison. The obtained priors provide an excellent guide to the labels of the dance segments.

Afterwards, the obtained proposal priors $P(\tilde{L})$ 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

The DD-MCMC is a Bayesian inference algorithm. Nonetheless, it can be used as a robust labeling method. The MAP label sequence are taken from the discovered posterior distribution $P(L|Z)$ where the label of 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. 9. It can be observed that DD-MCMC delivers solutions that concur very well with the ground truth. On the other hand, the approximate Viterbi labels at the bottom over-segments the data (insertion errors). The insertion errors of approximate Viterbi highlight one of the limitations of the class of deterministic algorithms for SLDS. In this respect, the proposed DD-MCMC inference method



(a) Sequence 1

(b) Sequence 2

Figure 9: DD-MCMC MAP, ground truth, Viterbi labels.

is shown to improve upon the Viterbi result and provide more robust labeling (inference) capabilities.

Some errors between the MAP labels and the ground truth occur 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. Considering that SLDSs are learned exclusively from the observation data, the results are fairly good.

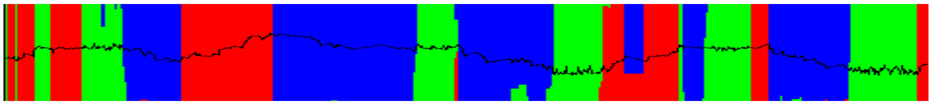


Figure 10: Posterior distribution $P(L|Z)$ is discovered from sequence 1. The heading angle of a bee is superimposed on the figure as an indicator of a dancer's dance mode. Key : waggle , right-turn , left-turn

To further analyze the inference capabilities of a standard SLDS withinin our application, we investigated the posterior distribution $P(L|Z)$ which is discovered from the first sequence using the proposed DD-MCMC inference, see Fig.10. The discovered posterior shows that most of the over-segmentations are induced due to the strong noise in the data. As an example of an extreme systematic noise, around the two fifths from the right in Fig.10, the tracked bee systematically side-walks to the left 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 : 70% turn-left and 30% turn-right roughly, and it results in the over-segmentation of the data where it appears at the top color strip in Figure 9(a).

4.4 Discussion

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

From the visuallized posterior in Fig.10, we notice two limitations of 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 a slight noise introduces an over-segmentation even though such noise appears only for a few frames. Secondly, 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 severely deviate from the ground truths. Moreover, it is certain that the global information can provide a better cue for overall labeling processes.

Accordingly, we introduce segmental SLDS and parametric SLDS as the robust extensions to resolve the problems mentioned above.

It should be noted that DD-MCMC method is still computationally demanding although it is an efficient solution in the space of MCMC methods. For example, it proposed approximately 4,000 samples to converge in one of the experiments above. As each proposed label sequence requires temporal smoothing step for Rao-Blackwellised inference, the computation required for every samples is approximately identical to that of an approximate Viterbi (VI) method. As a consequence, DD-MCMC method consumed approximately 4,000 times more computation than VI method. The models to be introduced in the following sections are shown to reflect the characteristics of the honey bee dance data more tightly and were able to produce satisfactory results using VI or a variational approximation method. Additionally, theoretical justification of computational complexity of data-driven MCMC methods is still an on-going area of research in spite of its success in hard computer vision problems. Hence, we plan to investigate the scaling issue of the proposed DD-MCMC method in the extended models in the future and adopt computationally less-demanding approximation inference methods in the following sections.

5 Segmental SLDS

We introduce the segmental SLDS (S-SLDS) model, which improves on the standard SLDS model by relaxing the Markov assumption at a time-step level to a coarser *segment level*. The development of S-SLDS model is motivated by the regularity in durations being exhibited by the honey bee dances. As discussed in Section 3.2, a dancer bee will attempt to stay in the waggle regime for a certain duration to effectively communicate a message. In such a case, the geometric distribution induced in standard SLDSs is not an appropriate choice to model the duration patterns. Fig. 5 shows that a geometric distribution accords the highest probability on the duration of only one time step. As a result, the inference in standard SLDSs is susceptible to over-segmentation due to the noise in data.

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 in modeling duration, and more robust inference capabilities than the standard SLDS. Nonetheless, we show that one can always convert a learned S-SLDS model into an equivalent standard SLDS, operating in a different label space. Hence, as a significant advantage we are able to resuse the large array of approximate inference and learning techniques developed for SLDSs.

