Nonparametric Bayesian Models for Unsupervised Activity Recognition and Tracking

Neil Dhir[1,†], Yura Perov[1] and Frank Wood[1]

***Abstract*— Human locomotion and activity recognition systems form a critical part in a robot's ability to safely and effectively operate in a environment populated with human end users. Previous work in this area relies upon strong assumptions about the labels in the training data; e.g. that are noise-free and that they exist at all. Our approach does not predefine the relevant behaviours or their number, as both are learned directly from observations, similar to real-world human-robot interactions, where labels are neither available. Instead we introduce models that make no assumptions about the state space, by presenting a fully unsupervised nonparametric Bayesian recognition approach, in which we leverage recent advances in state space modelling with automatic inference using probabilistic programming. We demonstrate the utility of full model optimisation using Bayesian optimisation and validate our approach on several challenging problems, using different feature modalities.**

I. Introduction

A generative model describes a process, usually one by which observable data is generated. Generative models represent knowledge about the causal structure of the world. It is possible to use deterministic generative models to describe possible ways a process could unfold, but due to sparsity of observations or actual randomness there will often be many ways that our observations could have been generated. We are interested in sequential data, from which we aim to infer meaningful states, along with the defining characteristics of each state. More often than not, such state discovery needs to be done in an unsupervised fashion, as the application domain commands little or no information about the latent state cardinality of the state space. One such domain is activity recognition systems for modern robotics.

Being able to determine the mode of locomotion of a human user or the system's place in its operating environment, is necessary for effective use of robotic systems for everyday usage; from aiding elderly users with their daily activities of living to providing accurate assistance to the user in an industrial and/or home setting. Although many previous works have focused on this topic, recognising complex activities endures as a challenging task which is still an open problem [1]. We contribute a fully unsupervised approach to the problem of locomotion and activity recognition, through the use of Bayesian nonparametric state-space models (SSM) that use probabilistic programming (PPS) general-purpose inference [2]. Specifically we demonstrate the utility of the stateful hierarchical Dirichlet process hidden Markov model.

[1]Department of Engineering Science, University of Oxford, Parks Road, OX1 3PJ Oxford, United Kingdom. [†]Corresponding author: neild@robots.ox.ac.uk

Nonparametric Bayesian SSMs, are a subset of the larger family of infinite hidden Markov models (iHMM) [3]. Parametric HMMs have been used with great success for approaching learning problems in sequential data, such as speech and finance. In the nonparametric Bayesian paradigm, inference is performed in models with an infinite number of states. By adopting this approach, as introduced in [4], we can perform activity recognition on datasets without using prior knowledge about the activities or their number.

The paper is organised as follows; in §II we give a thorough exegesis of infinite HMMs as well as an outline of our extensions. Section §III gives an brief overview of the employed inference schemes, as used in our PP framework, and §IV reviews BO. Finally, experiments and results are demonstrated for synthetic and human observations in sections §V and §VI respectively.

II. Bayesian nonparametric state-space models

Two of the most important examples of SSMs is the HMM in which the latent variables are discrete, and linear dynamical systems (LDS), in which the latent variables are Gaussian. In this paper we will restrict ourselves to the former model class, but note that our methodology is fully applicable to LDSs.

A. Hierarchical mixture models

In order to infer state cardinality from observations and to flexibly model the distribution of continuous data, we adopt Bayesian nonparametrics. It requires the specification of a prior model for continuous distributions. A fruitful and general approach for defining such a prior model was first suggested by [5] in terms of an infinite dimensional mixture model:

$$\begin{aligned} P &\sim \mathcal{P} \\ X_i \mid P &\overset{i.i.d.}{\sim} P && i = 1, 2, \dots \\ Y_i \mid X_i &\overset{ind.}{\sim} F(\cdot \mid X_i) && i = 1, 2, \dots \end{aligned} \tag{1}$$

where P is a discrete random probability measure (RPM) with distribution $\mathcal{P}$, $Y_{1:n}$ are a collection of continuous and possibly multivariate observations and $X_{1:n}$ are the corresponding collection of latent random variables from an exchangeable sequence directed by P. In which $F(\cdot \mid X_i)$ is some continuous distribution parametrised by X_i. The nonparametric hierarchical model (1) defines a mixture model (MM) with a potentially countably infinite number of components. Because the RPM in equation (1) is discrete, this means that the pair of consecutive values of X take on the the same value with a strictly positive probability.

This value is a mixture component. By setting the RPM to the Dirichlet process (DP) [6] we obtain the familiar DPMM. The Dirichlet process, denoted by $\mathcal{DP}(\gamma, H)$, is a stochastic process over countably infinite random measures on parameter space Θ. It is uniquely defined by a base measure H on Θ and a concentration parameter γ.

