Mixture of Switching Linear Dynamics to Discover Behavior Patterns in Object Tracks

Julian F. P. Kooij, Gwenn Englebienne, and Dariu M. Gavrila

Abstract—We present a novel non-parametric Bayesian model to jointly discover the dynamics of low-level actions and high-level behaviors of tracked objects. In our approach, *actions* capture both linear, low-level object dynamics, and an additional spatial distribution on where the dynamic occurs. Furthermore, *behavior classes* capture high-level temporal motion dependencies in Markov chains of actions, thus each learned behavior is a switching linear dynamical system. The number of actions and behaviors is discovered from the data itself using Dirichlet Processes. We are especially interested in cases where tracks can exhibit large kinematic and spatial variations, e.g. person tracks in open environments, as found in the visual surveillance and intelligent vehicle domains. The model handles real-valued features directly, so no information is lost by quantizing measurements into 'visual words', and variations in standing, walking and running can be discovered without discrete thresholds. We describe inference using Markov Chain Monte Carlo sampling and validate our approach on several artificial and real-world pedestrian track datasets from the surveillance and intelligent vehicle domain. We show that our model can distinguish between relevant behavior patterns that an existing state-of-the-art hierarchical model for clustering and simpler model variants cannot. The software and the artificial and surveillance datasets are made publicly available for benchmarking purposes.

Index Terms—Human behavior analysis, hierarchical non-parametric graphical model, switching linear dynamical systems

1 Introduction

Observing and reasoning about object motion patterns are key tasks in numerous real-world application domains, such as visual surveillance (e.g., [1], [2], [3], [4]), robotics (e.g., [5]), and intelligent vehicles (e.g., [6], [7], [8]). Computer vision and machine learning techniques can capture and analyze person trajectories, and detect anomalous movements that deviate from normative behavior found in the training data. In fixed-camera video surveillance, this could aid human operators to monitor many video streams, and focus their attention on possible incidents. In the intelligent vehicles domain, accurate object motion models enable sophisticated driver assistance systems to tracks multiple objects simultaneously, and perform path prediction for situation analysis and object avoidance.

A few issues arise when modeling complex object motion patterns from low-level observations. First, how is high-level behavior structured and composed of low-level actions; second, where in the observed space does specific behavior occur; third, how can temporal dynamics of behavior be exploited? Ideally, action decomposition, spatial context, and temporal dynamics should be inferred jointly from the training data. Some previous work [1], [2], [3] has modeled behavior at the image level to capture patterns that govern the whole scene (e.g., monitoring traffic flow at junctions). We, however, target individual behavior patterns where execution of the same movement may have large spatial and kinematic variations.

In this paper we propose a mixture of switching linear dynamic systems to discover normative actions and their temporal relations at the object level. Actions describe low-level motion dynamics occurring in a semantic region using tracked object locations as observations. We use continuous distributions in the feature space to capture variance in action execution. As Fig. 1 illustrates, our unsupervised approach segments tracks into sequences of common actions and jointly clusters the action sequences into distinct behavior classes. In our Bayesian inference scheme, the number of actions and the number of behaviors are not fixed but discovered from the data itself using Dirichlet process (DP) mixture models.

2 Previous Work

Motion models are essential to tasks as tracking [9], [10], [11], [12], path prediction [6], [7], [13], [14], and anomalous track detection [15], [16], found in a variety of application domains ranging from intelligent vehicles [6], [7], [8], [11], [12], [13], [17] to surveillance [10], [15], [16], [18], [19]. For instance, methods that build tracks from object detections [11], [20] use motion models to reduce false positives and resolve data association [11], [12].

When both the motion dynamics and observations can be expressed as a linear transformation of a latent state with added Gaussian noise, then the resulting model is a linear dynamical system (LDS). The Kalman filter (KF) [9] can then be used for efficient online inference, expressing uncertainty over the state by a single normal distribution. For

- *J. F. P. Kooij and G. Englebienne are with the Intelligent Systems Lab, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands. E-mail: julian.kooij@gmail.com, G.Englebienne@uva.nl.*
- *D. M. Gavrila is with the Intelligent Systems Lab, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands, and the Environment Perception Department, Daimler R&D, Ulm, Germany. E-mail: dariu@gavrila.net.*

Manuscript received 3 Oct. 2014; revised 20 Feb. 2015; accepted 22 Apr. 2015. Date of publication 9 June 2015; date of current version 13 Jan. 2016. Recommended for acceptance by S. Nowozin.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2443801

0162-8828 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

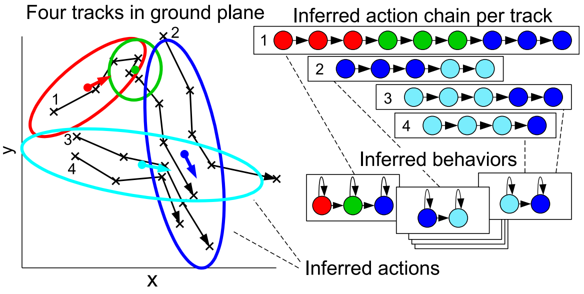


Fig. 1. Inferring a mixture of switching linear dynamic systems from tracks (see Section 1). Black crosses are observations. Tracks are segmented into actions, each action describing a semantic region (2D Gaussian) and motion as a linear dynamical system (mean motion shown as arrow). Behaviors cluster tracks with similar action chains. In this example, four actions and three behaviors are inferred. The red and green actions distinguish walking and standing in track 1. Tracks 3 and 4 are clustered in the same behavior, but track 2 has a behavior with different action order.

batch processing, the Kalman Smoother [21] extends the KF by a backward pass to incorporate information from future time steps. However, a LDS cannot represent alternating dynamics, such as a pedestrian changing from walking to standing [6], [7]. For such maneuvering targets, Switching Linear Dynamical Systems (SLDS) [22], [23] use a latent top-level Markov chain of switching states to select one of K possible linear dynamics each time step. The K switching states have their own prior and transition probabilities. Unfortunately, exact inference in a SLDS is intractable as the number of modes in the state distribution grows exponentially over time with K, the number of the switching states [22], [24].

For online real-time inference, assumed density filtering (ADF) [25], [26] approximates the state posterior at every time step with a simpler distribution in closed form. ADF can be applied to discrete state Dynamic Bayesian Networks, known as Boyen-Koller inference [27], and more generally to mixed discrete-continuous state spaces with conditional Gaussian posterior [28] such as the SLDS [6]. Interacting Multiple Model KF (IMM-KF) [9] mixes the states of several KFs running in parallel, which has lower computational cost than ADF for SLDS, and is commonly found in the intelligent vehicle domain [7], [13].

To learn parameters, however, complete sequences of switching states must be estimated jointly, e.g. using EM, Viterbi or Variational inference [24]. [22] presents a Markov chain monte carlo (MCMC) method to approximate the switching state posterior with samples, while integrating over the latent positions. MCMC sampling has also proven useful when extending the SLDS, e.g., to impose distributions on persistent state durations [29]. Note that one could learn model parameters using MCMC, and then perform online inference using IMM-KF [30]. This paper will focus on unsupervised offline learning of motion models, but similarly does not restrict how learned models are utilized for online inference.

Another key difficulty when learning SLDS states, and mixture models in general, is determining the number of components or states K. We therefore take a moment to review the use of Dirichlet process in non-parametric mixture models that learn the number of components. DP mixture models are often described as a *Stick-breaking* prior over any number of component weights, and a 'base' distribution over component parameters [23], [31]. Stick-breaking refers to the analogy of iteratively breaking off a part from a unit-length stick, so the lengths of all parts (plus remainder) always sum to one, and represent component weights. As this iterative process could in principle be repeated indefinitely, the resulting distribution is said to be *infinite* dimensional. A draw from a DP yields an existing component with probability proportional to its stick length, or a new one proportional to the remainder (in which case new parameters are sampled from the base distribution, and a part is broken of the remainder). A concentration parameter expresses prior preference for few large or many small clusters. The posterior on component weights is always finite, reflecting the number of components that the data actually supports.

DPs are therefore encountered in unsupervised learning of mixtures in Markov models [23], [32], and *topic models* for text document analysis [31] (which have also been applied to behavior modeling tasks in computer vision [1], [2], [3], [15]). For instance, the hierarchical Dirichlet process (HDP) [31] extends latent dirichlet allocation (LDA) [33], and discovers common topics in a corpus of unlabeled documents represented as a bag-of-words [33]. HDP regards each document as a DP mixture of topics, where topics (i.e. mixture components) are distributions over words. The document DPs use as base distribution one global DP on topic parameters, hence documents draw topics from a shared 'infinite' topic mixture. Multi-level mixtures are also used in the infinite hidden Markov model (HDP-HMM), which learns the number of states in the Markov chain [31], [32]. The HDP-HMM has been further extended to the HDP-SLDS [23] to learn switching states.

Previous work shows that models for multiple dynamics can benefit from identifying where dynamics occur (spatial context) and what the long-term behavior is (temporal context). For instance, [6] improves pedestrian path prediction in intelligent vehicles with a SLDS where state transitions are informed by contextual cues of scene layout and pedestrian attention (determined by head orientation). Online inference with ADF predicts various behaviors (e.g. crossing or halting at curb while (un)aware of the approaching vehicle) that are composed of two dynamics (walking and standing).

Unsupervised learning of behavior usually involves clustering track data, e.g with a pair-wise distance measures (such as Euclidean [34] or Hausdorff [35] distance), dynamic time warping (DTW) [36], or by longest common subsequence matching [7]. These methods are however not probabilistic [34], [35], [36] or computationally demanding [6], [7], and the complexity of clustering N trajectories is $O(N^2)$ (c.f. [15]). [37] clusters tracks with similar start and exit regions, and learns a LDS per cluster. These track clusters provide useful insight in the underlying behavior patterns, and are exploited for crowd simulation.

Instead of clustering full tracks, others cluster observations into semantic regions for activity modeling and anomaly detection [35], [38], [39]. [14] proposes a Markov decision process to model the influence of the scene layout onto future pedestrian actions, and predict track paths towards various possible endpoints. Here, labeled regions found by a semantic image segmentation step, inform where people tend to walk (e.g., road or sidewalk). Learned regions can also represent dynamics as local flow fields of the continuous position/velocity pairs using Lie algebra [40], or Gaussian Processes [5], but these methods do not consider alternating dynamics within a single track.

Dual-HDP [15] on the other hand is a topic model that uses multi-level DPs to *jointly* cluster unlabeled tracks into behavior classes, and decompose them into regions of motion. Each track is seen as a 'document' represented as a bag-of-words, 'words' being quantized position / motion pairs. Topics thus cluster co-occurring words, forming a region of motion, and tracks with similar topic distributions are clustered into behaviors classes. A trained HDP can detect unexpected movements in common areas (e.g. vehicle moving against traffic) or uncommon combinations of areas (e.g. u-turn at junction). However, quantization can lead to sparse bins, and learned models are not suited for applications relying on LDS [6], [9], [13].

