Interpreting Lion Behaviour with Nonparametric Probabilistic Programs

Neil Dhir[*], **Frank Wood**
Department of Engineering Science
University of Oxford
neild@robots.ox.ac.uk

Matthijs Vákár[*], **Andrew Markham**
Department of Computer Science
University of Oxford
matthijs.vakar@cs.ox.ac.uk

Matthew Wijers, Paul Trethowan,
Byron Du Preez, Andrew Loveridge,
David MacDonald
Department of Zoology
University of Oxford

Abstract

We consider the problem of unsupervised learning of meaningful behavioural segments of high-dimensional time-series observations, collected from a pride of African lions[1]. We demonstrate, by way of a probabilistic programming system (PPS), a methodology which allows for quick iteration over models and Bayesian inferences, which enables us to learn meaningful behavioural segments. We introduce a new Bayesian nonparametric (BNP) state-space model, which extends the hierarchical Dirichlet process (HDP) hidden Markov model (HMM) with an explicit BNP treatment of duration distributions, to deal with different levels of granularity of the latent behavioural space of the lions. The ease with which this is done exemplifies the flexibility that a PPS gives a scientist[2]. Furthermore, we combine this approach with unsupervised feature learning, using variational autoencoders.

1 INTRODUCTION

Animal accelerometer data allows ecologists to identify important correlates and drivers of animal behaviour. Use of accelerometers is widespread within animal biotelemetry as they provide a means of measuring an animal's activity in a meaningful and quantitative way, where direct observation is not possible. In *sequential* acceleration data there is a natural *dependence between observations* of movement or behaviour, a fact that has been largely ignored in most analyses [13]. Recordings are typically sampled at a high temporal resolution, sometimes for years at a time, using tri-axial accelerometer tags [15], which quickly results in terabytes of data that present various challenges regarding transmission, storage, processing and statistical modelling. The latter can be achieved by employing statistical classification methods, and entails observing the animal, *manually assigning labels* corresponding to *known* behaviours to segments of the data, and training a model using the labelled data in order to subsequently classify remaining unlabelled data based on certain *chosen acceleration features* deemed to be salient by domain experts.

Consider the recent work of Pagano et al. [20] wherein the authors use tri-axial accelerometers to identify wild polar bear behaviours. They note that identification of wild animals can be facilitated using captive counterparts, as their accelerometer signatures are generally assumed to be similar to those of their wild kin [20]. They use their captive bears as surrogates for wild ground-truth behaviour, upon which they model polar bear behaviour on sea ice and land, using random forests classification and hand-engineered features. Their results, though of good accuracy, rely on hand-engineered features, large assumptions about captive and wild behaviours and a fundamental need for labelled data to infer behaviour. Along this trail of thought, the work McClune et al. [17] is relevant for our discussion. Therein, the authors fitted tri-axial accelerometers to a tame and captive Eurasian badger, upon which it was allowed to roam free in an enclosure, whilst movements were video recorded and used as ground-truth for its behavioural states. Again, features were hand-engineered using, e.g., acceleration magnitude and principal component analysis. The k-nearest neighbour classifier and decision trees were used to automate classification of behaviours [17]. Their success ranged from 77.4% to 100% classification accuracy, though again deploying a highly laborious process, where a human is necessary for the extraction of the video ground-truth (and classification is conditioned on the ground truth existing at all). The work by Leos-

[*] Equal contribution

[1] You may be familiar with one of the more famous members of our study, the late Cecil the lion.

[2] Example code: https://goo.gl/14s8Sa

Barajas et al. [13] is faced with precisely this challenge, where they seek to measure an animal's activity in a meaningful and quantitative way where direct observation is not possible [13]. In doing so, they investigate a marine and an aerial system (sharks and eagles). They used the classical HMM to effect supervised and unsupervised learning of animal activity. In the latter case, an HMM is used to segment unlabelled acceleration data into a finite set of pre-specified categories [13]. They go on to show that the metrics that they derive from the learned eagle -states provide meaningful insight into activity levels and thus can lead to biologically interpretable states. Their study is similar to that of Phillips et al. [21], in which the authors applied HMMs in an unsupervised context to model the behaviour of free swimming tuna from vertical movement data collected by data-storage tags [13, §3.3].

As we have thus demonstrated some studies *do* employ models with temporal dependency (e.g., the Markov assumption) but the far more popular method is to deploy classification algorithms that do not, e.g., support vector machines (SVM) or random forests see, e.g., [3]. We, like Leos-Barajas et al. [13], note that disregarding the serial dependence in the acceleration data usually is not a realistic assumption. Moreover, independent and identically distributed statistics (i.i.d.) pose a particular risk if "inferential statistics are applied to the output of say a machine learning algorithm" [13].

In this study, we therefore focus on state-space models like HMMs to model accelerometer data. In doing so we take care to emphasise the two significant disadvantages of a simple HMM: (1) *state duration distributions are necessarily restricted to be of the geometric shape* $\mathbb{P}(d) = a^{d-1}(1-a)$ (and, in particular, monotonically decreasing), where d denotes the duration of a given state and a denotes its self-transition probability, which is not appropriate for many real-world problems like modelling animal behaviour (for instance, a human being usually sleeps for roughly 7-8 hours and, certainly, sleep durations are not monotonically decreasing), and (2) the *number of hidden states must be set a priori* while one of the main objectives of behavioural research is to discover new or more fine-grained behavioural patterns.

Recent work has addressed the latter issue by way of *Bayesian nonparametric HMMs*, which allow us to infer state cardinality from observations, and allows it to grow in a data-driven fashion. The former issue is addressed by variations of the HMM with non-geometric duration distributions like the Hidden semi-Markov model [18]. In the BNP paradigm, inference is usually performed in models with an infinite number of states. Additionally, we consider a BNP treatment of durations.