5.1 Conceptual view on the generative process of S-SLDS

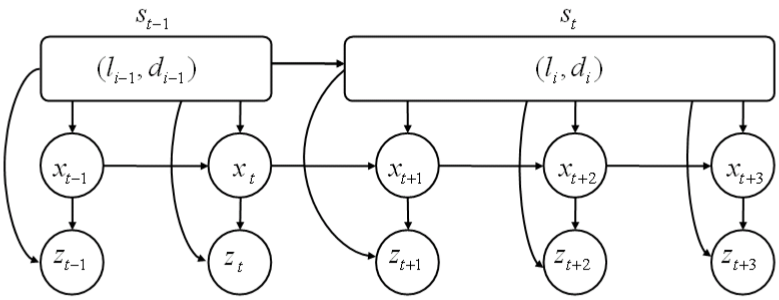


Figure 11: A schematic sketch of an S-SLDS with explicit duration models.

Conceptually, 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 of label l_i and 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, however, 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.11. 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 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 continuous states X.

5.2 Graphical Representation of S-SLDS

In this section we present a graphical representation of S-SLDSs, transforming the conceptual generative model described in Section 5.1 into a concrete model that uses conventional model switching 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.12, 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.

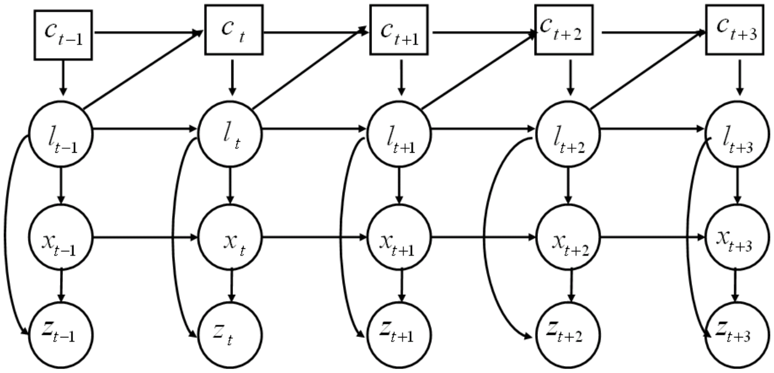


Figure 12: Graphical representation of an S-SLDS

The counter chain C maintains an incremental counter which evolves based on a set 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 B. If the counter simply increments, then the new label is set to be the current label l_t, i.e. $l_{t+1} \leftarrow l_t$.

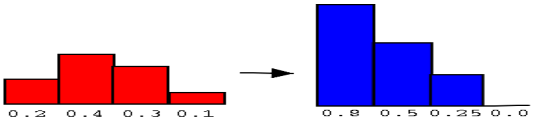


Figure 13: Evaluating an NSTF (right) from an explicit duration model (left).

While the explicit duration models D introduced in Section 5.1 are more understandable and readily obtained from the labeled data, it is necessary to transform the explicit duration models D into an equivalent NSTFs U to incorporate the knowledge in durations into a framework based on graphical models. To do this, we can observe that the explicit duration

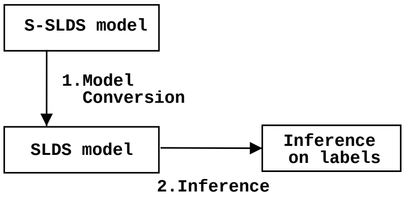


Figure 14: Inference in S-SLDS.

models D and the NSTFs U are analogous to the duration models of VD-HMMs [15] and NS-HMMs [26] respectively. Hence, we can exploit the duality between the VD-HMMs and NS-HMMs, which appeared in [11]. The equivalent NSTFs U are exactly evaluated from the explicit duration models D as follows :

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

Above, D_l^{max} denotes the maximum duration allowed for the l th model. Intuitively, the latter composite term on the r.h.s. denotes the probability to reset the counter variable c_{t+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.13 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 depends on the choice of duration models. An example is demonstrated in Section 7 where we learn the duration models in Gaussian forms from the honey bee dance sequences. However, Gaussian models encode proabilities for non-existing negative durations as well. Hence, 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 be adopted.

5.4 Inference for Segmental SLDSs

Below we demonstrate that an S-SLDS can be always converted to an equivalent SLDS. This is an important advantage as it allows us to readily reuse the large array of approximate inference algorithms discussed in Section 2.3. 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.

The overall idea of inference is depicted in Figure 14. 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 parameters of the equivalent standard SLDS are learned via EM, the obtained SLDS model is converted back to S-SLDS model 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 :

$$B'_{(l_i,c_i),(l_j,c_j)} = \begin{cases} U_{l_i}(c_i) & \text{increment} \\ \tilde{B}_{l_i,l_j}(1 - U_{l_i}(c_i)) & \text{reset} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

In Eq. 8, the three cases differ as follows : (increment) $l_i = l_j$ and $c_j = c_i + 1$. (reset) $c_j = 1$. (otherwise) 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 :

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

Nonetheless, it is important to note that the naive reuse of the learning and inference algorithms for SLDS to S-SLDS may induce substantial increase in computational overhead. The efficient implementation and increased computational overheads are discussed below.