While Dual-HDP learns long-term behavior of individuals from tracked object paths, there are also methods that combine temporal dynamics and topic models for scene-wide analysis of video. With quantized optical flow features as 'visual words', topics describe regions of co-occurring motion within frames. In [1] dynamics are modeled using a Markov chain on top of LDA. The state of the chain determines the topic distribution at each frame. This approach is used to learn models for traffic junctions where the Markov chain captures the dynamics of traffic flow. [2] extends this to an infinite mixture of infinite Markov chains with HDP, giving more flexibility. In [3] a combination of HDP with Probabilistic Latent Sequential Motifs [41] is used to sequential patterns of scene wide flow.

3 Proposed Approach

Our method processes track data, e.g., obtained from tracked or annotated object paths, where each track is an ordered list of 2D measured positions on the ground plane. In our unsupervised approach, tracks are clustered into *behaviors*, each behavior defining transition probabilities between *actions*, which we refer to as *topics* in the remainder of this paper, to follow the nomenclature of topic models. Each topic describes for a common action the spatial area, and the low-level motion dynamics with an LDS. Topics can be shared among behaviors, thus multiple behaviors may contain the same topic but use different topic transition probabilities. Since each behavior is a SLDS, the full model forms a Mixture of SLDS. The temporal dynamics capture topic duration with the self-transition probability, and help distinguish between behaviors with spatially overlapping actions. In Fig. 1 for instance, tracks 2 and 3 have the same actions but different behaviors.

3.1 Contributions

Our main contributions are (1) a hierarchical model to jointly discover what and how many low-level actions and high-level behavior classes are present in unlabeled track data, for which (2) we present an MCMC inference scheme. Our model (3) extends SLDS switching states with spatial distributions (spatial context), and (4) can discriminate behaviors with different action orders (temporal context). (5) Features are not quantized, but actions capture motion and variance directly in the continuous feature space.

Dual-HDP [15] also jointly discovers actions and behavior classes, but its bag-of-words representation does not capture temporal order, and quantizing position/motion pairs can lead to sparse data, even at low binning resolution. Like [15], we derive spatial and motion distributions from tracked 2D positions, but unlike [15] use Gaussians and LDSs to represent both in the continuous feature space.

While SLDS has been combined with HDP [23] before, the combination of SLDS with spatial distributions and multi-level track clustering is novel. Track clusters provide insight in the types of long-term behavior in the data, and induce higher-order dependencies between actions, as opposed to a single SLDS [23] that only models first-order dependencies.

This paper is based on our earlier conference papers [16] and [42]. In [16] we learned spatial distributions of where different states occur, while clustering the tracks that exhibit similar state transitions. Compared to this conference submission, we propose an improved sampling method, and extend our evaluation with more methodological comparisons on datasets from different application domains. We use data from [42], where no multi-level clustering was discussed.

3.2 Multi-Level Clustering

The data consists of J tracks, each being a sequence of T_j spatial 2D measurements x_{jt} with j the track index and t the time index. In our model the indicator variable $z_{jt} = k$ indicates that observation x_{jt} is generated by topic k. To simplify notation we define $\mathbf{x}_j = \{x_{j1}, \ldots, x_{jT_j}\}$, $\mathbf{z}_j = \{z_{j1}, \ldots, z_{jT_j}\}$, and we denote the superscript $-t$ in $\mathbf{z}_j^{-t}$ to indicate all z_{jt} of track j except t (throughout the paper we use boldface notation for vectors or sets that can be indexed). Measurements x_{jt} are used as both spatial and dynamic features which we denote as x_{jt}^{loc} and x_{jt}^{dyn} respectively. For clarity of exposition, this section first describes multi-level clustering in our model without the low-level motion dynamics. These will be included in Section 3.3.

Each topic k defines a probability distribution over the location x_{jt}^{loc} on the ground plane (i.e., a semantic region) as a 2D Gaussian parametrized by μ_k, Σ_k. Further, each track is assigned to a behavior, indexed by c_j, where a behavior, or *cluster*, c defines the topic transition probabilities using a vector $\tilde{\boldsymbol{\pi}}_c^k = \{\tilde{\pi}_c^{k(1)}, \tilde{\pi}_c^{k(2)}, \ldots\}$ for each k, i.e $p(z_{jt} = k'|z_{jt-1} = k, c) = \tilde{\pi}_c^{k(k')}$[1]. We add extra indices $z_{j0} = init$ for the start of the topic chain, such that $p(z_{j1}|init, c)$ is the initial topic distribution, and $z_{jT_j+1} = exit$ for track termination [37]. $\mathbf{z}_c$ denotes all z_j with $c_j = c$, and $\mathbf{x}^{-j}, \mathbf{z}^{-j}, \mathbf{c}^{-j}$ are respectively all observations, topic labels and behavior labels except those of track j.

The multinomial topic distributions $\tilde{\boldsymbol{\pi}}_c^k$ are sampled from a DP over a behavior-specific topic distribution $\boldsymbol{\pi}_c$. The various $\boldsymbol{\pi}_c$ are sampled from a DP over $\boldsymbol{\pi}_0$, which is the global topic distribution shared by all behaviors. The distribution $\boldsymbol{\pi}_0$ follows a Stick-breaking construction and thus represents a multinomial distribution over infinite topics, although at any time during inference only some K topics will actually

[1] We shall adopt the common notation for this categorical distribution as an instance of the multinomial distribution, since indicator variables can be written as 1-of-K vectors [25].

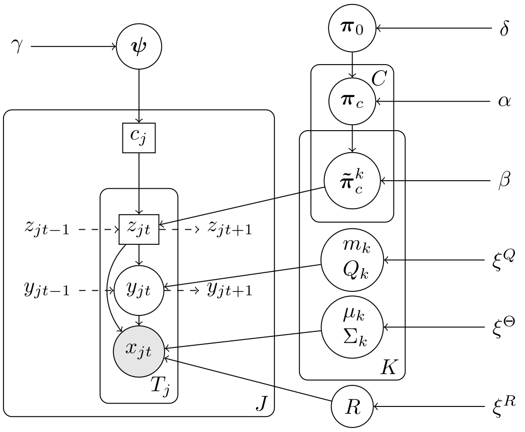


Fig. 2. The Mixture of SLDS, see Sections 3.2 and 3.3. Square nodes are discrete, dashed arrows are temporal dependencies. Any time during inference, K topics and C behaviors are represented. Track j consists of T_j observations x_{jt}, latent positions y_{jt}, topic label z_{jt} and has behavior label c_j drawn from distribution $\boldsymbol{\psi}$. $\boldsymbol{\pi}_c$ is the topic distribution in behavior c, and $\tilde{\boldsymbol{\pi}}_c^k$ are transition probabilities from topic k. Behaviors sample topics from global topic distribution $\boldsymbol{\pi}_0$. R is the observation noise. m_k, Q_k are the LDS and μ_k, Σ_k the spatial region (Gaussian) parameters of topic k. $\alpha, \beta, \gamma, \delta$ are DP concentration hyperparameters, ξ^R, ξ^Θ, ξ^R are priors on the continuous distributions.

be used. Note how a behavior c describes a HDP-HMM [31], where the $\{\tilde{\boldsymbol{\pi}}_c^k\}$ form the rows of the transition matrix. Further, the distribution $\boldsymbol{\pi}_c$ concentrates probability mass on a subset of the topics present in $\boldsymbol{\pi}_0$, and the transition matrices $\tilde{\boldsymbol{\pi}}_c$ are constrained to use those topics from $\boldsymbol{\pi}_c$. Without the $\boldsymbol{\pi}_c$ and corresponding prior, distinct behaviors that do not share topics tend to be merged.

The hierarchical model, which is represented graphically in Fig. 2, can be summarized as

$$\boldsymbol{\pi}_0 | \delta \sim \mathrm{Stick}(\delta) \tag{1}$$

$$\boldsymbol{\pi}_c | \boldsymbol{\pi}_0, \alpha \sim \mathrm{DP}(\alpha, \boldsymbol{\pi}_0) \tag{2}$$

$$\tilde{\boldsymbol{\pi}}_c^k | \boldsymbol{\pi}_c, \beta \sim \mathrm{DP}(\beta, \boldsymbol{\pi}_c) \tag{3}$$

$$z_{jt} | z_{jt-1}, c_j, \{\tilde{\boldsymbol{\pi}}_c^k\} \sim \mathrm{Mult}(\tilde{\boldsymbol{\pi}}_{c_j}^{z_{jt-1}}) \tag{4}$$

$$x_{jt}^{loc} | z_{jt}, \{\mu_k, \Sigma_k\} \sim \mathcal{N}(\mu_{z_{jt}}, \Sigma_{z_{jt}}) \tag{5}$$

$$\mu_k, \Sigma_k | \xi^\Theta \sim \mathcal{NW}^{-1}(\xi^\Theta), \tag{6}$$

where α, β and δ are concentration hyper-parameters. $\xi^\Theta = (\mu_0^\Theta, \kappa_0^\Theta, \nu_0^\Theta, \Psi_0^\Theta)$ are the hyper-parameters for the Normal-Inverse-Wishart (NIW) distribution [25], which is the conjugate prior for the mean and covariance of a multivariate Normal. Behavior labels c_j are sampled from the multinomial $\boldsymbol{\psi}$, which also follows a Stick-breaking distribution,

$$\boldsymbol{\psi} | \gamma \sim \mathrm{Stick}(\gamma), \quad c_j | \boldsymbol{\psi} \sim \mathrm{Mult}(\boldsymbol{\psi}). \tag{7}$$

3.3 Low-Level Dynamics

The hierarchical model is extended with a SLDS on the labels $\mathbf{z}_j$ by introducing latent 2D-position variables y_{jt}. Topics now not only define a distribution over the 2D space, but also the low-level dynamics of the position sequence. In fact, topic labels $\mathbf{z}_j$ form a Markov chain of switch variables which select the stochastic state dynamics.

$$y_{jt} = A y_{jt-1} + q_{jt} \qquad q_{jt} \sim \mathcal{N}(m_{z_{jt}}, Q_{z_{jt}}) \tag{8}$$

$$x_{jt}^{dyn} = C y_{jt} + r_{jt} \qquad r_{jt} \sim \mathcal{N}(0, R). \tag{9}$$

Matrices A and C are fixed and determine the type of kinematics used. In our experiments we set A and C to identity, resulting in a fixed-velocity model where the learned velocity is captured in the mean of the process noise, $m_{z_{jt}}$. Unlike in [16], we let the topics share the observation noise parameters R, since this noise is typically a systematic error that does not depend on the topic z_{jt}. Using appropriate priors on the noise components, the model now also defines

$$y_{jt} | y_{jt-1}, z_{jt}, \{m_k, Q_k\} \sim \mathcal{N}(A y_{jt-1} + m_{z_{jt}}, Q_{z_{jt}}) \tag{10}$$

$$x_{jt}^{dyn} | y_{jt}, R \sim \mathcal{N}(C y_{jt}, R) \tag{11}$$

$$m_k, Q_k \sim \mathcal{NW}^{-1}(\xi^Q), \qquad R \sim \mathcal{W}^{-1}(\xi^R) \tag{12}$$

with $\xi^Q = (\mu_0^Q, \kappa_0^Q, \nu_0^Q, \Psi_0^Q)$ and $\xi^R = (\nu_0^R, \Psi_0^R)$ the hyper-parameters for a NIW and Inverse-Wishart (IW) distribution respectively.