In this paper, we aim to introduce these *two innovations to the zoology toolbox*, while at the same time showing the value that *unsupervised (behaviour) learning* can have, without reliance on expensive and difficult manual data annotation, as well as *unsupervised feature learning*, reducing the reliance on domain experts to guess the most salient features for possibly unknown kinds of behaviour. We perform full unsupervised nonparametric time-series learning on the observations, resulting in a time-segmented set of waveforms. Such a segmentation can provide useful interpretations which allow us to quantify animal behaviour, energetic expenditure and deepen our insights into individual behaviour as a constituent of populations and ecosystems [13]. Further, we hope to convey the huge value that a *probabilistic programming language* like Anglican can have when exploring the space of possible new models for a novel application domain like ours, in enabling quick implementation and evaluation of many complex models and even design of new ones with minimal mathematical and coding burden to the scientist (e.g., no need to derive custom inference algorithms).

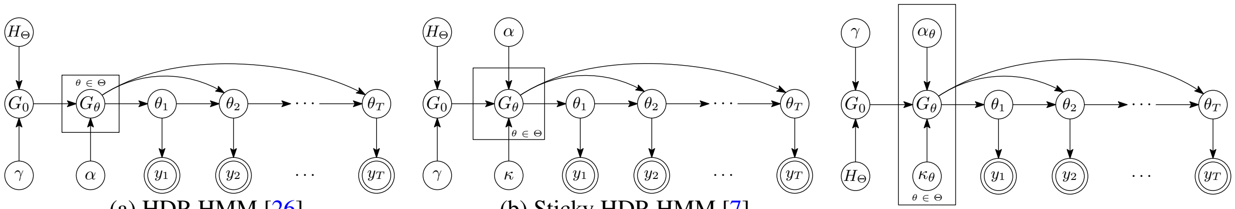
The paper is organised as follows; in section 2, we give an exegesis of BNP HMMs; in section 3, we present the infinite duration HMM and section 4 discusses our choice of using a probabilistic programming framework to implement our models. Experiments and results are presented for the understanding of lion behaviour ecology in section 5. In section 6, we conclude.

2 BAYESIAN NONPARAMETRICS

We motivate our approach by reviewing the principles underpinning BNP staet-space models (SSM), and use this foundation to motivate our contributions in the coming sections, taking our cue from Dhir et al. [5, §II].

We need to be able to infer state cardinality from observations, as well as discover new states as we acquire more data. The former consideration of determining model complexity motivates a Bayesian approach, while the latter suggests a nonparametric model. The defining difference between BNP methods and their parametric cousins is that the size of the representation can grow as we gather more data: for the HDP-HMM, the expectation and variance of the number of states grow logarithmically with the size of the dataset [24]. We can frame our prior through Lo et al. [14]'s suggestion of an infinite dimensional mixture model:

$$\begin{aligned} & G \sim \mathcal{P} & & \\ \theta_i \mid G & \overset{i.i.d.}{\sim} G & & i = 1, 2, \ldots \\ y_i \mid \theta_i & \overset{ind.}{\sim} F(\theta_i) & & i = 1, 2, \ldots \end{aligned} \tag{1}$$



(a) HDP-HMM [26].

(b) Sticky HDP-HMM [7].

(c) Stateful HDP-HMM [5].

Figure 1: Three types of Bayesian nonparametric hidden Markov models, with various structural prior encoded which, from left to right, induce higher probability of state persistence.

where G is a discrete random probability measure (RPM) with distribution $\mathcal{P}$, $y_{1:n}$ are a collection of continuous and possibly multivariate observations and $\theta_{1:n}$ are the corresponding collection of latent random variables from an i.i.d. sequence directed by G and $F(\theta_i)$ is some continuous distribution parametrised by θ_i [5]. The nonparametric hierarchical model in eq. (1) defines a mixture model (MM) with a potentially countably infinite number of components. Because the RPM in equation eq. (1) is discrete, this means that a pair of consecutive values of θ take on the same value with a strictly positive probability. This value defines a mixture component. By setting the RPM to the Dirichlet process (DP) [6] we obtain the familiar DP mixture model

$$\begin{aligned} &G \mid \gamma, H \sim \mathcal{DP}(\gamma, H), \\ &\theta_i \mid G \sim G, \\ &y_i \mid \theta_i \sim F(\theta_i) \qquad\qquad i = 1, 2, \ldots. \end{aligned} \tag{2}$$

The DP, denoted by $\mathcal{DP}(\gamma, H)$, is a distribution over random measures on a parameter space Θ with countably infinite support. It can be parameterised 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 [5]. Note also that, e.g., the Pitman-Yor process or any other distribution over discrete RPMs are valid alternatives to the DP.

In many scenarios we posit that *groups* of data are thought to be produced by related, yet distinct, generative processes [5]. Concurrently, 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 and can be thought of as generated by related but distinct processes [7]. Both of these scenarios describe the problem we face in this work, and both can be analysed through a *hierarchical* BNP approach. Viewed through this lens, observations can be subdivided into a countable collection of groups [5]. Then, we take it that 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 obtain the hierarchical DP (HDP) [26]. Teh et al. [26] explain that the HDP induces sharing of support among the random measures G_j since each inherits its support from the same G_0. This idea is used to develop HMMs with unknown, potentially infinite, state spaces.