5.5 Computational Considerations

As mentioned above, 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 $D_{max} \triangleq \max\{D_l^{max}\}_{l=1}^{n}$, i.e. the number of all meta variables. Thus, there is a considerable computational overhead, by a factor of $O(D_{max}^2)$. This increased asymptotic running time overhead applies to the approximate inference algorithms with HMM-type components in general, e.g. approximate Viterbi [37] and a variational method [37, 17, 33], as they require the computations between all possible state pairs from the previous time-step to the next time-step.

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.

In case of the presented DD-MCMC method, the complexity of the method in S-SLDS is a topic of on-going research. This is partly due to the fact that the straightforward use of DD-MCMC would not be the optimal choice, as the Markov Chain properties in S-SLDS has very regularized structure and the proposed DD-MCMC method does not exploit this fact. Hence we plan to present the research results on this issue elsewhere in the future.

6 The Parametric SLDS Model

The idea of parametric extension of SLDSs (P-SLDS) described in this section originally appeared in [32].

As discussed in Section 3.3, the standard SLDS does not provide the means to quantify global variations. Hence, the development of a framework which can decode the global parameters w.r.t. the canonical behavioral template is necessitated. 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 a hive, and the communication messages of dances are exactly the variables of interest in our bee application.

Moreover, it should be noted that the superior global parameter estimates, which are closer to ground truth, can provide improved labeling capabilities and vice versa, i.e. 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. Hence, a parametric SLDS (P-SLDS) which provides a means to quantify the global variables and solves both labeling and quantification problem in an iterative manner is introduced.

The resulting P-SLDS learns canonical behavioral templates from data with additional information on the associated global parameters. A P-SLDS effectively decodes the global parameters while it simultaneously labels the sequences. This is done using an expectation-maximization (EM) algorithm [10, 28], presented in detail below.

6.1 Graphical representation of P-SLDS

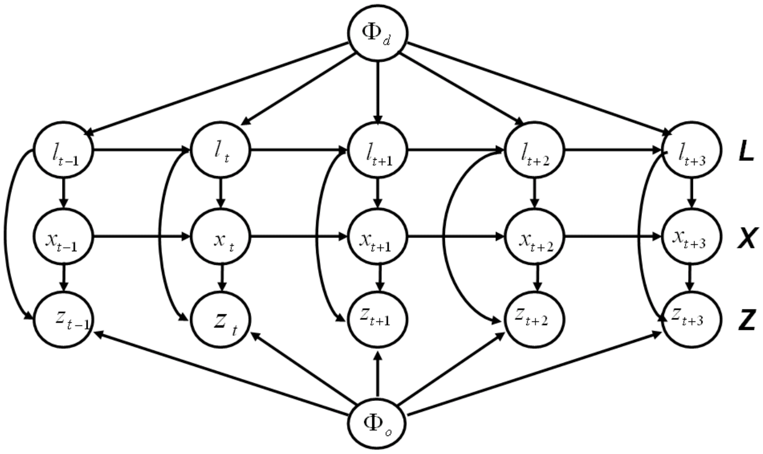


Figure 15: Parametric SLDS (P-SLDS)

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.15. 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. E.g., for the bee-dance, a food source that is far away leads a dancer bee to stay in each dance regime longer to make a larger dance, which will result in 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.

Additionally, the canonical parameters Θ represent the common underlying behavioral template. Note that the canonical parameters Θ are embedded in the conditional dependency arcs in Fig.15. 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.

In detail, it can be observed that the discrete state transitions at the top chain of Fig. 15 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 by solely 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, in a systematic manner. 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.

The graphical model of P-SLDS necessitates parameterized versions of an initial state distribution $P(l_1|\Theta, \Phi_d)$, a discrete state transition table $P(l_t|l_{t-1}, \Theta, \Phi_d)$ and an 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

Algorithm 1 EM1 for Learning in P-SLDS

- E-step 1: obtain the posterior distribution

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

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 :

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

function of Φ, and (c) no functional form for PDF is available. In the latter case, a neural network may be used as suggested in [46].

In the following Sections, 6.2 and 6.3, we discuss learning and inference in P-SLDS assuming that functional forms are available.

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 these parameters known in our training data.

Learning in P-SLDS entails estimating the P-SLDS canonical parameters Θ, given the data $\bar{D} \triangleq \left\{ \bar{\Phi} = \{\bar{\Phi}_d, \bar{\Phi}_o\}, \bar{L}, \bar{Z} \right\}$ 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 [10, 28] with the continuous states X as the only hidden variables to find an ML estimate of the canonical parameters $\hat{\Theta}$, as described before.