4 Bayesian Inference

Posterior inference in our model relies on two types of Markov chain monte carlo sampling, namely Gibbs and Metropolis-Hastings (MH). Generally, MCMC constructs a Markov chain of samples which has as stable distribution the joint posterior of a model's latent variables $\mathrm{s} = \{s_1, s_2, \ldots\}$. Gibbs sampling draws each s_j in turn from their posterior $p(s_j | \mathrm{s}^{-j}, \mathrm{x})$, while keeping the other s^{-j} fixed. However, in some cases this posterior is hard to obtain. The MH rule states that we can then instead sample a candidate $s_j^\star \sim q$ from a more convenient proposal distribution $q(s_j^\star | s_j, \mathrm{s}^{-j})$, and accept the proposal $s_j^\star$ with probability $\mathcal{A}$, or reject it and repeat the current s_j otherwise, where

$$\mathcal{A} = \min\left(\frac{p(s_j^\star | \mathbf{x}, \mathrm{s}^{-j}) \; q(s_j | s_j^\star, \mathrm{s}^{-j})}{p(s_j | \mathbf{x}, \mathrm{s}^{-j}) \; q(s_j^\star | s_j, \mathrm{s}^{-j})}, 1\right). \tag{13}$$

The ratio in this term effectively removes the bias that is introduced by sampling from the proposal distribution q instead of the true distribution p.

4.1 Dirichlet Process Hierarchy

As in the hierarchical Dirichlet process [31], the Stick-breaking distribution over $\boldsymbol{\pi}_0$ does not need to be implemented explicitly. It can be approximated by a Dirichlet distribution, using concentration parameter δ in a parameter *vector* of fixed length K [31],

$$\boldsymbol{\pi}_0 \sim \mathrm{Dir}([\delta/K, \ldots, \delta/K]). \tag{14}$$

This yields the same distribution as the Stick-breaking prior in the limit of $K \to \infty$. In practice, we fix K to an upper-limit on the actual number of clusters, which is discovered from the data, but reduces the burden on updating all data structures for new or deleted topics. The actual number of clusters is found by counting the number of non-zero elements in $\boldsymbol{\pi}_0$. The prior on each $\boldsymbol{\pi}_c$ can also be expressed [31] as a Dirichlet distribution $\boldsymbol{\pi}_c \sim \mathrm{Dir}(\alpha\boldsymbol{\pi}_0)$, and the prior on each row in a transition matrices as $\tilde{\boldsymbol{\pi}}_c^{k'} \sim \mathrm{Dir}(\beta\boldsymbol{\pi}_c)$.

We now first formulate some posteriors that will be used in the next sections. The transition matrix can be integrated out analytically to obtain the probability of a topic sequence $\mathbf{z}_j$ for given $\boldsymbol{\pi}_c$ (c.f. [31]),

$$p(\mathbf{z}_j|\boldsymbol{\pi}_c) = \prod_k \int p\big(\mathbf{m}_j^k|\tilde{\boldsymbol{\pi}}_c^k\big)p\big(\tilde{\boldsymbol{\pi}}_c^k|\beta\boldsymbol{\pi}_c\big)\,\mathrm{d}\tilde{\boldsymbol{\pi}}_c^k \tag{15}$$

$$= \prod_k \left[\frac{\Gamma(\sum_{k'} \beta\pi_c^{(k')})}{\Gamma(\sum_{k'}(\beta\pi_c^{(k')} + m_j^{k(k')}))} \prod_{k'} \frac{\Gamma(\beta\pi_c^{(k')} + m_j^{k(k')})}{\Gamma(\beta\pi_c^{(k')})}\right].$$

Here $m_j^{k(k')}$ are the number of state transitions from $z_{jt-1} = k$ to $z_{jt} = k'$ in track j. $m_c^{k(k')} = \sum_{j|c_j=c} m_j^{k(k')}$ are the total transition counts in behavior c.

If topics $\mathbf{z}_c^{-j}$ have been assigned to behavior c, the posterior $p(\mathbf{z}_j|\boldsymbol{\pi}_c, \mathbf{z}_c^{-j})$ is obtained by adding the corresponding topic occurrence counts to the prior [31], i.e., substituting $\beta\pi_c^{(k')}$ for $\beta\pi_c^{(k')} + m_{c^{-j}}^{k(k')}$ in Eq. (15). For a single transition of k at time $t-1$ to k' at t, as used in Eq. (22), this posterior reduces to just

$$p(z_{jt} = k'|z_{jt-1} = k, \boldsymbol{\pi}_c, \mathbf{z}_c^{-jt}) \propto \beta\pi_c^{(k')} + m_{c^{-j}}^{k(k')}. \tag{16}$$

As we see here, clustering is a result from common transitions obtaining a higher prior probability.

Next, note that the Dirichlet prior on $\boldsymbol{\pi}_c$ is not conjugate with Eq. (15), but we can employ the auxiliary variables scheme of [31] to obtain $\tau_c^{k(k')}$, which represent the number of times topic k' would have been drawn from the base distribution $\boldsymbol{\pi}_c$ in the DP of $\tilde{\boldsymbol{\pi}}_c^k$. So the posterior on $\boldsymbol{\pi}_c$ can then be written as

$$p(\boldsymbol{\pi}_c|\boldsymbol{\pi}_0, \mathbf{z}_c) = \mathrm{Dir}\left(\boldsymbol{\pi}_c|\alpha\boldsymbol{\pi}_0 + \sum_{k'} \boldsymbol{\tau}_c^{k'}\right). \tag{17}$$

Eq. (16) and (17) will be used in the next section. The auxiliary variable scheme is also used to obtain the posterior $p(\boldsymbol{\pi}_0|\{\boldsymbol{\pi}_c\})$, but then with auxiliary variable representing draws from $\boldsymbol{\pi}_0$. We use this posterior to Gibbs sample top-level topic distribution $\boldsymbol{\pi}_0$.

4.2 Assigning Tracks to Clusters

In [16] we also Gibbs sampled topic labels $\mathbf{z}_j$, while keeping cluster label c_j fixed, and in turn we sampled c_j while keeping $\mathbf{z}_j$ fixed. However, the behavior and topic labels are strongly correlated, and these Gibbs samples can mix slowly. To improve mixing of tracks between clusters, we here sample from $p(c_j|\mathbf{x}, \mathbf{c}^{-j}, \boldsymbol{\pi}_0)$, with $\mathbf{z}_j$ *not* fixed, by performing a MH step to propose a label c_j. The target distribution is

$$p(c_j|\mathbf{x}, \mathbf{c}^{-j}, \boldsymbol{\pi}_0) \propto p(c_j|\mathbf{c}^{-j})p(\mathbf{x}_j|c_j, \boldsymbol{\pi}_0). \tag{18}$$

In the first term, the prior $\mathbf{c}$ over ψ has been integrated out analytically.

Let n_c^{-j} be the occurrence count of behavior c in $\mathbf{c}^{-j}$, then due to the Stick-breaking prior

$$p(c_j = c|\mathbf{c}^{-j}) \propto \begin{cases} n_c^{-j} & \text{for existing class } c \\ \gamma & \text{for new class } c = c_{\text{new}}. \end{cases} \tag{19}$$

We use Eq. (19) also as the MH proposal distribution.

The MH accept-reject also requires the likelihood $p(\mathbf{x}_j|c_j, \boldsymbol{\pi}_0)$ of observing a track j in cluster c_j. Let $\mathbf{z}_c^{-j}$ be topics from other tracks assigned to c_j, then

$$\begin{aligned} p(\mathbf{x}_j|c_j = c, \boldsymbol{\pi}_0) &= \int_{\mathbf{z}_j\boldsymbol{\pi}_c} p(\mathbf{x}_j|\mathbf{z}_j)p(\mathbf{z}_j|\boldsymbol{\pi}_c, \mathbf{z}_c^{-j})p(\boldsymbol{\pi}_c|\boldsymbol{\pi}_0) \\ &= \mathbb{E}_{\mathbf{z}_j,\boldsymbol{\pi}_c|\mathbf{z}_c^{-j},c_j=c,\boldsymbol{\pi}_0}\big[p(\mathbf{x}_j|\mathbf{z}_j)\big], \end{aligned} \tag{20}$$

i.e. the expected data likelihood under the $\mathbf{z}_j$ distribution in cluster c_j (for the likelihood of $c_j = c_{\text{new}}$, $\mathbf{z}_c^{-j}$ is an empty set). In a SLDS, computing the data likelihood $p(\mathbf{x}_j|\mathbf{z}_j)$ of a track j with given state sequence $\mathbf{z}_j$ is done straightforward with the Kalman filter equations, as the linear dynamics and transition matrices at each time step are fully specified. Still, the expectation cannot be evaluated in closed form, but it could be estimated empirically by drawing joint samples for $(\mathbf{z}_j, \boldsymbol{\pi}_c)$ from the prior $p(\mathbf{z}_j|\boldsymbol{\pi}_c, \mathbf{z}_c^{-j})p(\boldsymbol{\pi}_c|\boldsymbol{\pi}_0)$, and averaging the obtained values for $p(\mathbf{x}_j|\mathbf{z}_j)$. However, most samples from this prior will not contain relevant values for $\mathbf{z}_j$ where the likelihood term has any mass.

Instead, we can use importance sampling, where we again utilize a proposal distribution q to obtain relevant samples $(\mathbf{z}_j, \boldsymbol{\pi}_c)$. We rewrite Eq. (20) as

$$p(\mathbf{x}_j|c_j, \boldsymbol{\pi}_0) = \mathbb{E}_q\left[p(\mathbf{x}_j|\mathbf{z}_j)\frac{p(\mathbf{z}_j|\boldsymbol{\pi}_c, \mathbf{z}_c^{-j})p(\boldsymbol{\pi}_c|\boldsymbol{\pi}_0)}{q(\mathbf{z}_j, \boldsymbol{\pi}_c)}\right] \tag{21}$$

and compute the expectancy empirically from the samples of q. We find that posteriors $p(\boldsymbol{\pi}_c|\boldsymbol{\pi}_0, \mathbf{z}_c)$ from Eq. (17) and $p(\mathbf{z}_j|\mathbf{x}_j, \mathbf{z}_c^{-j}, \boldsymbol{\pi}_c)$, (explained in the next section) provide good proposals $q(\mathbf{z}_j, \boldsymbol{\pi}_c)$.

In summary, we first obtain posterior samples for $\mathbf{z}_c$ and $\boldsymbol{\pi}_c$, similar to the Gibbs sampling scheme in [16]. However, [16] then constructed a full posterior to sample c_j, conditioned on $\mathbf{z}_j$ and $\boldsymbol{\pi}_c$. Here, we empirically integrate over $\mathbf{z}_j$ in Eq. (21) to estimate the posterior Eq. (18) for a proposed cluster label c_j within a MH accept-reject step.