2.1 INFINITE HIDDEN MARKOV MODELS

Dhir et al. [5] describe an HMM as 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 $\{G_\theta : \theta \in \Theta\}$ [25]. 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 [26] - see fig. 1a; an HMM with a countably infinite state space, with generative model

$$G_0 \mid \gamma, H \sim \mathcal{DP}(\gamma, H), \tag{3}$$

$$G_\theta \mid \alpha, G_0 \sim \mathcal{DP}(\alpha, G_0) \qquad \text{for } \theta \in \Theta, \tag{4}$$

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

$$y_t \mid \theta_t \sim F(\theta_t).$$

To frame this discussion, consider two alternatives to eq. (4) (see figs. 1b and 1c). The first extension is by Fox et al. [7]

$$G_\theta \mid \alpha, G_0, \kappa, \theta \sim \mathcal{DP}\left(\alpha + \kappa, \frac{\alpha G_0 + \kappa \delta_\theta}{\alpha + \kappa}\right) \tag{5}$$

and the second by Dhir et al. [5]

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

We shall consider both in turn, in league with our discussion and motivation of this approach. Hence, consider that each G_θ is a DP draw (this is true for eqs. (4) - (6)), 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 [5]. 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 [11, 5]. But, the rate at which state change 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 [11]. To alleviate this the *sticky* HDP-HMM (see fig. 1b) was introduced by Fox et al. [7], with transition kernel in eq. (5), that augment the HDP-HMM with an extra parameter $\kappa > 0$ which encourages self-transitions and thus longer state durations. The model, however, still shares its global self-transition bias with all the other states, and so it does not allow for learning state-specific duration information [11]. An attempt to deal with this restriction was introduced by Dhir et al. [5] with the *stateful* HDP-HMM (fig. 1c), in which the authors 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. Where the transition kernel is given in eq. (6).

Performing inference over these types of models seeks to infer the posterior distributions over the state sequence and, therefore, implicitly, over the persistence of each state. It is with respect to this latter domain, that we propose new methodology for inferring state-specific durations, drawn from an infinite set.

3 INFINITE DURATION HIDDEN MARKOV MODEL

As discussed, in the classic HMM, the duration of a given state has a geometric (in particular, monotonically decreasing) distribution, because of the Markov property. Geometric duration distributions have been found to be deficient not just in behavioural modelling but also in e.g. speech synthesis [1]. Explicit duration HMMs (EDHMM) have been developed [4] to make the duration distribution explicit and allow it to have a more general form. Put simply, in an EDHMM, during a (Markov) state transition, a duration is drawn explicitly from a specified duration distribution depending on the new state. After that, the probability of self-transition is one until the duration has elapsed. We prefer to consider the EDHMM over the hidden semi-Markov model (HSMM), which achieves a similar effect through different, less explicit means [11].

The HDP-HSMM was introduced by Johnson & Willsky [11] and has its graphical model structure shown in fig. 2c (see appendix E for further discussion). Contrast this to the infinite duration HMM (IDHMM), a BNP variant on the EDHMM, which we present herein (see fig. 2a). We posit that the IDHMM can be preferable as it gives a nonparametric rather than parametric treatment of duration distributions. The IDHMM models the relationship between state $\theta_t \in \Theta \subseteq \mathbb{N}$, duration $d_t \in \mathcal{D} \subseteq \mathbb{N}$ and observation $y_t \in \mathcal{Y} \subseteq \mathbb{R}^n$, $\forall t \in \mathcal{T} \triangleq \{1, \ldots, T\}$, whilst giving a nonparametric treatment of state cardinality and state duration. The base and group distributions, in the generative model, are drawn as

$$\begin{aligned}
&G_0 \mid \gamma, H_\Theta \sim \mathcal{DP}(\gamma, H_\Theta) \\
&D_0 \mid \gamma', H_\mathcal{D} \sim \mathcal{DP}(\gamma', H_\mathcal{D}) && (7) \\
&G_\theta \mid \alpha, G_0 \sim \mathcal{DP}(\alpha, G_0) \qquad \text{for } \theta \in \Theta \\
&D_\theta \mid \alpha', D_0 \sim \mathcal{DP}(\alpha', D_0) \qquad \text{for } \theta \in \Theta && (8)
\end{aligned}$$

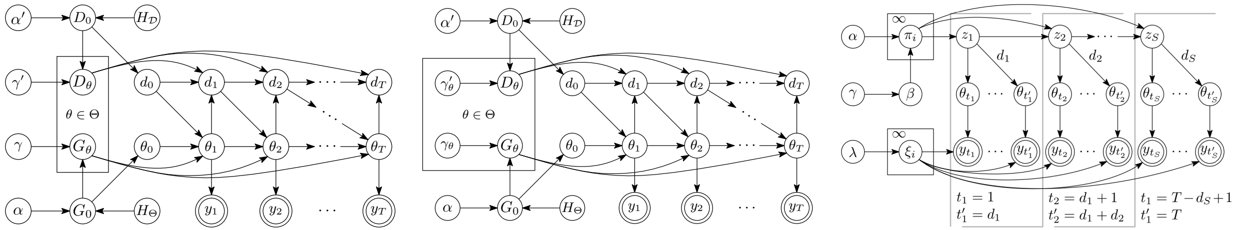
where G_0 and D_0 have support Θ and $\mathcal{D}$ respectively. Noting that G_θ, being a draw from a DP, is a discrete distribution, we can define $G_\theta[\mathbb{P}(\theta) = 0]$ as the measure obtained from G_θ by setting the probability of drawing θ to 0 and renormalising. Next, a sequence of states, durations and emissions are drawn as

$$\begin{aligned}
\theta_t \mid \theta_{t-1}, d_{t-1} &\sim \begin{cases} \delta_{\theta_{t-1}}, & \text{if } d_{t-1} > 1 \\ G_{\theta_{t-1}}[\mathbb{P}(\theta_{t-1}) = 0], & \text{otherwise} \end{cases} \\
d_t \mid d_{t-1}, \theta_t &\sim \begin{cases} \delta_{d_{t-1}-1}, & \text{if } d_{t-1} > 1 \\ D_{\theta_{t-1}}, & \text{otherwise} \end{cases} \\
y_t \mid \theta_t &\sim F(\theta_t).
\end{aligned}$$

where δ_a is a δ-distribution with all its mass on a [4]. That is, we keep the state fixed and decrease the duration with one or, if the duration would reach 0, we sample a *different* new state, depending on the old state, and a corresponding new duration, depending on the new state.