The DP is typically used as a prior on the mixture components θ, of a MM of unknown complexity resulting in the aforementioned DPMM. But there are many scenarios in which *groups* of data are thought to be produced by related, yet unique, generative processes. Indeed, a recurring problem in many areas of information technology is that of segmenting a signal into a set of time intervals that have a useful interpretation in some underlying domain. In such scenarios we can take a hierarchical Bayesian approach.

We posit that observations can be subdivided in a countable collection of groups. Groups of observations are modelled by considering a collection of DPs $\{G_j : j \in \mathcal{J}\}$, defined on a common space Θ, where $\mathcal{J}$ indexes the groups. By placing a global DP prior $\mathcal{DP}(\gamma, H)$ on the base distribution G_0, from whence we draw group specific distributions $G_j \sim \mathcal{DP}(\alpha, G_0)$, we receive the hierarchical DP (HDP). The HDP induces sharing of atoms among the random measures G_j since each inherits its set of atoms from the same G_0 [4]. This idea can be used to develop HMMs with unknown, potentially infinite, state spaces [7].

B. Hidden Markov models with infinite state spaces

Formally, an HMM is a doubly-stochastic Markov chain in which a state sequence $\{\theta_1, \ldots, \theta_T\}$ is drawn according to a Markov chain on a discrete state space Θ with transition kernels $\{G_\theta : \theta \in \Theta\}$ [8]. Corresponding observations $\{y_1, \ldots, y_T\}$, conditional on the state sequence, are drawn from a fixed emission distribution $y_t \mid \theta_t \sim F(\theta_t)\ \forall t \in \{1, \ldots, T\}$. By employing the HDP in an HMM setting, a prior distribution is defined on transition kernels, yielding the HDP-HMM [4] (see figure 1); an HMM with a countably infinite state space

$$G_0 \mid \gamma, H \sim \mathcal{DP}(\gamma, H), \tag{2}$$

$$G_\theta \mid \alpha, G_0 \sim \mathcal{DP}(\alpha, G_0) \qquad \text{for } \theta \in \Theta, \tag{3}$$

$$\theta_t \mid \theta_{t-1}, G_{\theta_{t-1}} \sim G_{\theta_{t-1}} \qquad \text{for } t = 1, \ldots, T, \tag{4}$$

$$y_t \mid \theta_t \sim F(\theta_t). \tag{5}$$

To properly qualify this nonparametric Bayesian approach to HMMs, consider that each G_θ is a DP draw, and is interpreted as the transition distribution over $\theta_t \mid \theta_{t-1}$. All transition distributions are linked by the same discrete measure G_0. Hence, in expectation $\mathbb{E}[G_\theta] = G_0,\ \forall\theta \in \Theta$. Thus, transition distributions *tend* to have their mass concentrated around a common set of states, providing the desired bias towards re-entering and re-using a consistent set of states [9].

C. Stateful representations

The rate at which re-entering and re-using states unfolds in the HDP-HMM is typically too fast for many real-world problems. The model construction furthermore encourages the creation of redundant states and rapid switching amongst these too [8]. To combat this, the *sticky* HDP-HMM [10] (see figure 2a) augments the HDP-HMM with an extra parameter $\kappa > 0$ that biases the process towards self-transitions and thus provides a method to encourage longer state durations. Hence transitions kernels, equation (3) above, are instead sampled as so:

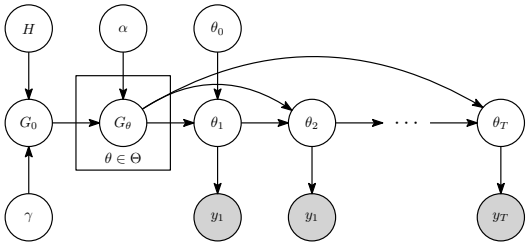


Fig. 1. Graphical model representation of the HDP-HMM [4].

$$G_\theta \mid \alpha, G_0, \kappa, \theta \sim \mathcal{DP}\left(\alpha + \kappa, \frac{\alpha G_0 + \kappa\delta_\theta}{\alpha + \kappa}\right). \tag{6}$$

where δ_θ is a point mass at θ.

