“© 2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

ADON HDP-HMM: an Adaptive Online Model for Segmentation and Classification of Sequential Data

Ava Bargi, Richard Yi Da Xu, Massimo Piccardi
Faculty of Engineering and IT, University of Technology Sydney
PO Box 123 Broadway NSW 2007 Australia
Ava.Bargi,YiDa.Xu,Massimo.Piccardi@uts.edu.au

***Abstract*—Recent years have witnessed an increasing need for the automated classification of sequential data such as activities of daily living, social media interactions, financial series and others. With the continuous flow of new data, it is critical to classify the observations on-the-fly and without being limited by a pre-determined number of classes. In addition, a model should be able to update its parameters in response to a possible evolution in the distributions of the classes. This compelling problem, however, does not seem to have been adequately addressed in the literature since most studies focus on offline classification over pre-defined class sets. In this paper, we present a principled solution for this problem based on an adaptive online system leveraging Markov switching models and hierarchical Dirichlet process priors. This adaptive online approach is capable of classifying the sequential data over an unlimited number of classes, while meeting the memory and delay constraints typical of streaming contexts. In this paper, we introduce an adaptive "learning rate" that is responsible for balancing the extent to which the model retains its previous parameters or adapts to new observations. Experimental results on stationary and evolving synthetic data and two video datasets, TUM Assistive Kitchen and collated Weizmann, show remarkable performance in terms of segmentation and classification, particularly for sequences from evolutionary distributions and/or those containing previously unseen classes.**

I. INTRODUCTION AND RELATED WORK

Segmentation and classification of sequential data are important problems that have attracted significant research in a variety of domains. They provide the core technology for applications as diverse as speaker diarisation, financial market monitoring, activity understanding, multimedia annotation and human-computer interaction. Many approaches have been proposed to date including sliding windows [1], the hidden Markov model (HMM) [2], conditional random fields (CRFs) [3] [4], and structural SVMs [5], covering the range of generative, discriminative and maximum-margin learning of sequential classifiers. Together with advancements in learning and inference, increasingly realistic datasets have helped bridge the gap between lab and real applications [6] [7].

Nevertheless, the important challenge of model adaptation to class distributions that evolve both in parameters and number remains significantly unresolved. In this paper, we address this limitation by an adaptive, online model that can accommodate an unlimited (theoretically infinite) number of classes. The foundation of this model is the use of a Bayesian nonparametric process, the *hierarchical Dirichlet process* (HDP), as the prior for a hidden Markov model (a model known as the HDP-HMM [8] [9]), together with an adaptive learning rate that governs the model's adaptation. The proposed model provides an adaptive, online learning approach for joint segmentation and classification of sequential data with varying class sets. We refer to it as ADON HDP-HMM in the following.

The ADON HDP-HMM inherits properties of Bayesian nonparametric models such as parameter adaptation for the existing classes as well as instantiation of new unseen classes. In a previous paper, we have presented the online scheme of this model [10]: the input data are processed in batches, segmentation and recognition are performed on-the-fly, and learning continues throughout the entire life of the application. The model leverages a relatively short initial "bootstrap phase" of supervised training, and after that it adapts in a fully unsupervised manner. This makes the model suitable for a very large span of real-life problems, but renders adaptation challenging. To tackle the problem of model adaptation, in this paper we propose the use of an adaptive learning rate controlling the balance between the current parameters ("memory") and the new observations ("adaptability"). The learning rate is modelled as an additional random variable within the Bayesian framework and estimated along the other parameters at every batch. Experiments over a diverse set of datasets give evidence to the effectiveness of the learning rate adaptation, especially in the case of evolving distributions and unseen classes.

The rest of this paper is organised as follows: in the rest of this section we present the related literature and the scope of this study. In Section II we describe the hierarchical Dirichlet process and the HDP-HMM. Section III presents the proposed online approach, expanding on the adaptive learning rate. Through the experiments and discussions in Section IV, we evaluate and compare the proposed variants with existing benchmarks, and present the conclusion in Section V.

A. Related work

The inference techniques commonly referred to as Bayesian nonparametrics offer a principled way to classify sets of samples into arbitrary numbers of classes by using stochastic processes as priors. Amongst them, the hierarchical Dirichlet process (HDP) is used to estimate the class distribution of group-conditional data, typically by Gibbs sampling [11] or variational inference [12]. The HDP has been used for a variety of applications, including the modelling of sequential data by integrating HDP priors into the HMM. In the resulting HDP-HMM [8] [9], the classes correspond to the discrete

states of a Markov chain and the data are explained by a state-conditional observation model. Given a set of samples, classification is performed by decoding the states while allowing their number to dynamically grow or shrink. The hierarchical Dirichlet process has found increasing application in domains as varied as activity recognition, trajectory classification, speaker diarisation, statistical genetics and financial modelling (see [13] [14] [15] [16] [17] [18] [19] [20] for some recent references).

In the literature on sequential classification, most works adopt an *offline* approach where the entire data set is presented at once during the learning stage [6] [7]. However, this approach cannot be applied with data that are streamed in turn and calls for some form of online learning. The term *online* has been given a variety of meanings in different contexts. Our use here is as *sequential processing of temporal data in mini-batches*, inspired by recursive Bayesian estimation [21]. This interpretation is distinct from that of works where online refers to a closed dataset that is processed incrementally and repeatedly, such as in online Bayesian nonparametrics [22] [23], stochastic optimisation [24] [25] [26] and works on formal bounds [27] [28].

While almost all the existing approaches consider closed, pre-defined sets of classes, in scenarios like long-term learning the number of classes is not precisely predictable. Additionally, as more data stream in, the distributions of the known classes may change due to the observation of a more comprehensive sample or a natural evolution over time. Therefore, models are expected to be able to update the parameters of the known classes and/or add new classes once they appear. A recent study [29] has proposed a sequential Monte Carlo algorithm for online inference of the states and parameters of an HDP-HMM. However, as pointed out in [30], for the steps of resampling, the algorithm seems to assume the availability of all past data points, which would not be feasible in a streaming context.

Unsupervised adaptation can be very challenging in non-stationary domains, where adaptation and drift[1] are hardly distinguishable. To our knowledge, a frequent assumption in online studies is to have access to periodic or ad-hoc feedback from the user (active learning [31] [27] [25]). This feedback allows the model to evaluate the *regret* and redress possible drifts and misclassifications. However, such information is hard or costly to obtain in many real application domains. In the absence of expert feedback, we elaborate more on the *learning rate* as a dynamic lever for balancing adaptability (Section III-A). Most previous studies approach this problem by assigning fixed weights to the past learning and the new observations. However, in more general cases the learning rate should be adapted during the lifetime of the system. Some authors have proposed the use of an adaptive learning rate with exponential decay [32], and, more recently, regret-based adaptation of the learning rate (i.e., the step size of gradient descent) [24] [23] [25]. However, such adaptation strategies are only suitable for finite training sets. In our solution, we introduce an adaptive learning rate that constantly adjusts to the statistics of the streaming data, without revision or supervision. For stationary problems where the parameters change only slightly, the learning rate tends to rely more on the past parameters. Conversely, under evolving distributions, the learning rate tends to increase to allow for prompter parameter adaptation. Adding to the complexity, many real-life problems require a mixture of both, i.e. a continuous spectrum for the learning rate to follow the dynamics of the observations at each point in time. In this work, we tackle this problem by a posterior re-estimation of the learning rate at every data batch, performed separately for each parameter in the model to select the most appropriate, individual level of adaptation.

[1]Defined as an undesirable deviation from the ideal model.

II. The hierarchical Dirichlet process

A Dirichlet process, $DP(\gamma, H)$, is a stochastic process that can be thought of as a distribution over discrete distributions with countably infinite categories. It is fully specified by a scalar parameter, γ, known as the concentration parameter, and a base measure, H, over a measurable space θ. A sample from a Dirichlet process, G_0, is a distribution over θ which differs from zero at only a countably infinite number of locations or atoms, $\theta_k, k = 1 \dots K$:

$$\begin{aligned} &G_0 \sim DP(\gamma, H): \\ &G_0 = \sum_{k=1}^{K} \beta_k \delta(\theta - \theta_k), \quad K \to \infty \\ &\theta_k \sim H, \quad \beta \sim GEM(\gamma) \end{aligned} \tag{1}$$

The discrete set of locations is obtained by repeatedly sampling the base measure, H, and the weight for each location, $\beta_k, k = 1 \dots K$, is set by a *stick-breaking process*, noted as GEM(γ) (from Griffiths, Engen and McCloskey) [33]. We refer to the weight vector simply as β. In turn, a *hierarchical Dirichlet process* (HDP) consists of (at least) two layers of Dirichlet processes, which are obtained with a similar construction:

$$\begin{aligned} &G_j \sim HDP(\gamma, \alpha, H): \\ &G_0 \sim DP(\gamma, H) \\ &G_j = \sum_{k=1}^{K} \pi_{jk} \delta(\theta - \theta_k) \quad K \to \infty \\ &\theta_k \sim H, \quad \pi_j \sim DP(\alpha, \beta), \quad \beta \sim GEM(\gamma) \end{aligned} \tag{2}$$

where γ and α are the concentration parameters of the top-level and lower-level Dirichlet processes, respectively. Since G_0 is discrete, the various G_j's ($j = 1 \dots J$) are also discrete. The construction in Equation 2 is equivalent to a hierarchical sampling of the G_j's from G_0 (Figure 1). As proven in [8], it allows us to bypass the explicit resampling of G_0 and to directly reuse its set of locations, $\theta_k, k = 1...K$, as locations for the G_j's with appropriate weights π_{jk}.

The DP and HDP are typically used to generate priors for the parameters of a data likelihood, $f(y|\theta)$. Given the generative model of the HDP, the joint distribution of its data and parameters factorises as $f(y|\theta)G_j(\theta)$. Typically, multiple

G_j's are sampled from the HDP to model data that belong to different groups. Yet, given the hierarchical structure of the HDP, all the G_j's will usefully share distributional properties. Examples of grouped data can be as diverse as words in different books or genetic markers across different populations.

A. The HDP-HMM

The HDP has also been used as prior distribution for the parameters of switching models such as the hidden Markov model and switching auto-regressive models [8] [14]. When applied to a Markov chain, $z_{1:T}$, $p(z_{1:T}) = p(z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})$, the HDP changes its interpretation significantly (Figure 2). In this case, each $\pi_j = \{\pi_{jk}\}$, $k = 1 \dots K$, is used as one row of the Markov chain's transition matrix, representing the probability of transitioning from state j in the previous time-step to any other states in the current time-step, $p(z_t|z_{t-1} = j)$. Thanks to the properties of the HDP, new states are created when the data are not adequately explained by the current set of states. In contrast to the conventional HDP, the index of the group of each observation, j, is not known explicitly anymore, but is instead inferred in sequential order from the chain. Therefore, in the case of the HDP-HMM, $z_t \sim p(z_t|z_{t-1} = j) = \pi_j, \quad y_t \sim f(y_t|\theta_{z_t})$. As a consequence, in the HDP-HMM the number of groups (J) and the number of indices in each π_j (K) coincide. At their turn, the parameters for the emissions, $\theta_k, k = 1 \dots K$ are sampled from a Normal-Inverse-Wishart distribution, $NIW(\theta_k|\lambda)$.

In previous work, it has been reported that the HDP-HMM tends to over-segment the data sequence [34]. Fox *et al*. have proposed adding a *'sticky'* prior (κ) to the transition matrix to emulate inertia towards changing states by updating the transition distribution: $\pi_j \sim DP(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_j}{\alpha+\kappa})$. Thus, in the probability of transition from state j, a κ amount is added to the j^{th} component to reinforce staying in the same state [35]. We utilise the *sticky* prior in this study (see Figure 2), yet still refer to the model as HDP-HMM for brevity.