Consider that each $G_\theta[\mathbb{P}(\theta) = 0]$ is obtained from 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 $\mathbb{E}[G_\theta[\mathbb{P}(\theta) = 0] \mid G_0] = G_0[\mathbb{P}(\theta) = 0]$, $\forall \theta \in \Theta$. Since transition distributions tend to have their mass concentrated around a common set of states, a bias towards re-entering and re-using a common of states is received. Similarly, our choice to extend the representational power of an HDP to durations as well means that the model reuses a common set of state durations.

We can also expose the IDHMM to the stateful representation, fig. 2b, as analogous to that in [5]. As for the stateful HDP-HMM, this adds greater heterogeneity in the state and dwell-time distributions of the inferred states. We will explain in section 4 how general purpose inference allows us to quickly build and experiment with models in this modular fashion, which can give us greater flexibility than using bespoke sampling algorithms such



(a) IDHMM. (b) Stateful IDHMM. (c) HDP-HSMM [11].

Figure 2: Bayesian nonparametric state-space models with explicitly modelled state dwell durations.

as those employed by, e.g., Fox et al. [7] and Johnson & Willsky [11], in settings like ours where the computational efficiency of the latter is outweighed by the programmer efficiency of the former.

4 GENERAL PURPOSE INFERENCE

As the optimal model structure, to nonparametrically and in an unsupervised fashion, infer behavioural states from accelerometer data was far from clear, we chose to use a PPS for our analyses, specifically *Anglican* [30]. The idea of a PPS is that it is a programming language with an implementation of many of the basic building blocks for both statistical models and (general purpose) inference algorithms. By composing these building blocks, even a non-expert should be able to quickly build complex statistical models. For instance, Anglican provides a primitive for a Chinese restaurant process which easily allowed us to implement our (H)DPs.

Importantly, Anglican (like many PPS:es) provides a separation between model specification and implementation of inference. In particular, it comes equipped with several general purpose inference algorithms, such as Sequential Monte Carlo (SMC), Markov chain Monte Carlo (MCMC) and Particle MCMC. These can be combined with any model, allowing a user to focus on the modelling task, without having to worry about the implementation of a complicated custom inference algorithm.

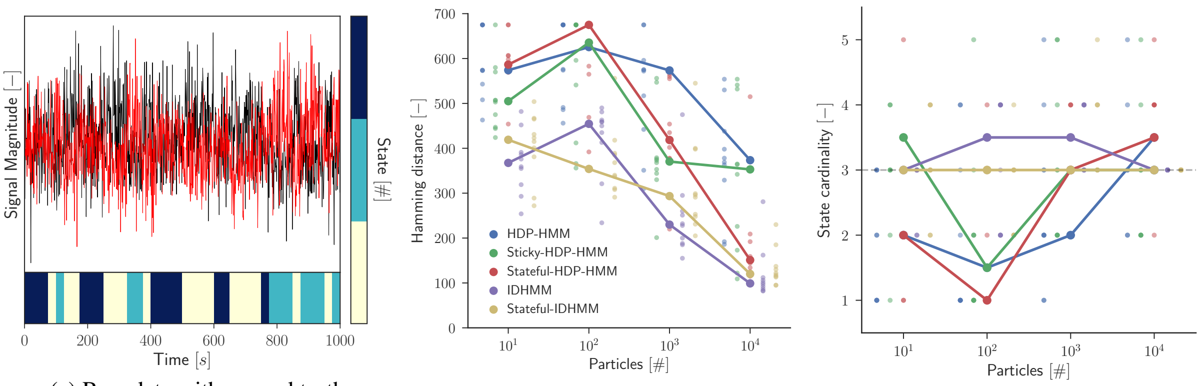
A model is a simplified representation of reality, and the simplifications are made to discard unnecessary detail and allow us to focus on the aspect of reality that we want to understand. Depending on the problem, it is important to assess the trade-offs between speed, accuracy, and complexity of different models and algorithms and find a model that works best for that particular problem. Consequently, the pairing of a suitable inference scheme to a model is a notoriously difficult problem; no one method is likely to generalise across the board [29]. A PPS allows us to quickly and accurately iterate over models and inference methods, to find the most optimal pair conditioned on the problem domain. In our case, this enabled us to quickly experiment with and evaluate many models for lion accelerometer data. The fact that we could quickly design and implement a new model, the IDHMM, speaks to the ease of use and flexibility of the probabilistic programming methodology. In practice, we chose to perform our analyses in Anglican with the SMC inference algorithm. SMC takes as a parameter “the number of particles”, where higher numbers increase the expected accuracy of inference.

5 LEARNING LION BEHAVIOUR

We qualify our models and inference methodology, by applying them, and contemporary state-of-the-art, to synthetic followed by lion observations.

5.1 SYNTHETIC OBSERVATIONS

We explore the relative performance between the five models introduced hitherto, by simulating observations $y_t \in \mathbb{R}^{D=2}$, from a very noisy three-state multivariate HSMM with Gaussian emissions – see fig. 3a. Consequently, from the inference perspective, the emission distribution has unknown mean and covariance parameters. The conjugate prior is the normal-inverse-Wishart distribution, denoted by $\mathsf{NIW}(\mu_0, \lambda_0, \Psi, \nu)$. Through conjugacy we seek the posterior distribution of $\{\mu_\theta, \Sigma_\theta\} \; \forall \theta \in \Theta$, where we index group-specific (i.e., behaviour-specific) parameter samples by θ, given a set of observations $y_t \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$. For convenience of notation let $\mathbf{Y} = [y_1, \ldots, y_T]^\mathsf{T}$. For brevity, results are only shown for SMC, chosen for its superior performance for this model class. In addition, particle Gibbs (iterated conditional SMC) and lightweight Metropolis-Hastings were all tested as part of our experiments [30]. Model and emissions parameter priors are shown in table 1 in appendix C. We place non-informative hyperpriors on model parameters and then condition the models on the observations and sample state trajectories; i.e. $\theta \to t, \; \forall \theta \in \Theta \wedge \forall t \in \{1, \ldots, T\}$. We use synthetic observations of size $T = 1000$ and use 100 samples for each particle count (see fig. 3).