This model shares the original HDP-HMM's restriction to geometric state durations, thus limiting the model's expressiveness regarding duration structure. More importantly, its global self-transition bias is shared among all states, and so it does not allow for learning state-specific duration information [9]. Instead, we propose that by allowing for group-specific self-transition biases κ_θ, greater heterogeneity can be achieved in the dwell-time distribution of the inferred states. We extended it further by allowing for group-specific concentrations: α_θ. Hence equation (6) becomes

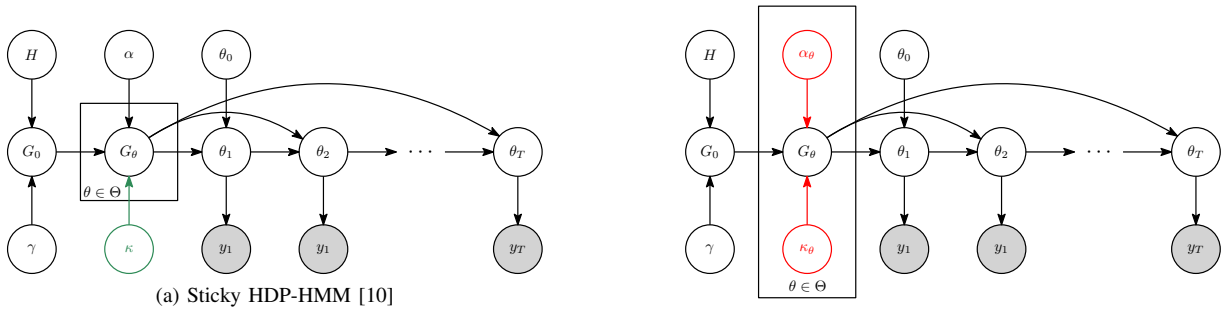
$$G_\theta \mid \alpha_\theta, G_0, \kappa_\theta, \theta \sim \mathcal{DP}\left(\alpha_\theta + \kappa_\theta, \frac{\alpha_\theta G_0 + \kappa_\theta\delta_\theta}{\alpha_\theta + \kappa_\theta}\right) \tag{7}$$

which we refer to as the *stateful* HDP-HMM (see figure 2b) - in reference to its pronounced usage of memoized groups and their statistics. In adopting this approach we imbue the original sticky HDP-HMM with more flexibility w.r.t. modelling the state duration more accurately. We allow for state-specific duration information to be encoded via κ_θ and also admit α_θ to determine the extent of the repetition of the values of G_θ.

III. Inference and Learning

Inference in HDP models is typically achieved using bespoke model-specific algorithms, usually using one of the various mathematical representations available for nonparametric models, including; stick-breaking representations, urn models and truncations [8]. The authors in [4] present three related Markov chain Monte Carlo (MCMC) sampling schemes for the hierarchical DP mixture model. The extension to HDP-HMMs can be done with Gibbs sampling or slice sampling [11], or by truncating the allowable state-space and then us the forwards-backwards algorithm. We instead adopt probabilistic programs for our inference.

Probabilistic programs are regular programs extended by two constructs [12]: (I) the ability to draw random values



(a) Sticky HDP-HMM [10]

(b) Stateful HDP-HMM

Fig. 2. Graphical model representations of the *sticky* and *stateful* HDP-HMM with state-persistence imbued in the generative process. HDP-HMM extensions are indicated by coloured nodes and arrows.

from probability distributions, and (II) the ability to condition values computed in the programs on probability distributions. A PPS unifies techniques for formal description of computation with the representation and use of uncertain knowledge. PPSs' main advantage is in separating the modelling and the inference problems which allows us to focus on the on the former without worrying about the latter. Throughout this paper, our weapon of choice will be a PPS *Anglican* [2]. Anglican has implementations of several importance sampling based methods such as sequential Monte Carlo (SMC) and particle Markov chain Monte Carlo [13]. For reference, w.r.t. to SMC; consider the generative model $p(x_{1:T}, y_{1:T})$ with hidden variables $x_{1:T}$ and observations $y_{1:T}$. In a PPS, we let the observing random variable y_t be the value of the t^{th} observe, and the hidden variables $\mathbf{x}_t = x_{1:t}$ be the execution trace before this observe. These methods then give us a particle estimate of our posterior $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})$ along with the estimate of the marginal likelihood $p(\mathbf{y}_{1:T})$:

$$\hat{p}(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}) = \frac{1}{\sum_{i=1}^{P} W_i} \sum_{j=1}^{P} W_j \delta_{\mathbf{x}_{1:T}^{(j)}}(\mathbf{x}_{1:T}) \tag{8}$$

$$\hat{p}(\mathbf{y}_{1:T}) = \sum_{j=1}^{P} W_j, \tag{9}$$

where $\{W_j\}_{j=1,\dots,P}$ are the unnormalised weights calculated by weighting the particles $\{\mathbf{x}_{1:T}^{(j)}\}_{j=1,\dots,P}$. Where $\mathbf{x}_{1:T}$ are our latent variables and $\delta_{\cdot}(\cdot)$ is the Dirac measure on sample $\mathbf{x}_{1:T}^{(j)}$ such that $\delta_{\mathbf{x}_{1:T}^{(j)}} = 0$ if $\mathbf{x}_{1:T}^{(j)} \notin \mathbf{x}_{1:T}$ and 1 otherwise.