The E-step in Eq.9 is equivalent to inference in an LDS model. In more detail, as 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, the inference over the continuous hidden states X in E-step can be performed by Kalman-smoothing [4]. Given the posterior distribution $f_L^i(X)$ in Eq. 9 we then update the parameters Θ^{i+1}.

In case the parameterized dependencies are linear functions of the global parameters Φ, M-step in Eq.10 can be analytically solved. However, in case the parametric dependencies are non-linear, an exact M-step is often infeasible and needs to 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 the inference on the label sequence L (labeling), given exclusively 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 distinct dataset relying on an inference where neither the global parameters Φ nor the label sequence L is known.

As in [46], we use EM to quantify the optimal global parameters $\hat{\Phi}$ as shown in Algorithm 2. Note that we perform EM1 in Section 6.2 'Learning' to learn the canonical model parameters Θ, while EM2 in Section 6.3 'Inference' is performed to estimate the global parameters Φ with simultaneous inference on the labels L. More details on EM2 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 known to be intractable [35, 36]. Thus, we need to rely on the approximate inference methods. Here, we present a derivation of E-step based on approximate Viterbi (VI) method [35]. Note that the choice of VI method does not harm the generality of the framework although it delivers more concise derivation. At every ith EM iteration,

Algorithm 2 EM2 for Inference in P-SLDS

- E-step 2 : obtain the posterior distribution:

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

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:

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

the joint posterior over the hidden variables L and X is approximated by a peaked posterior over the X with the obtained pseudo-optimal label sequence $\hat{L}^i$:

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

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

Note that the implicit conditional dependence on the fixed canonical parameters Θ is omitted 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:

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

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 $\mathcal{LLH}$ $\mathcal{L}^i(\Phi)$ (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 $\mathcal{L}(\Phi_o)$:

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

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

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 shown in Eq.19 and 20 where the term $f_I^i(x_t)$ denotes the marginal on x_t from the full posterior $f_I^i(X)$, i.e. $f_I^i(x_t) \triangleq \int_{X/x_t} f_I^i(X)$.

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

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

In case we are modeling data with parametric S-SLDS models, the global dynamic parameters Φ_d are associated with the duration models of S-SLDSs, 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-SLDSs is equally evaluated as in Eq. 20. The detailed machinery in the M-step will depend on the application domain. In case the parametric forms are linear in the global parameters Φ, the M-step is analytically feasible. 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

The goal in our application is to build a vision system that can reliably label the motion sequences with simultaneous quantification. To achieve this, we model the honey bee dances using a 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 regimes where $n = 3$. The global observation parameter Φ_o is set to be the angle orientation of the bee dance.

7.1 Canonical parameters

The canonical parameters in honey bee dances are defined to be a tuple of a set of initial label distribution π, semi-Markov segmental Markov transition matrix B, LDS model parameters M and variances in durations in each behavioral modes Σ, i.e., $\Theta \triangleq \left\{\pi, B, M \triangleq \{M_l|1 \leq l \leq n\}, \Sigma \triangleq \{\Sigma_l|1 \leq l \leq n\}\right\}$. 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 validated from the background knowledge on honey bee dances. For example, it is reasonable to assume that the initial label distribution π and the segment label transition matrix B between different dance regimes do not vary across the dance sequences. In addition, the dancer bees try to regulate its waggle durations to convey correct dance messages. Hence, the degrees of variations in durations of each dance regime are assumed to be constant. Hence, they are learned and represented as the variances Σ.

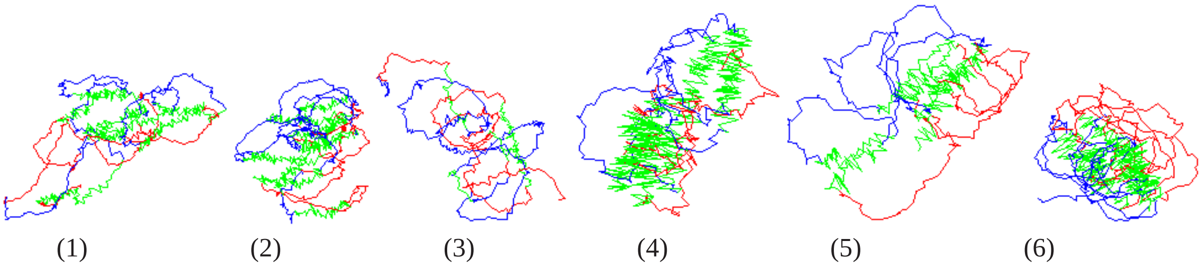


Figure 16: Bee dance sequences used in the experiments. Each dance trajectory is the output of a vision-based tracker. Tables 1 and 2 give the global motion parameters for each of the numbered sequences. Key : waggle, right-turn, left-turn.