(a) Raw data with ground truth

(b) Hamming distance

(c) State cardinality

Figure 3: Results from experiments on multivariate synthetic Gaussian observations with SMC inference. A three-state HSMM with non-geometric duration distributions was sampled to create the observation set in fig. 3a. All models under discussion are compared and contrasted in fig. 3b and fig. 3c.

Performance is measured with the Hamming distance, a common clustering metric [16]. This distance metric works by mapping the randomly chosen indices of the estimated state sequence, to the set of indices that maximise the overlap with the true sequence [7]. In other words the distance between two sequences, is the number of positions at which the corresponding symbols are different. Hence, the lower, the better.

The results in fig. 3 demonstrate the advantage that our proposed model structure can have for modelling phenomena with non-geometric duration distributions. First, as demonstrated by [5], stateful models yield a clear benefit compared to their contemporaries. This is because we increase the heterogeneity of the dwell-time distribution of the inferred states of the model, by making the model statistics group specific. This is shown in fig. 3b and fig. 3c where the stateful HDP-HMM outperforms the HDP-HMM and sticky HDP-HMM for some particle counts. The benefits of a stateful representation are less clearcut for the IDHMM, but both IDHMMs outperform other state-of-the-art BNP SSMs.

By extending the HDP-HMM and stateful HDP-HMM, by drawing upon explicit-duration semi-Markovianity [11], as was done in the HDP-HSMM, one allows for the *parametric* construction of highly interpretable models which admit prior information on the state durations [11]. We make that model more flexible still by giving a nonparametric treatment treatment to the state cardinality *and* durations. By levering the statistical strength of HDPs for durations, we can model complex duration phenomena, which may be more difficult to treat in a parametric setting. We demonstrate, in appendix D, in fig. 9b and fig. 9c the flexibility that our model structure provides. By employing the duration distributions shown in fig. 9a, we gain approximately the same utility in terms of the Hamming distance, when using a uniform-discrete duration prior. However because we are targeting a duration process of the form shown by the red bars in fig. 9a, it is more appropriate to focus our duration prior density on those regions. Hence, we see that by employing two poorly-specified mixture duration distributions (fig. 9b: $\mathcal{N}(35,15)+\mathcal{N}(85,15)$ and fig. 9c: $\text{Pois}(35)+\text{Pois}(85)$ – with equal mixing proportions), the posterior state cardinality is better specified.

5.2 LION BEHAVIOUR

We demonstrate our methods by applying them to lion behaviour segmentation, to better understand their ecology. Hence, biologger observations are becoming increasingly popular tools for animal behaviour research. The number of studies using accelerometers in particular, has increased rapidly over the last 15 years due to the advantages offered over methods relying solely on direct observation [2]. While direct observation may be the only viable means of studying animal behaviour in certain cases, it can pose several difficulties which may include biases suffered as a result of observer presence [9] or the inability to continuously observe the focal animal if it is an elusive species, or a species that occurs in inaccessible habitats. The African lion is an example of a species for which behavioural research can benefit from accelerometer data-loggers due the challenges associated with keeping study individuals in sight continuously while avoiding influencing their behaviour. However, with the ability to record continuously at sampling frequencies as high as 10,000Hz [2], accelerometers generate extremely large datasets which are impossible to classify manually, which is why unsupervised learning could help.

The majority of studies which make use of machine learning to classify large accelerometry datasets, tend to rely on supervised learning techniques [2] or very coarse quantification of activity [19]. While such techniques have proven effective for many studies focussing on select, broad behavioural states such as 'stationary', 'mobile' and 'feeding' as shown by [8] on the cheetah, they are potentially limiting for those aimed at developing detailed activity budgets, as detailed manual labelling can be labour intensive and difficult. Hence, in this study we seek to dwell deeper by investigating, on a per-second basis, tri-axial accelerometry $(\ddot{x}, \ddot{y}, \ddot{z})$ and magnetometer $(\mathbf{B}_x, \mathbf{B}_y, \mathbf{B}_z)$ observations, collected from a male lion using a 32Hz sampling rate. *Unlike* previous studies, we shall employ a novel form of feature engineering, for our time-series observations, using the recent development of the variational autoencoder [12].

5.3 UNSUPERVISED FEATURE LEARNING

The study by Rahman et al. [22] presented an application of autoencoders (AE) to temporal tri-axial accelerometry observations. They used it to effect unsupervised feature learning, later used for supervised classification of cattle behaviour. This data-driven approach is one which we shall espouse too with the difference that we instead use *variational* AEs (VAE), whose generative nature has the advantage that the learned features are easier to interpret. For further details on the VAE structure and our implementation see appendix B.

5.4 FUZZY GROUND TRUTH

Ground truth (GT), as we have already alluded to, is an elusive property in the zoological domain. In section 1, we noted how, e.g., video was used as a form of GT. In our case, GT, or labelled observations, is received via sound [27]. The collar of each lion, apart from being equipped with sensors that log physical variations (e.g., acceleration), also log the audio of each animal. This enables the zoologist to get a measure of the animal's activity at time t. It also means that in order to ascertain a dataset that can be used for statistical learning, an exceptionally expensive process takes place where a human listens to an audio recording. For it to be of any use though, that recording has to span not hours, but days. Consequently this type of labelling is prohibitive due to its huge cost in man hours, but it also needs to be performed by the same person, as to remove as much bias as possible (e.g., our dataset contains 'trot' and 'walking' - two activities that are perceptively similar). This is another reason why unsupervised learning could prove preferable, being solely observation-driven. Furthermore, whilst it is true that a human does listen to the lion for a significant portion of time, she does *not* listen to the whole recording (which again can span days). Instead, recordings are sub-sampled, where, e.g., every other minute (or five minutes in some cases) are monitored and the inferred label (based on the zoologists' interpretation of the lion's *audible* activity at that point in time) is interpolated until the next sampling point. All of which leads to a form of semi ground truth.