IV. Bayesian optimisation

Selecting model hyperparameters θ is a common problem within machine learning and indeed statistics. The appropriateness of the hyperparameters, or the fitness of the model, can be, for example, modelled by the marginal likelihood of the model. To this end, we use the aforementioned particle-based inference algorithms in Anglican to provide us with a noisy estimates of $\mathbb{P}(y_{1:T})$. As this is expensive since inference must be performed, often on large datasets, it fits well with the Bayesian optimisation (BO) framework which allows us to find the global maximum of some expensive black-box function $f : \mathbb{R}^d \to \mathbb{R}$ [14]. This is particularly true for PPS where the objective function is often expensive to evaluate, as is typically takes the form of an intractable integral such as the log marginal likelihood: $\log \mathbb{P}(y_{1:T})$. Formally, BO seeks to find the global maximum, on a d-dimensional space with bounds B:

$$\theta^* = \underset{\theta \in B \subset \mathbb{R}}{\arg\max} f(\theta) \tag{10}$$

where we may only be able to evaluate f noisily. BO models the objective function as a random function and uses this model to determine informative sample locations. A popular approach is to model the underlying function as a Gaussian process (GP), fully specified by its mean μ and covariance function K. We incorporate prior beliefs about f by placing a prior measure over the space of such possible objectives. By conditioning f on the available data $\mathcal{D}_n = \{\theta_i, y_i\}_{i=1}^n$, the posterior over functions $\mathbb{P}(f \mid \mathcal{D}_n)$ is retrieved. This allows estimation of the expected value and uncertainty in $f(\theta),\ \forall\theta \in \mathbb{R}^d$. BO calculates this posterior and uses it to define an acquisition function $a(\cdot)$, which assigns a utility to evaluating f at particular θ, based on the trade off between exploration and exploitation in finding the maximum. Each evaluation yields an additional training point (θ_i, y_i). After updating the GP with the latter, BO repeats the cycle until convergence or an upper bound on the total number of evaluations. By interleaving optimisation of the acquisition function, evaluating f at the suggested point and updating the surrogate, BO forms an efficient global optimisation algorithm, in the number of function evaluations, whilst naturally dealing with noise in the outputs [15].

V. Synthetic experiments

We explore the relative performance between the three models introduced hitherto, by simulating data from a very noisy three-state HMM with Gaussian emissions, see figure 3. We treat the hyperparameters of the models as unknown quantities and perform full Bayesian inference over these quantities. We use the conjugate prior to the multivariate Gaussian emission distribution, namely the normal inverse Wishart prior $\mathsf{NIW}(\boldsymbol{\mu}_0, \lambda, \nu, \boldsymbol{\Psi})$. Through conjugacy we seek the posterior distribution of $\{\mu_j, \Sigma_j\}\ \forall j \in \mathcal{J}$, where we index group-specific parameter samples by j, given a set of observations $y_t \sim \mathcal{N}(\mu_j, \Sigma_j)$. The parameters of the conjugate prior are set as follows: $\boldsymbol{\mu}_0 = \overline{Y}$ (the empirical mean), $\nu = D + 2$, $\lambda = 0.01$ and $\boldsymbol{\Psi} = S \times \mathrm{Cov}(Y)$. Where $Y = [y_1, y_2, \dots, y_T]^\top$ and $y_t \in \mathbb{R}^D$. Where λ are the pseudo

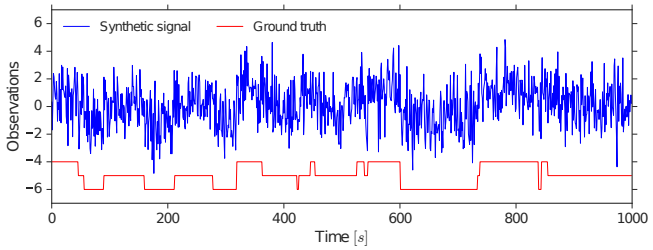


Fig. 3. Raw observations with ground truth.