7.2 Dynamics model

We set the global dynamic parameter of 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 models D_l of S-SLDSs are modeled as Gaussian distributions as follows :

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

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 :

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

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

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

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

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 respectively denote the 2D coordinates and the heading angle of a tracked dancer bee at time t. The angle of zero corresponds to the direction of positive x-axis and the angle increases clock-wisely. The triangular function elements in the observations were introduced to make the system be able to learn the location-invariant rotating motions. Note that the observed temporary heading angle θ_t differs from the global dance angle Φ_o.

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

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

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:

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

Given all this, we obtain the observation $\mathcal{LLH}$ $\mathcal{L}^i(\Phi_o) \equiv$

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

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, as the non-linearities are involved due to a rotation, there is no analytical solution for this maximization problem where Eq.27 involves quadratic triangular function terms, e.g., $sin(\Phi_o)^2$. Thus, we perform 1D gradient ascent on the obtained functional.

8 Experimental Results

The experimental results show that PS-SLDS provides reliable global parameter quantification capabilities along with improved recognition abilities over the standard SLDS. The six dancer bee tracks obtained from the videos are shown in Fig.16. The vision-based tracker we used [21] is shown in Fig.1(b) where the dancer bee is automatically being tracked in the green rectangle.

We performed experiments with 6 video sequences[1] with length 1058, 1125, 1054, 757, 609 and 814 frames, respectively. Once the sequence observations Z are obtained, the whole trajectories were preprocessed in such a way that the mean of each track is located at (100,100). Note from Fig.16 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 the comparison and learning purposes. The dimensionality of the continuous hidden states was set to be four.

Given the relative difficulty of obtaining this data, which has to be labeled manually to allow for a ground-truth comparison, we adopted a leave-one-out (LOO) strategy. 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 are performed using both PS-SLDS and the standard SLDS, switching the test data sequence. 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 at the inferred "waggle" segments.

[1] The experimental data used in this work are available at : `www.cc.gatech.edu/~borg/ijcv_psslds.`

8.1 Learning from training data

The parameters of both PS-SLDS and standard SLDS are learned from the data sequences depicted in Fig. 16. The standard SLDS model parameters were learned as described in Section 2.3. The canonical parameters tuple described in Section 7.1 are all learned solely based on the observations Z without any restriction on the parameter structures. 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 waggle directions head to the same direction based on the evaluated ground truth waggle angles. Such preprocessing was performed to allow 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 via LOO approach and the global angle of the test sequence is not known in the learning phase a priori. 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

In the test phase, the learned parameter set was used to infer the labels of the left-out test sequence. An approximate Viterbi (VI) method [35, 37] and variational approximation (VA) methods [17, 35, 37, 33] were used to infer the labels in standard SLDSs. The initial probability distributions for the VA method were initialized based on the VI labels. Simply, VI labels were trusted by a probability of 0.8 and the other two labels at every time-step are assigned probability of 0.1 respectively.

For the inference in PS-SLDS, VI method was used due to its simplicity and speed. Our initial experiments indicated that it is rather difficult to measure the per-computation benefit of DD-MCMC method over VI or VA method when PS-SLDS model is adopted. As described in Sec. 4.4, the scalability issue of DD-MCMC method in PS-SLDS remains as future work and we devote the experimental result section to compare SLDS and PS-SLDS models based on computationally less-demanding VI and VA methods. In addition, the inference results on labels which are obtained via VA method complicates the update of global duration model parameters described in Eq. 24 and hence omitted in the experiments for PS-SLDS.

8.3 Qualitative Results

The experimental results show 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.17. The four color-coded strips in each figure represent SDLS VI, SLDS VA, PS-SLDS VI and the ground-truth (G.T.) labels from the top to the bottom. The x-axis represents time flow and the color is the label at that corresponding video frame.