5.5 EXPERIMENTS

Before describing the minutia of our experiments, it is important to re-iterate the purpose of this exercise, and the potential value it could have for zoological studies. Whilst a human will need to be in the semi-GT extraction-loop, we propose that the methods within can function as a conduit for *behaviour discovery*. Differently put, we posit that our methodology can function as a useful tool for zoologist, as the unsupervised segmentation will allow them to hone in on regions of interest, and consequently will allow them to more intelligently choose regions of interest for their work. Upon which their semi-GT can be used in any of the standard supervised classification methods which hitherto have proved their worth in this domain (when good GT is available).

For our experiments, we used 10 hours of labelled[3] observations for one lion, part of a pride currently being studied by our institution. After experimenting with several window sizes, we settled on $3s$ as providing a useful level of granularity. As such, each observation $y \in \mathbb{R}^{96\times 6}$ (see appendix B, fig. 8b), was normalised, passed to a VAE, where a low-dimensional latent representation $z \in \mathbb{R}^3$ was extracted. The sequence of latent representations is what we used as input for our models, all of which were written as probabilistic programs (see appendix appendix F, for an example), to which we applied black-box SMC inference. We used the same conjugate prior as in section 5.1. For details see table 2 in appendix C. In summary, the purpose of these methods is not to segment the signal into 'correct' features (given that no true form of the ground truth exists). Rather, the purpose, given limited and noisy information, is to detect regions of interest (as opposed to, e.g., large regions of a resting behaviour). Results are shown in fig. 4.

5.6 DETAILED ANALYSIS: A HUNT

In fig. 5, we demonstrate the utility of the IDHMM and the stateful IDHMM, on a smaller segment of fig. 4. In this experiment, we sought to understand if our methods could accurately segment fast-changing animal behaviour, specifically that related to hunting. Panel four,

[3] In the sense in which this exposition is framed.

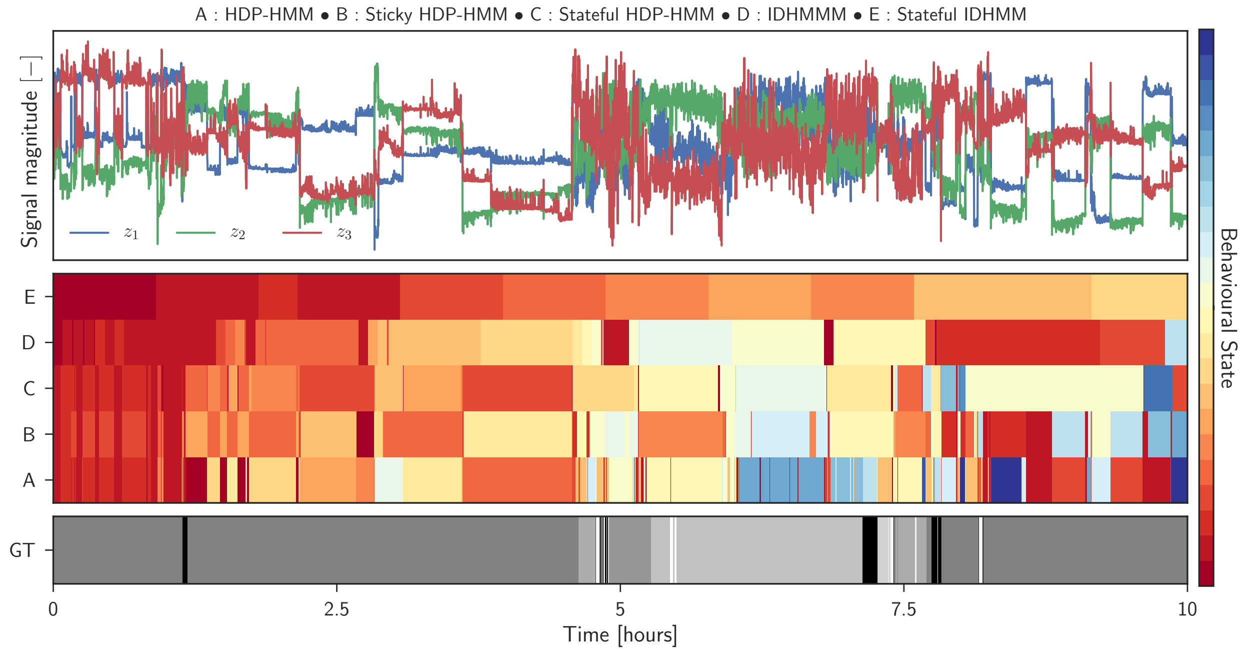


Figure 4: The **top panel** displays the feature-set over which inference was performed. The **middle panel** shows the inferred state trajectories, with the highest log-marginal likelihood $\log \mathbb{P}(y_{1:T})$ for all models. The **bottom panel** displays the manually labelled ground-truth which serves as a comparison to our unsupervised labelling. The colorbar maps the numbers of inferred states to each model heatmap in the middle panel.