B. Inference and Learning

In the HDP-HMM, inference and learning are typically performed simultaneously by estimating the joint posterior distribution of the states, parameters and hyper-parameters conditioned on the observations. Deriving such an extensive joint posterior analytically is intractable, hence it is mainly inferred using Gibbs sampling or variational inference. Gibbs sampling is a simple yet effective method that can estimate complex posteriors with significant accuracy, yet it may converge slowly or be trapped in a poor local minimum (*poor mixing*). Variational inference is usually faster to compute, however it requires prior derivation of analytical approximations and the accuracy can suffer due to the approximations. In this paper, we adopt Gibbs sampling following [8] [14] and we show that a relatively short, initial supervised training leads to rapid convergence and accurate distributions.

After inferring the class labels, $z_{1:T}$, we establish the correspondence between predicted classes and ground-truth classes by assigning each predicted class to the closest ground-truth class in terms of the emission probabilities' means. For non-stationary ground-truth distributions, we use the means current at time t.

III. The Adaptive Online HDP-HMM

The proposed AdOn HDP-HMM uses an initial phase of T_b frames of supervised learning (*bootstrap*) for initialisation, and then starts operating as unsupervised, adaptive online inference (Figure 3). While an extensive supervised phase is desirable, its extent will depend on the annotation costs of the specific domain. During the bootstrap, class variables $z_{1:T_b}$ are fixed to their ground-truth values, and the model's parameters are sampled for a given number of iterations. After conclusion of the bootstrap phase, the data are processed in successive batches, and the posterior probabilities of both the class variables and the parameters are estimated iteratively over each batch.

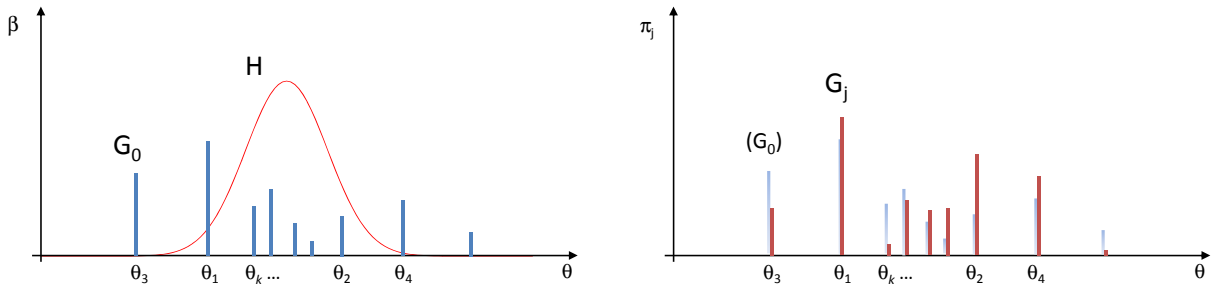
Considering a generic stream of data, $y_{1:t}$, the posterior probability of the parameters can be written as $p(\phi|y_{1:t}) \propto f(y_{1:t}|\phi)\ p(\phi)$, where ϕ indicates the parameter vector of the HDP-HMM in Figure 2. The parameter vector breaks down as $\phi = \{\theta, \pi, \beta\}$ where θ are the parameters of the emission densities, π are the transition probabilities (and weights of the G_j's), and β are the weights of G_0. Further, since we assume normal densities, we have $\theta = \{\mu, \Sigma\}$, with μ and Σ the usual mean and covariance parameters. The online version leverages posterior adaptation, using the posterior computed up to time t, as the prior for the next batch of data, $y_{t+1:t+\Delta t}$:

$$\begin{aligned} p(\phi_{n+1}|y_{1:t+\Delta t}) &\propto f(y_{t+1:t+\Delta t}|\phi_n, y_{1:t})\ p(\phi_n|y_{1:t}) \\ &\approx f(y_{t+1:t+\Delta t}|\phi_n)\ q(\phi_n|\Lambda) \end{aligned} \tag{3}$$

where Λ denotes the hyper-parameters of ϕ including $\gamma, \kappa, \alpha, \lambda$ and n is the batch number (Figure 4). Given that the updated posterior embeds the distributional properties of the observations up to the current time, observations $y_{1:t}$ in Equation 3 can be discarded after adaptation. It implies that the accumulated sufficient statistics of previous data are propagated parametrically as $q(\phi_n|\Lambda)$, rather than through the previous data samples i.e. $p(\phi_n|y_{1:t})$, allowing the model to operate from a limited observation buffer. The nonparametric nature of the model is related to the inference method of the current data batch. While batch processing may come at a price of reduced accuracy, it is the only viable approach for unbounded streaming data.

A. Adaptive learning rate

In the proposed adaptive system, a variable learning rate is applied over the prior and noted as τ in the following. In each batch, τ is responsible for setting the weight of the prior distributions over the model's parameters (θ, π, β). In other words, our target is to balance the impact of the current observations with the learning accumulated along the previous batches. This can increase or weaken the posterior learning 'inertia' in 'adapting' to the current data (likelihood term), as opposed to retaining 'memory' (prior term):



(a) Dirichlet process

(b) Hierarchical Dirichlet process

Fig. 1. The Dirichlet process (a) and the hierarchical Dirichlet process (b) constructions. The parameter space has been kept to one dimension for ease of visualisation.

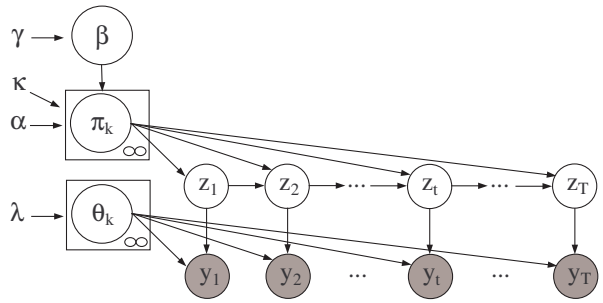


Fig. 2. The HDP-HMM graphical model. The box notation is used to show replication.

$$p(\phi|y,\tau) \propto p(y|\phi)p(\phi)^{\tau} \tag{4}$$

It is important to note that the length of the current batch compared to the number of past samples plays a role in their relative influence on the posterior parameters (see Appendix A for more details). Accordingly, τ can be articulated as a scaling factor to the number of 'pseudo-observations' in the prior to balance with the respective number for the current batch[2].

For prior distributions belonging to the exponential family, it is easy to integrate the learning rate in the posteriors thanks to the properties of canonical parameters. Accordingly, we use exponential family likelihoods and priors for easier integration of the learning rate into the model. Here, we focus on the prior in Equation 4 and its hyperparameters, translating them into exponential family notations. The standard parameters, ϕ, are converted into the corresponding canonical parameters, Θ, and we make their dependence on hyper-parameters, η, explicit:

$$\begin{aligned} p(\Theta|\eta)^{\tau} = p(\Theta|\eta,\tau) &= f(\eta)^{\tau} g(\Theta)^{\tau} \exp\left(\Theta^T \eta\right)^{\tau} \\ &= f'(\eta) \exp\left(\tau \Theta'^T \eta'\right), \\ \Theta' = [\ln(g(\Theta)); \Theta], \quad & \eta' = [1; \eta] \end{aligned} \tag{5}$$

[2]For convenience, in this paper we have constrained all batches to be of the same length and explored the variable-length alternative in [10].

Adding the learning rate, τ, as an exponent to this prior does not alter the type of distribution for members of the exponential family. Rather, it updates the canonical parameters of the prior, ultimately affecting its weight in the resulting posterior. Please note that we only need to derive a proportional posterior for sampling purposes. Hence, the τ exponent on any term independent from Θ (such as $f(\eta)$) can be ignored thanks to the proportionality. The normalisation coefficient $g(\Theta)^{\tau}$ can be merged into the sufficient statistics, assuring that its τ exponent is absorbed into the scaled canonical parameter ($\tau\eta$). [3]

In general terms, the posterior distribution of τ given Θ in the presence of N data samples in $\mathbf{Y}$ can be inferred as follows:

$$p(\tau|\Theta,\mathbf{Y},\eta) \propto p(\Theta|\tau,\mathbf{Y},\eta)p(\tau) \tag{6}$$

In our case, Θ represents the parameters of the HDP-HMM (μ, Σ, β and π) and their distributions are a Normal-Inverse-Wishart distribution over μ and Σ, and Dirichlet distributions over π and β. Given that both the NIW distribution and the Dirichlet distribution are members of the exponential family, Equation 7 shows a unified way of inferring the posterior parameters in canonical form [36, pp. 116–117]:

$$\begin{aligned} p(\Theta|\mathbf{Y},\tau^*,\eta^*) &\propto p(\mathbf{Y}|\Theta,\tau,\eta)p(\Theta|\eta,\tau) \\ p(\Theta|\mathbf{Y},\tau^*,\eta^*) &\propto \left[\left(\prod_{n=1}^{N} h(y_n) g(\Theta)\right) \exp\left(\Theta^T \sum_{n=1}^{N} u(y_n)\right)\right] \\ &\quad \left[f(\eta,\tau) g(\Theta)^{\tau} \exp\left(\tau \Theta^T \eta\right)\right] \end{aligned} \tag{7}$$

removing the constants with respect to Θ:

$$p(\Theta|\mathbf{Y},\tau^*,\eta^*) \propto g(\Theta)^{\tau+N} \exp\left(\Theta^T \left(\sum_{n=1}^{N} u(y_n) + \tau\eta\right)\right)$$

$$\tau^* = \tau + N, \qquad \eta^* = \sum_{n=1}^{N} u(y_n) + \tau\eta$$

[3]As in: $g(\Theta)^{\tau} \exp(\tau \Theta^T \eta) = \exp\left(\tau \ln(g(\Theta)) + \tau \Theta^T \eta\right)) = \exp\left(\tau \left[\ln(g(\Theta)); \Theta\right]^T [1; \eta]\right)$

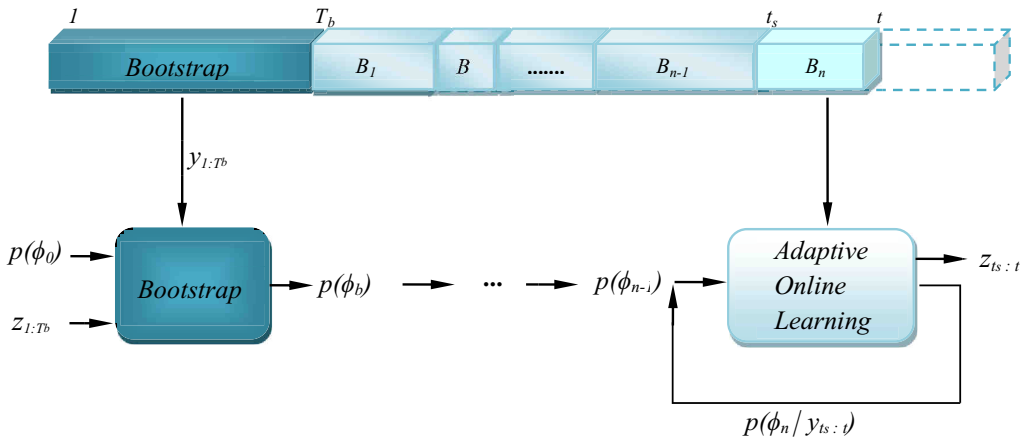


Fig. 3. The adaptive online learning flowchart: initialised by a supervised bootstrap, learning continues unsupervised over streaming data split into batches. The figure shows a general case with batches of variable size. For simplicity, in this paper we assume all batches to have the same size (we have explored the variable alternative in [10]). The posterior over ϕ in each batch is passed to the next batch as the prior distribution.

