Inertial Hidden Markov Models: Modeling Change in Multivariate Time Series

George D. Montañez
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA USA
gmontane@cs.cmu.edu

Saeed Amizadeh
Yahoo Labs
Sunnyvale, CA USA
amizadeh@yahoo-inc.com

Nikolay Laptev
Yahoo Labs
Sunnyvale, CA USA
nlaptev@yahoo-inc.com

Abstract

Faced with the problem of characterizing systematic changes in multivariate time series in an unsupervised manner, we derive and test two methods of regularizing hidden Markov models for this task. Regularization on state transitions provides smooth transitioning among states, such that the sequences are split into broad, contiguous segments. Our methods are compared with a recent hierarchical Dirichlet process hidden Markov model (HDP-HMM) and a baseline standard hidden Markov model, of which the former suffers from poor performance on moderate-dimensional data and sensitivity to parameter settings, while the latter suffers from rapid state transitioning, over-segmentation and poor performance on a segmentation task involving human activity accelerometer data from the UCI Repository. The regularized methods developed here are able to perfectly characterize change of behavior in the human activity data for roughly half of the real-data test cases, with accuracy of 94% and low variation of information. In contrast to the HDP-HMM, our methods provide simple, drop-in replacements for standard hidden Markov model update rules, allowing standard expectation maximization (EM) algorithms to be used for learning.

Introduction

"Some seek complex solutions to simple problems; it is better to find simple solutions to complex problems." - Soramichi Akiyama

Time series data arise in different areas of science and technology, describing the behavior of both natural and man-made systems. These behaviors are often quite complex with uncertainty, which in turn require us to incorporate sophisticated dynamics and stochastic models to model them. Furthermore, these complex behaviors can *change* over time due to some external event and/or some internal systematic change of dynamics/distribution. For example, consider the case of monitoring one's physical activity via an array of accelerometer body sensors over time. A certain pattern emerges on the time series of the sensors' readings while the person is walking; however, this pattern quickly changes to a new one as she begins running. From the data analysis perspective, it is important to first detect these *change points* as they are quite often indicative of an "interesting" event or an anomaly in the system. We are also interested in characterizing the new *state* of the system (e.g. running vs. walking) which reflects its mode of operation. Change point detection methods (Kawahara, Yairi, and Machida 2007; Xie, Huang, and Willett 2013; Liu et al. 2013; Ray and Tsay 2002) have been proposed to answer the first question while Hidden Markov Models (HMM) can answer both.

One crucial observation in many real-world systems, natural and man-made, is the behavior changes are typically infrequent; that is, the system takes some (unknown) time before it changes its behavior to a new modus operandi. For instance, in our earlier example, it is unlikely for a person to rapidly fluctuate between walking and running, making the durations of different activities over time relatively long and highly variable. We refer to this as the *inertial property*, alluding to the physical property of matter that ensures it will continue along a fixed course unless acted upon by an external force. Unfortunately, classical HMMs are not equipped with sufficient mechanisms to capture this property and often result in a high rate of state transitioning and subsequently false positives in terms of detecting change points.

Few solutions exist in the literature to address this problem. In the context of Markov models, Fox *et al.* (Fox et al. 2011; Willsky et al. 2009) have recently proposed the *sticky hierarchical Dirichlet process hidden Markov model (HDP-HMM)* which uses a Bayesian non-parametric approach with appropriate priors to promote self-transitioning (or *stickiness*) for HMMs. Despite its elegant theoretical foundation, the sticky HDP-HMM is not a practical solution in many real-world situations. In particular, the performance of the HDP-HMM tends to degrade as the dimensionality of the problem increases beyond ten dimensions. Moreover, due to iterative Gibbs sampling for its learning, the sticky HDP-HMM can become computationally prohibitive. In practice, the most significant drawback of the sticky HDP-HMM originates with its non-parametric Bayesian nature: due to the existence of many hyperparameters, the search space for initial tuning is exponentially large and significantly affects the learning quality for a given task.

In this paper, we propose a regularization-based framework for HMMs, called *Inertial hidden Markov models* (Inertial HMMs), which are biased towards the inertial state-transition property. Similar to the sticky HDP-HMM, our framework is based on theoretically sound foundations, yet is much simpler and more intuitive than the HDP-HMM. In particular, our framework has only two initial parameters for which we have developed intuitive initialization techniques that significantly minimize the effort needed for parameter tuning. Furthermore, as we show later, our proposed methods in practice boil down to upgraded update rules for standard HMMs. This allows one to easily upgrade exist-

ing HMM libraries to take advantage of our methods, while still preserving the computational efficiency of the standard HMM approach. By performing rigorous experiments on both synthetic and moderate dimensional real datasets, we show that Inertial HMMs are not only much faster than the sticky HDP-HMM, but also produce significantly better detection, suggesting Inertial HMMs as a more practical choice in comparison to the current state-of-the-art.