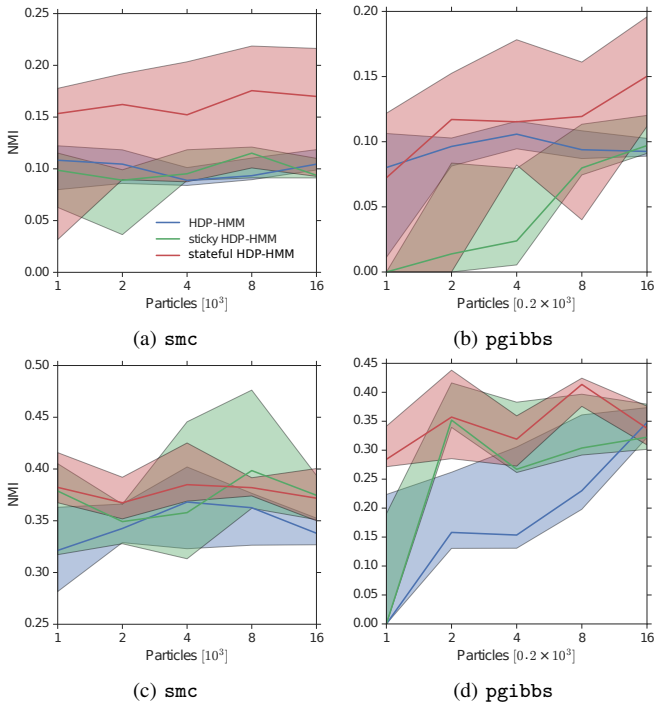
counts, $\boldsymbol{\mu}_0$ is the mean, ν the degrees of freedom, $\boldsymbol{\Psi}$ the scale matrix and S is a scaling factor.

Performance is measured with the maximum likelihood estimate of the normalised mutual information – a clustering metric [16]. For all models we place a prior of $\Gamma(1, 0.01)$ on the concentration parameter γ, of the base measure G_0 and use the same discrete measure H for all models. We place non-informative hyperpriors on the space of α and κ and perform inference using sequential Monte Carlo (`smc`) and particle Gibbs (an iterated and conditional form of `smc`) (`pgibbs`) [2]. We condition the models on the data and sample state trajectories. We set the emission prior to be $\mathsf{NIW}(\overline{\mathbf{Y}}, 3, 0.01, 0.75 \times \mathrm{Cov}(\mathbf{Y}))$.

For experiments *with* BO we seek to maximise $\log \mathbb{P}(y_{1:T})$ by optimising the model hyperparameters used to sample state trajectories s.t. $\theta \triangleq \{\alpha_\alpha, \alpha_\beta, \gamma, \lambda, C, \nu\}$, where α_α and β_α are the shape and rate parameters of the gamma distribution respectively, which is the prior on the concentration parameter α. For the sticky and stateful HDP-HMM we extend this set to $\{\alpha_\alpha, \beta_\alpha, \alpha_\kappa, \beta_\kappa, \gamma, \lambda, C, \nu\}$. We use the expected improvement [15] as our acquisition function a_{EI}. For all BO experiments we use 1000 particles for inference and use the samples to optimise a_{EI}. Once θ^* is recovered, these hyperparameters are used for the respective model, and experiments rerun for the full particle set $\{k \mid k \in \{0, \ldots, 4\} \wedge 2^k \times 10^3\}$.

Further we make use of two kernels; the radial basis function K_{RBF} and the Matérn 3/2 -kernel K_{M32} (see [17] for details). These are popular choices in the literature, we explore both because K_{RBF} is infinitely differentiable, which means that the GP with this covariance function has mean square derivatives of all orders, and is thus very smooth [17]. Smoothness this strong is typically unrealistic for many physical processes, hence K_{M32} is explored as well. It is only once differentiable and therefore only makes weak assumptions about the smoothness of f.

Performing full Bayesian inference on the model parameters, for all three models, yields the results in the top row of figure 4. It is clear from this instance that NMI increases with particle count, but so too does computational cost. Equally model structure plays a large role where the stateful HDP-HMM demonstrates better clustering ability than the other two models. At the same time neither model, for this low number of particles, performs well, and performs best under `smc` inference. Instead by optimising the hyperparameters demonstrates a clear gain – even for a signal as noisy as the test case (figure 3), for all models.