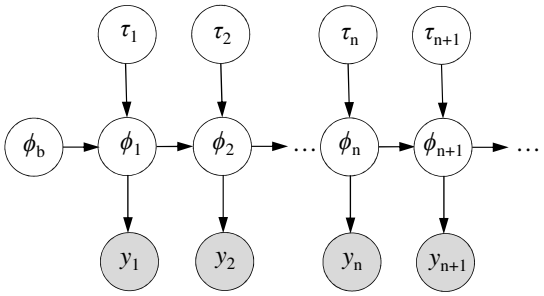


Fig. 4. Graphical model for the proposed adaptive online model. ϕ can be any of the parameters in Figure 2 (θ, π, β); τ is the respective learning rate (a positive, continuous random variable) and n represents the batch number.

B. Derivation of prior distributions and learning rate adaptation

In the previous sections, we have made repeated reference to a generic parameter vector (Θ) and learning rate (τ). The derivation of the posterior distributions of all parameters under the influence of their learning rate is obtained by Equation 7. In the following subsections, we derive the key steps for the inference: i) the specific prior distribution of each parameter (Σ, μ, β and π) under its own learning rate, and the ii) posterior distributions of their respective learning rates ($\tau_\Sigma, \tau_\mu, \tau_\beta, \tau_\pi$). The derivation of the posteriors for the τ's uses, where possible, the distributions over the parameters (Inverse-Wishart, normal, and Dirichlet distributions) as likelihoods, and a conjugate prior. For the case of τ_Σ, where conjugacy is not available, we have adopted a heuristic based on a simplifying assumption.

1) Inference of the covariance matrix, Σ: We infer μ and Σ in the Normal-Inverse-Wishart prior by first sampling Σ using an Inverse-Wishart (IW) distribution, and then using Σ to sample μ from a Normal distribution [37].

$$\begin{aligned} &p(\mu,\Sigma)^\tau = p(\mu|\Sigma)^\tau p(\Sigma)^\tau : \\ &p(\Sigma) = IW(\Sigma|\Psi,\nu), \quad p(\mu|\Sigma) = \mathcal{N}(\mu|\mu_0,\Sigma) \end{aligned} \tag{8}$$

As mentioned earlier, the addition of a positive learning rate as exponent on the IW prior does not alter the type of distribution and can be merged into the hyper-parameters. Below, we convert the hyper-parameters $\phi_{IW} = \{\Psi, \nu\}$ into their natural form (η) to show the impact of τ more clearly. Eventually, they are converted back to standard form (ϕ) to show the linear transformation caused by the learning rate; p denotes the number of dimensions.

$$\begin{aligned} &\phi_{IW} = (\Psi,\nu) \rightarrow \eta_{IW} = \left(-\frac{1}{2}\Psi, -\frac{\nu+p+1}{2}\right), \\ &\eta'_{IW} = \tau_\Sigma \eta_{IW} = \left(-\frac{\tau_\Sigma}{2}\Psi, -\frac{\tau_\Sigma(\nu+p+1)}{2}\right) \\ &\rightarrow \phi'_{IW} = (\tau_\Sigma\Psi, \tau_\Sigma(\nu+p+1)-p-1) \end{aligned} \tag{9}$$

Inference of τ_Σ

To derive a posterior for τ_Σ, we would ideally like to exploit a conjugate prior and analytically obtain the posterior's parameters from those of the prior and the sufficient statistics of the current data. For clarity, in the following we refer to the parameters of these distributions as "hyper-parameters" since they are hyper-parameters of the HDP-HMM. A candidate conjugate prior for the IW distribution is the Gamma distribution. However, the Inverse-Wishart is only conjugate to the Gamma as the prior for the scale parameter (or a scaling coefficient for the scale parameter, Ψ, in the multivariate cases). Hence, a Gamma distribution cannot be used as a conjugate prior for deriving the posterior of τ_Σ.

Therefore, we propose a heuristic procedure to derive the posterior hyper-parameters for τ_Σ. The posterior for τ_Σ is

modelled using an Inverse-Gamma (IG) distribution, the univariate correspondent of the Inverse-Wishart. The samples of IG are positive real values, suitable for the scalar learning rate τ_Σ. The distributions are displayed below, introducing a univariate version of IW in the second line, where matrix symbols are replaced with scalar versions: Σ as σ, Ψ as ψ. This intermediate representation helps show the compatibility of parameters in the IW and IG distributions.

$$\begin{aligned} IW(\Sigma|\Psi,\nu) &= \frac{|\Psi|^{\frac{\nu}{2}}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}|\Sigma|^{-\frac{\nu+p+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}(\Psi\Sigma^{-1})\right) \\ IW(\sigma|\psi,\nu) &= \frac{\psi^{\frac{\nu}{2}}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}\sigma^{-\frac{\nu+2}{2}}\exp\left(-\frac{\psi}{2\sigma}\right) \quad \textit{where } p=1 \\ IG(\tau_\Sigma|\beta,\alpha) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\tau_\Sigma^{-\alpha-1}\exp\left(-\frac{\beta}{\tau_\Sigma}\right) \end{aligned} \tag{10}$$

A simple way to derive this heuristic is to restrict variable Σ in the first line of Equation 10 to be spherical (let us say, with diagonal value k). Under this assumption, function $\operatorname{tr}(\Psi\Sigma^{-1})$ becomes equal to $1/k\ \operatorname{tr}(\Psi)$. By well-known properties, at its turn $\operatorname{tr}(\Psi)$ is equal to the sum of the eigenvalues of Ψ. This justifies its use as the parameter for a univariate variance. In our experimental results, we have found that the largest eigenvalue provides a better heuristic than the sum of the eigenvalues, most likely because of the noise over smaller eigenvalues. According to the above discussion and comparing the univariate IW and IG distributions in Equation 10, we derive the posterior hyper-parameters as:

$$\begin{aligned} &IG(\tau_\Sigma|\beta^*,\alpha^*) \leftarrow IW(\Sigma|\Psi,\nu) \\ &\beta^* = \frac{f(\Psi)}{2}, \alpha^* = \frac{\nu}{2} \end{aligned} \tag{11}$$

where $f(\Psi)$ is the largest eigenvalue of Ψ.

So far, we have established a way to infer τ_Σ from the parameters of a single IW distribution. However, the HDP-HMM has multiple classes and we wish to merge the influence of all the class distributions into a tied τ_Σ value. This is done by a weighted average of $f(\Psi_k), k = 1 \dots K$, where we use the *degrees of freedom* parameter (ν) of the IW distributions as weights:

$$\begin{aligned} &IG(\tau_\Sigma|\beta^*,\alpha^*) \leftarrow IW(\Sigma|\Psi,\nu) \\ &\alpha^* = \frac{\sum_{k=1}^K \nu_k}{2K}, \ \ \beta^* = \frac{1}{2}\frac{\sum_{k=1}^K \nu_k \ \max(\operatorname{eig}(\Psi_k))}{\sum_{k=1}^K \nu_k} \end{aligned} \tag{12}$$

2) Inference of the mean, μ: Having inferred Σ, the next step is to derive the multivariate mean, μ, in the NIW prior. Let us consider a generic multivariate Normal distribution $N = (\mu|\mu_0,\Sigma)^T$ with known covariance. To observe the impact of the learning rate, we convert its parameters $\phi_\mu = (\mu_0,\Sigma)$ into the natural form and multiply them by the learning rate τ_μ, and ultimately revert them back to the standard form:

$$\begin{aligned} &\phi_\mu = (\mu_0,\Sigma) \rightarrow \eta_\mu = \left(\Sigma^{-1}\mu_0, \frac{1}{2}\Sigma^{-1}\right)^T \\ &\eta'_\mu = \tau_\mu\eta_\mu = \left(\tau_\mu\Sigma^{-1}\mu_0, \frac{\tau_\mu}{2}\Sigma^{-1}\right)^T \\ &\rightarrow \phi'_\mu = \left(\mu_0, \frac{1}{\tau_\mu}\Sigma\right), \quad N(\mu|\eta'_\mu) = \mathcal{N}(\mu|\mu_0, \frac{1}{\tau_\mu}\Sigma) \end{aligned} \tag{13}$$

Inference of τ_μ

Posterior sampling of τ_μ is conducted with a similar approach to τ_Σ, but it enjoys prior conjucacy. The distribution of the means for the K classes are assumed multivariate normal distributions of parameters μ_{0k} and Σ_k, added with scaling parameter τ_μ at the denominator of the covariance. Let us now draw a sample, $\mu_k, k = 1 \dots K$, from each of the distributions and place a Gamma prior over τ_μ. In Appendix B, we prove that this prior is conjugate and we derive the (hyper-)parameters for the posterior over τ_μ. In addition, we again choose to weigh these K samples as a single sample by using the ν_k parameters derived in Equations 8-9 as weights. The resulting values are reported in Equation 14 below.

$$\begin{aligned} &G(\tau_\mu|\alpha^*,\beta^*): \\ &\alpha^* = \alpha + 1/2, \\ &\beta^* = \beta + \frac{1}{2}\frac{\sum_{k=1}^K \nu_k(\mu_k-\mu_{0k})^T\Sigma_k^{-1}(\mu_k-\mu_{0k})}{\sum_{k=1}^K \nu_k} \end{aligned} \tag{14}$$

3) Inference of the HDP transition parameters, β and π: Thus far, we have presented the prior distributions and learning rate for the emission parameters. The other main set of parameters in our ADON HDP-HMM are the HDP's β and π parameters that jointly and hierarchically cater for the transition probabilities. The distributions of these parameters are shown in Equation 15, where n_{jk} and $m_k, k = 1 \dots K$, are the sufficient statistics of the HDP-HMM. The former represents the number of transitions from class j to k in the state sequence, while the latter is an intermediate variable that represents "clusters" of transitions to class k and is used for sampling the top level of the HDP hierarchy. A full derivation can be found in [8] and we omit it here for brevity.

$$\begin{aligned} &\beta \sim Dir(\gamma/K + m_1, \dots, \gamma/K + m_K) \\ &\pi_j \sim Dir(\alpha_1\beta_1 + n_{j1}, \dots, \alpha_j\beta_j + \kappa + n_{jj}, \dots, \alpha_K\beta_K + n_{jK}) \end{aligned} \tag{15}$$

In Equation 16 we illustrate the impact of the learning rates, τ_β and τ_π, on the parameters of the Dirichlet distributions in Equation 15. Resembling the previous parameters, we first convert them from standard to canonical form [38], multiply them by τ and convert them back to standard form:

Posterior for β :

$$\phi_\beta = (\gamma/K + m_1, \ldots, \gamma/K + m_K)$$
$$\rightarrow \eta_\beta = (\gamma/K + m_1 - 1, \ldots, \gamma/K + m_K - 1)$$
$$\eta'_\beta = \tau_\beta \eta_\beta = (\tau_\beta(\gamma/K + m_1) - \tau_\beta, \ldots, \tau_\beta(\gamma/K + m_K) - \tau_\beta)$$

$$\beta \sim Dir(\tau_\beta(\gamma/K + m_1 - 1) + 1, \ldots, \tau_\beta(\gamma/K + m_K - 1) + 1)$$

Posterior for π :

$$\phi_\pi = (\alpha_1\beta_1 + n_{j1}, \ldots, \alpha_j\beta_j + \kappa + n_{jj}, \ldots, \alpha_K\beta_K + n_{jK}))$$
$$\rightarrow \eta'_\pi = (\tau_\pi(\alpha_1\beta_1 + n_{j1}) - \tau_\pi, \ldots, \tau_\pi(\alpha_j\beta_j + \kappa + n_{jj}) - \tau_\pi$$
$$, \ldots, \tau_\pi(\alpha_K\beta_K + n_{jK}) - \tau_\pi)$$

$$\pi_j \sim Dir(\tau_\pi(\alpha_1\beta_1 + n_{j1} - 1) + 1, \ldots, \tau_\pi(\alpha_j\beta_j + \kappa + n_{jj} - 1) + 1$$
$$, \ldots, \tau_\pi(\alpha_K\beta_K + n_{jK} - 1) + 1)) \tag{16}$$

Inference of τ_β and τ_π

To the best of our knowledge, there are no conjugate priors over the scaling factor τ in the Dirichlet distributions in Equation 16 due to the presence of the offset ("+1"). Hence, we estimate the next batch's learning rate using a Metropolis-Hastings (MH) jump. This approach is used in several other studies (e.g., [39] [40]) and is a valid MCMC move. For the MH step, one can choose a suitable proposal distribution (Q) and its samples are accepted with probability of acceptance $p(\xi \rightarrow \xi^*) \propto \min(1, p(\xi^*)Q(\xi \rightarrow \xi^*)/p(\xi)Q(\xi^* \rightarrow \xi))$.