Problem Statement

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ denote a d-dimensional multivariate time series, where $\mathbf{x}_t \in \mathbb{R}^d$. Given such a time series, we seek to segment $\mathbf{X}$ along the time axis into *segments*, where each segment corresponds to a subsequence $\mathbf{X}_{i\ldots i+m} = \{\mathbf{x}_i, \ldots, \mathbf{x}_{i+m}\}$ and maps to a predictive (latent) state $\mathbf{z}$, represented as a one-of-K vector, where $|\mathbf{z}| = K$ and $\sum_{i=1}^{K} z_{t,i} = 1$. For simplicity of notation, let $\mathbf{z}_t = k$ denote $z_{t,k} = 1$ and let $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$ denote the sequence of latent states. Then for all $\mathbf{x}_t$ mapping to state k, we require that

$$\begin{aligned}\Pr(\mathbf{x}_{t+1}|\mathbf{X}_{1\ldots t}, \mathbf{z}_t = k) &= \Pr(\mathbf{x}_{t+1}|\mathbf{z}_t = k) \\ &= \Pr(\mathbf{x}_{t'+1}|\mathbf{z}_{t'} = k) \\ &= \Pr(\mathbf{x}_{t'+1}|\mathbf{X}_{1\ldots t'}, \mathbf{z}_{t'} = k).\end{aligned}$$

Thus, the conditional distribution over futures at time t conditioned on being in state k is equal to the distribution over futures at time t' conditioned on being in the same state. Thus, we assume conditional independence given state, and stationarity of the generative process.

We impose two additional criteria on our models. First, we seek models with a small number of latent states, $K \ll T$, and second, we desire state transition sequences with the inertial property, as defined previously, where the transitioning of states does not occur too rapidly.

The above desiderata must be externally imposed on our model, since simply maximizing the likelihood of the data will result in $K = T$ (i.e., each sample corresponds a unique state/distribution), and in general we may have rapid transitions among states. For the first desideratum, we choose the number of states in advance as is typically done for hidden Markov models (Rabiner 1989). For the second, we directly alter the probabilistic form of our model to include a parameterized regularization that reduces the likelihood of transitioning between different latent states.

Inertial Hidden Markov Models

Hidden Markov models (HMMs) are a class of long-studied probabilistic models well-suited for sequential data (Rabiner 1989). As a starting point for developing our inertial HMMs, we begin a standard K-state HMM with Gaussian emission densities. HMMs trained by expectation maximization (locally) maximize the likelihood of the data, but typically do not guarantee slow inertial transitioning among states. The number of states must be specified in advance, but no other parameters need to be given, as the remaining parameters are all estimated directly from the data.

To accommodate the inertial transition requirement, we derive two different methods for enforcing state-persistence in HMMs. Both methods alter the probabilistic form of the complete data joint likelihood, which result in altered transition matrix update equations. The resulting update equations share a related mathematical structure and, as is shown in the Experiments section, have similar performance in practice.

We will next describe both methods and provide outlines of their derivations.

Maximum A Posteriori (MAP) Regularized HMM

Following (Gauvain and Lee 1994), we alter the standard HMM to include a Dirichlet prior on the transition probability matrix, such that transitions out-of-state are penalized by some regularization factor. A Dirichlet prior on the transition matrix $\mathbf{A}$, for the jth row, has the form

$$p(A_j; \eta) \propto \prod_{k=1}^{K} A_{jk}^{\eta_{jk}-1}$$

where the η_{jk} are free parameters and A_{jk} is the transition probability from state j to state k. The posterior joint density over $\mathbf{X}$ and $\mathbf{Z}$ becomes

$$P(\mathbf{X}, \mathbf{Z}; \theta, \eta) \propto \left[\prod_{j=1}^{K}\prod_{k=1}^{K} A_{jk}^{\eta_{jk}-1}\right] P(\mathbf{X}, \mathbf{Z} \mid \mathbf{A}; \theta)$$

and the log-likelihood is

$$\begin{aligned}\ell(\mathbf{X}, \mathbf{Z}; \theta, \eta) \propto &\sum_{j=1}^{K}\sum_{k=1}^{K}(\eta_{jk} - 1)\log A_{jk} + \log P(\mathbf{z}_1; \theta) \\ &+ \sum_{t=1}^{T}\log P(\mathbf{x}_t|\mathbf{z}_t; \theta) + \sum_{t=2}^{T}\log P(\mathbf{z}_t|\mathbf{z}_{t-1}; \theta).\end{aligned}$$