4.3 Linear Dynamics Information Filter

For the proposal distribution $q(\mathbf{z}_j, \boldsymbol{\pi}_c)$ in Eq. (21), we need to sample a track's latent topic chain $\mathbf{z}_j$ from the posterior $p(\mathbf{z}_j|\mathbf{x}_j, \mathbf{z}_c^{-j}, \boldsymbol{\pi}_c)$ for a candidate cluster c. Instead of sampling values for the hidden positions $\mathbf{y}_j$ [23], we shall sample z_{jt} sequentially from the distribution $p(z_{jt}|\mathbf{x}_j, \mathbf{z}_c^{-jt}, \boldsymbol{\pi}_c)$ [16], [22], [23]. In [22] it is shown that for a SLDS the Kalman information filter can be used to compute efficiently in $O(T_j \times K)$,

$$\begin{aligned} p(z_{jt}|\mathbf{x}_j, \mathbf{z}_c^{-jt}, \boldsymbol{\pi}_c) &\propto p(x_{jt}|z_{jt}, \mathbf{x}_{j1:t-1}) \\ &\times p\big(z_{jt}|z_{jt-1}, \boldsymbol{\pi}_c, \mathbf{z}_c^{-jt}\big)p\big(z_{jt+1}|z_{jt}, \boldsymbol{\pi}_c, \mathbf{z}_c^{-jt}\big) \\ &\times \int p(\mathbf{x}_{jt+1:T_j}|y_{jt}, \mathbf{z}_{jt+1:T_j})p(y_{jt}|\mathbf{x}_{j1:t}, \mathbf{z}_{j1:t})\mathrm{d}y_{jt}. \end{aligned} \tag{22}$$

The first term is the predictive distribution of the observation, found by the forward Kalman filter, the second and third terms are topic transition probabilities from Eq. (16). The integral can be analytically solved and expressed in term of the forward and backward filtering statistics (c.f. [22]). The filter also provides Kalman Smoother estimates $p(y_{jt}|\mathbf{x}_j^{-t}, \mathbf{z}_j) = \mathcal{N}(y_{jt}|\cdot)$, which will be used in Section 4.4.

There is again strong correlation between the subsequent topic labels, and sampled sequences can get stuck in a single mode. To explore alternative modes within a behavior, we first reinitialize all $\mathbf{z}_j$ randomly, and then perform five iterations over the sequence with Eq. (22) to obtain a consistent $\mathbf{z}_j$ sample (typically, the $\mathbf{z}_j$ converge after only a few iterations).

4.4 Sampling SLDS Parameters

The SLDS parameters are Gibbs sampled from their estimated posteriors. For instance, the prior on dynamics (m_k, Q_k) is NIW, see Eq. (12). Due to conjugacy with Eq. (8), its posterior is also NIW with parameters $\xi^Q_{+k} = (\mu^Q_{+k}, \kappa^Q_{+k}, \nu^Q_{+k}, \Psi^Q_{+k})$. Given ξ^Q, and the N_k motion vectors $q_{jt|z_{jt}=k}$ associated to k by labels $\mathbf{z}$, we could compute posterior parameters ξ^Q_{+k} from the sample mean, $\hat{q}_k$, and scatter matrix $\hat{S}_k = \sum_{jt|z_{jt}=k}((q_{jt} - \hat{q}_k)(q_{jt} - \hat{q}_k)^\top)$ (see [25]),

$$\kappa^Q_{+k} = \kappa^Q_0 + N_k \qquad \nu^Q_{+k} = \nu^Q_0 + N_k \tag{23}$$

$$\mu^Q_{+k} = \left(\kappa^Q_0/\kappa^Q_{+k}\right)\mu^Q_0 + \left(N_k/\kappa^Q_{+k}\right)\hat{q}_k \tag{24}$$

$$\Psi^Q_{+k} = \Psi^Q_0 + \hat{S}_k + \frac{\kappa^Q_0 N_k}{\kappa^Q_{+k}}\left(\hat{q}_k - \mu^Q_0\right)\left(\hat{q}_k - \mu^Q_0\right)^\top. \tag{25}$$

However, the true motion vectors $q_{jt} = y_{jt} - Ay_{jt-1}$ (Eq. (8)), and observation noise vectors $r_{jt} = x^{dyn}_{jt} - Cy_{jt}$ (Eq. (9)) depend on the values of all latent $\mathbf{y}_j$. While we do not know the true $\mathbf{y}_j$, recall that the Kalman Information Filter (Section 4.3) also provided posterior distributions over these latent positions as normals, i.e., the output of a Kalman Smoother. One could then sample *pseudo*-true positions $\mathbf{y}_t$ from those posteriors [23], but this is computationally demanding and may lead to slower convergence. Instead, we use the smoothed estimates of the joint distributions $\mathcal{N}(y_{jt-1}, y_{jt}|\mathbf{x}_j, \mathbf{z}_j)$ to compute the distributions over the motion vectors $p(q_{jt}|\mathbf{x}_j, \mathbf{z}_j)$, which are normals too and parameterized as $\mathcal{N}(q_{jt}|\bar{\mu}_{jt}, \bar{\Sigma}_{jt})$. In Eq. (24)-(25) we then use the expected values of $\hat{q}_k$ and $\hat{S}_k$,

$$\mathbb{E}[\hat{q}_k] = \frac{1}{N_k}\sum_{jt|z_{jt}=k}\left(\mathbb{E}[q_{jt}]\right) = \frac{1}{N_k}\sum_{jt|z_{jt}=k}\left(\bar{\mu}_{jt}\right) \tag{26}$$

$$\mathbb{E}[\hat{S}_k] = \sum_{jt|z_{jt}=k}\Big((\bar{\mu}_{jt} - \mathbb{E}[\hat{q}_k])(\bar{\mu}_{jt} - \mathbb{E}[\hat{q}_k])^\top + \bar{\Sigma}_{jt}\Big). \tag{27}$$

The same strategy can be applied to estimate the IW posterior parameters ξ^R_+ on the observations noise R, using expectation and variance of the noise vectors r, and ξ^Θ_{+k} for the spatial distributions (μ_k, Σ_k).

4.5 Unusual Trajectory Detection

Given normative training data $\mathbf{x}^{-j}$, anomaly detection requires a measure of 'normality' for unseen tracks $\mathbf{x}_j$ to rank these tracks, or to set a threshold to isolate unusual from normal tracks. The indicative measure proposed by [15] is the normalized log-likelihood, which is $\log(p(\mathbf{x}_j|\mathbf{x}^{-j}))/T_j$. Since the log-probability is expected to decrease linearly with T_j, the normalization compensates for track length.

For Dual-HDP a Variational Bayesian (VB) approximation [33] is used to integrate over latent topic labels $\mathbf{z}_j$ [15]. Following [33], this approximation introduces free variational parameters which are optimized with respect to the lower-bound of $\log p(\mathbf{x}_j|\mathbf{x}^{-j})$. For our model the same measure can be used, but we average over N MCMC samples $\{\mathbf{z}^{-j(n)}, \{\boldsymbol{\pi}^{(n)}_c\}, \mathbf{c}^{-j(n)}\}$ from the model posterior given the training data:

$$p(\mathbf{x}_j|\mathbf{x}^{-j}) = \frac{1}{N}\sum_n\left[\sum_{c_j} p(c_j|\mathbf{c}^{-j(n)})\sum_{\mathbf{z}_j} p(\mathbf{x}_j|\mathbf{z}_j)p\big(\mathbf{z}_j|\mathbf{z}^{-j(n)}_{c_j}, \boldsymbol{\pi}^{(n)}_{c_j}\big)\right]. \tag{28}$$

However, averaging over the number of observations will reduce the effect of a single unlikely observation. Since we also model the temporal order of topics, we would like to detect unlikely topic transitions too, even if such a transition occurs only once. We therefore propose a different measure $\mathbb{E}_{c_j}[\min_t p(x_{jt}|\mathbf{x}^{-j})]$ which penalizes unusual temporal transitions more. It is not necessary to average this measure over the number of observations.

For Dual-HDP we estimate this term from the VB approximation's lower-bound, taking the minimum over the likelihoods of the quantized words for the $\mathbf{x}_j$. For our model we again use the MCMC samples,

$$\mathbb{E}_{c_j}[\min_t p(x_{jt}|\mathbf{x}^{-j})] = \frac{1}{N}\sum_n \left[\sum_{c_j} p(c_j|\mathbf{c}^{-j(n)})\left(\min_t \sum_{\mathbf{z}_j} p(x_{jt}|\mathbf{z}_j)p(\mathbf{z}_j|\mathbf{z}^{-j(n)}_{c_j}, \boldsymbol{\pi}^{(n)}_{c_j})\right)\right]. \tag{29}$$

5 Split and Merge Moves

The MCMC sampling scheme from Section 4 samples cluster labels c_j one track at a time. However, in [16] we noted that, sometimes, assigning several similar tracks to a new behavior class would have high probability, but creating the new behavior first for a single track has low probability. The sampler may therefore fail to explore the possibility of a new behavior. This problem in DP mixture models has also been noted by other authors [43], [44], [45]. In [16] we presented an ad-hoc solution to address this issue, but here we explore split and merge moves to change cluster labels of multiple tracks simultaneously.

Sequentially-Allocated Merge-Split [43] (SAMS) takes a 'data-driven' approach to propose splits and merges moves of track clusters. Each iteration of the SAMS sampler, two track indices j_a and j_b are drawn uniformly, which will be termed *anchors*. If the tracks belong to the same cluster, $c_{j_a} = c_{j_b}$, a split of the cluster is considered, otherwise a merge of clusters c_{j_a} and c_{j_b} is proposed. In case we split a cluster c, two subclusters $m_a = \{j_a\}$ and $m_b = \{j_b\}$ are initialized with the anchors. Then, restricted Gibbs sampling

[43] is used ('restricted' since we ignore all other clusters) to sample one-by-one a subcluster label $c_j \in \{a, b\}$ for the remaining tracks in the cluster, updating the likelihood statistics of a subcluster after every assignment. This result in a split proposal, which is accepted or rejected using the Metropolis-Hastings rule. To merge two clusters, we need to apply the same procedure to estimate reversed splitting move, see [43] for details.

To generate a SAMS proposal, the likelihood of a track occurring together with some other tracks needs to be evaluated multiple times. Since estimating the data likelihood of Eq. (21) is relatively expensive, we instead estimate $p(\mathbf{z}_j|\mathbf{z}_c^{-j}, \boldsymbol{\pi}_0)$, the likelihood of the last sampled $\mathbf{z}_j$, by importance sampling with $q(\boldsymbol{\pi}_c) = p(\boldsymbol{\pi}_c|\boldsymbol{\pi}_0, \mathbf{z}_c)$ (Eq. (17)) as proposal distribution,

$$p(\mathbf{z}_j|\mathbf{z}_c^{-j}, \boldsymbol{\pi}_0) = \mathbb{E}_{\boldsymbol{\pi}_c \sim q}\left[\frac{p(\mathbf{z}_j|\boldsymbol{\pi}_c, \mathbf{z}_c^{-j})p(\boldsymbol{\pi}_c|\boldsymbol{\pi}_0)}{q(\boldsymbol{\pi}_c)}\right]. \tag{30}$$

All terms can be computed analytically, see Section 4.1, and we find one importance sample suffices.

