Multi-Aspect Target Detection via the Infinite Hidden Markov Model

Kai Ni, Yuting Qi and Lawrence Carin*

Electrical and Computer Engineering Department

Duke University

Durham NC 27708-0291

(Dated: October 17, 2006)

Abstract

A new multi-aspect target detection method is presented based on the infinite hidden Markov model (iHMM). The scattering of waves from a target is modeled as an iHMM with the number of underlying states treated as infinite, from which a full posterior distribution on the number of states associated with the targets is inferred and the target-dependent states are learned collectively. A set of Dirichlet processes (DPs) are used to define the rows of the HMM transition matrix and these DPs are linked and shared via a hierarchical Dirichlet process (HDP). Learning and inference for the iHMM are based on a Gibbs sampler. The basic framework is first demonstrated using synthetic data, followed by a detailed analysis of measured acoustic scattering data.

PACS numbers: 43.60.-c, 43.60.Bf, 43.60.Cg

I. INTRODUCTION

In many acoustic sensing scenarios the target is observed from multiple target-sensor orientations (or aspects), and the underlying acoustic scattered waveforms are highly aspect-dependent. It is often difficult to achieve reliable identification based on a single target view, because the waveforms emitted from two different targets may be similar at certain angles and easily confused. This motivates using a sequence of aspect-dependent looks at a single target, with this *sequential* information offering the potential to substantially improve identification performance.

The data observed when sensing a target from T target-sensor orientations are represented as a sequence of T scattered waveforms. Each of the waveforms is typically represented in terms of a set of features, and in the work presented here physics-based matching-pursuits feature extraction¹⁻³ is performed. Hidden Markov models (HMMs) have been successfully applied to model such sequential data for target detection and classification problems^{1,2,4}. Each HMM “state” represents a set of generally contiguous target-sensor orientations over which the signal statistics are relatively invariant. The T feature vectors characteristic of the sequential data implicitly sample a sequence of T target states. The probability of transitioning from one state to another on consecutive measurements is modeled as a Markov process. Furthermore, because the target is usually concealed or distant, the target-sensor orientation is unknown and the actual sampled state sequence is hidden, motivating an HMM.

In the context of target detection and classification using HMMs, a key issue is to develop a methodology for defining an appropriate set of states. In some problems the state decomposition may be inferred from the underlying physics. For example, acoustic scattering from an elastic target is typically categorized by diffractive⁵⁻⁷, resonant⁸, and chirp-like (dispersive) scattering⁹; hence, ideally an HMM state may be defined by a set of angles for which a particular class of physics dominates. However, in many problems it is

*{kn6,yuting,lcarin}@ee.duke.edu

not possible to easily find a relationship between the physics and the underlying state. In previous work the state decomposition has been performed in an *ad hoc* manner^{1,2,4}, requiring trial and error to manually select the model structure, (e.g., number of states). In the work reported here we investigate the idea of an *infinite* hidden Markov model (iHMM), which by construction has an infinite number of hidden states. We use a set of Dirichlet processes (DPs)¹⁰ to build the iHMM. The Dirichlet process is often used to define mixture models with an infinite number of components^{11,12}, but the simple DP form is insufficient for an iHMM; in an iHMM each row of the transition matrix is modeled as a DP and multiple rows are linked because they share the same set of hidden states. Therefore we use a hierarchical Dirichlet process (HDP) to constitute the iHMM. Proposed by Teh *et al.*^{13,14}, the HDP is a nonparametric Bayesian model for sharing clusters among related groups. However, the HDP assumes a fixed partition of groups of data, while the group partition is random in an iHMM. In addition, the iHMM must utilize the assumed underlying temporal information of the sequential data, which is not considered in the original HDP.

Another limitation of many target detection and classification algorithms is that they usually require substantial training data, assumed to be similar to the data on which the algorithm is tested. Unfortunately, in many sensing applications one often has limited training data. In acoustic target detection problems, for instance, one may have multiple sets of limited scattering data, with each data set collected from different but related targets. Rather than building models for each target individually, as adopted in^{1,2,15}, it is desirable to appropriately share the information among these related targets, thus offering the potential to improve overall identification performance. The iHMM effectively solves this problem using the HDP framework.

Teh *et al.*^{13,14} give a brief introduction to applying HDP for the iHMM. Here we perform a more expansive study of information sharing between multiple HMMs using the iHMM, and provide a detailed inference algorithm. We show results and analysis for continuous HMMs, and give examples of a new application, in particular multi-aspect target detection using the iHMM. We demonstrate how the iHMM may be used to build a single model

for a class of targets, and how the learned model yields information about the physical characteristics and the relationships between the targets within the class. These ideas are demonstrated using measured acoustic scattering data, with comparison to more traditional techniques.

The remainder of the paper is organized as follows. In Section II we briefly review the basic technical definition of DPs and discuss the DP mixture model. Section III introduces the HDP and provides its interpretation. With these representations as background, we describe in Section IV the iHMM and the corresponding Markov chain Monte Carlo (MCMC) algorithm for inference. We present experimental results on both synthetic data and measured underwater scattering data in Section V. Conclusions are summarized in Section VI.

II. DIRICHLET PROCESSES

Let H be a measure on a space Θ , and let γ be a positive real number. A Dirichlet process (DP)¹⁰ is a distribution for a random density function $G(\theta)$, denoted by $G \sim DP(\gamma, H)$; while most distributions are defined on random variables, we emphasize that $DP(\gamma, H)$ is a distribution on the distribution G , the latter defined for an associated set of random variables θ . The “base” distribution H provides the prior information of G with $E[G] = G_0$, and the concentration parameter γ controls how similar G is to H .