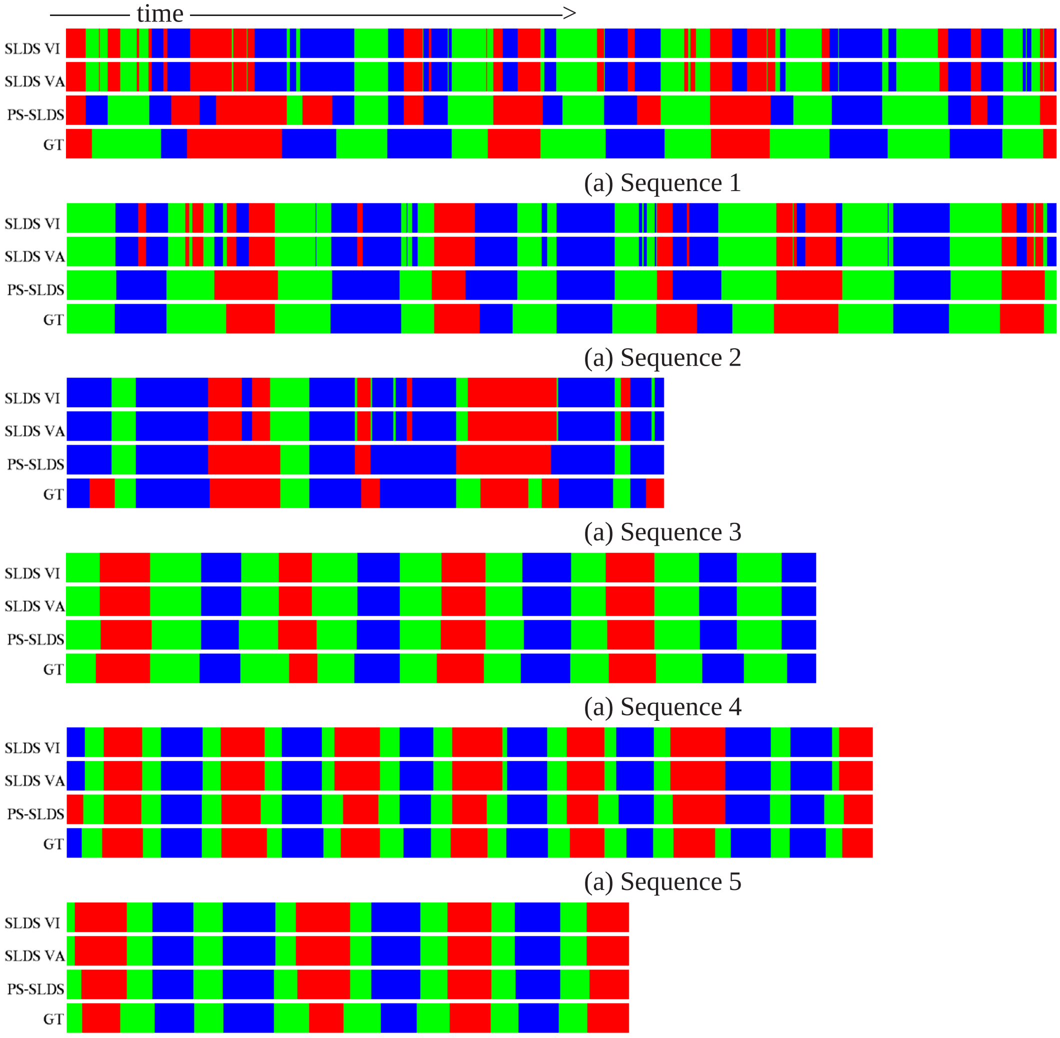
The superior recognition abilities of PS-SLDS can be observed from the presented results. The PS-SLDS results are closer to ground truth or comparable to SLDS results in all sequences. Especially, the sequences 1, 2 and 3 are challenging. The tracking results obtained from the vision-based tracker were more noisy. In addition, the patterns of switching in dance modes and the durations in each 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 the sequence 1 and 3. However, assuming that even an expert human can introduce labeling noise, the labeling capabilities of PS-SLDS are fairly good.

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 the errors of PS-SLDS estimate, SLDS estimates based on VI and VA methods and the GT angle, from top to the bottom. The best estimates are accented in bold fonts. The SLDS estimates are obtained via heuristics where we averaged the heading angles in the sequences with corresponding labels inferred as waggle. All the error values are the difference between estimated results and known G.T. values.

Among six tests, PS-SLDS and SLDS shows comparable waggle angle estimation capabilities. There is no distinguishable gap in performance between VI and VA methods. In addition, 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.



(a) Sequence 1

(a) Sequence 2

(a) Sequence 3

(a) Sequence 4

(a) Sequence 5

(a) Sequence 6

Figure 17: 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.

Sequence	1	2	3	4	5	6
PS-SLDS	-0.09	**0.01**	0.03	-0.11	**0.11**	**-0.06**
SLDS VI	**-0.05**	-0.03	**-0.02**	**-0.09**	0.18	-0.09
SLDS VA	**-0.05**	-0.03	**-0.02**	**-0.09**	0.18	-0.09
GT	-0.30	-0.25	1.13	-1.33	-2.08	-0.80

Table 1: Errors in the global rotation angle estimates from PS-SLDS and SLDS in radians. Last row contains the GT rotation angles. Sequence numbers refer to Fig. 16.