MAP estimation is then used in the M-step of the expectation maximization (EM) algorithm to update the transition probability matrix. Maximizing, with appropriate Lagrange multiplier constraints, we obtain the update equation for the transition matrix,

$$A_{jk} = \frac{(\eta_{jk} - 1) + \sum_{t=2}^{T}\xi(z_{(t-1)j}, z_{tk})}{\sum_{i=1}^{K}(\eta_{ji} - 1) + \sum_{i=1}^{K}\sum_{t=2}^{T}\xi(z_{(t-1)j}, z_{ti})} \quad (1)$$

where $\xi(z_{(t-1)j}, z_{tk}) = \mathbb{E}[z_{(t-1)j} z_{tk}]$.

Given our prior, we can control the probability of self-transitions among states, but this method requires that we choose a set of K^2 parameters for the Dirichlet prior. However, since we are solely concerned about increasing the probability of self-transitions, we can reduce these parameters to a single parameter λ governing the amplification of self-transitions. We therefore define $\eta_{jk} = 1$ when $j \neq k$ and $\eta_{kk} = \lambda \geq 1$ otherwise, and the transition update equation becomes

$$A_{jk} = \frac{(\lambda - 1)\mathbb{1}(j = k) + \sum_{t=2}^{T}\xi(z_{(t-1)j}, z_{tk})}{(\lambda - 1) + \sum_{i=1}^{K}\sum_{t=2}^{T}\xi(z_{(t-1)j}, z_{ti})} \quad (2)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

Inertial Regularization via Pseudo-observations

Alternatively, we can alter the HMM likelihood function to include a latent binary random variable, V, indicating that a self-transition was chosen at random from among all transitions, according to some distribution. Thus, we view the transitions as being partitioned into two sets, self-transitions and non-self-transitions, and we draw a member of the self-transition set according to a Bernoulli distribution governed by parameter p. Given a latent state sequence $\mathbf{Z}$, with transitions chosen according to transition matrix $\mathbf{A}$, we define p as a function of both $\mathbf{Z}$ and $\mathbf{A}$. We would like p to have two properties: 1) it should increase with increasing $\sum_k A_{kk}$ (probability of self-transitions) and 2) it should increase as the number of self-transitions in $\mathbf{Z}$ increases. This will allow us to encourage self-transitions as a simple consequence of maximizing the likelihood of our observations.

We begin with a version of p based on a penalization constant $0 < \epsilon < 1$ that scales appropriately with the number of self-transitions. If we raise ϵ to a large positive power, the resulting p will decrease. Thus, we define p as ϵ raised to the number of non-self-transitions, M, in the state transition sequence, so that the probability of selecting a self-transition increases as M decreases. Using the fact that $M = (T-1) - \sum_{t=2}^{T}\sum_{k=1}^{K} z_{(t-1)k}z_{tk}$, we obtain

$$\begin{aligned} p = \epsilon^M &= \epsilon^{\sum_{t=2}^{T} 1 - \sum_{t=2}^{T}\sum_{k=1}^{K} z_{(t-1)k}z_{tk}} \\ &= \epsilon^{\sum_{t=2}^{T}\sum_{k=1}^{K} z_{(t-1)k} - \sum_{t=2}^{T}\sum_{k=1}^{K} z_{(t-1)k}z_{tk}} \\ &= \prod_{t=2}^{T}\prod_{k=1}^{K} \epsilon^{z_{(t-1)k} - z_{(t-1)k}z_{tk}}. \end{aligned} \tag{3}$$

Since ϵ is arbitrary, we choose $\epsilon = A_{kk}$, to allow p to scale appropriately with increasing probability of self-transition. We therefore arrive at

$$p = \prod_{t=2}^{T}\prod_{k=1}^{K} A_{kk}^{z_{(t-1)k} - z_{(t-1)k}z_{tk}}.$$

Thus, we define p as a computable function of $\mathbf{Z}$ and $\mathbf{A}$. Defining p in this deterministic manner is equivalent to choosing the parameter value from a degenerate probability distribution that places a single point mass at the value computed, allowing us to easily obtain a posterior distribution on V. Furthermore, we see that the function increases as the number of self-transitions increases, since $A_{kk} \leq 1$ for all k, and p will generally increase as $\sum_k A_{kk}$ increases. Thus, we obtain a parameter $p \in (0, 1]$ that satisfies all our desiderata. With p in hand, we say that V is drawn according to the Bernoulli distribution, Bern(p), and we observe $V = 1$ (i.e., a member of the self-transition set was chosen).