The density function G drawn from $DP(\gamma, H)$ is discrete with probability one¹⁰, a property made explicit by the stick-breaking construction¹⁶

$$G(\theta) = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^*}, \quad (1)$$

where $\delta_{\theta_k^*}$ is a discrete measure concentrated at θ_k^* . The countably infinite random parameters $\{\theta_k^*\}_{k=1}^{\infty}$ are independently sampled from the base measure H , which is assumed nonatomic (i.e., $H(\theta^*) = 0$ for any θ^*). Once the set of θ^* ’s (the support of G) is instantiated, a θ drawn from G takes value θ_k^* with probability β_k . The weights $\{\beta_k\}_{k=1}^{\infty}$ are defined by a Beta

distribution that partition a unit-length “stick” as follows

$$\beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) \quad \beta'_k \sim \text{Beta}(1, \gamma), \quad \text{for } k = 1, 2, \dots \quad (2)$$

The discreteness of G implies that samples drawn from G have a clustering property. Let $(\theta_1, \theta_2, \dots, \theta_n)$ be i.i.d random variables sampled from G . If we integrate over G , we have the conditional distribution of θ_i as

$$\theta_i | \boldsymbol{\theta}^{-i}, \gamma, H \sim \sum_{k=1}^M \frac{n_k}{i-1+\gamma} \delta_{\theta_k^*} + \frac{\gamma}{i-1+\gamma} H, \quad (3)$$

where $\theta_1^*, \dots, \theta_M^*$ denote the distinct values (clusters) taken by the set $\boldsymbol{\theta}^{-i} = \{\theta_1, \theta_2, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n\}$ and n_k is the number of θ 's equal to θ_k^* ¹⁷. Equation (3) indicates that the DP framework encourages parameter sharing, and the number of clusters is flexible and inferred from the data (through the likelihood function used to yield the posterior for the parameters¹¹). The hyperparameter γ controls how often a new cluster is created. In the limit as $\gamma \rightarrow \infty$, each θ_i is a cluster and drawn independently from H , resulting in $G = H$.

A Dirichlet process is commonly used as a nonparametric prior distribution for a mixture model with unbounded number of components¹¹. Assume the observation x_i is generated from a distribution $F(\theta)$ with parameter θ_i . The density function on the θ 's is G , which is assumed to be drawn from $DP(\gamma, H)$, and thus we have a Dirichlet process mixture model. A graphical representation of a DP mixture model is given in Fig. 1(a), where G is represented by the stick-breaking construction (1). Indicator variable z_i denotes the mixture component generating the data point $x_i \sim F(\theta_{z_i}^*)$, i.e., $\theta_i = \theta_{z_i}^*$. Unlike standard mixture models with which we have to set a fixed initial number of mixture components, with DP as a prior on G the number of mixture components is treated as infinite and the actual (finite) components used by the mixture model are inferred automatically from the data. For example, if each mixture component $F(\theta)$ is a Gaussian distribution with θ representing the mean and variance, using DP as a prior yields the infinite Gaussian mixture model¹⁸.

III. HIERARCHICAL DIRICHLET PROCESSES

In many problems there are several groups of data and each group may be described by a mixture model; moreover, we assume that some of the groups are related, such that data from one group may be shared with data from another, to consequently improve the models of each. However, we know neither the number of mixture components used in each group, nor the sharing structure between the groups. Teh *et al.*^{13,14} proposed a nonparametric Bayesian approach, namely, the hierarchical Dirichlet process (HDP) to describe a set of coupled mixture models. Assume we have J groups of data. To construct an HDP, a global probability measure G_0 is first drawn from a Dirichlet process $G_0 \sim DP(\gamma, H)$, and then G_j 's are drawn independently from $DP(\alpha, G_0)$ for each group:

$$\begin{aligned} G_j &| \alpha, G_0 \sim DP(\alpha, G_0) \\ G_0 &| \gamma, H \sim DP(\gamma, H). \end{aligned} \tag{4}$$

The discreteness of G_0 (as shown in (1)) guarantees that the G_j 's will reuse the same set of shared mixture components defined in G_0 but with different proportions¹³:

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^*}, & \{\theta_k^*\}_{k=1}^{\infty} &\sim H \\ G_j &= \sum_{k=1}^{\infty} w_{jk} \delta_{\theta_k^*}, & \mathbf{w}_j &\sim DP(\alpha, \boldsymbol{\beta}). \end{aligned} \tag{5}$$

where $\boldsymbol{\beta} = \{\beta_1, \beta_2, \dots\}$ is an infinite-dimensional probability mass function (pmf) defined in (2) and $\mathbf{w}_j = \{w_{j1}, w_{j2}, \dots\}$ is also an infinite-dimensional pmf.

The graphical model of an HDP mixture is shown in Fig. 1(b). Assume θ_{ji} is the parameter associated with datum x_{ji} in group j , where each group has N_j data samples. Similar to the DP mixture model, let $\{\theta_k^*\}_{k=1}^{\infty}$ be the set of shared mixture components defined in G_0 and z_{ji} is the indicator variable for which $\theta_{ji} = \theta_{z_{ji}}^*$, and each data point is generated independently from $x_{ji} \sim F(\theta_{z_{ji}}^*)$. According to (5), the HDP mixture model can

be written as

$$\begin{aligned}
x_{ji} &| z_{ji}, \{\theta_k^*\}_{k=1}^\infty \sim F(\theta_{z_{ji}}^*) \\
\{\theta_k^*\}_{k=1}^\infty &| H \sim H \\
z_{ji} &| \mathbf{w}_j \sim \text{Mult}(\mathbf{w}_j) \\
\{\mathbf{w}_j\}_{j=1}^J &| \alpha, \boldsymbol{\beta} \sim \text{DP}(\alpha, \boldsymbol{\beta}) \\
\boldsymbol{\beta} &| \gamma \sim \text{Stick}(\gamma).
\end{aligned} \tag{6}$$

where $\text{Stick}(\gamma)$ are the weights from the stick-breaking process with parameter γ and $\text{Mult}(\mathbf{w}_j)$ is a multinomial distribution with parameter $\mathbf{w}_j$.