To sample τ_β, we have selected the prior over the learning rate, $G(\tau_\beta|\alpha, \beta)$, as the proposal distribution. The new sample (τ^*_β) is accepted with the probability in Equation 17, updating τ_β for the current batch with the accepted sample. An identical approach can be taken for τ_π by replacing τ_π for τ_β in Equation 17.

$$p(\tau_\beta \rightarrow \tau^*_\beta) \propto \min\left(1, \frac{p(\tau^*_\beta|\alpha,\beta)Q(\tau_\beta \rightarrow \tau^*_\beta)}{p(\tau_\beta|\alpha,\beta)Q(\tau^*_\beta \rightarrow \tau_\beta)}\right)$$
$$\frac{p(\tau^*_\beta|\alpha,\beta)Q(\tau_\beta \rightarrow \tau^*_\beta)}{p(\tau_\beta|\alpha,\beta)Q(\tau^*_\beta \rightarrow \tau_\beta)} \propto \frac{Dir(\beta|\alpha,\tau^*_\beta)G(\tau^*_\beta)G(\tau_\beta)}{Dir(\beta|\alpha,\tau_\beta)G(\tau_\beta)G(\tau^*_\beta)} \tag{17}$$
$$= \frac{Dir(\beta|\alpha,\tau^*_\beta)}{Dir(\beta|\alpha,\tau_\beta)}$$

Algorithm 1 summarises the main steps of the inference for the proposed model. The final inference of the class labels, $z_1 \ldots z_T$, is performed as $\text{argmax}\, p(z_t|z_{t-1}, \pi_{t-1})$ since it empirically improved accuracy. In writing the algorithm, the length of all batches are assumed to be equal to T merely for notational simplicity, i.e. the ADON HDP-HMM can work with variable batch lengths as well.

Algorithm 1: ADON HDP-HMM: main inference steps.

Input: HDP-HMM hyperparameters (Table 2), observation batches Y_0 (bootstrap), $Y_1(y_{1,1} \ldots y_{1,T}) \ldots Y_N(y_{N,1} \ldots y_{N,T})$
Output: Frame label batches $Z_1(z_{1,1} \ldots z_{1,T}) \ldots Z_N(z_{N,1} \ldots z_{N,T})$
Initialise: All variables $(\mu, \Sigma, \tau_{(\mu,\Sigma,\pi,\beta)}, \pi, \beta, z_1 \ldots z_T)$ by their priors and input
for *batch* = 0 *to* N **do**
  for *iteration* = 1 *to* 1000 **do**
    Step 1: Sample all variables $(\mu, \Sigma, \tau_{\mu,\Sigma,\pi,\beta}, \pi, \beta, z_1 \ldots z_T)$ from their posterior distributions, given Y_{batch} and their prior distributions at $batch - 1$ (Eq. 7 and Section III-B)
    Step 2: Select $z_t, t = 1 \ldots T$ as $\text{argmax}\, p(z_t|z_{t-1}, \pi_{t-1})$
  end
  $Z_{batch} = z_1 \ldots z_T$
end
return $Z_1 \ldots Z_N$

C. Discussion on the learning rates

In the above sections, the learning rates for each parameter are inferred separately to allow their independent adaptation to the changes in the underlying distributions. The empirical results in Section IV confirm the validity of this choice. The impact of the learning rate on the prior distribution of the mean can also be explained in intuitive terms: as shown in Equation 13, the learning rate does not change the mean of this distribution (μ_0), but inversely impacts its covariance. Accordingly, for all cases when $0 \leq \tau < 1$, the prior distribution has a larger covariance and will allow the mean to drift more. Conversely, for $\tau > 1$, the covariance is tighter and the mean will follow the prior mean more closely. Appendix C discusses the similar impact of the learning rate on the prior distribution of the covariance. In the following experiments, the adaptation of τ with respect to the data is explored extensively.

IV. EXPERIMENTS

The experiments aim to explore the effectiveness of the proposed ADON HDP-HMM for time segmentation and classification of sequential data in a variety of scenarios. To closely examine the adaptability of the model, we have designed several synthetic datasets with stationary and evolutionary distributions. This also allows us to investigate the effects of using adaptive learning rates for enhancing adaptability. Following with two video datasets, we assess the performance of the proposed model on various challenging sequences with noisy data, abrupt changes and new classes in the test data. It is important to notice that it is not easy to compare the degree of challenge of the synthetic experiments with those on the video data, due to differences in the nature of the signals, noise and, most importantly, degree of evolution that is stronger by design in the synthetic data. Hence, both categories of experiments can shed more light on the adaptability of the ADON HDP-HMM in various contexts.

To evaluate the results more comprehensively, metrics for both classification and time segmentation performance are introduced. For classification accuracy, we have used the frame-level comparison of the decoded classes with the ground

truth. To evaluate time segmentation, the metrics of precision and recall are used to quantify the accuracy at detecting boundaries between segments: recall denotes what percentage of real segment boundaries is detected by the model, and precision represents the ratio of the estimated boundaries that are compatible with the ground-truth boundaries. A true boundary is regarded as correctly detected if a change of state is decoded within an interval of $\pm\Delta t$ frames from the ground-truth location, where Δt is set to 10 percent of the average segment length. Any additional detected boundaries are counted as false positives. We also report the difference between the overall number of actions detected in the test sequence and the number of actions in the ground truth (noted as *cardinality*, with an ideal value of zero).

The empirical results are quantitatively reported in tables, also visualised in colour plots of ground-truth vs. estimated labels. In each illustration (for instance, Figure 6), the horizontal axis is the time and the estimated labels are plotted on top of the true labels, providing a qualitative appreciation of the segmentation and classification performance. These plots are best viewed in colour.

A. Synthetic data

The basic framework of the synthetic dataset is generated from an HMM with 5 states and univariate normal emissions with means ($\mu = [100, 200, 300, 400, 500]$) and unit variance, and a Dirichlet-distributed transition matrix ($\alpha = [3, 3, 3, 3, 3]$). This generative model is similar to the AdOn HDP-HMM, but not an exact replicate, due to the absence of the HDP prior and adaptation of τ in the generative process.

1) Stationary distributions: Given the above base configuration, the stationary experiments are run over 3 sequences of length = 100. The model is trained using leave-one-out cross-validation, i.e. training with two sequences and testing on the third one. Hence, the distributions of training and test samples are the same. The test sequence is split into six batches with an approximate size of 16 time units. To provide adaptation, the inferred parameters of each batch are propagated into the next batch as priors.

The proposed AdOn HDP-HMM is able to recognise and segment this basic version with 100 percent accuracy, with or without adaptation of the learning rate. To probe the model further, we add a significant noise to the above model by increasing the standard deviation to 50, thereby causing over 30% overlap between the distributions of any two adjacent states (Figure 5). Despite this substantial noise, the model with the adaptive learning rate retains an average of 76.3% frame-level accuracy (Table I). Repeating this experiment with fixed learning rate ($\tau = 1$) shows a noticeable decrease in accuracy of 3 percentage points and undesirable extra states. The first two rows of Table I show the detailed accuracy figures in terms of precision, recall and cardinality.

2) Evolutionary distributions: A more advanced experiment is performed by training the model on synthetic data with evolving distributions, either forcing gradual shifts to the means of each class or including new unseen classes. The model is trained on a sequence of length = 200 with stationary parameters and tested on another sequence of length = 100 with an evolving distribution, split into six batches as above. The standard deviation of the generated sequences in these experiments is set to $\sigma = 10$, an intermediate value between the $\sigma = 1$ and $\sigma = 50$ used for the previous stationary scenarios.

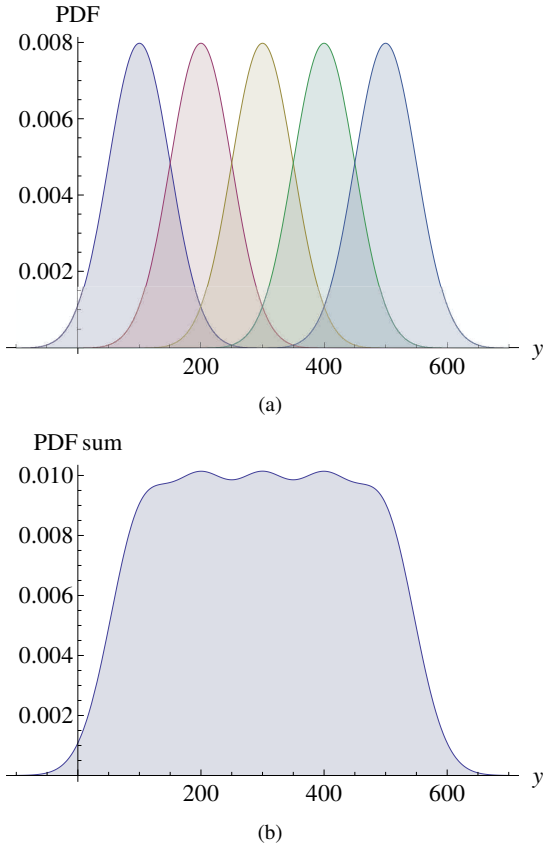


Fig. 5. (a) The distribution of the state emission densities in the noisy synthetic data set. Note that due to the large standard deviation ($\sigma = 50$) there is significant overlap between adjacent states, making classification challenging; (b) The sum of the density functions of all states.