To gain greater control over the strength of regularization, let λ be a positive integer and $\mathbf{V}$ be an λ-length sequence of pseudo-observations, drawn i.i.d. according to Bern(p). Since $P(V = 1|\mathbf{Z}; \mathbf{A}) = p$, we have

$$P(\mathbf{V} = \mathbf{1}|\mathbf{Z}; \mathbf{A}) = \left[\prod_{t=2}^{T}\prod_{k=1}^{K} A_{kk}^{z_{(t-1)k} - z_{(t-1)k}z_{tk}}\right]^{\lambda}$$

where $\mathbf{1}$ denotes the all-ones sequence of length λ.

Noting that $\mathbf{V}$ is conditionally independent of $\mathbf{X}$ given the latent state sequence $\mathbf{Z}$, we maximize (with respect to A_{jk}) the expected (with respect to $\mathbf{Z}$) joint log-density over $\mathbf{X}$, $\mathbf{V}$, and $\mathbf{Z}$ parameterized by $\theta = \{\pi, \mathbf{A}, \phi\}$, which are the start-state probabilities, state transition matrix and emission parameters, respectively. Using appropriate Lagrange multipliers, we obtain the regularized maximum likelihood estimate for A_{jk}:

$$A_{jk} = \frac{B_{j,k,T} + \mathbb{1}(j = k)C_{j,k,T}}{\sum_{i=1}^{K} B_{j,i,T} + C_{j,j,T}} \tag{4}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, $\gamma(z_{tk}) = \mathbb{E}[z_{tk}]$ and

$$\begin{aligned} B_{j,k,T} &= \sum_{t=2}^{T} \xi(z_{(t-1)j}, z_{tk}), \\ C_{j,k,T} &= \lambda\left[\sum_{t=2}^{T}[\gamma(z_{(t-1)k}) - \xi(z_{(t-1)j}, z_{tk})]\right]. \end{aligned} \tag{5}$$

The forward-backward algorithm can then be used for efficient computation of the γ and ξ values, as in standard HMMs (Bishop 2007).

Ignoring normalization, we see that

$$A_{jk} \propto \begin{cases} B_{j,k,T} + C_{j,j,T} & \text{if } j = k \\ B_{j,k,T} & \text{otherwise.} \end{cases}$$

Examining the $C_{j,j,T}$ term (i.e., Equation (5)), we see that λ is a multiplier of additional mass contributions for self-transitions, where the contributions are the difference between $\gamma(z_{(t-1)j})$ and $\xi(z_{(t-1)j}, z_{tj})$. These two quantities represent, respectively, the expectation of being in a state j at time $t-1$ and the expectation of remaining there in the next time step. The larger λ or the larger the difference between arriving at a state and remaining there, the greater the additional mass given to self-transition.

Parameter Modifications

Scale-Free Regularization In Equation 2, the strength of the regularization diminishes with growing T, so that asymptotically the regularized estimate and unregularized estimate become equivalent. While this is desirable in many contexts, maintaining a consistent strength of inertial regularization becomes important with time series of increasing length, as is the case with online learning methods. Figure 1 shows a regularized segmentation of human accelerometer data (discussed later in the Experiments section), where the regularization is strong enough to provide good segmentation. If we then increase the number of data points in each section by a factor of ten while keeping the same regularization parameter setting, we see that the regularization is no longer strong enough, as is shown in Figure 2. Thus, the λ parameter is sensitive to the size of the time series.

We desire models where the regularization strength is scale-free, having roughly the same strength regardless of how the time series grows. To achieve this, we define the

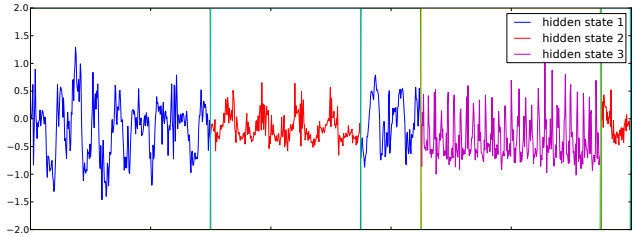

Figure 1: Human activities accelerometer data, short sequence. Vertical partitions correspond to changes of state.

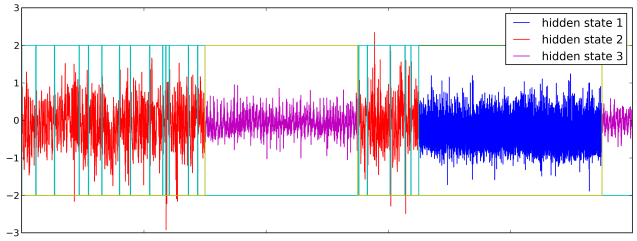

Figure 2: The long sequence human activities accelerometer data using regularization parameter from short sequence.