The generative process underlying HDP can be understood as the Chinese restaurant franchise (CRF)¹³. In this setup, there is a globally shared menu G_0 across J restaurants. The customers x_{ji} (associated with θ_{ji}) in restaurant j sitting at table t share a single dish θ_{jt}^* . G_j defines the j^{th} restaurant sharing structure defined by the customer sitting plan. Integrating out the G_j using (3), we have

$$\theta_{ji} | \boldsymbol{\theta}_j^{-i}, \alpha, G_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{N_j - 1 + \alpha} \delta_{\theta_{jt}^*} + \frac{\alpha}{N_j - 1 + \alpha} G_0, \tag{7}$$

where n_{jt} is the number of customers in restaurant j sitting at table t and sharing parameter (dish) θ_{jt}^* and $\boldsymbol{\theta}_j^{-i} = \{\theta_{j1}, \dots, \theta_{ji-1}, \theta_{ji+1}, \dots, \theta_{jN_j}\}$. The globally shared menu G_0 plays the role of the distribution over dishes θ_{jt}^* across the franchise. If we integrate out G_0 , we have the sharing structure among the restaurants, which links all the dishes in the franchise

$$\theta_{jt}^* | \boldsymbol{\Theta}_1^*, \dots, \boldsymbol{\Theta}_{j-1}^*, \boldsymbol{\Theta}_j^{*(-t)}, \gamma, H \sim \sum_{k=1}^M \frac{m_k}{\sum_k m_k + \gamma} \delta_{\theta_k^*} + \frac{\gamma}{\sum_k m_k + \gamma} H, \tag{8}$$

where $\boldsymbol{\Theta}_r^* = \{\theta_{r1}^*, \dots, \theta_{rT_r}^*\}$ indicates all the dishes in a particular restaurant r , $\boldsymbol{\Theta}_j^{*(-t)}$ indicates all the dishes in restaurant j except the dish θ_{jt}^* , m_k is the number of restaurant dishes (double subscriptions) equal to θ_k^* in *all* restaurants and M denotes the number of existing kinds of global dishes across the franchise. In the CRF, customers prefer sitting at the table t serving dish θ_{jt}^* where many other people n_{jt} are already seated, as in (7), but sometimes choose a new table. When a new table is created, popular kinds of dishes already

served among the franchise are more likely to be ordered, while a new kind of dish (not used by any restaurant thus far) can also be introduced from H , represented by the right-most term in (8). The dishes (mixture components) in HDP are shared not only within a group, but also across different groups.

IV. INFINITE HIDDEN MARKOV MODEL

A. Hidden Markov models

The idea of using HDP to describe coupled mixture models with unbounded components can be applied to solve fundamental problems in more complex models, such as the hidden Markov model (HMM). Hidden Markov models have been used widely for modeling sequential data, such as in speech recognition¹⁹ and target classification^{1,2,20}. An HMM includes a sequence of observations $\mathbf{o} = \{o_1, o_2, \dots, o_T\}$ and a sequence of hidden underlying states $\mathbf{s} = \{s_1, s_2, \dots, s_T\}$ which follow a Markov process. In many applications, the first order Markov chain is used, i.e., the current state variable only depends on the previous state regardless of the history $P(s_t | s_1, s_2, \dots, s_{t-1}) = P(s_t | s_{t-1})$; the observation is then independently generated given the state.

Typically maximum-likelihood HMM parameter learning is performed via the EM algorithm¹⁹, with the number of states assumed known. Assume we use an HMM with M states and C possible observations (later we generalize to the continuous HMM, for which the observations are not discretized). The parameters of the model are $(\pi^M, A^{M \times M}, B^{M \times C})$, with π being the initial-state probability, A the transition matrix of $P(s_t | s_{t-1})$ and B the observation matrix of $P(o_t | s_t)$. Figure 2 shows the HMM model.

An M -state HMM can be considered as a set of M coupled mixture models, where the mixture components are shared among all the models. Given the hidden state (random variable) $s_{t-1} = j$, we have the joint distribution of s_t and o_t as $p(o_t, s_t = i | s_{t-1} = j) = a_{ji} b_i(o_t)$, where a_{ji} is the i^{th} element of j^{th} row in A , and $b_i(o_t)$ is the i^{th} row distribution of

B. Summing over s_t we have

$$p(o_t | s_{t-1} = j) = \sum_{i=1}^M a_{ji} b_i(o_t). \quad (9)$$

This is a finite mixture model where the row of A is the mixture weights and B defines the mixture components. In an M -state HMM there are M such mixture models. Each row of A defines distinct mixture weights for each model and all the models share the same set of mixture components (M components) defined in B . For datum o_t , the random variable s_{t-1} selects the mixture model to be used and the random variable s_t selects the component from which o_t is generated. The form of $b_i(\cdot)$ varies depending on the properties of the data: if the data are discretized, we use a multinomial distribution for $b_i(\cdot)$; and if continuous features are adopted, we use a Gaussian distribution for each $b_i(\cdot)$ and the continuous HMM may be viewed as M coupled Gaussian mixture models.

B. HMM for multi-aspect target sensing

Using HMMs for multi-aspect target detection and classification has been described in^{1,2,4}. Here we provide a brief summary. We use this discussion to motivate the need for the iHMM, explaining for example that the number of underlying HMM states is typically unknown *a priori*. Moreover, the HMM is only a model for the target scattering, and therefore there may not be a single “correct” number of states, motivating the need for inferring a full posterior for the underlying number of states.

The data collected when sensing a target from T target-sensor orientations are represented as a sequence of T scattered waveforms. Although the scattered waveforms are strongly dependent on the (hidden) target-sensor orientation, there is a set of “states” that represent a set of angles over which the scattering physics is relatively stationary. Within each state, the signal statistics characterize the variability in the scattering physics at the corresponding set of target-sensor orientations.

To a good approximation the transitions from one state to another on consecutive measurements can be modeled as a Markov process. Moreover, the target is usually distant or

concealed and the actual target-sensor orientation (sampled state) is unknown. This motivates an HMM for representation of each target. The waveforms are typically represented in terms of sets of features, which are here extracted using physics-based matching-pursuits¹⁻³.