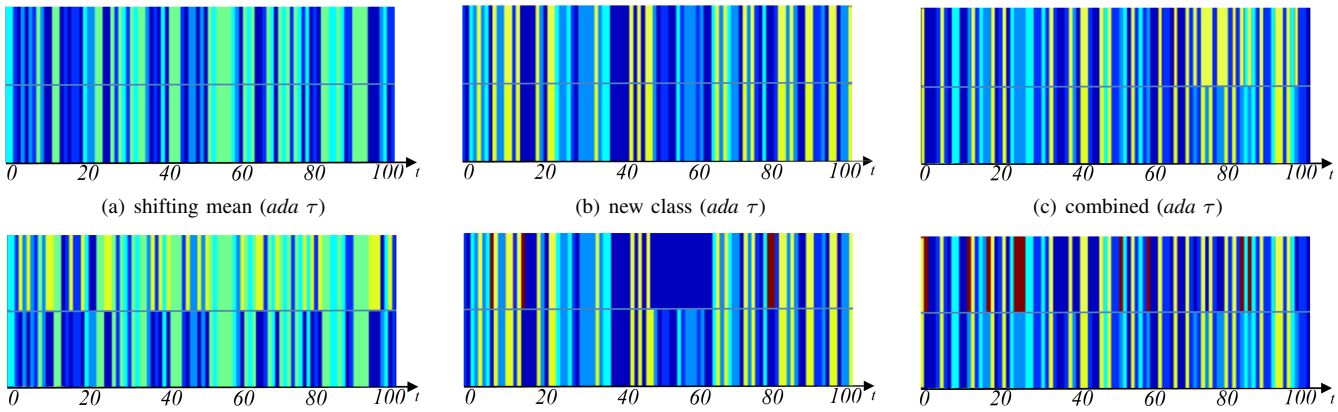
Shifting class means: To examine the adaptability of the model, we drift the class means by $\delta = 0.5$ at each time step. Therefore, an instance appearing at $t = 10$ in the test sequence is generated from a distribution with its mean shifted by 5 units. For a non-adaptive model and given the synthetic generation scheme, such data can cause significant classification errors after a few tens of time units. However, the results of the AdOn HDP-HMM with the adaptive learning rate demonstrate smooth adaptation and excellent accuracy over the evolving sequence (Figure 6). There are a few misclassifications towards the end of the sequence which are due to the heavy distributional drift. Conversely, the results with fixed learning rates show a significant drop of 26 percentage points in accuracy and one spurious new class (see Table I, third and fourth rows).

New classes: In this experiment, distributions do not shift, yet one new class appears around $\mu = 600$ with the same σ as the other classes. The model is able to create a new state (shown with a random new colour in Figure 6), and

TABLE I

FRAME-LEVEL ACCURACY, SEGMENTATION RECALL, PRECISION AND CARDINALITY ERROR FOR THE SYNTHETIC EXPERIMENTS. EACH TABLE SECTION INCLUDES THE RESPECTIVE RESULTS OF EXPERIMENT SECTIONS, COMPARING THE PERFORMANCE WITH ADAPTIVE AND FIXED LEARNING RATES: I) STATIONARY (*Sta*) WITH HIGH NOISE REPORTING AVERAGE RESULTS ON 3 SEQUENCES, II) EVOLUTIONARY (*Evo*) WITH SHIFTING MEANS, III) WITH NEW STATES, IV) WITH COMBINED SHIFTING AND NEW STATES.

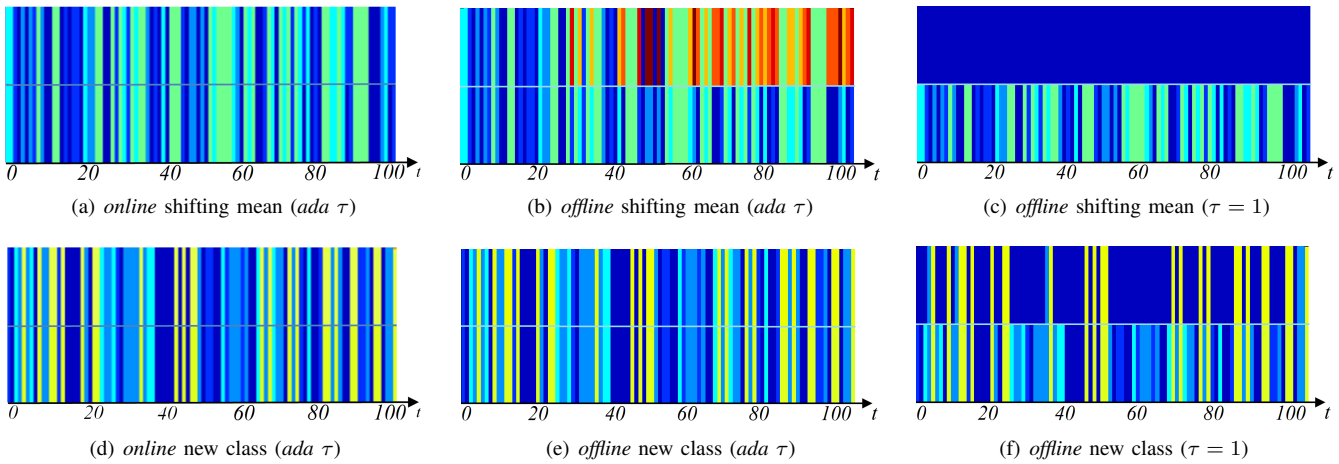
	Accuracy	**Recall**	**Precision**	**Cardinality**
Sta, Noisy (*ada* τ)	**0.76 ± 0.01**	**0.92 ± 0.02**	0.92 ± 0.02	**0.33**
Sta, Noisy ($\tau = 1$)	0.74 ± 0.04	0.89 ± 0.03	**0.93** ± 0.02	1.7
Evo, shifting mean (*ada* τ)	**0.97**	0.97	0.99	**0**
Evo, shifting mean ($\tau = 1$)	0.71	**0.99**	**1**	1
Evo, new class (*ada* τ)	**1.00**	**1.00**	**1.00**	**0**
Evo, new class ($\tau = 1$)	0.86	0.86	0.98	0
Evo, combined (*ada* τ)	**0.93**	**1.00**	**1.00**	**1**
Evo, combined ($\tau = 1$)	0.81	0.95	0.97	2



(a) shifting mean (*ada* τ) (b) new class (*ada* τ) (c) combined (*ada* τ)

(d) shifting mean ($\tau = 1$) (e) new class ($\tau = 1$) (f) combined ($\tau = 1$)

Fig. 6. Segmentation and classification results for evolutionary synthetic data, using fixed ($\tau = 1$) and adaptive (*ada* τ) learning rates. Top half of the stripes: *predicted states*; bottom half: *ground truth*. The horizontal axis is the time which is in turn proportional to deviations from the original means, $\delta = 0.5t$. This figure should be viewed in colour. (a,d) Shifting means: without the adaptation effects of the learning rate, shifting means can be misclassified as new classes (yellow) in d. (b,e) New class: the new class shown in yellow in the ground truth, is recognised and learned in both cases. (c,f) Combined: adding both challenges causes a slight decrease in accuracy and an extra new class. Nevertheless, in most cases the performance improvement with the adaptive τ is still visible.



(a) *online* shifting mean (*ada* τ) (b) *offline* shifting mean (*ada* τ) (c) *offline* shifting mean ($\tau = 1$)

(d) *online* new class (*ada* τ) (e) *offline* new class (*ada* τ) (f) *offline* new class ($\tau = 1$)

Fig. 7. Segmentation and classification results for evolutionary synthetic data, comparing offline and online runs of the model using fixed ($\tau = 1$) and adaptive (*ada* τ) learning rates. This figure should be viewed in colour. (a,b,c) Compare shifting means in offline and online settings: the best results are obtained with the online model with adaptive τ. Due to the evolution of the true parameters, the offline model only performs well initially. After δ increases beyond a certain extent, it starts to confuse the instances of the drifted classes as new classes. (d,e,f) New class: the new class shown in yellow in the ground truth, is recognised and learned in all cases. The existing classes are also correctly recognised in both offline and online cases, as the parameters are static along the sequence. The worst performance in all cases belongs to the offline model with fixed τ, where all classes collapse into one. For clarity, all the plots in Figure 6 are from the online model.

consistently recognises it in the later batches without distorting the parameters of the existing classes. The overall accuracy of 100 percent for this experiment is mostly owing to the contribution of the adaptive learning rate in adjusting the variances of each class with respect to the degree of adaptation. Using fixed learning rates substantially reduces the accuracy (14 percentage points) due to the drift in the means of the existing classes (Table I). It also exhibits an extra class and decreased recall and precision.

Combination of the two: Combining the two evolutionary scenarios above, we test the proposed model on a sequence with a *new* class that needs to be distinguished among the existing *shifting* classes. The challenge is two-fold: i) the shifting modes are prone to being misclassified as new classes, and ii) the new class might be merged into one of the existing shifted modes. This experiment is the closest to real-world scenarios where new states are likely to appear while the distributions can change over time. Given the combined challenge, the AdOn HDP-HMM proves highly accurate (93%), exhibiting a considerable improvement on the accuracy (12 percentage points) and cardinality thanks to the adaptive learning rate. The performance of the AdOn HDP-HMM is not significantly perturbed by these challenges since the learning rate adjusts the adaptability of the parameters with respect to the observed data. In an evolutionary scenario, the likelihood of the observations given the current parameters is low. This causes the learning rate over the covariance (τ_Σ) to increase, keeping the variance close to its prior. This, in turn, prevents a drift of the variance towards large values and encourages the mean to adjust.

Even with the fixed learning rate, the model still learns and recognises the new state thanks to the properties of the HDP. However, the overall performance deteriorates. On the one hand, new undesirable classes appear in response to the drift. On the other, some of the existing classes collapse into a single one, due to the considerable increase in the variance caused by the class shifts. In essence, with the fixed learning rate the predominant effect is an increase in the variance which eventually leads the model to merge some of the neighbouring states into a single class with a large variance (Figure 6(e,f)).

In a last experiment, we compare the proposed model with an *offline* learning mode, i.e. using the whole test sequence as a single batch for comparison with the online scheme. Figure 7, first row, shows the behaviours with evolutionary distributions: 1) the offline mode with fixed τ fails decoding completely; 2) the offline mode with adaptive τ performs well initially, but when δ increases beyond a certain extent, it starts to confuse the instances of drifted classes with new classes; 3) conversely, the proposed online model is capable of following the evolution of the underlying distributions. Figure 7, second row, shows the behaviours with fixed distributions and a new class. Since the ground-truth classes do not change over time, both the online and offline modes with adaptive τ perform similarly. Again, the offline mode with fixed τreports a much worse performance. Overall, this experiment confirms the principled advantages of both the online approach and the adaptive learning rate.

TABLE II
AdOn HDP-HMM hyper-parameters. In the hyper-parameters used for resampling λ, we set: $\nu = 1000, p =$ dimensionality of each observation.

Parameter	Resampled	Distribution
γ (Eq. 1)	Yes	Gamma(0.02, 0.01)
α (Eq. 2)	Yes	Gamma(1,0.01)
κ (p. 3)	No	0.01
λ (p. 3)	Yes	NIW($[0]_p, \kappa, I_p(\nu - p - 1), \nu$)
τ_μ, τ_Σ	Yes	Gamma(9,0.01)
τ_β, τ_π	Yes	Gamma (9,100)

B. Activity recognition datasets

In this section, we use two video datasets to assess the performance of the proposed model in activity recognition scenarios. The model parameters are all resampled based on the hyper-parameters listed in Table II.

1) Collated Weizmann dataset: The Weizmann dataset contains 93 single-action videos from a set of 10 classes performed by 9 different actors. While the recognition accuracy on the original dataset is saturated [41] [42], some studies have collated its individual actions into (unsegmented) sequences to experiment with time segmentation [5]. In a similar way, we have created 4 sequences, each consisting of 12 random actions selected from the provided action classes. Each sequence consists of approximately 900 frames. As feature set, we have used the position of the actor's centroid in the image plane and the distances between the centroid and the actors' contour along five given directions [43].