λ parameter to scale with the number of transitions, namely $\lambda = (T-1)^{\zeta}$, and our scale-free update equation becomes

$$A_{jk} = \frac{((T-1)^{\zeta}-1)\mathbb{1}(j=k) + \sum_{t=2}^{T} \xi(z_{(t-1)j}, z_{tk})}{((T-1)^{\zeta}-1) + \sum_{i=1}^{K}\sum_{t=2}^{T} \xi(z_{(t-1)j}, z_{ti})}. \tag{6}$$

This preserves the effect of regularization as T increases, and ζ becomes our new regularization parameter, controlling the strength of the regularization. For consistency, we also re-parameterize Equation (5) using $\lambda = (T-1)^{\zeta}$.

Towards Parameter-Free Regularization Although our methods require specifying the strength of regularization in advance, one can avoid this requirement by making an assumption concerning the distribution of segment lengths. Namely, by assuming that most of the segment lengths are of roughly the same order-of-magnitude scale, then we can automatically tune the regularization parameter, as follows.

We first define a range of possible regularization parameter values (such as $\lambda \in [0, 75]$), and perform a search on this interval for a value that gives *sufficient regularization*. Sufficient regularization is defined with respect to the Gini ratio (Gini 1936; Wikipedia 2014), which is a measure of statistical dispersion often used to quantify income inequality. For a collection of observed segment lengths $L = \{l_1, \ldots, l_m\}$, given in ascending order, the Gini ratio is estimated by

$$G(L) = 1 - \frac{2}{m-1}\left(m - \frac{\sum_{i=1}^{m} i l_i}{\sum_{i=1}^{m} l_i}\right).$$

We assume that the true segmentation has a Gini ratio less than one-half, which corresponds to having more equality among segment lengths than not. One can perform a binary search on the search interval to find the smallest ζ parameter for which the Gini ratio is at least one-half. This increases the time complexity by a factor of $O(\log_2(R/\epsilon))$, where R is the range of the parameter space.

Experiments

We perform two segmentation tasks on synthetic and real multivariate time series data, using our scale- and parameter-free regularized inertial HMMs. For comparison, we present the results of applying a standard K-state hidden Markov model as well as the sticky HDP-HMM of (Fox et al. 2011). We performed all tasks in an unsupervised manner, with state labels being used only for evaluation.

Datasets

The first (synthetic) multivariate dataset was generated using a two-state HMM with 3D Gaussian emissions, with transition matrix

$$\mathbf{A} = \begin{pmatrix} 0.9995 & 0.0005 \\ 0.0005 & 0.9995 \end{pmatrix},$$

equal start probabilities and emission parameters $\mu_1 = (-1,-1,-1)^{\top}$, $\mu_2 = (1,1,1)^{\top}$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \text{diag}(3)$. Using this model, we generated one hundred time series consisting of ten-thousand time points each. Figure 3 shows an example time series from this synthetic dataset.

The second dataset was generated from real-world forty-five dimensional human accelerometer data, recorded for users performing five different activities, namely, playing basketball, rowing, jumping, ascending stairs and walking in a parking lot (Altun, Barshan, and Tunçel 2010). The data were recorded from a single subject using five Xsens MTx™ units attached to the torso, arms and legs. Each unit had nine sensors, which recorded accelerometer (X, Y, Z) data, gyroscope (X, Y, Z) data and magnetometer (X, Y, Z) data, for a total of forty-five signals at each time point.

We generated one hundred multivariate time series from the underlying dataset, with varying activities (latent states) and varying number of segments. To generate these sets, we first uniformly chose the number of segments, between two and twenty. Then, for each segment, we chose an activity uniformly at random from among the five possible, and selected a uniformly random segment length proportion. The selected number of corresponding time points were extracted from the activity, rescaled to zero mean and unit variance, and appended to the output sequence. The final output sequence was truncated to ten thousand time points, or discarded if the sequence contained fewer than ten thousand points or fewer than two distinct activities. Additionally, prospective time series were rejected if they caused numerical instability issues for the algorithms tested. The process was repeated to generate one hundred such multivariate time series of ten thousand time ticks each, with varying number of segments, activities and segment lengths. An example data sequence is shown in Figure 4 and the distribution of the time series according to number of activities and segments is shown in Figure 5.

Experimental Methodology

We compared performance of four methods on the two datasets described in the previous section: a standard K-

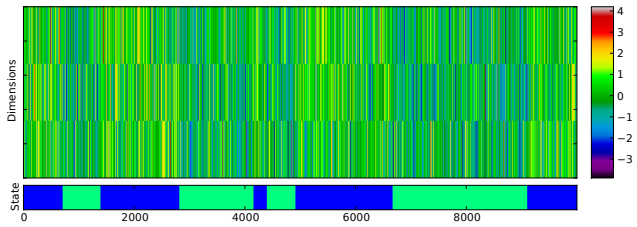

Figure 3: Synthetic data example. Generated from two-state HMM with 3D Gaussian emissions and strong self-transitions.