Sequence	1	2	3	4	5	6
PS-SLDS	**+13.7**	**+0.91**	**+1.9**	**-0.22**	**0.4**	**5.6**
SLDS VI	+40.8	+28.9	+11.1	-0.44	3.6	8
SLDS VA	+40.7	+28.9	+11.1	-0.44	3.6	8
GT	51.6	46.6364	21.4	41.1	19.4	32.6

Table 2: Errors in the Average Waggle Duration (AWD) estimates for PS-SLDS and SLDS in frames. Last row contains the GT AWD. Sequence numbers refer to Fig. 16.

Sequence	1	2	3	4	5	6
PS-SLDS	**75.9**	**92.4**	**83.1**	**93.4**	**90.4**	**91.0**
SLDS VI	71.6	82.9	78.9	92.9	89.7	89.2
SLDS VA	71.9	82.8	78.9	92.9	89.7	89.2

Table 3: Accuracy of label inference in percentage. Sequence number refer to Fig. 16.

The quantitative results on average waggle duration (AWD) quantification show that PS-SLDS can robustly quantify the global dynamics parameters as well. AWD is an indicator of the distance to the food source from the hive and is a valuable data to the insect biologists. Table.2 shows the errors of PS-SLDS estimate, SLDS estimates of VI and VA methods and the GT AWDs, from top to the bottom. 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-truths closely. Especially, PS-SLDS AWD estimates are impressively correct for the sequence 2, 3, 4 and 5. In contrast, it is observed that the SLDS estimates are inaccurate. More specifically, the estimates deviate far from the ground truths in most cases except for the sequence 4. The reliability of AWD estimates of PS-SLDS model is based on the robust duration modeling and the canonical parameters supported by the enhanced models.

Finally, Table.3 shows the overall accuracy of the inferred labels in percentage, statistics from PS-SLDS and SLDS VI and SLDS VA results from top to the bottom. It can be observed that PS-SLDS provides very accurate labeling results w.r.t. the ground truth. Moreover, PS-SLDS consistently improves on 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.

8.5 Discussion

It can be ambiguous to choose the right dimensionality for the hidden continuous states X. In our experiments, dimension less than four resulted in poor classification. It is conjectured that such a small dimensions does not provide rich hypothesis space to represent the motion patterns of dancer bees. On the other hand, some experiments with higher dimensions (>10) suffered from over-segmentations when the model was trained based on the provided set of training data. The dimension of four showed the best performance so far. However, the generative power of the model was tested by simulating the dance trajectories with fourth dimensional continous states and did not result in realistic trajectories yet. It is expected that the more realistic trajectories can be generated with continuous states in higher imensions (≥ 4) although the limited amount of training data is the main bottleneck that does not provide enough generalization power now. The analysis on sensitivity on dimensions and generative power of the model are planned to be reported elsewhere.

9 Conclusion

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

In our work, we encounter three challenges in proposing a practical system based on a standard SLDS model : (1) intractability of inference, (2) limited duration modeling, and (3) absence of systematic means for quantification. All these three issues were addressed by introducing three solutions where they can be used in pairs depending on the problems faced.

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 on the hidden labels even in the presence of intractability. The observation on the discovered posterior leads us to investigate two limitations of SLDSs for some practical problems : limited duration modeling and lack of systematic means to quantify the global parameters.

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 is not supported by the standard SLDSs. Nonetheless, we also demonstrated that the proposed S-SLDS model can be converted into an equivalent standard SLDS model by introducing meta variables. Such an equivalency guarantees that the large array of approximate inference algorithms developed for standard SLDSs are readily reused in S-SLDSs.

Third, parametric SLDS (P-SLDS) is introduced as an extension to provide 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 presented EM algorithm.

Finally, the experimental results on real-world honey bee dance sequences were presented 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 proposed enhanced SLDS model can robustly infer the labels and global parameters based on the learned model. A large number of over-segmentations in labeling which appeared in standard SLDSs disappear in PS-SLDS results. In addition, the results on the quantification abilities of PS-SLDS show that PS-SLDS can reliably 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 improve upon standard SLDSs for the honey bee dance data and suggest that the three pieces of development in this work might be worth being tried for challenging applications individually or in pairs, e.g., PS-SLDS.

We hope that the presented DD-MCMC method, S-SLDS model and P-SLDS model are valuable additions to the researches related to motion modeling, behavior recognition and SLDSs. The experimental results suggest that the proposed methods are promising and provide a concrete framework for a variety of vision problems where the motions that are being exhibited in the video are too complex to be modeled by HMMs.