(a) `smc` (b) `pgibbs`

(c) `smc` (d) `pgibbs`

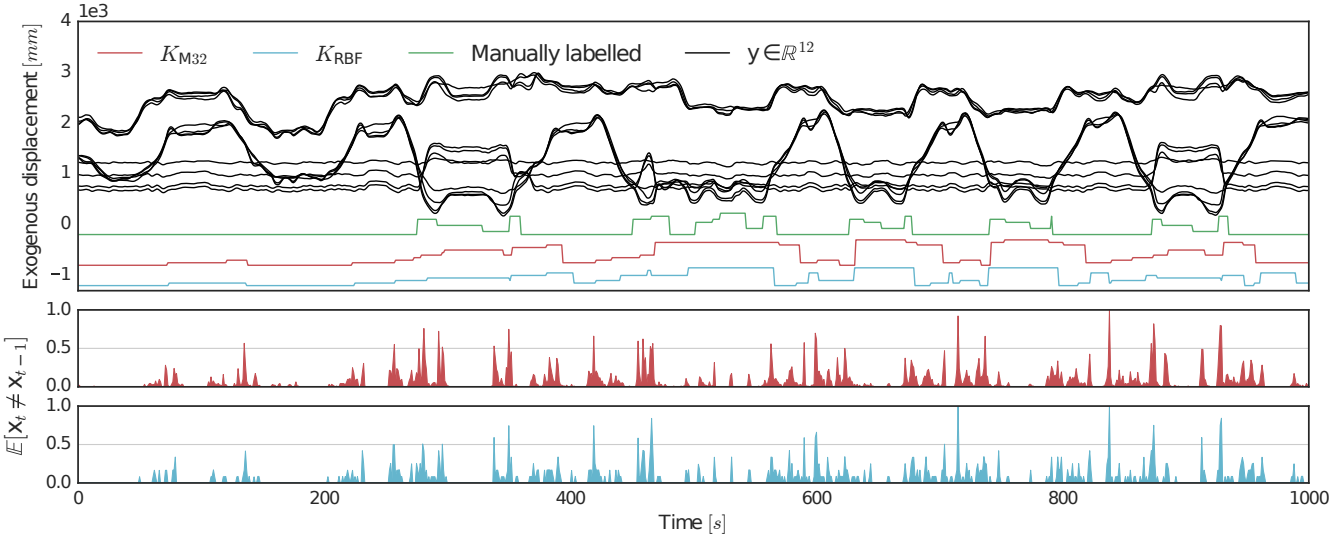
Fig. 4. The median (thick line), the 25^{th} and 75^{th} NMI quantile comparison of inferred latent state sequences, and the ground truth, using `smc` and `pgibbs`. The first row depicts baseline results without optimised hyperparameters, and the second row results optimised parameters.

VI. Human Locomotion Recognition

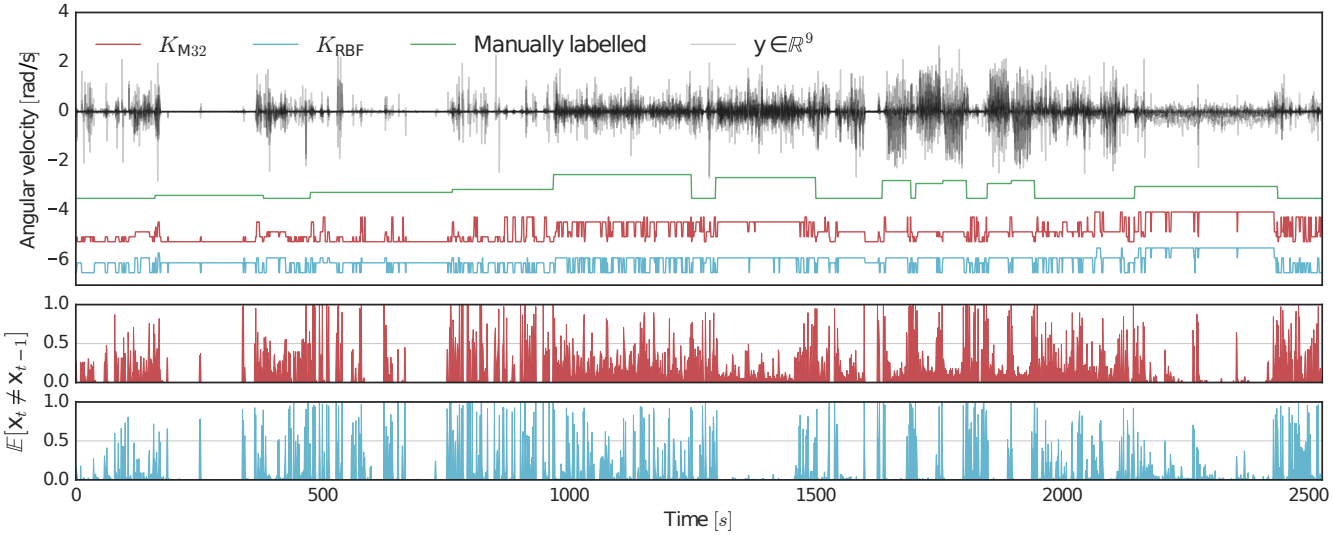
In this section we apply our methodology and specifically the stateful HDP-HMM, to two challenging labelled human locomotion datasets. It is important to note that labelling data is very subjective and tedious work. Sometimes annotators make mistakes which add noise into the labels. Treating these noisy labels as the ground truth is typically harmful for most learning methods [1]. Steps are sometimes taken to alleviate this labelling bias, e.g. [18] suggested a method that models each label as a multinomial distribution rather than deterministic. In [1], the authors treat all of the labels as noisy data, and add minor probability mass to incorrect labels enabling the model to converge to a better representation of the actions. Hence, with this in mind we use the labels with caution, in inferring statistical properties of the observations.