For target detection and classification, assume target i is assigned a distinct HMM_i , trained on the labeled sequential data associated with the target. Once all the target-dependent HMMs are built, the likelihood that testing data $\mathbf{o} = \{o_1, o_2, \dots, o_T\}$ is associated with target i is

$$p(\mathbf{o} \mid \theta_i) = \sum_{\mathbf{s}} \pi_{s_1} \cdot \prod_{t=2}^T a_{s_{t-1}s_t} \cdot \prod_{t=1}^T b_{s_t}(o_t), \quad (10)$$

where the a 's and b 's are elements of the matrices A and B . The testing data sequences are deemed to be associated with the training target for which the corresponding HMM yields the highest likelihood (i.e., we assume a prior that states that all targets are equally probable). If the likelihoods of all HMMs are below a prescribed threshold, the unknown object is classified as clutter, i.e., as none of the targets associated with the testing data.

The traditional method for estimating HMM parameters is the Expectation Maximization (EM) method, specified by the Baum-Welch algorithm¹⁹. However, in this estimation procedure the model structure must be specified in advance, i.e., the number of states is fixed. Rather than performing model selection to select a fixed number of states²¹, we employ a fully Bayesian approach in which the number of states is not fixed *a priori*. The nonparametric HDP is employed to implicitly construct an HMM with an infinite number of states (iHMM), with the posterior on the number of states usually peaked about a finite number of states characteristic of the sequential data used for model training.

C. iHMM via HDP

As described in Section IV.A, an M -state HMM can be considered as a set of M coupled finite mixture models (each with M mixture components). The value of state s_{t-1} indexes a specific row of the transition matrix serving as the mixture weights for choosing current state s_t . Given the value of s_t , the observation o_t is sampled from the observation model

(mixture component) indexed by s_t . To consider an infinite number of states, it is natural to use a set of state-specific DPs, one for each value of the state. Furthermore, these DPs must be shared because they use the same set of mixture components defined in the observation matrix. This is similar to the HDP mixture model but with a key difference being that the data (observations) to group partition is fixed in the HDP while the group partition is random (indexed by the hidden previous state) in the iHMM.

The HDP construction of the iHMM for multi-aspect target sensing is shown in Fig. 3, with parameters defined as

$$\begin{aligned}
o_{jt} \mid s_{jt}, \{\theta_k^*\}_{k=1}^\infty &\sim F(\theta_{s_{jt}}^*) \\
\{\theta_k^*\}_{k=1}^\infty \mid H &\sim H \\
s_{jt} \mid s_{jt-1}, \{\mathbf{w}_n\}_{n=1}^\infty &\sim \text{Mult}(\mathbf{w}_{s_{jt-1}}) \\
\{\mathbf{w}_n\}_{n=1}^\infty \mid \alpha, \boldsymbol{\beta} &\sim \text{DP}(\alpha, \boldsymbol{\beta}) \\
\boldsymbol{\beta} \mid \gamma &\sim \text{Stick}(\gamma),
\end{aligned} \tag{11}$$

where $\mathbf{w}_n$ corresponds to the row of transition matrix A , and $F(\theta_k^*)$ corresponds to the observation model $b_k(\cdot)$. Each sampled scattered waveform (observation) is represented with an L -dimensional feature vector $o_{jt} = [o_{jt}^1, \dots, o_{jt}^L]$. Each feature vector is assumed to be generated from a model F . When considering a continuous HMM, F is an L -dimensional Gaussian and θ represents the mean vector and covariance matrix. As discussed further below, the base distribution H is selected to be conjugate to F for better computational speed. For a discrete HMM θ represents the parameters of a multinomial distribution.

Considering the iHMM from the Chinese restaurant franchise view of the HDP, without loss of generality, we write the sequence of observations as $\{o_1, \dots, o_T\}$ and the corresponding hidden state sequence as $\{s_1, \dots, s_T\}$; the customer is o_t and o_t enters the restaurant indexed by the state s_{t-1} . The customer chooses table s_t according to $A(s_{t-1}, :)$, the transition probability from s_{t-1} . If the customer sits at an existing table, she eats the same dish $b_{s_t}(\cdot)$ on that table; the parameters associated with $b_{s_t}(\cdot)$ are the mean and covariance for continuous observations, and a probability mass function for discrete observations. If a new

table is chosen, a dish is selected according to its popularity among all the restaurants and sometimes a new kind of dish is chosen from H . Note that for customer o_t , the previous state s_{t-1} is the restaurant (group) index and the current state s_t is the table (component) index in that restaurant, while for customer o_{t+1} , the s_t becomes a restaurant (group) index. Thus the group index and the component index are coupled. This is a key difference between the HDP and the iHMM. Furthermore, the group assignment is decided by the hidden state, which is from a countably infinite set, thus there are also a countably infinite number of groups. A comparison of the HDP and the iHMM is given in Fig. 4.

D. Inference and learning

We utilize a Markov chain Monte Carlo (MCMC) sampling scheme for the iHMM. Specifically, the iHMM learning is based on an extension of the HDP Gibbs sampler^{13,14} with temporal information considered.

Assume we have a sequence of observations $\mathbf{o} = \{o_1, \dots, o_T\}$ associated with a hidden state sequence $\mathbf{s} = \{s_1, \dots, s_T\}$. The group partition of o_t is indexed by the previous hidden state s_{t-1} . Given the initialization of the hidden state sequence $\mathbf{s}$, we have the group partition plan of the observations. Let x_{ji} (corresponding to an o_t) denote the i^{th} observation in group j , and x_{ji} is drawn from $F(\theta_{z_{ji}}^*)$, with $z_{ji} = k$ denoting the mixture component associated with x_{ji} . Thus, we have two equivalent views of the data: the sequential data view $(\mathbf{o}, \mathbf{s})$ and the group data view $(\mathbf{x}, \mathbf{z})$, where the groups are coupled via the underlying state sequence. Further let n_{jk} denote the number of data points in group j sharing the same component k , and m_{jk} denote the number of clusters using component k in group j .

In the Chinese restaurant franchise, the data are first assigned to a table t_{ji} , then the tables are assigned to some mixture component k_{jt} . This indirect association from data to mixture component, viewed by Beal *et al.*²², complicates the iHMM because the group partition of the iHMM itself is random. We propose a method for which the mixture components are assigned directly to the data. Considering an instantiation of the global DP

G_0 , we write the mixture weights as $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_M, \beta_u\}$, where β_1 through β_M correspond to the M existing mixture components used by the data, and β_u represents all of the infinite components which are not used yet. The variables of interest in this sampling scheme are $\boldsymbol{\beta}$, $\mathbf{z}$ and $\mathbf{m}$ mentioned above. The hyperparameters are α and γ that are assumed to be fixed at this point.