6 Experiments

We target scenarios with people standing and walking in open spaces. In a surveillance setting, people may enter and exit the scene at different locations, though the system has no prior knowledge about these. Pedestrian tracks observed from a moving vehicle also exhibit spatial structure, e.g., one typically walks parallel to the road on the sidewalk and does not stand still in front of a vehicle. In contrast to cars in driving lanes [1], [2], [3], [15] for instance, people in open spaces can walk to the same destination along different parallel routes (i.e., spatial variation) and move at specific or varying speeds such as standing, walking, or running (i.e., kinematic variation). Furthermore, we are interested in long-term behavior [6], [15], [37], and in localizing regions with particular dynamics [5], [15], [40].

Our baseline is therefore Dual-HDP [15], which jointly discovers low-level topics and high-level behavior clusters, but relies on quantized motion features. We compare our Mixture of SLDSs (MoSLDS) to Dual-HDP on several artificial and real-world datasets.[2] The Dual-HDP codebook is created by dividing the spatial extent of the data into a 10 × 10 grid, and motion into five bins (four directions, as in [15], and a no-motion bin for people standing still), resulting in $B = 500$ unique 'words'. The prior topic distribution is a symmetric Dirichlet, with all B weights set to $100/B$. We also compare the mixture to a single SLDS [23] (with the spatial distributions), by limiting our sampler to only use one behavior but using the further the same hyperparameters. For all compared models, we start with a random assignment of topic and cluster labels, generate 500 MCMC samples, use 200 for burn-in, and keep every 25th sample. We test 15 split-merge proposals per sample.

As default, we set hyperparameters $\alpha = \beta = \gamma = \delta = 1$, spatial prior ξ^Θ with expected center on the averaged position in the data, and a motion prior ξ^Q with expected zero mean. The spatial, system and observation noise hyperparameters are set manually per experiment based on intended use (e.g., Ψ_0^Θ controls large/small areas, low κ_0^Θ spreads the spatial means w.r.t. their size), and rough estimates for the data (e.g., magnitude of observation/system noise). We report parameters Ψ_0 in terms of expected std. dev. σ_0, under the NIW and IW distributions, $\Psi_0 = \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_0^2 \end{bmatrix} \times \nu_0$.

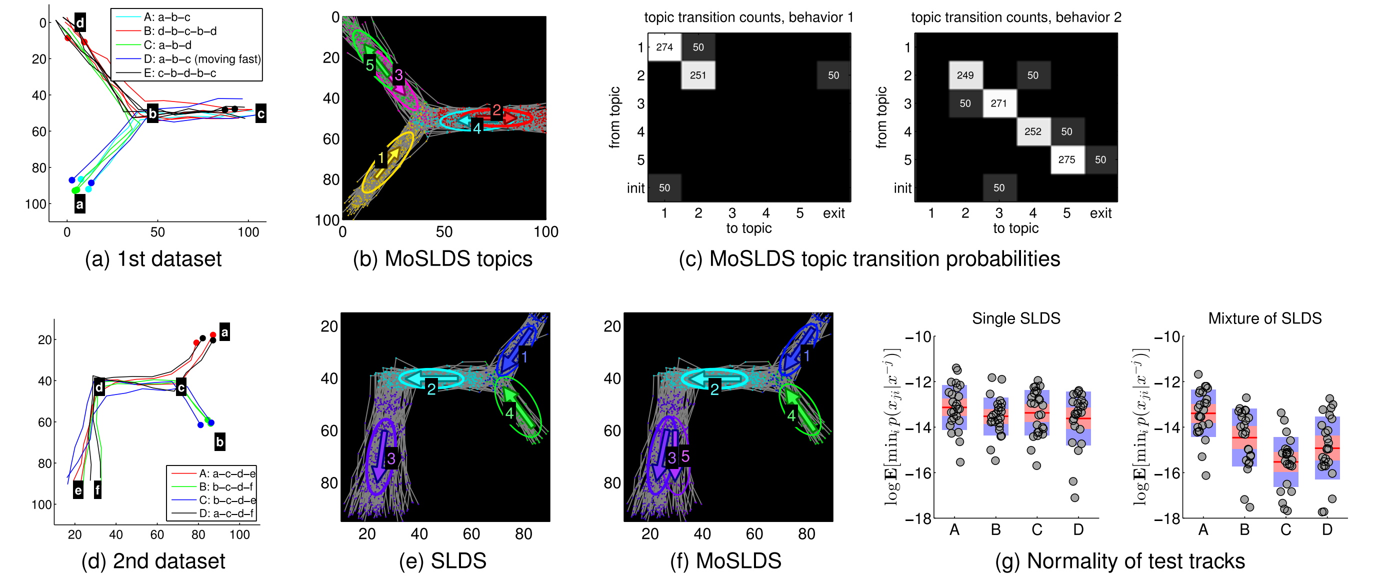
6.1 Artificial Datasets

We begin by evaluating our approach on artificially generated data from [16]. In the first artificial scenario, we define four waypoints (labeled a to e) on the ground plane, and five behavior classes (labeled A to E) as ordered lists of these waypoints, see Fig. 3a. The training data contains 50 tracks from A and B each. The test data contains 20 tracks from each of the five behaviors. Note that behavior C coincides partially with both A and B, D follows the same route as A but moves faster, and E moves in reverse direction of B. For a behavior class tracks are generated by first sampling observations at the waypoints with added Gaussian noise ($\Sigma = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$). Then, intermediate observations are created along this track at a speed of 10 units per time step (15 for behavior D), adding again Gaussian noise at each step ($\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$).

We set hyperparameters $(\kappa_0^Q, \nu_0^Q, \sigma_0^Q) = (.1, 50, 2)$, $(\nu_0^R, \sigma_0^R) = (10, .3)$, and $(\kappa_0^\Theta, \nu_0^\Theta, \sigma_0^\Theta) = (.1, 10, 25)$. Both MoSLDS and Dual-HDP infer two behavior classes from the training data, corresponding to the tracks from A and B. Fig. 3b shows the topics from our model, and Fig. 3c the corresponding topic transition counts. For instance, behavior 1 corresponds to topic chain $init - 1 - 2 - exit$, and matches A. Table 1 shows the normality ratings on the test data per behavior class. As in [16] (where the sampler necessitated other parameters), our model assigns high likelihood to the novel tracks from the normative classes A and B, but anomalous tracks from C, D and E have low likelihood and could be separated by a threshold, especially when using the expected minimum log-likelihood. Dual-HDP can also distinguish A and B from C, but not A and B from D and E, for the following reasons: first, as motion is only quantized into discrete directions [15], the speed differences between D and A are not recognized. This could be improved by quantizing motions at different speeds too, but this introduces again the problem of determining appropriate bins and increases the codebook size. Second, with the bag-of-words approach the unusual temporal order of E cannot be distinguished from B.

With a stronger prior for the limited dynamic variance, $(\kappa_0^Q, \nu_0^Q, \sigma_0^Q) = (.1, 100, 1)$, the MoSLDS finds the same topics and behaviors as before, but if trained on *all* test tracks, it places A and D into distinct behaviors classes, using topics with distinct mean velocities. Without doubling the codebook, however, Dual-HDP only distinguishes three behaviors in the test tracks: one for A+D, one for B+E, and a third for C.

[2] Our software, and the artificial and surveillance datasets, are available for non-commercial, research purposes. Please visit `http://www.gavrila.net` and follow the links (Datasets, etc.).



(a) 1st dataset (b) MoSLDS topics (c) MoSLDS topic transition probabilities

(d) 2nd dataset (e) SLDS (f) MoSLDS (g) Normality of test tracks

Fig. 3. First artificial dataset, see Section 6.1: (a) A few example tracks generated from the five different behaviors. The lower-case letters in the plot indicate the waypoints that define the behaviors. (b) Topics found by MoSLDS when training on only A and B. For each topic the topic label, Gaussian semantic area, and mean system motion direction (as arrows) are shown. (c) The topic transition counts for the two found behaviors in the MoSLDS. Second dataset: (d) Example tracks. (e) Topics found by single SLDS. (f) Topics found by Mixture of SLDS. (g) Expected min. likelihood of test tracks (as dots) for SLDS and MoSLDS (mean in red, std. dev. in blue).

Next, we compare the MoSLDS to a single SLDS when subtle motion differences occur depending on behavior. A second artificial dataset, seen in Fig. 3d, contains two behavior classes for training (A and B), and two more (C and D) for testing only. Similar to the first dataset, track data is generated using the shown waypoints, but now the added Gaussian noise has $\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$, except at the exits e and f which use $\Sigma = \begin{bmatrix} 20 & 0 \\ 0 & 5 \end{bmatrix}$. This results in a 'spread' at the end which is correlated to the initial motion of a track. Training tracks in behavior A start at waypoint a, then move via c to d, and end in e. Tracks in behavior B start at b, then also move via c to d, but end in f. Anomalous behaviors C and D are variations of A and B where the start and end waypoints have been exchanged. The training data consists of 100 tracks of which 70 belong to behavior A, and 30 to behavior B. The test data contains 25 tracks from all four behaviors, and we use the same parameters as for the previous dataset.

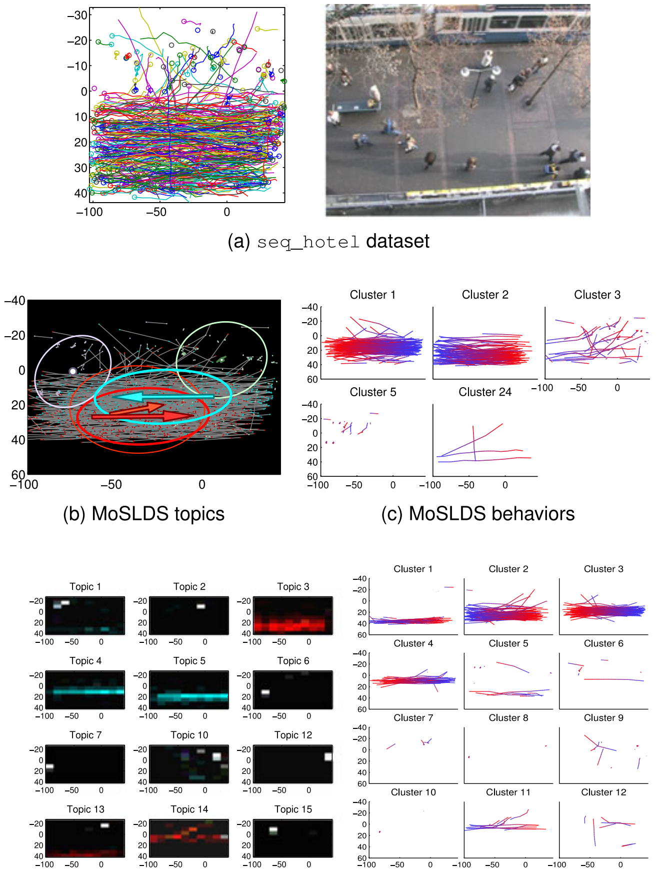
TABLE 1
Track Ratings per Ground Truth Behavior, after Training on Only Tracks from A and B (See Section 6.1 and Fig. 3a), and Testing on 20 Tracks from All Five Behaviors

		$\log(p(x_j \mid \mathbf{x}^{-j}))/T_j$				$\log \mathbb{E}_{c_j}[\min_i p(x_{ji} \mid \mathbf{x}^{-j})]$			
		Mean	Var	Min	Max	Mean	Var	Min	Max
Mixture of SLDS									
normal	A	−11.3	0.2	−12.3	−10.7	−14.8	3.3	−18.8	−12.9
	B	−11.5	0.1	−12.6	−10.9	−15.1	1.8	−17.9	−13.2
anom.	C	−15.1	0.1	−15.7	−14.4	−20.8	0.9	−22.5	−19.8
	D	−22.8	4.8	−26.8	−18.1	−35.4	26.4	−46.8	−25.2
	E	−12.3	0.1	−12.6	−11.8	−20.5	1.0	−22.5	−19.3
Dual-HDP									
normal	A	−3.4	0.1	−4.0	−3.0	−7.7	2.5	−10.5	−5.7
	B	−3.9	0.0	−4.5	−3.7	−9.0	2.4	−11.1	−6.8
anom.	C	−4.9	0.0	−5.2	−4.6	−12.7	0.7	−14.2	−11.5
	D	−3.3	0.1	−4.6	−3.1	−6.4	0.8	−8.7	−5.4
	E	−3.9	0.0	−4.4	−3.7	−8.5	1.7	−10.9	−6.9

While both models can rate C as unusual, MoSLDS assigns low likelihood to D and E too.