A. TUM Kitchen Everyday Manipulation Activities

The TUM-Kitchen dataset [19] is recorded in a home-care scenario where subjects perform daily activities in a kitchen. The kitchen is equipped with a set of ambient sensors and four static overhead cameras. The full body joints are tracked with a motion capture system. Labels are provided for the left and right hand, and the trunk of the subject. We use joint positions as they are a common feature set for locomotion segmentation [20]. Where $\mathcal{D} = \{y_i\}_{i=1}^n$ and $y_i \in \mathbb{R}^{28\times 3}$ s.t. there are 28 tracked 3D joints. We select a subset consisting of the left arm (upper arm, forearm, hand and fingers) s.t. $y_i \in \mathbb{R}^{4\times 3}$ and $n = 1000$ (total of 12,000 datapoints). Results are shown in figure 5a.



(a) TUM-Kitchen



(b) PAMAP2

Fig. 5. Top panel show experimental results with the raw data, the manually labelled state sequence (●) and the two inferred segmentation sequences with kernels K_{RBF} (●) and K_{M32} (●) respectively, for the highest log-likelihoods. The bottom panels show the expected state switching probability under the two kernels.

It is clear that our unsupervised segmentation is different to the ground truth. This was expected as has been discussed. Of greater interest is that the correct number of activities has been inferred (eight) for the K_{M32} kernel, whereas using K_{RBF} inferred a state cardinality of ten. What is worth noting is that the manually labelled sequence is segmented into highly discrete activities such as "reaching" or "carrying while locomoting" [19]. Locomotion, however, is not that discrete and instead consists of atomic motions, the combination of which serves to create larger locomotion behaviours. Tellingly, this periodic behaviour is indeed what is demonstrated in the inferred latent state sequences for the left hand. It being indicative of the periodicity and swing, often exhibited in human locomotion. That being said, it is also clear that our segmentation demonstrably fails to register certain activities. Indeed, consider the bottom two panels of figure 5a, where the expected probability of a state switch is displayed; $\mathbb{E}[\mathbf{x}_t \neq \mathbf{x}_{t-1}]$. For example there is a clear region in the middle of dataset $\sim$ 500s where the subject is grasping and reaching for objects, that is evidently not being registered with our methods. This is most likely due to an unrepresentative feature set. On the other hand the expectation plots demonstrate in more detail the dynamic switching behaviour of the dataset. From it, it would not be unreasonable to suggest that the labelling provided for this dataset is too coarse. This becomes evident when cross-validating with video evidence. Despite this, we achieve an NMI score of 0.54 and 0.51 for K_{M32} and K_{RBF} respectively.

B. PAMAP2 Physical Activity Monitoring

The PAMAP2 Physical Activity Monitoring dataset [21] contains observations of 18 different physical activities such as running, cycling and walking (each subject performs a smaller subset of these) performed by nine subjects wearing

three inertial measurement units (IMU) and a heart rate monitor. Where $y_i \in \mathbb{R}^{3\times 17}$. Where each IMU recorded large and small scale 3D acceleration, 3D gyroscopic readings and 3D magnetometer measurements. We select the gyroscopic observations s.t. $y_i \in \mathbb{R}^9$ and $n = 2528$ (total of 22,752 datapoints) as they are perceptively the more complicated segmentation case. Note that the authors have provided a special class for "transient motion" such as switching between the performance of different activities. These regions are of particular interest since the switching behaviour of the observations serves to inform the segmentation of atomic motions (if any) which may or may not be present in the observations. Results are shown in figure 5b.