1. Gibbs sampling of random variables

Sampling $\mathbf{z}$. The conditional distribution of the index variable z_{ji} given the remaining variables can be derived from (7) and (8). Let s_t be the hidden state in the sequential data view and z_{ji} be the equivalent variable in the group data view. Let s_{t+1} , associated with some other z , be the next state value, then z_{ji} can be sampled as follows

$$p(z_{ji} = k \mid \mathbf{z}^{-ji}, \mathbf{m}, \boldsymbol{\beta}) \propto \begin{cases} (n_{jk}^{-ji} + \alpha\beta_k) f_k^{-x_{ji}}(x_{ji}) \frac{\alpha\beta_{s_{t+1}} + n_{ks_{t+1}}}{\alpha + \sum_{k'=1}^M n_{kk'}}, & \text{if } k \in \{1, \dots, M\}; \\ \alpha\beta_u f_{k^{new}}^{-x_{ji}}(x_{ji}) \beta_{s_{t+1}}, & \text{if } k = k^{new}, \end{cases} \quad (12)$$

where the bold variable represents the collection of that kind of variables and the superscript refers to the element left out from the set of variables. For example, $\mathbf{z} = (z_{ji} : \text{all } j, i)$ is the set of all the index variables and $\mathbf{z}^{-ji}$ means the element z_{ji} is removed from the index variables set $\mathbf{z}$.

The posterior distribution of z_{ji} taking value k consists of three parts: incoming probability, observation likelihood and outgoing probability.

- The incoming probability $(n_{jk}^{-ji} + \alpha\beta_k)$ is derived from an HDP prior. The first term in the sum is the number of customers eating dish k in restaurant j , which comes from a single-restaurant DP, and the second term accounts for choosing a new table. A new table is chosen with a probability proportional to α and with probability β_k the dish served on new table is the same as a previously used dish k , and with probability β_u a totally new dish k^{new} is chosen. Note that when a new component (dish) is sampled, we draw $a \sim \text{Beta}(1, \gamma)$ and set $\beta_{k^{new}} = a\beta_u$ and $\beta_u^{new} = (1 - a)\beta_u$. This procedure can be understood as follows: once a new component is instantiated, it breaks the

remaining stick with probability a . Using the fact that the sequence of stick-breaking weights is a length-biased permutation of the weights, as shown in²³, the weight a will have the same distribution as any other stick-breaking weight, i.e., $\text{Beta}(1, \gamma)$.

- The data likelihood component refers to evaluation of the observation x_{ji} given the parameter $\theta_{z_{ji}=k}^*$. Assuming the base measure H is conjugate to data distribution F , we can analytically integrate out the mixture component parameters θ_k^* . Let $f(\cdot|\theta)$ and $h(\cdot)$ denote the density function of F and H respectively, the data likelihood is calculated as

$$f_k^{-x_{ji}}(x_{ji}) = \frac{\int f(x_{ji}|\theta_k^*) \prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'}|\theta_k^*) h(\theta_k^*) d\theta_k^*}{\int \prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'}|\theta_k^*) h(\theta_k^*) d\theta_k^*}. \quad (13)$$

If a new component k^{new} is selected, the data likelihood is simply

$$f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) = \int f(x_{ji}|\theta_k^*) h(\theta_k^*) d\theta_k^*. \quad (14)$$

- The outgoing probability component arises from the sequential property in the Markov chain. Since we assume a first-order Markov chain in our HMM, the conditional distribution of s_t (equivalent to z_{ji}) given all the other states is only dependent on the previous state s_{t-1} and the next state s_{t+1} . If the state s_t takes value k , then the outgoing probability is proportional to the transitions that already happened from k to s_{t+1} in group k , counted as $n_{ks_{t+1}}$, as well as the transitions counted from selecting a new table, given by $\alpha\beta_{s_{t+1}}$. In addition, the counting must be normalized within each group as different groups have different total outgoing transitions. If s_t is the last state of the sequence, then a uniform distribution is adopted for the outgoing probability.

If, as a result of updating z_{ji} , some component k becomes unallocated, i.e., $n_{jk} = 0$ for all j , then the corresponding β_k will become zero and the probability that this component will be used again in the future is zero, since that is always proportional to $n_{jk} + \beta_k$. Therefore we may delete this mixture component as well as adjust the global weight vector β .

Sampling m. The tables in this sampling scheme are represented in terms of m_{jk} , denoting the number of tables serving dish k in restaurant j . Given all the other variables,

the distribution of m_{jk} is a DP with concentration parameter $\alpha\beta_k$ and number of customers n_{jk} . Given n_{jk} customers in restaurant j eating dish k , m_{jk} may be understood as the number of tables occupied by those customers. Recall that a customer x_{ji} will sit at an existing table serving dish k with probability proportional to n_{jk}^{-ji} and choose a new table along with dish k with a probability proportional to $\alpha\beta_k$, thus the number of tables m_{jk} can be regarded as the number of tables in a special balcony which only holds the customers who are eating dish k in restaurant j . This turns out to be a DP with parameter $\alpha\beta_k$ and n_{jk} customers. The exact conditional distribution over m_{jk} has been shown by Antoniuk²⁵

$$p(m_{jk} = m \mid \mathbf{z}, \mathbf{m}^{-jk}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{jk})} s(n_{jk}, m) (\alpha\beta_k)^m, \quad (15)$$

where $s(n, m)$ are unsigned Stirling numbers of the first kind. By definition $s(0, 0) = s(1, 1) = 1$, $s(n, 0) = 0$ for $n > 0$ and $s(n, m) = 0$ for $m > n$. Other entries can be computed as $s(n + 1, m) = s(n, m - 1) + ns(n, m)$.