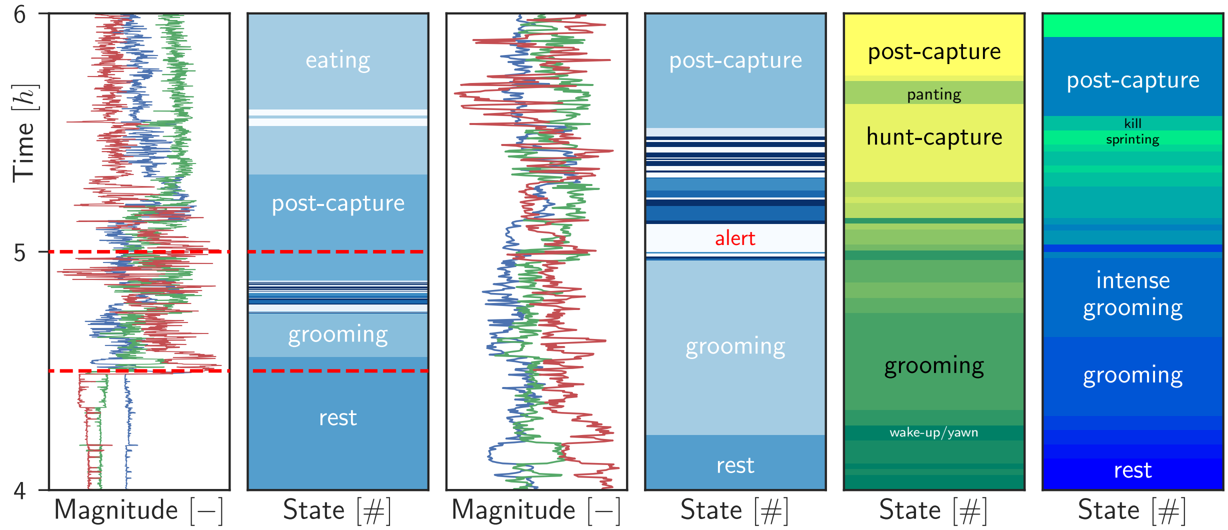


Figure 5: Detailed depiction of a two hour segment which features hunting behaviour, using the IDHMM and the stateful IDHMM. The **first two panels** (from the left) show a two hour segment with 'ground truth', the **second pair of panels** show a zoomed in 30min period in which multiple behaviours exist (some of which have been annotated). The **final two panels** show the inferred behaviour sequences using the IDHMM and the stateful IDHMM (last panel).

from the left, in fig. 5 contains the manual labelling of this segment, followed by the IDHMM and stateful IDHMM inferred state trajectories. For model and conjugate prior details, see table 3 in appendix C.

Post-analysis showed that the models are reasonably successful in segmenting the time-series from a zoological point of view; subtle and short behaviours can be picked out from the audio, which the models 'accurately' recognise from the input sequences. Granted, post-analysis is subject to human error and bias, but is still valuable, if only to validate that *something* has been found and should be studied in more detail. As an example, consider panel four again (from the left), the large top segment labelled 'post-capture' is immediately preceded by a long (minutes) sequence of alert and walking behaviours, all of which eventually lead to a chase and kill. The human labelling of this segment is good, but could be better at picking out subtle behaviours such as the sprint that leads to the kill (behaviours currently folded into the 'post-capture' and 'capture' labels, the latter of which is not shown on the plot). The stateful IDHMM successfully captures this, while the IDHMM does not.

6 DISCUSSION AND CONCLUSION

From the middle panel of fig. 4, a clear trend emerges regarding the nature and behaviour of the models w.r.t. to the observations. The IDHMMs variants (models E and D in fig. 4) allow the practitioner to employ specific domain knowledge regarding the duration distribution of the phenomena being studied. Hence as shown, the model samples from a bespoke duration distribution, where, in this instance, we have employed a simple mixture of discrete-uniform distributions that reflect the duration content as seen in the feature space. In fig. 4, the light grey area preceding the five-hour mark, constitutes an area of less frequent behaviour (as labelled by the zoologists); a hunt (labelled as 'capture'), followed by a kill, followed by post-kill behaviour such as eating and drinking. It is clear that all models segment the onset of this activity sequence, but then differ in the duration properties and number of activities present in this event segment. The HDP-HMM and the sticky HDP-HMM both capture the fast switching dynamics. The other models do not, the IDHMMs do not by design, as they are primed to find *large* regions of interest, with statistical observation similarity. This points towards a scenario where both types of models are used jointly, as the strength of their sum is greater than their individual parts. Viewed this way, we can apply the models of fig. 4 top to bottom. State-space models that better deal with coarse state-space granularity, and observations with non-geometric duration distributions, are labelled top to bottom in the middle panel, according to how much granularity they offer the user for this task. Conversely the IDHMM can be tuned to model bursts of activity as demonstrated in fig. 5.

Having ascertained where large regions of interest are located, we can turn to models A-C of the middle panel in fig. 4. The sticky HDP-HMM, not being as state-persistent as the stateful version, does not smooth out the activity labelling as much, but still quickly introduces new labels for surprising features. The inferred state cardinalities for the HDP-HMM, sticky HDP-HMM and the stateful HDP-HMM were 23, 18 and 20 respectively. The activity set, as labelled by the zoologists consisted of 14. That should not be taken as evidence that these models are converging to the right number. Critical analysis must still be maintained as there are many minor activities, which should be differentiated, such as 'trot' and 'walk' which, from a feature point of view are almost identical.

The unsupervised learning methodology demonstrated in this paper, holds promise when used in conjunction with supervised methods as no prior behavioural states need to be specified, thereby allowing for the recognition of less obvious or unknown behavioural states that may be missed through limited observation. Prior classification of states used in supervised learning may be subject to confirmation bias where an observer may oversimplify a chosen state based upon their expectations and thus exclude a separate, and perhaps more subtle, behavioural class [28]. Moreover, there are many ventures for further exploration from the modelling side, such as training the models in a semi-supervised fashion and then using those models, to segment other regions. We, furthermore, suggest that the methods demonstrated within can be particularly valuable for lion behavioural ecology as the last detailed activity budget for the species was compiled more than four decades ago by [23], where unobservable behaviours may not have been recognised.

Finally, we hope to have conveyed that PPS:es like Anglican hold promise to accelerate modelling innovation in scientific domains like zoology. In areas like these, programmer/scientist time often is a scarcer resource than computation time and the flexibility and ease of use of general purpose inference already weighs up against the downside of extra computation time. With computing power increasing, but our brain power remaining fixed, we expect this to become even more true in the future, particularly when more mature PPS:es are developed.

Acknowledgements

We would like to thank the shareholders of Bubye Valley Conservancy for allowing us to conduct field research on the conservancy. We would also like to thank the management staff of Bubye Valley Conservancy for assistance in the field. Collar design was supported by the John Fell Fund. N. D. is supported by the RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation). M.V. gratefully acknowledges support from the EPSRC and Clarendon Fund. We would like to honour Cecil the lion for his service to science.