Fig. 3e shows sampled topics for the single SLDS approach. The discovered topics reflect the first-order action dynamics, since the motion between c and d holds no information on the motion from d onward. Therefore, there is a single topic at the end of the tracks (i.e., topic 3) which mostly resembles motion of behavior A (which has more training samples). The MoSLDS can distinguish distinct actions at the end of the tracks, see Fig. 3f, since these represent the motion within each behavior better. This effect can be controlled with the prior by weighing topic counts $\mathbf{z}_c$ more heavily in the behavior topic distribution $\boldsymbol{\pi}_c$, i.e. high β. E.g. $\beta = 500$ results in only posterior samples that separate these behaviors. Indeed, due to joint inference, the high-level behavior clustering can inform the low-level action clustering.

The plots in Fig. 3g show the normality measure on the test behaviors of both approaches. Both models assign lower probability to tracks from B than those in A, since in the training data behavior B has a lower prior probability. However, on average the SLDS also assigns high probability to tracks from behavior C and D, since their motions at the start or end correspond to behavior A. The Mixture of SLDS is better equipped to distinguish the typical and anomalous tracks, due to the higher-order dependencies introduced by behaviors. On average, tracks from C have lower probability than the other behaviors, since the motion towards to exit is unlikely given the initial motion. Tracks from A are overall most probable, as their actions and behavior are most common in the training data.



(a) seq_hotel dataset

(b) MoSLDS topics

(c) MoSLDS behaviors

(d) Dual-HDP topics

(e) Dual-HDP behaviors

Fig. 4. See Section 6.2. (a) Tracks (in random colors) and screenshot from seq_hotel dataset [46]. (b) MoSLDS topics. Red topics have motion to the right, blue to the left. Bright topics without arrow have near zero mean motion. (c) MoSLDS behaviors (tracks start blue, end red). (d)(e) Topics and behaviors found by Dual-HDP.

6.2 BIWI Walking Pedestrians Dataset

Next, we make a qualitative comparison of our approach and Dual-HDP [15] on real-world track data. The publicly available *BIWI Walking Pedestrians dataset* [46] contains tracks top-down filmed pedestrians. seq_hotel contains pedestrians walking along a sidewalk and some waiting for a tram. seq_eth shows people entering and exiting the building from where the video was shot. We place the same prior on system and observation noise, $(\kappa_0^Q, \nu_0^Q, \sigma_0^Q) = (.1, 100, 5)$, $(\nu_0^R, \sigma_0^R) = (100, 5)$, and look for large spatial areas, $(\kappa_0^\Theta, \nu_0^\Theta, \sigma_0^\Theta) = (.001, 50, 40)$. Conclusions for seq_eth and seq_hotel are similar (as found in [16]), so we only discus the latter here.

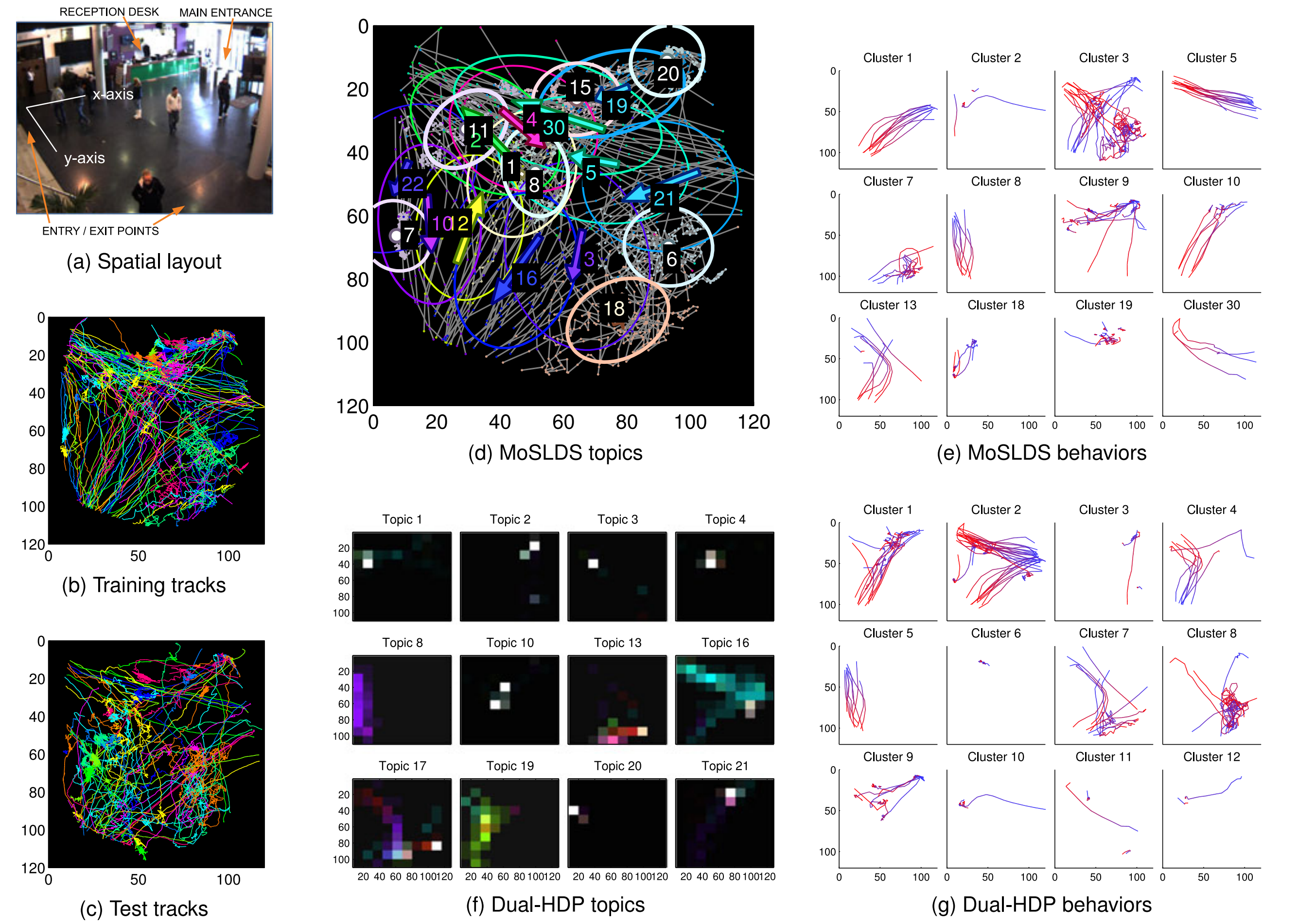
Both Dual-HDP and MoSLDS discover topics and behaviors that correspond to people walking in straight lines or standing and waiting, see Fig. 4. While the benefit of capturing dynamics within behaviors is minimal on these tracks (most tracks use only a single topic in either method), we do nevertheless observe some clear differences, especially on the found low-level topics. First, in Dual-HDP, waiting people are represented in the topics as 'no-motion' at one or few spatial cells, as can be seen by the white dots in the topics of Fig. 4d. These topics thus capture waiting at exactly those spatial positions, but do not generalize over people waiting in the near vicinity. The spatial distributions of our model on the other hand do generalize waiting areas, as seen in the topics in Fig. 4b. This also results in less fragmented behavior clustering, see Fig. 4c versus Fig. 4e. Second, our model finds spatially overlapping topics for people walking at different speeds and directions. These differences are found because we have a prior on the variance of low-level motion, not due to specifying a speed threshold for 'slow' or 'fast' movement.

6.3 Surveillance Dataset

In [16] we presented a novel dataset recorded at the central hall of a large building, see Fig. 5a, using actors to perform various roles that commonly occur there at a normal working day. Tracks are obtained using a multi-view tracker [4], though due to many occlusions not everybody is correctly tracked all the time. The training data (118 tracks, 2,737 observations, see Fig. 5b) contains normative behavior only, including employees entering at the main entrance and walking to one of the exits, and visitors that register at the reception and wait to be picked up by an employee. The test data of 64 tracks, see Fig. 5c, contains mostly normative behavior, but with some exceptions: in the scene a 'terrorist' scouts the environment, walking in and out of view. Later, a second 'terrorist' joins him and mixes with the visitors. At some point, the second 'terrorist' shoots, and bystanders run for safety.

We set weak SLDS priors, $(\kappa_0^Q, \nu_0^Q, \sigma_0^Q) = (1, 10, 2)$, $(\nu_0^R, \sigma_0^R) = (10, .3)$, and use $(\kappa_0^\Theta, \nu_0^\Theta, \sigma_0^\Theta) = (.1, 50, 25)$. As in [16], the MoSLDS discovers various actions in the training data, shown in Fig. 5d. For instance, we interpret topic 20 as the action 'waiting at reception desk', and topic 16 as 'walking to lower exit'. Fig. 5e shows tracks from the 12 behaviors present in this sample, corresponding to workers walking straight from the main entrance to an exit, visitors entering at the reception and waiting in various parts in the hall, and workers picking up visitors. Here we have used $\beta = 10$ because, interestingly, we observed that with $\beta = 1$ tracks with just a few topics in common tend to be clustered together, especially when behaviors have few tracks. So again, increasing β results in behaviors with more specific topic chains. With $\beta = 1$ we find about seven behaviors typically, e.g., behaviors 10 and 13 in Fig. 5e are merged. Note how behavior clusters can merge fragmented tracks with similar topics (e.g., clusters 9, 13), but systematic fragmentation may lead to new behaviors (e.g., cluster 19 with waiting people).