Sampling for β simply follows the definition of the DP weights

$$(\beta_1, \dots, \beta_M, \beta_u) \mid \mathbf{z}, \mathbf{m} \sim \text{Dir}\left(\sum_j m_{j1}, \dots, \sum_j m_{jM}, \gamma\right). \quad (16)$$

2. Hyperparameters sampling

In the above section we assumed the hyperparameters were fixed for sampling the random variables of interest. Here we relax this assumption by sampling the posterior distribution of hyperparameters given $\mathbf{z}$, $\mathbf{m}$ and $\boldsymbol{\beta}$.

For a given group, the index variables $z_j = \{z_{j1}, \dots, z_{jn_j}\}$ are multinomial distributed with parameter $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_M\}$ associated with that group.

$$p(z_{j1}, \dots, z_{jn_j} \mid \boldsymbol{\pi}) = \prod_{k=1}^M \pi_k^{n_{jk}}, \quad (17)$$

with n_{jk} denoting the number of data points under mixture component k and n_j the total number of data in group j . Since the group proportion $\boldsymbol{\pi}$ is a DP with concentration parameter $\alpha\boldsymbol{\beta}$

$$p(\boldsymbol{\pi} \mid \alpha, \boldsymbol{\beta}) \sim \text{Dir}(\alpha\beta_1, \dots, \alpha\beta_M) = \frac{\Gamma(\alpha)}{\prod_{k=1}^M \Gamma(\alpha\beta_k)} \prod_{k=1}^M \pi_k^{\alpha\beta_k - 1}, \quad (18)$$

we can analytically integrate out $\boldsymbol{\pi}$ to yield

$$p(z_j|\alpha, \boldsymbol{\beta}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_j)} \prod_{k=1}^M \frac{\Gamma(\alpha\beta_k + n_{jk})}{\Gamma(\alpha\beta_k)}. \quad (19)$$

Here we place a vague Gamma prior on the hyperparameter $\alpha \sim \mathcal{G}(a, b) = b^a/\Gamma(a)\alpha^{a-1}e^{-b\alpha}$ and the posterior can be derived from (19)

$$p(\alpha|\mathbf{z}, \mathbf{m}, \boldsymbol{\beta}) \propto \mathcal{G}(a_\alpha, b_\alpha) \prod_{j=1}^M \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_j)} \prod_{k=1}^M \frac{\Gamma(\alpha\beta_k + n_{jk})}{\Gamma(\alpha\beta_k)}. \quad (20)$$

Similarly we obtain the posterior of hyperparameter γ with

$$p(\gamma|\mathbf{z}, \mathbf{m}, \boldsymbol{\beta}) \propto \mathcal{G}(a_\gamma, b_\gamma) \frac{\Gamma(\gamma)}{\Gamma(\gamma + \sum_{jk} m_{jk})} \prod_{k=1}^M \frac{\Gamma(\gamma\beta_k + \sum_j m_{jk})}{\Gamma(\gamma\beta_k)}. \quad (21)$$

Note that γ controls the expected number of mixture components used among all the groups and α governs the distribution over the number of components used in each group independently.

E. Testing likelihoods

At any iteration of Gibbs sampling, we can build an HMM based on the posterior sample of the hidden state sequence $\mathbf{s}$ (equivalent to $\mathbf{z}$). The transition matrix A is obtained by counting the number of transitions given the sampled values of the hidden state sequence, and the observation model is built on all the observed data generated from a particular state. To compute the initialization state probabilities $\boldsymbol{\pi}$, we create a virtual state $s_0 = 0$ with transitions only to s_1 for each of the sequences, therefore $\boldsymbol{\pi}$ is computed from the transitions in the special group 0. In the iHMM, a set of P independent posterior samples (associated with P HMMs) are obtained to represent the posterior distribution over the hidden state sequence, and the testing likelihood for any testing sequence is the average likelihood of the P models.

V. EXPERIMENT RESULTS

A. Discovering the underlying model

We first apply the iHMM to a discrete observation data set generated from a 4-state, 3-symbol HMM with the following transition matrix A and observation matrix B

$$A = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}. \quad (22)$$

The data set contains $K = 100$ data sequences and each sequence contains $T = 10$ observations. The initial state probability is uniform, i.e., $\boldsymbol{\pi} = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$. The iHMM is initialized with random hidden state sequences containing M distinct states. The observation model for the discrete HMM is assumed to be a multinomial distribution and the conjugate base distribution H is a Dirichlet distribution. In Fig. 5 we show the evolution trajectories of the number of underlying states versus the iterations of Gibbs sampling. We start from a variety of initial numbers of states M and in all cases the iHMM posterior for the number of states is peaked at four. Note that only the stable states with enough information (transitions from the state) are plotted in Fig. 5. If a sampled state has a very few transitions from it, that state is not plotted in Fig. 5 but is still considered in the Gibbs sampling procedure.

Given a posterior sample of hidden state sequences, we may obtain the HMM parameters as described in Section IV.E. In particular, we show one posterior sample of model parameters $(A, B, \boldsymbol{\pi})$ obtained by the iHMM in Fig. 6, based on the last Gibbs iteration for initializations $M = [2, 4, 10]$. We compare the iHMM with the conventional HMM in which the EM method was used for estimating the model parameters. The number of states for the EM method is fixed to be four and the initial state probability is set to be the uniform distribution, while the transition matrix and observation matrix are randomly initialized. Note that the iHMM provides a posterior distribution on the HMM parameters rather than fixed parameters which are obtained by the EM method. For the inference we use a set of

P independent samples to compute the testing likelihood, but we use one posterior sample of parameters for illustration purpose in the sequel.

Hinton diagrams²⁶ are used to represent each matrix with a black square representing zero and white square representing a positive value for that entry. The bigger the white square, the larger the value of that entry compared with other entries in that matrix. From Fig. 6 we see that if we reorder the state index, the iHMM is able to discover the correct underlying model regardless of the initial model selected (number of states). In contrast, the EM method requires knowledge of the true number of hidden states. Furthermore, the EM method may converge only to a local minimum, which means that a bad initialization of parameters will result in poor parameter estimation (fifth column in Fig. 6), while the iHMM is guaranteed in theory to yield the global minimum after Gibbs sampling converges.