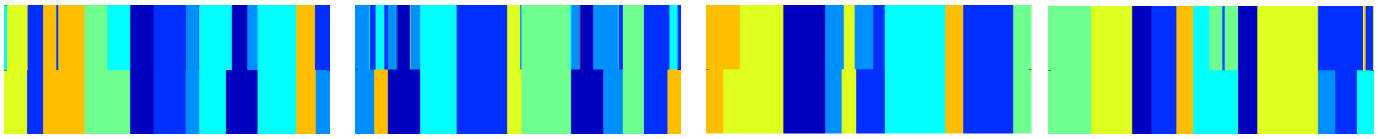
The estimated states of the AdOn HDP-HMM variants over the above sequences are visualised in Figure 8, showing remarkable qualitative accuracy in segmentation and classification. The quantitative results are reported in Table III, including from the offline variant with a single batch for the whole test sequence. In addition, we report the results from an offline max-margin approach [5]. However, its results are not directly comparable for two reasons: a) the datasets used here and in [5] are similar in conception, yet different in sequence collation, and b) the classifier in [5] operates over a closed set of classes, as opposed to ours which allows for an unlimited number of classes.

The results with the fixed learning rate show a similar trend to the adaptive, and only a slightly lower average accuracy. This can be due to the stationary nature of the dataset, as training and test sequences are drawn from similar distributions and adaptation is not significant. In addition, the accuracy with the online processing does not show any noticeable deterioration over the full, offline processing.

2) TUM kitchen dataset: The TUM kitchen dataset is a human assistive dataset, consisting of unsegmented sequences of everyday activities performed in a typical kitchen environment [7]. The dataset contains multi-modal data, annotated separately for the actors' left and right hands (9 classes) and torso (2 classes). The features are 28D vectors of joint coordinates for the torso and the relevant hands. The main actions include 'Reaching', 'Releasing Grasp Of Something', 'Taking An Object', 'Reaching Upward', 'Lowering An Ob-

TABLE III
FRAME-LEVEL ACCURACY, SEGMENTATION F1 SCORE, AND CARDINALITY ERROR FOR THE ADAPTIVE ONLINE HDP-HMM VARIANTS AND STATE-OF-THE-ART STUDIES ON THE COLLATED WEIZMANN DATASET.

	Accuracy					**F1 score**					**Cardinality**			
Method	**S1**	**S2**	**S3**	**S4**	**avg ± std**	**S1**	**S2**	**S3**	**S4**	**avg ± std**	**S1**	**S2**	**S3**	**S4**
Online HDP-HMM (*ada* τ)	**0.82**	**0.76**	0.89	**0.81**	0.82 ± 0.04	0.92	0.66	0.95	0.80	0.83 ± 0.11	**0**	**0**	**0**	**0**
Online HDP-HMM ($\tau = 1$)	0.81	0.70	**0.95**	0.80	0.81±0.09	0.92	0.66	0.89	0.80	0.82±0.10	0	-1	0	-1
Offline HDP-HMM	*0.78*	*0.76*	*0.95*	*0.81*	0.82±0.07	*0.91*	*0.66*	*0.95*	*0.81*	0.83±0.11	*0*	*0*	*0*	*0*
Offline Max-margin [5]	*0.87 (avg)*					-					-			



(a) Sequence 1 (b) Sequence 2 (c) Sequence 3 (d) Sequence 4

Fig. 8. Estimated states for the collated Weizmann dataset. Action labels are represented by colours. Top stripes: predicted states; bottom stripes: ground truth.

ject', opening and closing doors and drawers and 'Carrying While Locomoting', the distinctions between which are quite subtle at times even for human annotators. The main feature of this dataset compared to the collated Weizmann is that the transitions between actions occur naturally, hence time segmentation is more challenging. In our experiments, we have performed segmentation and classification of the actions of the left and right hands, separately. All the sequences provided by the 3D motion capture sensors are used in leave-one-out cross validation tests. Experiments are run for both the basic sequences (denoted as '*robotic*', taking objects one by one), and the more probing ones ('*complex*', including sequences with multiple objects moved together, in arbitrary order and repeatedly).

To analyse performance in detail, we report results for each of these sequences. The main differences in the sequences come from the different height and size of the actor and the frequencies of action occurrences. The experiment is repeated with fixed and adaptive learning rates and results are compared in Table IV. The overall accuracy with fixed and adaptive learning rates is roughly comparable, but the adaptive learning rate achieves a remarkable decrease in cardinality error. To test statistical significance, we have performed a paired *t*-test and report the *p*-values in Table IV; the *p*-values show that the difference in cardinality error is statistically significant, and so is the slightly better accuracy of the adaptive learning rate on the LH data. For visual evaluation, a few sequences are colour-plotted in Figure 9. It is worth noting that classes in this dataset are hard to segment. For instance, the boundary between 'putting an object' on the table and 'leaving grasp' of it is unclear and these classes are hard to distinguish (an example is highlighted with a box in Figure 9(e)).

To specifically observe the adaptive behaviour, we have trained the model on the *robotic* sequences and tested it on the *complex* ones. Although the emission parameters might not drastically change in this scenario, the transition probabilities need to adapt due to changes in the order of actions in the *complex* set. Table V shows the remarkable contribution of the adaptive learning rate to improving cardinality and overall accuracy. Similar to the synthetic results, the model with adaptive learning rates is able to prevent an excessive increase of the variance and avoid the collapse of neighbouring classes into one (the phenomenon observed when $\tau = 1$ in Figures 9(d-e)).

To evaluate the ability to recognise new classes, we have taken the first 4 sequences and removed the observations related to 'Lowering an object' (shown as a dotted pattern in Figure 9(f)) in all but the first sequence. We have then trained the model on sequences 2-4 and tested on the sequence containing the new action. The AdOn HDP-HMM is able to recognise a new action (brown in Figure 9(f)) and learn its parameters with consistent future recognition. This significant property of the model is inherent to the HDP approach and the behaviour is similar, irrespective of whether or not the learning rate is adapted.

The closest work using the TUM kitchen dataset is based on a CRF [7]. This method is not directly comparable since the AdOn HDP-HMM is online, adaptive and based on a dynamic class set. To create a closer match, we have run the *Offline* variant of the AdOn HDP-HMM, the results of which are similar to the CRF's and outperform them for complex sequences. This finding supports our main claim that adaptability leads to remarkable improvements when the test distributions are different from the training. The distribution of τ_π and τ_β (the transition-related learning rates) for these experiments mainly peak around 0.1, indicating that the learning rates encourage the model to rely on the observed data when inferring the HDP transition probabilities, which translates into faster adaptation.

C. Sampling efficiency and computational time

We next examine the Gibbs sampler's mixing rate and execution time for the above experiments. To gain an overall understanding of parameter mixing (emission and transition) the log-likelihood for the first batch of a Weizmann sequence is shown in Figure 10(e). Since most of the sampled variables

TABLE IV

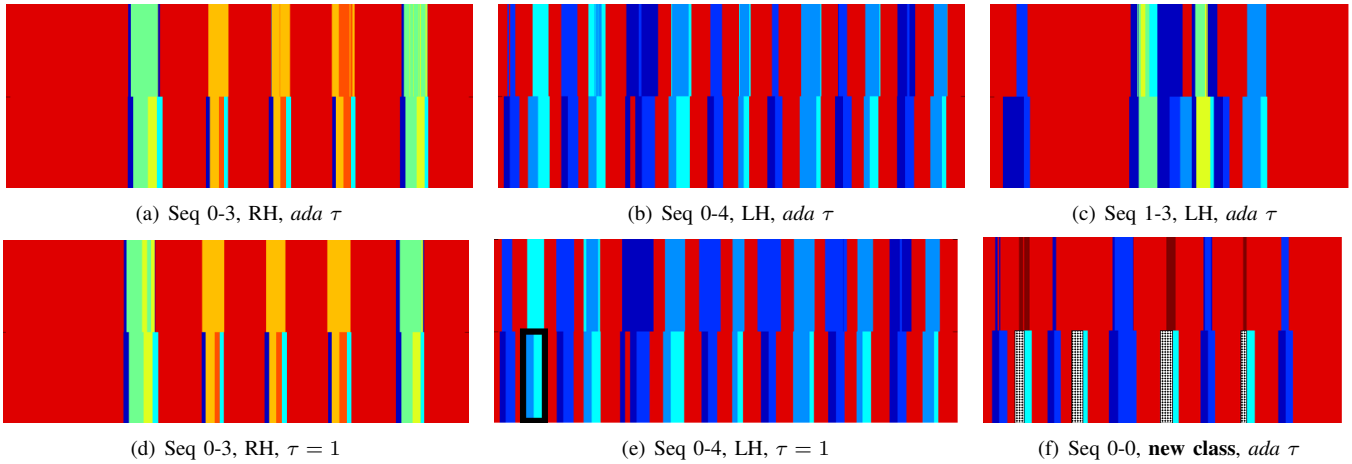
FRAME-LEVEL ACCURACY AND CARDINALITY ERROR FOR THE ADON HDP-HMM ON ALL TUM KITCHEN SEQUENCES. THE COMPARISON BETWEEN SIMILAR SEQUENCES WITH WITH ADAPTIVE ($ada\tau$) AND FIXED ($\tau = 1$) SHOWS INCREMENTAL IMPROVEMENT ON FRAME-LEVEL ACCURACY AND SIGNIFICANT DECREASE OF THE CARDINALITY ERROR.

		Accuracy				**Cardinality Error**			
Sequences		**RH**		**LH**		**RH**		**LH**	
		ada τ	$\tau = 1$	*ada* τ	$\tau = 1$	*ada* τ	$\tau = 1$	*ada* τ	$\tau = 1$
robotic	*Online* Seq 0-0	0.79	0.81	0.73	0.71	0	-1	-1	-1
	Online Seq 0-1	0.79	0.82	0.75	0.75	-2	-1	0	-1
	Online Seq 0-3	0.84	0.84	0.67	0.69	0	-2	0	-1
	Online Seq 0-4	0.70	0.69	0.71	0.72	1	-1	-2	-3
	Online Seq 0-6	0.51	0.48	0.56	0.55	-3	-6	-1	-3
	Online Seq 0-7	0.45	0.48	0.57	0.55	-3	-4	-1	-3
	Online Seq 0-8	0.64	0.68	0.62	0.63	-1	-3	-2	-2
	Online Seq 0-9	0.73	0.71	0.70	0.69	0	-2	-1	-2
	Online Seq 0-10	0.79	0.79	0.68	0.70	0	-2	0	-1
	Online Seq 0-11	0.70	0.76	0.63	0.63	-5	-4	-2	-3
	Online Seq 1-0	0.65	0.69	0.68	0.69	-1	-2	-3	-4
	Online Seq 1-1	0.71	0.69	0.65	0.62	-1	-1	-2	-3
	Online Seq 1-2	0.63	0.63	0.76	0.74	-1	-1	0	-1
	Online Seq 1-3	0.14	0.14	0.66	0.65	-6	-6	0	0
	Online Seq 1-6	0.67	0.67	0.61	0.61	0	0	-2	-2
	Online Seq 1-7	0.69	0.68	0.60	0.58	-1	-1	-1	-2
complex	*Online* Seq 0-2	0.76	0.70	0.78	0.75	-2	-1	-1	-1
	Online Seq 0-12	0.64	0.64	0.58	0.55	-1	-2	-3	-5
	Online Seq 1-4	0.64	0.67	0.74	0.71	-4	-5	0	-1
Overall Average		0.656	0.661	0.667	0.659	-1.68	-2.37	-1.15	-2.05
Paired T-test		*p-value*		*p-value*		*p-value*		*p-value*	
		0.43		**0.04**		**0.01**		**0.00**	
Robotic sequences		*average accuracy*				*average absolute cardinality error*			
Online		**0.65**	0.66	**0.70**	**0.66**	**1.56**	**2.31**	**1.12**	**2.00**
Offline		**0.65**	**0.68**	**0.70**	**0.66**	1.57	2.50	1.15	2.10
Offline CRF [7]		0.83				-			
Complex sequences		*average accuracy*				*average absolute cardinality error*			
Online		**0.68**	0.67	**0.70**	**0.67**	2.33	2.67	**1.33**	**2.33**
Offline		**0.68**	**0.68**	0.69	0.66	**2.21**	**2.58**	1.43	**2.33**
Offline CRF [7]		0.63				-			

TABLE V

ADAPTABILITY EXPERIMENT: FRAME-LEVEL ACCURACY AND STATE CARDINALITY ERROR FOR THE ADON HDP-HMM TRAINED WITH THE *robotic* SEQUENCES AND TESTED ON THE *complex* ONES. THE COMPARISON BETWEEN SIMILAR SEQUENCES WITH ADAPTIVE ($ada\tau$) AND FIXED ($\tau = 1$) LEARNING RATES SHOWS NOTICEABLE IMPROVEMENT OF THE FRAME-LEVEL ACCURACY AND SIGNIFICANT DECREASE OF THE CARDINALITY ERROR.