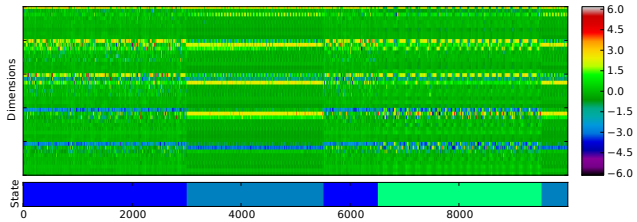

Figure 4: Human activities accelerometer data. Three state, 45-dimensional.

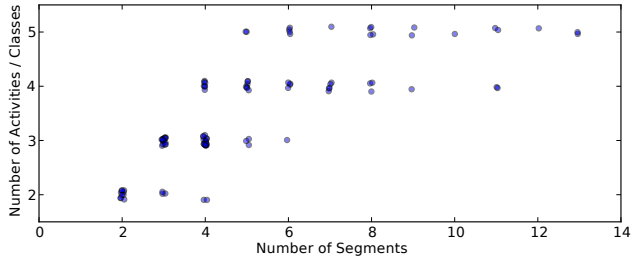

Figure 5: Distribution of Accelerometer Time Series Data.

state hidden Markov model, the sticky HDP-HMM and both inertial HMM variants. The task was treated as a multi-class classification problem, measuring the minimum zero-one loss under all possible permutations of output labels, to accommodate permuted mappings of the true labels. We measured the normalized variation of information (Meilă 2003) between the predicted state sequence and true state sequence, which is an information metric capturing the distance between two partitionings of a sequence. We also considered the ratio of predicted number of segments to true number of segments, which gives a sense of whether a method over- or under-segments data, and the absolute segment number ratio (ASNR), which is defined as ASNR $= \max(S_t, S_p)/\min(S_t, S_p)$, where S_t is the true number of segments in the sequence and S_p is the predicted number of segments, and quantifies how much a segmentation method diverges from the ground truth in terms of relative factor of segments. Lastly, we tracked the number of segments difference between the predicted segmentation and true segmentation and how many segmentations we done perfectly, giving the correct states at all correct positions.

Parameter selection for the inertial HMM methods was done using the automated parameter selection procedure described in the Parameter Modifications section. For faster evaluation, we ran the automated parameter selection process on ten randomly drawn examples, averaged the final ζ parameter value, and used the fixed value for all trials. The final ζ parameters are shown in Tables 1 and 2.

To evaluate the sticky HDP-HMM, we used the publicly available HDP-HMM toolbox for MATLAB, with default settings for the priors (Fox and Sudderth 2009). The Gaussian emission model with normal inverse Wishart (NIW) prior was used, and the truncation level L for each example was set to the true number of states, in fairness for comparing with the HMM methods developed here, which are also given the true number of states. The "stickiness" κ parameter was chosen in a data-driven manner by testing values of $\kappa = 0.001$, 0.01, 0.1, 1, 5, 10, 50, 100, 250, 500, 750 and 1000 for best performance over ten randomly selected examples each. The mean performance of the 500th Gibbs sample of ten trials was then taken for each parameter setting, and the best κ was empirically chosen. For the synthetic dataset, a final value of $\kappa = 10$ was chosen by this method. For the real human accelerometer data, a value of $\kappa = 100$ provided the best accuracy and relatively strong variation of information performance. These values were used for evaluation on each entire dataset, respectively.

To evaluate the HDP-HMM, we performed five trials on each example in the test dataset, measuring performance of the 1000th Gibbs sample for each trial. The mean performance was then computed for the trials, and the average of all one hundred test examples was recorded.

Synthetic Data Results

As seen in Table 1, the MAP regularized HMM had the strongest performance, with top scores on all metrics. The inertial pseudo-observation HMM also had strong performance, with extremely high accuracy and low variation of information. The standard HMM suffered from over-

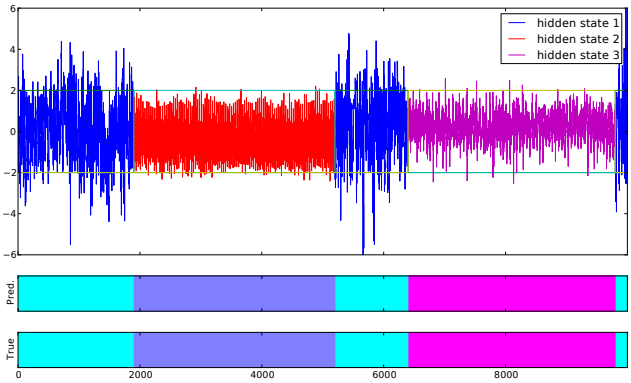

Figure 6: Example segmentation of human activities accelerometer data using inertial (MAP) HMM.