B. Discovering task relationship

In the previous experiment we demonstrated that the iHMM is able to simultaneously find the proper model structure as well as the model parameters. In this section we consider an HMM mixture model, and demonstrate that the iHMM can share information of the underlying states between different HMMs.

We simulate a mixture of two continuous HMMs $d = \sum_{i=1}^2 w_i HMM_i$, in which there is an overlap of the HMM-dependent states across the two HMMs. The first HMM has four underlying states with transition matrix and initial-state probability

$$A_1 = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \end{pmatrix} \quad \pi_1 = [0.25, 0.25, 0.25, 0.25]. \quad (23)$$

The observations for each state are assumed to be continuous and drawn from a 2-D Gaussian distribution with the same diagonal covariance matrix with unit variances. The Gaussian means for each state are $[-3, 3]$, $[3, 0]$, $[-3, 0]$ and $[0, 0]$. The conjugate base prior H is

assigned to be Normal-inverse Gamma distribution.

Similarly, the second HMM consists of five underlying states with parameters

$$A_2 = \begin{pmatrix} 0.5 & 0 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0 & 0.5 \end{pmatrix} \quad \pi_2 = [0.2, 0.2, 0.2, 0.2, 0.2]. \quad (24)$$

The Gaussian means for the observation models are $[-3, 0], [0, 0], [3, -3], [0, -3], [-3, -3]$, and the covariance matrices are the same as for the first HMM. The observations drawn from different states are overlapped, complicating the learning task. Using a traditional approach, one need first fix the number of states in each HMM then learn the HMM-dependent states separately; the iHMM can automatically find the proper number of states and share the state information between HMMs.

We simulate $K_1 = 100$ data sequences from the first HMM and $K_2 = 125$ data sequences from the second HMM, with each sequence containing $T = 10$ observations. Therefore the mixture weight for the first HMM is $w_1 = 4/9$ and the weight for the second HMM is $w_2 = 5/9$.

Starting from the simplest model with initialization $M = 2$, the iHMM discovers an overarching model peaked around seven distinct states, as shown in Fig. 7. The iHMM reveals there are two states shared between the HMMs, resulting in a total of seven states instead of nine. The model parameters obtained from the last posterior sample are shown in Fig. 8. We observe that the transition matrix obtained by the iHMM may be decoupled into two HMMs if we consider the third and the fourth states are shared between the two HMMs. The decoupled transition matrices are close to the true transition matrices. The initial state probability for the overarching model is $\pi = [0.1022, 0.1244, 0.2311, 0.2178, 0.1200, 0.1200, 0.0844]$. If we decouple π based on the two shared states, we obtain the initial state probability for the first HMM $\pi'_1 = [0.2266, 0.2759, 0.2562, 0.2414]$ (with normalization) and the initial state probability for the second HMM $\pi'_2 = [0.2105, 0.1984, 0.2186, 0.2186, 0.1538]$,

both of which are close to the true π_1 and π_2 . Moreover, if we sum up all the terms associated with each HMM in π , we get the mixture weights $w'_1 = 0.4511$ and $w'_2 = 0.5489$, as anticipated from the underlying mixture weights w_1 and w_2 discussed above.

The observation models learned from the iHMM are depicted in Fig. 9 with black ellipses indicating the true Gaussian models and red ellipses indicating the observation models derived by the iHMM. The contours are plotted for twice the standard deviation. The Gaussian models learned by the iHMM are similar to the true underlying models, which indicates that the iHMM has found both the sharing structure (two HMMs, one with four states and the other with five states, and two states are shared) and the corresponding model parameters (transition matrices and Gaussian observation models).

C. Modeling for acoustic-scattering from real targets

We consider multi-aspect acoustic scattering data from five elastic shell targets. Details on the targets, scattering data and features are described in^{1,2}. We give a brief summary here for completeness. The physical description of the five targets are shown in Fig. 10. The scattered fields are observed as a function of angle with data collected in 1° increments. Features for each target are then extracted from the time-domain scattered fields using matching pursuits¹⁻³. The feature vectors across all the target-sensor orientations and all types of targets are concatenated to perform vector quantization (VQ)²⁷. Therefore the observation is a discrete element mapped from the feature vector to a code within a VQ codebook of size 20.

The scattering data of the five targets are shown in Fig. 11. We plot the magnitude of the frequency response as a function of sensing angle (over 360° target-sensor orientations) for each target. In previous work^{1,2}, five distinct states were defined manually over a segment of 90° , and the remaining states at other angles are manifested using the symmetry of the target reflected in Fig. 10. Here we apply the iHMM to discover the number of states from the training data, without setting the model structure in advance. Note that for this real

data the HMM is simply a model, and therefore there may not be a clear (single) number of states for the observed data. Importantly, the iHMM will provide an estimate of the full posterior on the number of states for each target.

In the training process we use a small number of angular samples from the data. Each data sequence consists of 5 observations that are sampled in 5° increments. We use 114 data sequences from each target for training the model. Figure 12 shows the state evolution trajectories using the iHMM on the five shell targets. We consider iHMM initializations with $M = 2$ and $M = 10$ states and 10000 Gibbs iterations are performed for each case.

From Fig. 12 we observe that the iHMM represents each target with approximately five states, except the third target which is modeled with four hidden states. It is interesting to note that the results obtained from the iHMM (number of states) are similar to the results in². In that work, the model structure is defined manually based upon the physical statistics of the elastic target and each target is represented by five distinct states and the resulting models have been successfully applied in target classification. However, the real targets may not be modeled by a single HMM, and fixing model structure leads to under- or over-fitting. Instead, the iHMM gives a full posterior distribution on the number of states. An example for Target 5 (initialized with $M = 2$) is shown in Fig. 13, running 100000 Gibbs iterations after a 5000 burnin, collecting 400 posterior samples with a spacing of 250 samples. This posterior distribution gives the information of the proper number of states and eliminates the problem of model selection.