References

[1] Bilmes, Jeff A. What HMMs can do. *IEICE TRANSACTIONS on Information and Systems*, 89(3):869–891, 2006.

[2] Brown, Danielle D, Kays, Roland, Wikelski, Martin, Wilson, Rory, and Klimley, A Peter. Observing the unwatchable through acceleration logging of animal behavior. *Animal Biotelemetry*, 1(1):1, 2013.

[3] Carroll, Gemma, Slip, David, Jonsen, Ian, and Harcourt, Rob. Supervised accelerometry analysis can identify prey capture by penguins at sea. *Journal of Experimental Biology*, 217(24):4295–4302, 2014.

[4] Dewar, Michael, Wiggins, Chris, and Wood, Frank. Inference in hidden markov models with explicit state duration distributions. *IEEE Signal Processing Letters*, 19(4): 235–238, 2012.

[5] Dhir, Neil, Perov, Yura, and Wood, Frank. Nonparametric bayesian models for unsupervised activity recognition and tracking. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pp. 4040–4045. IEEE, 2016.

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

[7] Fox, Emily B, Sudderth, Erik B, Jordan, Michael I, and Willsky, Alan S. An HDP-HMM for systems with state persistence. In *Proceedings of the 25th international conference on Machine learning*, pp. 312–319. ACM, 2008.

[8] Grünewälder, Steffen, Broekhuis, Femke, Macdonald, David Whyte, Wilson, Alan Martin, McNutt, John Weldon, Shawe-Taylor, John, and Hailes, Stephen. Movement activity based classification of animal behaviour with an application to data from cheetah (acinonyx jubatus). *PloS one*, 7(11):e49120, 2012.

[9] Gutzwiller, Kevin J, Wiedenmann, Richard T, Clements, Krista L, and Anderson, Stanley H. Effects of human intrusion on song occurrence and singing consistency in subalpine birds. *The Auk*, pp. 28–37, 1994.

[10] Huggins, Jonathan H and Wood, Frank. Infinite structured hidden semi-markov models. *arXiv preprint arXiv:1407.0044*, 2014.

[11] Johnson, Matthew J. and Willsky, Alan S. The hierarchical dirichlet process hidden semi-markov model. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI, pp. 252–259, 2010.

[12] Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] Leos-Barajas, Vianey, Photopoulou, Theoni, Langrock, Roland, Patterson, Toby A, Watanabe, Yuuki Y, Murgatroyd, Megan, and Papastamatiou, Yannis P. Analysis of animal accelerometer data using hidden markov models. *Methods in Ecology and Evolution*, 8(2):161–173, 2017.

[14] Lo, Albert Y et al. On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, 12(1):351–357, 1984.

[15] Lush, L, Ellwood, S, Markham, A, Ward, AI, and Wheeler, P. Use of tri-axial accelerometers to assess terrestrial mammal behaviour in the wild. *Journal of Zoology*, 2015.

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

[17] McClune, David W, Marks, Nikki J, Wilson, Rory P, Houghton, Jonathan DR, Montgomery, Ian W, McGowan, Natasha E, Gormley, Eamonn, and Scantlebury, Michael. Tri-axial accelerometers quantify behaviour in the eurasian badger (meles meles): towards an automated interpretation of field data. *Animal Biotelemetry*, 2(1):5, 2014.

[18] Murphy, Kevin P. Hidden semi-markov models HSMMs. *unpublished notes*, 2, 2002.

[19] Noonan, Michael J, Markham, Andrew, Newman, Chris, Trigoni, Niki, Buesching, Christina D, Ellwood, Stephen A, and Macdonald, David W. Climate and the individual: inter-annual variation in the autumnal activity of the european badger (meles meles). *PLoS One*, 9(1): e83156, 2014.

[20] Pagano, AM, Rode, KD, Cutting, A, Owen, MA, Jensen, S, Ware, JV, Robbins, CT, Durner, GM, Atwood, TC, Obbard, ME, et al. Using tri-axial accelerometers to identify wild polar bear behaviors. *Endangered Species Research*, 32:19–33, 2017.

[21] Phillips, Joe Scutt, Patterson, Toby A, Leroy, Bruno, Pilling, Graham M, and Nicol, Simon J. Objective classification of latent behavioral states in bio-logging data using multivariate-normal hidden markov models. *Ecological Applications*, 25(5):1244–1258, 2015.

[22] Rahman, Ashfaqur, Smith, Daniel, Hills, James, Bishop-Hurley, Greg, Henry, Dave, and Rawnsley, Richard. A comparison of autoencoder and statistical features for cattle behaviour classification. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 2954–2960. IEEE, 2016.

[23] Schaller, George B. *The Serengeti Lion: A Study of Predator-prey Relations. Drawings by Richard Keane*. University of Chicago Press, 1974.

[24] Teh, Yee Whye. Dirichlet process. In *Encyclopedia of machine learning*, pp. 280–287. Springer, 2011.

[25] Teh, Yee Whye and Jordan, Michael I. Hierarchical bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1, 2010.

[26] Teh, Yee Whye, Jordan, Michael I, Beal, Matthew J, and Blei, David M. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.

[27] Trethowan, Paul, Fuller, Andrea, Haw, Anna, Hart, Tom, Markham, Andrew, Loveridge, Andrew, Hetem, Robyn, Preez, Byron, and Macdonald, David W. Getting to the core: Internal body temperatures help reveal the ecological function and thermal implications of the lions mane. *Ecology and evolution*, 7(1):253–262, 2017.

[28] van Wilgenburg, Ellen and Elgar, Mark A. Confirmation bias in studies of nestmate recognition: a cautionary note for research into the behaviour of animals. *PloS one*, 8 (1):e53548, 2013.

[29] Wolpert, David H and Macready, William G. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82, 1997.

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

[31] Wood, Frank. Nips probabilistic programming tutorial 2015, December 2015.