Dual-HDP discovers behavior clusters and regions of motion resembling those of MoSLDS, see Figs. 5f and 5g. However, the phenomena discussed in Section 6.2, where standing people yield very specific topics (e.g., topic 3 in Fig. 5f), again occur. Both models are used to rank the test tracks, where we discern three relevant classes: normal tracks, the 'terrorist' suspects, and fleeing visitors. Figs. 6a and 6b show the normality measures for Dual-HDP and MoSLDS respectively. With Dual-HDP, 33 out of 64 test tracks contain words from the codebook not seen in the training data, even at the low spatial resolution of 10 × 10 bins. Evaluation on the test data is still possible, since the unseen words have non-zero probability due to the prior. Still, the most anomalous test tracks are those with many words that did not occur in the training data, as found in both the normative and the suspect class. The running



(a) Spatial layout

(b) Training tracks

(c) Test tracks

(d) MoSLDS topics

(e) MoSLDS behaviors

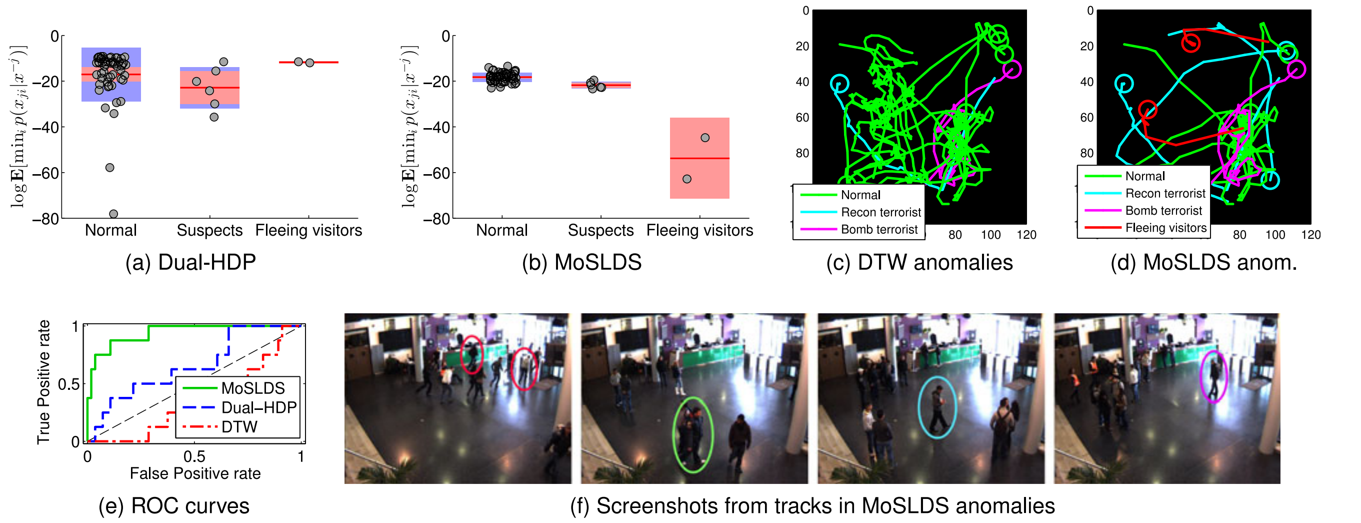
(f) Dual-HDP topics

(g) Dual-HDP behaviors

Fig. 5. (a) Screenshot from the surveillance dataset, Section 6.3. (b) Tracks (in random colors) in the normative training data. When people stand still their tracks form dense spots. (c) Tracks in the test data, containing normal people, suspicious individuals, and people running after a gunshot. (d) Topics found by MoSLDS, with their id, Gaussian semantic region, and mean motion. (e) Several found behaviors (tracks start blue, end red). (f) Dual-HDP topics (↑ green, ↓ purple, ← blue, → red, no-motion white) and (g) behaviors.

visitors have relatively high likelihood, since the velocity magnitude is disregarded.

Test tracks are also compared to all training tracks using dynamic time warping [36] with Euclidean distance measure (this gives better results than the implementation in [16]), and the lowest DTW distance is used as normality measure. We classify tracks while varying a normality threshold and see from the ROC curves, Fig. 6e, that



(a) Dual-HDP

(b) MoSLDS

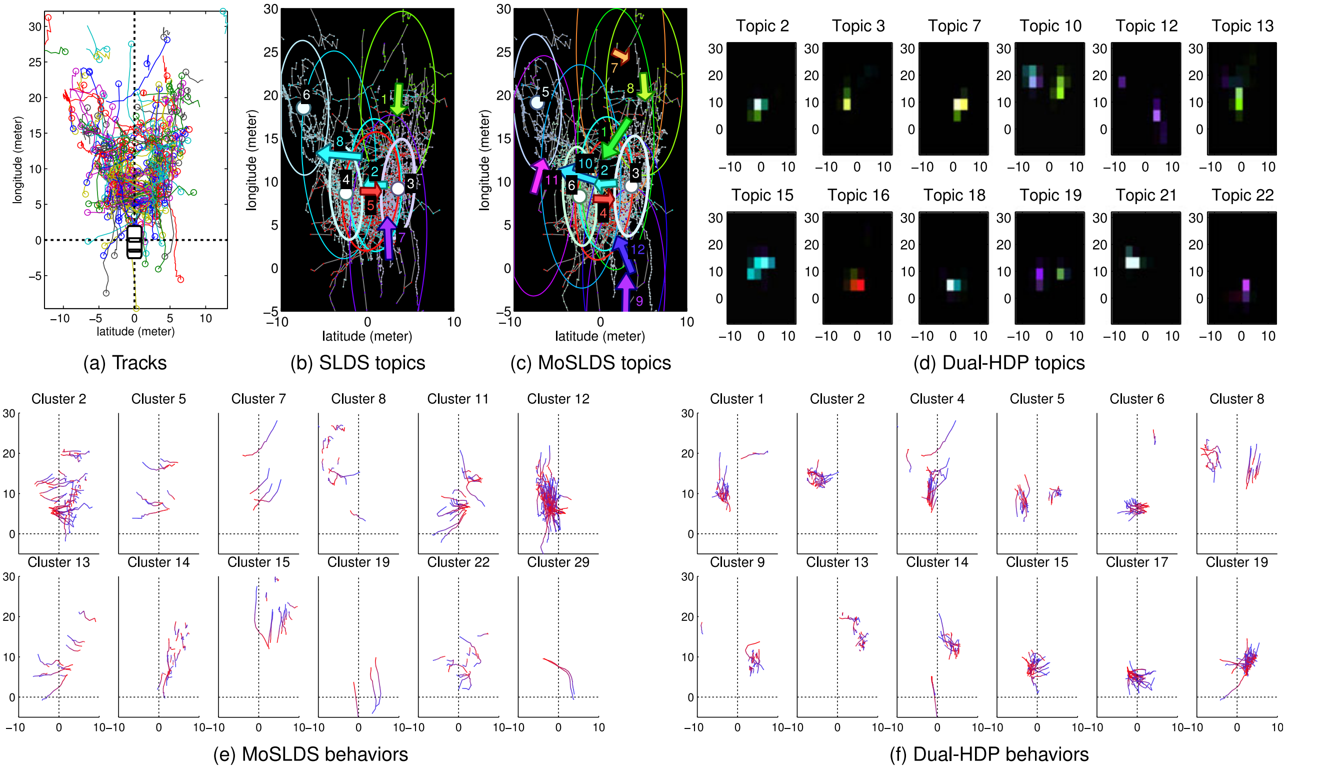
(c) DTW anomalies

(d) MoSLDS anom.

(e) ROC curves

(f) Screenshots from tracks in MoSLDS anomalies

Fig. 6. (a) (b) Expected min. likelihood for Dual-HDP and MoSLDS on surveillance test tracks, see Section 6.3. (c),(d) Most unusual tracks (circles mark track start) found by (c) dynamic time warping, and (d) MoSLDS. Green tracks are false positives. (e) ROC curves by varying normality threshold. (f) Screenshots of the unusual tracks found by MoSLDS: fleeing visitors; wandering visitor (false positive); two 'terrorists' exploring the space.



(a) Tracks (b) SLDS topics (c) MoSLDS topics (d) Dual-HDP topics

(e) MoSLDS behaviors (f) Dual-HDP behaviors

Fig. 7. (a) Tracks from the intelligent vehicle pedestrian dataset, Section 6.4. (b)-(f) Samples of learned dynamics and behavior clusters for an SLDS, MoSLDS, and Dual-HDP. In (e) and (f), tracks start blue, end red. Tracks have been aligned such that the longitudinal axis is the road, so crossing pedestrians move across this axis.

MoSLDS ranks the tracks best. Fig. 6c shows the eight most anomalous tracks for DTW, but the only true positives are at rank 6 and 8. Matching tracks globally does not consider the common spatial variance in the training data. The eight most anomalous MoSLDS tracks, Fig. 6d, have only two *false* positives at rank 4 and 7. The two most unusual tracks are running visitors, whose dynamics have low likelihood. Other tracks belong to the 'terrorists' who walk unusually between waiting areas. False positives occur when normal visitors walks to a different area to chat with others, which did not occur in the training data. Fig. 6f shows several screenshots of these events.

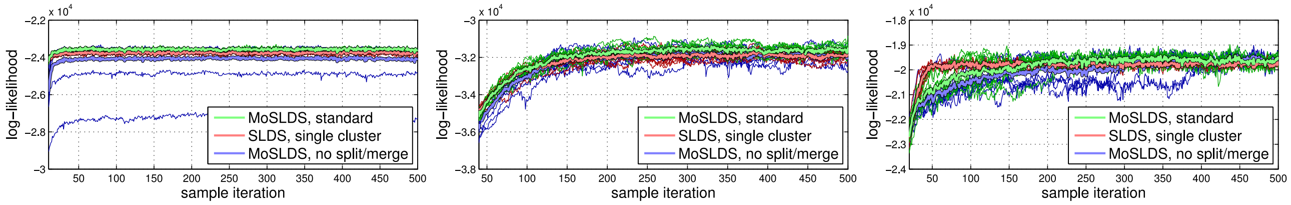
6.4 Intelligent Vehicle Pedestrians Dataset

Next, we experiment with pedestrian tracks as observed from the perspective of a moving vehicle. The dataset, presented in [42], consists of pedestrian trajectories recorded in and around various European cities, with an embedded stereo vision system in a car. Unoccluded pedestrians have been manually annotated with bounding boxes. Using depth estimation from stereo vision, the pedestrian's distance to the vehicle could be measured, resulting in 2D top-down measurement (latitude and longitude) relative to the camera coordinate system. Obtained tracks are compensated for the vehicle ego-motion as estimated from on-board sensors. Finally, tracks are aligned in a coordinate system with the vehicle at the origin traveling along the vertical axis [42], such that pedestrians on the sidewalk move in longitudinal direction, and lateral to cross the road, see Fig. 7a. The resulting data has 423 tracks, and 4,103 observations in total. As more longitudinal than lateral variance is expected, we set $(\kappa_0^Q, \nu_0^Q, \sigma_0^{Qx}, \sigma_0^{Qy}) = (.5, 50, .25, .5)$, $(\nu_0^R, \sigma_0^{Rx}, \sigma_0^{Ry}) = (50, .25, .5)$, $(\kappa_0^\Theta, \nu_0^\Theta, \sigma_0^{\Theta x}, \sigma_0^{\Theta y}) = (1, 100, 5, 15)$.