Acknowledgments

The authors would like to thank Zia Khan and Grant Schindler for providing the bee tracking data. This work is supported in part by Samsung Lee Kun Hee scholarship awarded to Sang Min Oh, NSF Award IIS-0133779 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 helpful comments.

Appendix A. Data-Driven MCMC for SLDS

A. 1. Metropolis Hastings

Data-driven MCMC method adopts Metropolis-Hastings (MH) framework [29, 19] to generate samples from arbitrary distributions. The pseudo-code for the MH algorithm is shown in Algorithm 3 (adapted from [18]).

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*
       $$a = \frac{P(L^{(r)'}|Z)}{P(L^{(r)}|Z)} \frac{Q(L^{(r)}; L^{(r)'})}{Q(L^{(r)'}; L^{(r)})} \qquad (28)$$
   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. We put a fixed-sized window on the data and obtain cues by looking inside it. For example, the change of angles are collected as temporal cues in the bee application, as illustrated in Fig. 6.

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 for inference phase to construct proposal priors.

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.6,7,8 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*, 2005. Accepted for publication.

[3] T. Balch, Z. Khan, and M. Veloso. Automatically tracking and analyzing the behavior of live insect colonies. In *Proc. Autonomous Agents 2001*, 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] Matthew Brand and Aaron Hertzmann. Style machines. In Kurt Akeley, editor, *Siggraph 2000, Computer Graphics Proceedings*, pages 183–192. ACM Press / ACM SIGGRAPH / Addison Wesley Longman, 2000.

[8] Christoph Bregler. Learning and recognizing human dynamics in video sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1997.

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

[10] 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.

[11] P. M. Djuric and J-H. Chun. An MCMC sampling approach to estimation of nonstationary hidden Markov modles. *IEEE Trans. Signal Processing*, 50(5):1113–1123, 2002.

[12] A. Doucet and C. Andrieu. Iterative algorithms for state estimation of jump markov linear systems. *IEEE Trans. Signal Processing*, 49(6), 2001.

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

[14] A. Feldman and T Balch. Representing honey bee behavior for recognition using human trainable models. *Adaptive Behavior*, 2004.

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

[16] B. Frey and N. Jojic. Transformation-invariant clustering and dimensionality reduction using em. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(1):1–17, January 2003.

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

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

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

[20] A. Howard and T. Jebara. Dynamical systems trees. In *Conf. on Uncertainty in Artificial Intelligence*, pages 260–267, Banff, Canada, 2004.

[21] Z. Khan, T. Balch, and F. Dellaert. A Rao-Blackwellized particle filter for EigenTracking. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[22] C.-J. Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60, 1994.

[23] M.W. Lee and I. Cohen. Human upper body pose estimation in static images. In *Eur. Conf. on Computer Vision (ECCV)*, 2004.

[24] U. Lerner and R. Parr. Inference in hybrid networks: Theoretical limits and practical algorithms. In *Proc. 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 310–318, Seattle, WA, 2001.

[25] U. Lerner, R. Parr, D. Koller, and G. Biswas. Bayesian fault detection and diagnosis in dynamic systems. In *Proc. AAAI*, Austin, TX, 2000.

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

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

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

[29] 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.

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

[31] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Data-driven MCMC for learning and inference in switching linear dynamic systems. In *AAAI Nat. Conf. on Artificial Intelligence*, 2005.

[32] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Learning and Inference in Parametric Switching Linear Dynamic Systems. In *Intl. Conf. on Computer Vision (ICCV)*, 2005.

[33] 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, College of Computing, 2005.

[34] 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.

[35] V. Pavlović and J.M. Rehg. Impact of dynamic model learning on classification of human motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2000.

[36] V. Pavlović, J.M. Rehg, T.-J. Cham, and K. Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Intl. Conf. on Computer Vision (ICCV)*, 1999.

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

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

[39] 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.*, 2005.

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

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

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

[43] S. Soatto, G. Doretto, and Y.N. Wu. Dynamic Textures. In *Intl. Conf. on Computer Vision (ICCV)*, pages 439–446, 2001.

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

[45] R. Vidal, A. Chiuso, and S. Soatto. Observability and identifiability of jump linear systems. In *Proceedings of IEEE Conference on Decision and Control*, 2002.

[46] Andrew D. Wilson and Aaron F. Bobick. Parametric hidden Markov models for gesture recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 21(9):884–900, 1999.

[47] Y.Li, T.Wang, and H-Y. Shum. Motion texture : A two-level statistical model for character motion synthesis. In *SIGGRAPH*, 2002.

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