Table 1: Results from quantitative evaluation on 3D synthetic data. Statistical significance is computed with respect to MAP results.

Method	Acc.	SNR	ASNR	SND	VOI	Per.
HDP-HMM ($\kappa = 10$)	0.85*	0.59*	3.50*	2.79*	0.56*	0/100
Standard HMM	0.87*	172.20*	172.20*	765.91*	0.62*	0/100
MAP HMM ($\zeta = 2.3$)	**0.99**	**0.96**	**1.13**	**0.51**	**0.07**	**2/100**
PsO HMM ($\zeta = 8.2$)	**0.99**	0.87‡	1.43‡	1.15*	0.14†	1/100

Acc. = Average Accuracy (value of 1.0 is best)
SNR = Average Segment Number Ratio (value of 1.0 is best)
ASNR = Average Absolute Segment Number Ratio (value of 1.0 is best)
SND = Average Segment Number Difference (value of 0.0 is best)
VOI = Average Normalized Variation of Information (value of 0.0 is best)
Per. = Total number of perfect/correct segmentations

paired t-test: † $< \alpha = .05$, ‡ $< \alpha = .01$, * $< \alpha = .001$

segmentation of the data (as reflected in the high SNR, ASNR, and SND scores), while the sticky HDP-HMM tended to under-segment the data. All methods were able to achieve fairly high accuracy.

Human Activities Accelerometer Data Results

Table 2: Results from real 45D human accelerometer data.

Method	Acc.	SNR	ASNR	SND	VOI	Per.
HDP-HMM ($\kappa = 100$)	0.60*	0.75‡	4.68*	5.03‡	0.95*	0/100*
Standard HMM	0.79*	134.59*	134.59*	584.16*	0.38*	9/100*
MAP HMM ($\zeta = 33.5$)	**0.94**	1.28	1.43	2.62	**0.14**	**48/100**
PsO HMM ($\zeta = 49.0$)	**0.94**	**1.03**†	**1.29**	**1.29**	0.15	**48/100**

paired t-test: † $< \alpha = .05$, ‡ $< \alpha = .01$, * $< \alpha = .001$

Results from the human accelerometer dataset are shown in Table 2. Both the MAP HMM and inertial pseudo-observation HMM achieved large gains in performance over the standard HMM model, with average accuracy of 94%. Furthermore, the number of segments was close to correct on average, with a value near one in both the absolute and simple ratio case. The average normalized variation of information was low for both the MAP and pseudo-observation methods. Figure 6 shows an example segmentation for the MAP HMM, displaying a single dimension of the multivariate time series for clarity.

In comparison, a standard hidden Markov model performed poorly, strongly over-segmenting the sequences in many cases. Even more striking was the improvement over the sticky HDP-HMM, which had an average normalized variation of information near 1 (i.e., no correlation between the predicted and the true segment labels). The method tended to *under*-segment the data, often collapsing to a single uniform output state, reflected in the SNR having a value below one, and may struggle with moderate dimensional data, as related by Fox and Sudderth through private correspondence. The sticky HDP-HMM suffers from slow mixing rates as the dimensionality increases, and computation time explodes, being roughly cubic in the dimension. As a result, the one hundred test examples took several days of computation time to complete, whereas the inertial HMM methods took a few hours.

Discussion

Our results demonstrate the effectiveness of inertial regularization on HMMs for behavior change modeling in multivariate time series. Although derived in two independent ways, the MAP regularized and pseudo-observation inertial regularized HMM converge on a similar maximum likelihood update equation, and thus, had similar performance.

The human activity task highlighted an issue with using standard HMMs for segmentation of time series with infrequent state changes, namely, over-segmentation. Incorporating regularization for state transitions provides a simple solution to this problem. Since our methods rely on changing a single update equation for a standard HMM learning method, they can be easily incorporated into HMM learning libraries with minimal effort. This ease-of-implementation gives a strong advantage over existing persistent-state HMM methods, such as the sticky HDP-HMM framework.