	Accuracy				**Cardinality**			
Sequences	**RH**		**LH**		**RH**		**LH**	
	ada τ	$\tau = 1$	*ada* τ	$\tau = 1$	*ada* τ	$\tau = 1$	*ada* τ	$\tau = 1$
Online Actor1, complex	**0.73**	0.72	0.65	**0.68**	**-2**	-3	**2**	**-2**
*Online*Actor3, complex	**0.55**	0.54	**0.52**	0.49	**-1**	-3	**-4**	-6
*Online*Actor1, repetitive	0.45	**0.48**	**0.57**	0.55	**-3**	-4	**-1**	-3



(a) Seq 0-3, RH, *ada* τ (b) Seq 0-4, LH, *ada* τ (c) Seq 1-3, LH, *ada* τ

(d) Seq 0-3, RH, $\tau = 1$ (e) Seq 0-4, LH, $\tau = 1$ (f) Seq 0-0, **new class**, *ada* τ

Fig. 9. Estimated states for the TUM kitchen dataset using the ADON HDP-HMM. LH and RH stand for left and right hand. The two consecutive actions highlighted in (e) are hardly distinguishable even for human annotators, hence mostly decoded as one action by the model. The first two columns show *robotic* sequences, whereas the third column includes *complex* ones. (c) is a sequence with altered orders of actions performed spontaneously and (f) contains a new action shown as a dotted pattern in the ground-truth and recognised by the model in a random brown colour. In most cases, using the adaptive τ causes noticeable improvements on both the performance and cardinality of the inferred states.

contribute to the likelihood calculation, the well-mixed results indicate general mixing efficiency in the model. Additionally, mixing trends of the learning rates ($\tau_\mu, \tau_\Sigma, \tau_\pi, \tau_\beta$) for a generic evolutionary run are shown in Figure 10(a-d), to both monitor mixing and support the experiments' discussion. The large values of τ_Σ prevent the model from increasing the variance to fit the changing distributions. Rather, the model allows for the means to evolve, by converging to small values of τ_μ. The similarly small values of τ_π and τ_β ensure adaptability of the model towards changing state transitions for HDP-HMM. Through the orchestration of these parameters, the proposed model adapts to changes in the streaming batches and retains greater classification accuracy.

Eventually, the computational time per frame for runs on an Intel Xeon E5 2.90 GHz processor over the Weizmann and TUM kitchen datasets are shown in Figure 10(f). The boxplot includes online and offline variants with fixed and adaptive learning rates to explore how the learning rates and the online scheme affect the computational time. Based on the elapsed time (in seconds), the offline run is the fastest since all the data are processed in a single batch. The adaptive online runs occur in 3-4 batches of 1000 iterations each, showing an increase of about 5-10 ms in execution time. Adapting the learning rate causes between 3-10 ms delay, yet given the discussed benefits, particularly for evolving sequences, this latency seems quite reasonable. It is important to mention that given the initial bootstrap training, the Gibbs algorithm converges rapidly allowing for the model to run in acceptable time. Overall, using the adaptive learning rate ensures multiple improvements without imposing an excessive computational load.

V. Conclusion

In this paper, we have presented a novel, adaptive online model - the AdOn HDP-HMM - suited for on-the-fly time segmentation and recognition of sequential data from incremental and variable class sets. The main contribution of the proposed model is the unsupervised posterior adaptation of the parameters over the successive data batches, accomplished through an adaptive learning rate that continuously balances the impact of the current batch with the memory accumulated so far. This is a suitable solution for online sequential estimation problems requiring adaptation over evolving distributions.

The performance of the proposed model is evaluated via a number of experiments including *stationary* and *evolutionary* scenarios. We have tested the general segmentation and classification accuracies in addition to the ability to detect the correct number of classes. The results are reported over variants of synthetic data and two activity recognition video datasets (collated Weizmann and TUM Assistive Kitchen). The proposed model achieves remarkable accuracy in all cases, and the improvement is considerable in the evolutionary scenarios.

Thanks to the unsupervised adaptive online estimation and the ability to learn over infinite class sets, the proposed AdOn HDP-HMM can be a solution for sequential estimation in a number of scenarios which have to date received relatively little attention in the literature. Not relying on human intervention, revision or correction of estimated labels, this model can be a suitable candidate for streaming applications. In addition, although mainly designed for evolutionary distributions, its accuracy over stationary data has proved higher than or equal to that of the most comparable results, and without a major overhead in computational load.

VI. Appendix A: τ for balance

In this appendix, we address the posterior inference of parameters and explore how the prior and likelihood distributions convey the knowledge of the previous and current observations. Considering the online HDP-HMM model with parameters ϕ, observations Y and learning rate τ, the posterior for the parameters in the n^{th} batch is given as:

$$\begin{aligned} p(\phi_n|Y_n, \phi_{n-1}, \tau) &\propto p(Y_n|\phi_{n-1})p(\phi_{n-1})^\tau \\ &\propto \underbrace{\prod_{i=1}^{N} p(y_{n,i}|\phi_{n-1})}_{N} \underbrace{\left(p(\phi_0) \prod_{j=1}^{n-1}\prod_{i=1}^{N} p(y_{j,i}|\phi_{j-1}) \right)^\tau}_{N(n-1)\tau} \end{aligned} \tag{18}$$

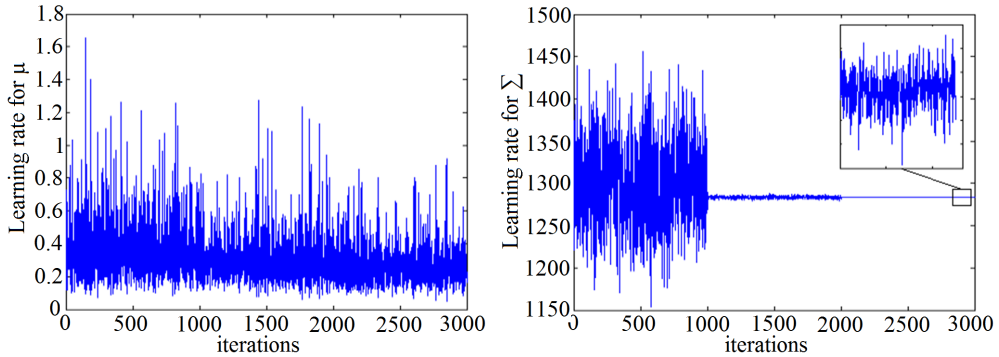
where $y_{n,i}$ represents the i^{th} observation $(i = 1 \dots N)$ in the n^{th} batch. As more batches stream in (i.e. n increases), the weight of the prior is accumulated and adaptability to new data decreases. The learning rate, however, can be used as an equaliser that controls the balance between the prior and the likelihood, and tunes the model's adaptability. For positive values of $\tau < 1$, the model discounts the impact of the accumulated previous data and allows for more adaptability. However, when $\tau > 1$, posterior ϕ_n is inclined to follow the prior more strictly. In other words, τ can be seen as the scaling coefficient for the number of 'pseudo-observations' in the prior.

VII. Appendix B: Conjugacy for τ_μ

Let us consider K multivariate normal distributions of parameters μ_k and Σ_k added with an additional strictly positive scaling parameter, τ_μ, at the denominator of the covariance (Eq. 19). Let us assume to have drawn a sample, $A_k, k = 1 \dots K$, from each of the distributions and to have a Gamma prior over τ_μ. In this appendix, we show that the Gamma prior is conjugate even though these samples are not identically distributed, and we compute the parameters for the Gamma posterior. The resulting parameters can be easily extended to the case of multiple samples from each distribution.

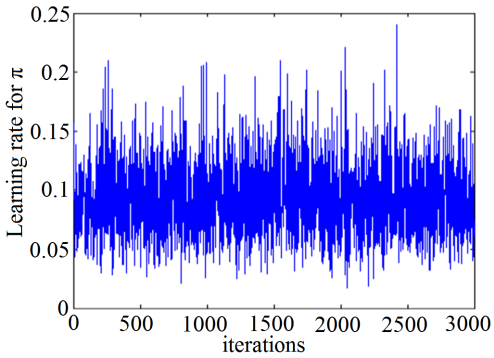
$$\begin{aligned} G(\tau_\mu|A_{1:K}, \alpha^*, \beta^*) &= \prod_{k=1}^{K} N(A_k|\mu_k, \frac{\Sigma_k}{\tau_\mu}) G(\tau_\mu|\alpha, \beta) \\ &\propto \prod_{k=1}^{K} \frac{\tau_\mu^{1/2}}{|\Sigma_k|^{1/2}} \exp(-\frac{\tau_\mu}{2}(A_k - \mu_k)^T \Sigma_k^{-1}(A_k - \mu_k)) \\ &\quad \frac{\beta^\alpha}{\Gamma(\alpha)} \tau_\mu^{\alpha-1} \exp(-\beta\tau_\mu) \end{aligned} \tag{19}$$

Discarding the terms that are independent of random variable τ_μ, we have:

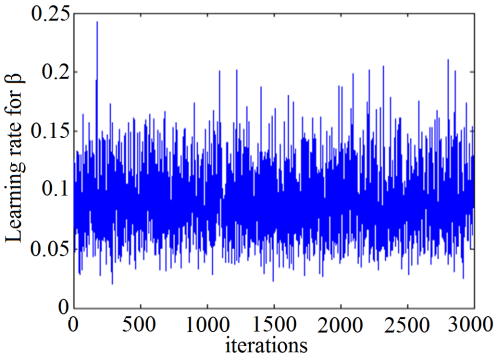


(a) τ_μ

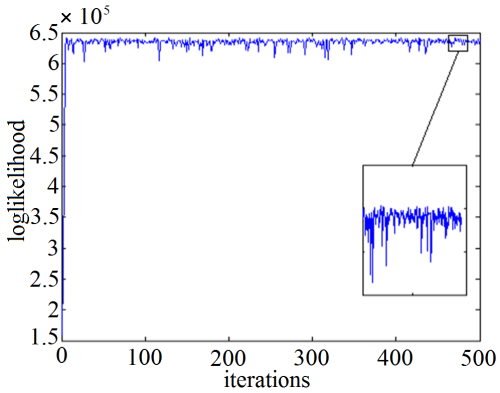
(b) τ_Σ



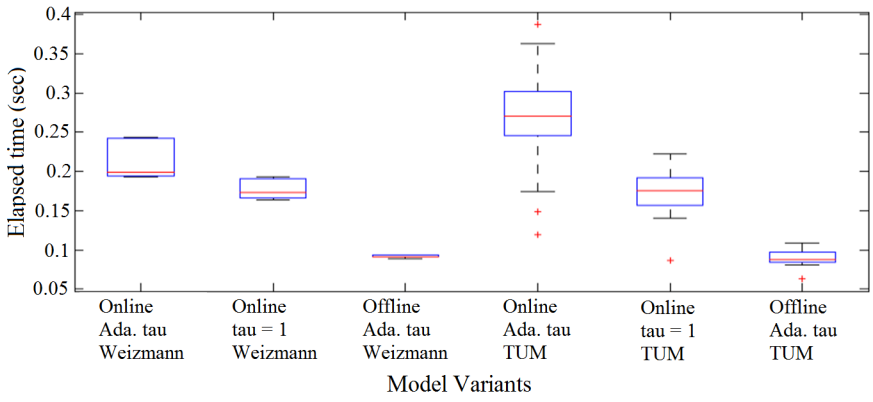
(c) τ_π



(d) τ_β



(e) Log-likelihood



(f) Computational time

Fig. 10. Sampling efficiency and computational time: (a-d) Sample mixing for all the four learning rates across three online batches and (e) The log-likelihood plot for the first batch of a Weizmann run shows good mixing and convergence both for the learning rates and for all other parameters involved in the likelihood calculation. The zoomed boxes show that even after conversion there is considerable variation within the sampled variables. (f) Computational time per frame (seconds) for the Weizmann and TUM kitchen datasets over the online and offline variants, with adaptive and fixed learning rates.

$$
\begin{aligned}
&G(\tau_\mu|A_{1:K},\alpha^*,\beta^*)\\
&\propto \tau_\mu^{K/2}\tau_\mu^{\alpha-1}\exp(-\beta\tau_\mu-\frac{\tau_\mu}{2}\sum_{k=1}^{K}(A_k-\mu_k)^T\Sigma_k^{-1}(A_k-\mu_k))\\
&\propto \tau_\mu^{\alpha+K/2-1}\exp(-\tau_\mu(\beta+\frac{1}{2}\sum_{k=1}^{K}(A_k-\mu_k)^T\Sigma_k^{-1}(A_k-\mu_k)))
\end{aligned}
\tag{20}
$$

Thereby, such terms are proportional to a Gamma distribution with the following parameters:

$$
\begin{aligned}
\alpha^* &= \alpha + K/2\\
\beta^* &= \beta + \frac{1}{2}\sum_{k=1}^{K}(A_k-\mu_k)^T\Sigma_k^{-1}(A_k-\mu_k)
\end{aligned}
\tag{21}
$$

As a further assumption, let us assume that each of the K class distributions is weighed by an exponent $\lambda_k = \nu_k / \sum_{k=1}^{K}\nu_k$ that represents our confidence in the class. Under this assumption, Equation 20 modifies as:

$$
\begin{aligned}
&G(\tau_\mu|A_{1:K},\alpha^*,\beta^*)\\
&\propto \tau_\mu^{\alpha+\sum_k\lambda_k/2-1}\exp(-\tau_\mu(\beta+\frac{1}{2}\sum_{k=1}^{K}\lambda_k(A_k-\mu_k)^T\Sigma_k^{-1}(A_k-\mu_k)))
\end{aligned}
\tag{22}
$$

and we eventually obtain the parameters for the Gamma posterior of Equation 14:

$$
\begin{aligned}
\alpha^* &= \alpha + \sum_{k=1}^{K}\lambda_k/2 = \alpha + 1/2\\
\beta^* &= \beta + \frac{1}{2}\sum_{k=1}^{K}\lambda_k(A_k-\mu_k)^T\Sigma_k^{-1}(A_k-\mu_k)
\end{aligned}
\tag{23}
$$

□

VIII. Appendix C: Impact of τ on the parameters of the Inverse-Wishart

In this appendix, we study the impact of τ on the mean and covariance of the Inverse-Wishart distribution. We can see that:

$$
\begin{aligned}
IW(\Sigma|\Psi,\nu)^\tau &\propto \left(|\Sigma|^{-\frac{\nu+p+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}(\Psi\Sigma^{-1})\right)\right)^\tau\\
&= |\Sigma|^{-\frac{\tau(\nu+p+1)}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}(\tau\Psi\Sigma^{-1})\right)\\
&\approx |\Sigma|^{-\frac{\tau\nu+p+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}(\tau\Psi\Sigma^{-1})\right)\\
&\Rightarrow IW(\Sigma|\Psi,\nu)^\tau \propto IW(\Sigma|\tau\Psi,\tau\nu)
\end{aligned}
\tag{24}
$$

This approximation only holds where $\nu >> p+1$ which is often the case in our experiments since $\nu = 1000$ and the largest p is 28.

The following equations show the effect of τ on the mean and variance of Σ (noted as M and V to avoid confusion):

$$
\begin{aligned}
&\text{mean of } \Sigma\sim IW(\Psi,\nu): && M_\Sigma = \frac{\Psi}{\nu+p+1},\\
&\text{mean of } \Sigma\sim IW(\tau\Psi,\tau\nu): && M_\Sigma^{(\tau)} = \frac{\tau\Psi}{\tau\nu+p+1}\approx M_\Sigma\\
&\text{variance of } \Sigma\sim IW(\Psi,\nu): && V_\Sigma \approx \frac{\Psi_{ij}^2}{\nu^3}\\
&\text{variance of } \Sigma\sim IW(\tau\Psi,\tau\nu): && V_\Sigma^{(\tau)} \approx \frac{\tau^2\Psi_{ij}^2}{\tau^3\nu^3}\approx V_\Sigma/\tau
\end{aligned}
\tag{25}
$$

As can be seen, the resulting Σ samples are drawn approximately around the same mean, but with a scaled variance. When $0 \le \tau_\Sigma < 1$ the variance increases, whereas for $\tau_\Sigma > 1$ the distribution is more peaky. In other words, the posterior samples of Σ in the former case are allowed to move away from the IW mean and tend to have greater adaptability towards the current observed data. Conversely, in the latter case the posterior samples concentrate around the prior mean and thereby discourage covariance adaptation.

References

[1] H. Bunke, P. J. Dickinson, M. Kraetzl, and W. D. Wallis. *A graph-theoretic approach to enterprise network dynamics*. Birkhäuser, 2007.

[2] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *CVPR*, pages 379–385, 1992.

[3] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2-3):210–220, 2006.

[4] D. L. Vail, M. M. Veloso, and J. D. Lafferty. Conditional Random Fields for Activity Recognition. In *Int. Conf. on Autonomous Agents and Multi-Agent Systems*, 2007.

[5] M. Hoai, Z-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *CVPR*, 2011.

[6] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, and A.F. Smeaton. TRECVID 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID 2011*. NIST, USA, 2011.

[7] 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 ICCV 2009*, 2009.

[8] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[9] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *NIPS*, pages 577–584, 2001.

[10] A. Bargi, R. Y. D. Xu, and M. Piccardi. An infinite adaptive online learning model for segmentation and classification of streaming data. In *ICPR*, pages 3440–3445, 2014.

[11] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721 –741, 1984.

[12] Y. W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. In *NIPS*, volume 20, pages 1481–1488, 2007.

[13] M. Zanotto, D. Sona, V. Murino, F. Papaleo, and H. Kjellstrom. Dirichlet process mixtures of multinomials for data mining in mice behaviour analysis. In *ICCV Workshops*, pages 197–202, 2013.

[14] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Bayesian Nonparametric Inference of Switching Dynamic Linear Models. *IEEE Transactions on Signal Processing*, 59(4):1569–1585, 2011.

[15] C. Zhang, E. Henrik, X. Gratal, F. Pokorny, and H. Kjellstrom. Supervised hierarchical Dirichlet processes with variational inference. In *ICCV Workshops*, pages 254–261, 2013.

[16] W. Fan and N. Bouguila. Online learning of a Dirichlet process mixture of Beta-Liouville distributions via variational inference. *IEEE Transactions on Neural Networks and Learning Systems*, 24(11):1850–1862, 2013.

[17] K. Bousmalis, S. Zafeiriou, L. Morency, and M. Pantic. Infinite hidden conditional random fields for human behavior analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 24(1):170–177, 2013.

[18] T. Taniguchi, K. Hamahata, and N. Iwahashi. Unsupervised segmentation of human motion data using a sticky hierarchical Dirichlet process-hidden Markov model and minimal description length-based chunking method for imitation learning. *Advanced Robotics*, 25(17):2143–2172, 2011.

[19] A. Bargi, R. Y. D. Xu, and M. Piccardi. An online HDP-HMM for joint action segmentation and classification in motion capture data. In *CVPR Workshops*, pages 1–7, 2012.

[20] V. Bastani, L. Marcenaro, and C. Regazzoni. Unsupervised trajectory pattern classification using hierarchical Dirichlet process mixture hidden Markov model. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2014.

[21] H. W. Sorenson and D. L. Alspach. Recursive Bayesian estimation using Gaussian sums. *Automatica*, 7(4):465 – 479, 1971.

[22] K. R. Canini, L. Shi, and T. L. Griffiths. Online Inference of Topics with Latent Dirichlet Allocation. In *AISTATS*, volume 5, pages 65–72, 2009.

[23] L. I. Kuncheva and C. O. Plumpton. Adaptive learning rate for online linear discriminant classifiers. In *2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, SSPR & SPR '08, pages 510–519, 2008.

[24] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011.

[25] T. Schaul, S. Zhang, and Y. LeCun. No more pesky learning rates. *CoRR abs/1206.1106v2*, 2013.

[26] L. Bottou. Stochastic learning. *Advanced Lectures on Machine Learning*, LNAI 3176:146–168, 2004.

[27] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[28] A. Pentina and C. H. Lampert. A PAC-bayesian bound for lifelong learning. In *ICML*, pages 991–999, 2014.

[29] A. Rodriguez. On-line learning for the infinite hidden Markov model. *Communications in Statistics - Simulation and Computation*, 40(6):879–893, 2011.

[30] C. Wang, J. W. Paisley, and D. M. Blei. Online variational inference for the hierarchical Dirichlet process. In *AISTATS*, pages 752–760, 2011.

[31] C. C. Loy, T. M. Hospedales, T. Xiang, and S. Gong. Stream-based joint exploration-exploitation active learning. In *CVPR*, pages 1560–1567, 2012.

[32] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3), 1951.

[33] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[34] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Developing a tempered HDP-HMM for systems with state persistence. Technical report, MIT Laboratory for Information and Decision Systems, 2007.

[35] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. An HDP-HMM for systems with state persistence. In *ICML*, pages 312–319, 2008.

[36] C.M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

[37] E. B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Ph.D. thesis, MIT, Cambridge, MA, 2009.

[38] J. Juang. Maximum likelihood estimation of Dirichlet distribution parameters. Technical report, CMU, 2005.

[39] D. Knowles and Z. Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534–1552, 2011.

[40] A. Bargi, R. Y. D. Xu, Z. Ghahramani, and M. Piccardi. A non-parametric conditional factor regression model for multi-dimensional input and response. *AISTATS*, 33:77–85, 2014.

[41] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *CVPR*, pages 1–8, 2007.

[42] L. Nanni, S. Brahnam, and A. Lumini. Combining different local binary pattern variants to boost performance. *Expert Systems with Applications*, 38(5):6209 – 6216, 2011.

[43] Z. Moghaddam and M. Piccardi. Deterministic initialization of hidden Markov models for human action recognition. In *DICTA*, pages 188–195, 2009.