Fig. 7b shows topics found by the single SLDS approach, as in [42], and Fig. 7c topics found with MoSLDS. In both cases, we see clear lateral motion in front of the car of people crossing the road (in red and blue) to and from areas located left and right of the car. These have near zero mean motion (colored white) but large motion variance, since pedestrians at those locations move on the sidewalk in both longitudinal directions at low velocity. We find that the SLDS and MoSLDS topics are mostly similar here, but the MoSLDS behaviors additionally capture which topics co-occur. E.g. in Fig. 7e cluster 11 transitions from topic 4 to 3 (i.e., reaching the sidewalk after crossing) but then moving from 3 to 2 (i.e. crossing back) is improbable. Cluster 2 does contain topics 2 and 3, but not topic 4. In this data, people only stand still in specific spatial regions, and we now see no Dual-HDP topics for isolated standing individuals (Fig. 7d), but do see regions with lateral (crossing) and longitudinal (sidewalk) motion similar to our MoSLDS. Dual-HDP does divide tracks at the sidewalk and crossing areas into many small behavior clusters with little overlap, each containing specific motion at a specific location, see Fig. 7f (e.g., clusters 15 and 17 for crossing).

6.5 Runtime and Sampler Comparison

Finally, we compare convergence of our full MoSLDS sampler, the sampler without split-merge jumps (Section 5), and the single SLDS sampler. Our code is written in Matlab, and C++ to sample the $\mathbf{z}_j$ (Section 4.3). On the surveillance data, a run of 500 samples takes ~1 hour for the full sampler, ~30 min. without split-merge, and ~15 min. using only one



(a) Artificial dataset (b) Surveillance dataset (c) Intelligent vehicle dataset

Fig. 8. Log likelihood of samples, plotted against sampling iterations, see Section 6.5. Thin lines are plotted for each run of a sampler. Thick lines show the average over the all runs. First iterations are omitted for clarity.

SLDS. With J tracks of length T, K topics, and C clusters, the MH proposals runs in $O(J \times T \times K)$ (Section 4.2, 4.3), Gibbs sampling $\boldsymbol{\pi}_0$ and all $\boldsymbol{\pi}_c$ in $O(K \times C)$ (Section 4.1), Gibbs sampling SLDS parameters in $O(J \times T + K)$ (Section 4.4), and each SAMS proposal in $O(J \times T \times K)$. So sampling scales linearly with track count and length, but longer tracks may require more information filter runs to obtain good $\mathbf{z}_j$, and more clusters reduces the chance of good label c_j proposals, or good SAMS proposals.

We perform 10 MCMC runs on several discussed datasets. Fig. 8 shows how the log probability of the MCMC samples converges, where thin lines are individual runs, thick lines show the average. The figure illustrates that without the split/merge moves there is a risk of the sampler getting stuck in a suboptimal region for a long time. However, this problem appears to be more prevalent in data with some often used topics that overlap spatially with others, such as the artificial and intelligent vehicle data. In the latter for instance, the crossing motions have quite some spatial overlap with the neutral areas at the side of the road, and initially such motions are not placed in distinct topics. The results also show that this is only a problem in some runs, and, on real-world data, is resolved after sufficient iterations. The single SLDS sampler converges more quickly as there are fewer latent variables to estimate without the behavior classes.

7 Conclusions

We have presented a novel model that uses a Mixture of SLDS to jointly infer common actions and behaviors from track data. Evaluation on test data from artificial, surveillance, and intelligent vehicle scenarios, shows that our MCMC sampler can discover meaningful dynamics and track clusters. We find that on some datasets, additional split-merge moves help the sampler to avoid local optima more reliably, but on our real-world data all samplers eventually convergence to similar solutions.

Behavior clusters can capture higher order dependencies on the topics, and provide insight in the behavioral structure of the data. If, however, one is only interested in the low-level topics, it may suffice to use our proposed SLDS with spatial areas and a single behavior cluster, as the experiments on real-world data show. For Dual-HDP, the spatial variance of walking and standing people results in sparse spatial bins, and different motions can only be distinguished if appropriate thresholds are set in advance. Our MoSLDS does not rely on feature quantization, but learns relevant motions types in the continuous feature space, using priors on the variance within a motion class, which is especially beneficial when testing on new tracks. It is capable of inferring informative low-level motions even when tracks have spatial and kinematic variations, and available training data is limited.

References

[1] T. Hospedales, S. Gong, and T. Xiang, "A Markov clustering topic model for mining behaviour in video," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1165–1172.

[2] D. Kuettel, M. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? Discovering spatio-temporal dependencies in dynamic scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2010, pp. 1951–1958.

[3] R. Emonet, J. Varadarajan, and J. Odobez, "Multi-camera open space human activity discovery for anomaly detection," in *Proc. 8th IEEE Int. Conf. Adv. Video Signal-Based Surveillance*, Aug. 2011, p. 6.

[4] M. Liem and D. M. Gavrila, "Multi-person localization and track assignment in overlapping camera views," in *Proc. 33rd Int. Conf. Pattern Recog.*, 2011, pp. 173–183.

[5] J. Joseph, F. Doshi-Velez, A. S. Huang, and N. Roy, "A Bayesian nonparametric approach to modeling motion patterns," *Auton. Robots*, vol. 31, no. 4, p. 383–400, 2011.

[6] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 618–633.

[7] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? A study on pedestrian path prediction," *IEEE Trans. Intell. Transportation Syst.*, vol. 15, no. 2, pp. 494–506, Apr. 2014.

[8] Z. Chen, D. Ngai, and N. Yung, "Pedestrian behavior prediction based on motion patterns for vehicle-to-pedestrian collision avoidance," in *Proc. 11th Int. IEEE Conf. Intell. Transportation Syst.*, 2008, pp. 316–321.

[9] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Norwood, MA, USA: Artech House, 1999.

[10] M. Liem and D. M. Gavrila, "Multi-person tracking with overlapping cameras in complex, dynamic environments," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 199–218.

[11] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[12] J. Tao and R. Klette, "Tracking of 2d or 3d irregular movement by a family of unscented Kalman filters," *J. Inf. Commun. Convergence Eng.*, vol. 10, no. 3, pp. 307–314, 2012.

[13] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive Bayesian filters: A comparative study," in *Proc. 35th German Conf. Pattern Recog.*, 2013, pp. 174–183.

[14] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 201–214.

[15] X. Wang, K. T. Ma, G. W. Ng, and W. E. Grimson, "Trajectory analysis and semantic region modeling using a nonparametric Bayesian model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[16] J. F. P. Kooij, G. Englebienne, and D. M. Gavrila, "A non-parametric hierarchical model to discover behavior dynamics from tracks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 270–283.

[17] V. Romero-Cano, J. I. Nieto, and G. Agamennoni, "Unsupervised motion learning from a moving platform," in *Proc. IEEE Symp. Intell. Veh.*, 2013, pp. 104–108.

[18] P. Scovanner and M. F. Tappen, "Learning pedestrian dynamics from the real world," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 381–388.

[19] G. Antonini, S. V. Martinez, M. Bierlaire, and J. P. Thiran, "Behavioral priors for detection and tracking of pedestrians in video sequences," *Int. J. Comput. Vis.*, vol. 69, no. 2, pp. 159–180, 2006.

[20] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[21] H. E. Rauch, C. Striebel, and F. Tung, "Maximum likelihood estimates of linear dynamic systems," *AIAA J.*, vol. 3, no. 8, pp. 1445–1450, 1965.

[22] A.-V. Rosti and M. J. F. Gales, "Rao-Blackwellised Gibbs sampling for switching linear dynamical systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 1, pp. I-809–I-812.

[23] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1569–1585, Apr. 2011.

[24] V. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 981–987.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 1. New York, NY, USA: Springer, 2006.

[26] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 17th Conf. Uncertainty Artif. Intell.*, 2001, pp. 362–369.

[27] X. Boyen and D. Koller, "Tractable inference for complex stochastic processes," in *Proc. 14th Conf. Uncertainty Artif. Intell.*, 1998, pp. 33–42.

[28] S. L. Lauritzen, "Propagation of probabilities, means, and variances in mixed graphical association models," *J. Amer. Statist. Assoc.*, vol. 87, no. 420, pp. 1098–1108, 1992.

[29] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert, "Learning and inferring motion patterns using parametric segmental switching linear dynamic systems," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 103–124, May 2008.

[30] E. Fox, "Bayesian nonparametric learning of complex dynamical phenomena," Ph.D. Thesis, MIT, Cambridge, MA, USA, 2009.

[31] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[32] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, vol. 14, pp. 577–584.

[33] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[34] Z. Fu, W. Hu, and T. Tan, "Similarity based vehicle trajectory clustering and anomaly detection," in *Proc. IEEE Int. Conf. Image Process.*, 2005, vol. 2, pp. II-602–II-605.

[35] X. Wang, K. Tieu, and E. Grimson, "Learning semantic scene models by trajectory analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 110–123.

[36] E. Keogh and M. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2000, pp. 285–289.

[37] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2871–2878.

[38] J. Fernyhough, A. Cohn, and D. Hogg, "Generation of semantic regions from image sequences," in *Proc. Eur. Conf. Comput. Vis.*, 1996, pp. 475–484.

[39] D. Makris and T. Ellis, "Automatic learning of an activity-based semantic scene model," in *Proc. IEEE Int. Conf. Adv. Video Signal-Based Surveillance*, 2003, pp. 183–188.

[40] D. Lin, E. Grimson, and J. Fisher, "Learning visual flows: A lie algebraic approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 747–754.

[41] J. Varadarajan, R. Emonet, and J. M. Odobez, "Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 117.1–117.11.

[42] J. F. P. Kooij, N. Schneider, and D. M. Gavrila, "Analysis of pedestrian dynamics from a vehicle perspective," in *Proc. IEEE Symp. Intell. Veh.*, 2014, pp. 1445–1450.

[43] D. B. Dahl, "Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models," Department of Statistics, Texas A&M University, Tech. Rep., 2005.

[44] M. C. Hughes, E. B. Fox, and E. B. Sudderth, "Effective split-merge monte carlo methods for nonparametric models of sequential data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1304–1312.

[45] J. Chang and J. W. Fisher III, "Parallel sampling of DP mixture models using sub-cluster splits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 620–628.

[46] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 261–268.

Julian F. P. Kooij received the MSc degree in artificial intelligence from the University of Amsterdam in 2008. He is currently working toward the PhD degree at the University of Amsterdam, after having been employed in part at Daimler Research and Development, Ulm, Germany. His research interests include Machine Learning and Bayesian methods for computer vision, with a focus on modeling trajectory dynamics and unsupervised discovery of motion patterns.

Gwenn Englebienne received the PhD degree in computer science from the University of Manchester in 2009. He has since focused on automated analysis of human behavior at the University of Amsterdam, where he has developed computer vision techniques for tracking humans across large camera networks and machine learning techniques to model human behavior from networks of simple sensors. His main research interests are in models of human behavior, especially their interaction with other humans, with the environment, and with intelligent systems.

Dariu M. Gavrila received the PhD degree in computer science from the University of Maryland at College Park in 1996. Since 1997, he has been with Daimler R&D in Ulm, Germany, where he is currently a principal scientist. In 2003, he was named professor at the University of Amsterdam, in the area of intelligent perception systems (part time). Over the past 15 years, he has focused on visual systems for detecting human presence and activity, with application to intelligent vehicles, smart surveillance and social robotics. He co-received the IEEE ITS Outstanding Application Award 2014 for his role in transferring pedestrian detection technology into Mercedes-Benz vehicles. His personal web site is www.gavrila.net.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.