While the sticky HDP-HMM performed moderately well on the low-dimensional synthetic dataset, the default parameters produced poor performance on the real-world accelerometer data. It remains possible that different settings of hyperparameters may improve performance, but the cost of a combinatorial search through hyperparameter space combined with lengthy computation time prohibits an exhaustive exploration. The results, at minimum, show a strong dependence on hyperparameter settings for acceptable performance. In contrast, the inertial HMM methods make use of a simple heuristic for automatically selecting the strength parameter ζ, which resulted in excellent performance on both datasets without the need for hand-tuning several hyperparameters. Although the sticky HDP-HMM has poor performance on the two segmentation tasks, there exist tasks for which it may be a better choice (e.g., when the correct number of states is unknown).

Related Work

Hidden Markov models for sequential data have enjoyed a long history, gaining popularity as a result of the widely influential tutorial by Rabiner (Rabiner 1989). Specific to the work presented here, the use of regularization for HMM parameters received a general treatment in (Gauvain and Lee 1994), for both transition and emission parameters. Our work details a more specific version of the regularization, useful for state persistence. Neukirchen and Rigoll (Neukirchen and Rigoll 1999) studied the use of regularization in HMMs for reducing parameter overfitting of emission distributions due to insufficient training data, but

without an emphasis on inertial transitioning between states. Similarly, Johnson (Johnson 2007) proposed using Dirichlet priors on multinomial hidden Markov models as a means of enforcing sparse emission distributions.

Fox *et al.* (Fox et al. 2011) specifically developed a Bayesian sticky HMM to provide inertial state persistence. They presented a method capable of learning a hidden Markov model without specifying the number of states or regularization-strength beforehand, using a hierarchical Dirichlet process and truncated Gibbs sampling. As discussed, their method requires a more complex approach to learning the model and specification of several hyperparameters for the Bayesian priors along with a truncation limit. In contrast, our models only require the specification of two parameters, K and ζ, whereas the sticky HDP-HMM requires analogous truncation level L and κ parameters to be chosen, in addition to the hyperparameters on the model priors.

Conclusions

For modeling changes in multivariate time series data, we derive two modified forms of hidden Markov model that effectively enforce state persistence. Although the derived methods are simple, they perform well and are computationally tractable. We have shown that inertial models are easily implemented, run efficiently, add almost no additional computation cost, and work well on data with moderate dimensions. Their simplicity is thus a feature and not a bug.

Furthermore, a simple method was developed for automated selection of each regularization parameter. Our experiments on synthetic and real-world data show the effectiveness of inertial HMMs, giving large improvements in performance over standard HMMs and the sticky HDP-HMM.

The simplicity of our models pave the way for natural extensions, such as incremental parameter learning and changing the form of the class conditional emission distributions to incorporate internal dynamics. Such extensions are the focus of future work.

References

Altun, K.; Barshan, B.; and Tunçel, O. 2010. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recogn.* 43(10):3605–3620.

Bishop, C. M. 2007. *Pattern Recognition and Machine Learning*. Springer. 616–625.

Fox, E. B., and Sudderth, E. B. 2009. HDP-HMM Toolbox. `https://www.stat.washington.edu/~ebfox/software.html`. [Online; accessed 20-July-2014].

Fox, E. B.; Sudderth, E. B.; Jordan, M. I.; Willsky, A. S.; et al. 2011. A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics* 5(2A):1020–1056.

Gauvain, J.-l., and Lee, C.-h. 1994. Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing* 2:291–298.

Gini, C. 1936. On the measure of concentration with special reference to income and statistics. In *Colorado College Publication*, number 208 in General Series, 73–79.

Johnson, M. 2007. Why doesn't EM find good HMM POS-taggers. In *In EMNLP*, 296–305.

Kawahara, Y.; Yairi, T.; and Machida, K. 2007. Change-point detection in time-series data based on subspace identification. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, 559–564. IEEE.

Liu, S.; Yamada, M.; Collier, N.; and Sugiyama, M. 2013. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks* 43:72–83.

Meilă, M. 2003. Comparing clusterings by the variation of information. In Schölkopf, B., and Warmuth, M., eds., *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 173–187.

Neukirchen, C., and Rigoll, G. 1999. Controlling the complexity of HMM systems by regularization. *Advances in Neural Information Processing Systems* 737–743.

Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Ray, B. K., and Tsay, R. S. 2002. Bayesian methods for change-point detection in long-range dependent processes. *Journal of Time Series Analysis* 23(6):687–705.

Wikipedia. 2014. Gini coefficient — Wikipedia, the free encyclopedia. [Online; accessed 8-June-2014].

Willsky, A. S.; Sudderth, E. B.; Jordan, M. I.; and Fox, E. B. 2009. Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, 457–464.

Xie, Y.; Huang, J.; and Willett, R. 2013. Change-point detection for high-dimensional time series with missing data. *Selected Topics in Signal Processing, IEEE Journal of* 7(1):12–27.