The PAMAP2 datasets is significantly more challenging than TUM Kitchen, particularly as we have chosen a difficult modality. Starting with the inferred state cardinality, using K_{M32} yields seven and K_{RBF} six. Compared to true value of nine. Perhaps the biggest problem with the observation is that they are very noisy, and as such, the clustering becomes very challenging, as there is little to distinguish inbetween features; consequently assigning the wrong labels to activities. It is worth noting that activities with relatively little noise such as 'walking' (start: ~ 2200s) and high noise such as 'vacuum cleaning' (start: ~ 1300s) are labelled with ease since these features are easy to distinguish from noise and/or other similar activities (in feature space). We achieve a similar NMI score of 0.54 and 0.41 for K_{M32} and K_{RBF} respectively. This is to be expected given that the smoothness assumption the latter kernel makes, are inappropriate for this labelling task. Further, by considering the expected switching probability in the bottom two panels of figure 5b, we can use these expectations as a different form of labelling. The trajectories that we display in the main panel of figure 5b, is indeed the maximum log-marginal-likelihood sequence under the observations. However we use all sampled sequences to generate the expectation plots. As is shown, they do indeed demonstrate more certain segmentation as they are weighted probability functions. Thus, we take uncertainty into account; although state sequences are poor compared to the ground truth, we can estimate validity of the inferred state switches in the sequence, using the calculated expectations.

VII. Conclusion

By using little prior knowledge of the state space and the data at hand, we demonstrated through the use of Nonparametric SSMs and Bayesian optimisation, a methodology for sampling complete activity sequences. We have shown that quality of the feature set greatly influences the utility and accuracy of the recognition. Moreover, by using probabilistic programming, it is possible to leverage powerful inference methodologies, in a black-box manner. Married with BO, these methods define a powerful new way in which activity recognition can be induced in an almost automatic manner. Since we do not need to preselect the state space cardinality nor model hyperparameters.

There are number of different ways in which results can be improved. The most obvious, to start, is to pick a feature set that captures the full modalities of human locomotion in some setting. Secondly inference algorithmic development is ever ongoing, and will become more adept at performing inference in high-dimensional state spaces. Finally, BO has been used throughout this work, but usually only with default settings and standard kernels. Kernels are the most important item in BO, and should be chosen with care. Or better yet, their structured learned from observations as well.

References

[1] N. Hu, G. Englebienne, Z. Lou, and B. Krose, "A hierarchical representation for human activity recognition with noisy labels," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 2517–2522.

[2] F. Wood, J. W. van de Meent, and V. Mansinghka, "A new approach to probabilistic programming inference," in *Proceedings of the 17th International conference on Artificial Intelligence and Statistics*, 2014.

[3] J. H. Huggins and F. Wood, "Infinite structured hidden semi-markov models," *arXiv preprint arXiv:1407.0044*, 2014.

[4] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the american statistical association*, vol. 101, no. 476, 2006.

[5] A. Y. Lo *et al.*, "On a class of bayesian nonparametric estimates: I. density estimates," *The annals of statistics*, vol. 12, no. 1, pp. 351–357, 1984.

[6] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[7] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden markov model," in *Advances in neural information processing systems*, 2001, pp. 577–584.

[8] Y. W. Teh and M. I. Jordan, "Hierarchical bayesian nonparametric models with applications," *Bayesian nonparametrics*, vol. 1, 2010.

[9] M. J. Johnson and A. S. Willsky, "Bayesian nonparametric hidden semi-markov models," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 673–701, 2013.

[10] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 312–319.

[11] J. Van Gael, Y. Saatci, Y. W. Teh, and Z. Ghahramani, "Beam sampling for the infinite hidden markov model," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1088–1095.

[12] A. D. Gordon, T. A. Henzinger, A. V. Nori, and S. K. Rajamani, "Probabilistic programming," in *Proceedings of the on Future of Software Engineering*. ACM, 2014, pp. 167–181.

[13] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 3, pp. 269–342, 2010.

[14] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[15] T. Rainforth, J.-W. van de Meent, and F. Wood, "Bayesian optimization for probabilistic programs," *1st NIPS Workshop on Black Box Learning and Inference*, 2015.

[16] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[17] C. E. Rasmussen and C. K. I. Williams, "Gaussian processes for machine learning," 2006.

[18] N. Hu, Z. Lou, G. Englebienne, B. Kröse, *et al.*, "Learning to recognize human activities from soft labeled data," 2014.

[19] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition," in *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009*, 2009.

[20] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from rgbd images." *plan, activity, and intent recognition*, vol. 64, 2011.

[21] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Wearable Computers (ISWC), 2012 16th International Symposium on*. IEEE, 2012, pp. 108–109.