D. Information sharing among multiple targets

We have already shown in Section V.B that the iHMM is able to share the data from different HMMs to learn the HMM-dependent states. By using the data from multiple targets (usually each target is modeled by an HMM) collectively, the training data for each target are strengthened and the generalized classification results are improved. In the next experiments, we consider a target detection problem where five shell targets are considered

as “targets of interest”, and six clutter items are considered as false targets. The six false targets are a small missile-like object, a plastic container, a 55-gallon drum, two rocks and a log respectively, with associated data shown in Fig. 14. The details of the shell targets and the clutter are described in^{1,2,20}. The goal is to test if a new target is in the family of shell targets or not. We train on four of the shell targets and the item under test is the remaining one shell target (label 1) or one of the six false targets (label 0). All six of the false targets and the held-out shell target are not seen when training. When testing and evaluating performance, we consider each possible test item, as viewed from all possible initial angles (although only a small subset of angles are considered for a particular testing example). Both the training sequences and the testing sequences were of length 5, and the angular sampling between consecutive measurements was 5° .

We compare development of single HMMs (SHMM) designed for each of the four training targets as well as one overarching HMM designed for all four training targets (class-based iHMM), and in both cases the iHMM is used for model design. We use 40 posterior samples (with a spacing of 250 iterations) to compute each HMM likelihood. The testing likelihood for the SHMM is the maximum of the four individual HMM likelihoods. We note that for the SHMM, there are counterpart models that may be designed using the conventional EM algorithm, one model for each shell. The number of HMM states may be inferred for the target-dependent HMMs using physical considerations (note that for the EM solution the number of states is set *a priori*, and fixed). By contrast, for the class-based iHMM (single iHMM for all four training targets), it is difficult to have an analog computed using the EM algorithm, because the proper number of states is difficult to determine (some states associated with the single targets may be shared, and others not shared, and this decomposition is difficult to determine *a priori*). A significant advantage of the iHMM is that it adaptively determines a posterior on the proper number of states, automatically learning the appropriate state sharing structure across the four training targets. We also note that this class-based iHMM, by sharing data across the training targets, will not model any one of the training targets, but will rather infer relationships between the multiple

training targets within this class. This generalization is beneficial when testing on targets not seen while training, as long as the testing targets of interest are within the class of training data.

This generalization is reflected in the results presented in Fig. 15, for which the class-based iHMM consistently outperforms the iHMMs designed for each of the training targets separately. The results are measured by Receiver Operation Characteristic (ROC) curves. We emphasize that the iHMM does not find a single HMM but rather a full posterior distribution on the model structure and we show in Fig. 16 the posterior distribution on the number of states associated with the class-based iHMM (for case 1 in Fig. 15), running 100000 Gibbs iterations after a 5000 burnin, collecting 400 posterior samples with a spacing of 250 samples.

The performance gain using the class-based iHMM for case 2 and case 3 are smaller compared with other cases. This is due to the fact that shell target 2 and target 3 possess relatively similar scattered field (Fig. 11). In case 2 and 3, the SHMM detects the held-out testing shell target well because of having similar targets in the training set and leaving a small margin for improvement using the class-based iHMM. In other cases, the class-based iHMM provides the ability for sharing state information between shell targets and giving better generalization for unseen shell target.

To see the relationship of target-dependent states across the shell targets, we also trained a class-based iHMM on the collection of *all* five shell targets and plot the hidden state distribution based on the last Gibbs iteration, with the result depicted in Fig. 17. We observe that the underlying states are shared over all targets, showing the utility of using the class-based iHMM as compared to the single HMMs.

As mentioned above, the HMM “state” corresponds to a sector of target-sensor orientations over which the signal statistics are invariant. Due to the target symmetry presented in Fig. 11. The set of states are uniquely defined in a 90° segment. In Fig. 18, we plot the physical angles in relation to the sampled states from 90° to 180° of a target. It can be seen that at certain angles the targets have similar scattering signal, thus they share the statistics

for these angles, while at some other angles, the targets are discriminative. The class-based iHMM explores this sharing structure and strengthens the generative model learning for targets. These types of results have significant potential for learning relationships in the aspect-dependent scattering physics between different targets, solely based on the data.

VI. CONCLUSION

We have considered an extensive study of the infinite hidden Markov model (iHMM), which provides a principled approach to solve the fundamental problem of model selection in hidden Markov models, here applied to multi-aspect acoustic target detection. The iHMM is implemented using a nonparametric tool, namely the hierarchical Dirichlet process (HDP), which infers a full posterior distribution on the number of underlying states from the training data. After discussing the difficulties in extending HDP to iHMM, we have provided a detailed procedures for performing Gibbs sampling for iHMM inference.

Promising iHMM results have been obtained using both synthetic and measured data from real elastic targets. It is shown that the iHMM not only discovers the correct number of hidden states for modeling the data, but also yields the corresponding model parameters in both discrete and continuous HMMs. In addition, our results indicate that the iHMM implicitly shares the underlying state information among different targets, as appropriate. The results on real data demonstrate improved detection performance using the iHMM relative to using other competing methods.

References

- ¹ P. Runkle, P. K. Bharadwaj, L. Couchman, and L. Carin, “Hidden Markov models for multi-aspect target classification,” *IEEE Transactions on Signal Processing*, vol. 47, pp. 2035–2040, 1999.
- ² P. Runkle, L. Carin, L. Couchman, T. J. Yoder, and J. A. Bucaro, “Multi-aspect identification of submerged elastic targets via wave-based matching pursuits and hidden Markov

- models,” *Journal of the Acoustical Society of American*, vol. 104, pp. 937–946, 1999.
- ³ M. McClure and L. Carin, “Matching pursuits with a wave-based dictionary,” *IEEE Transactions on Signal Processing*, vol. 45, pp. 2912–2927, 1997.
 - ⁴ P. K. Bharadwaj, P. Runkle, and L. Carin, “Target identification with wave-based matching pursuits and hidden Markov models,” *IEEE Transactions on Antennas Propagation*, vol. 47, no. 3, pp. 1543–1554, 1999.
 - ⁵ R. Varriere and R. L. Moses, “High-resolution radar target modeling using a modified prony estimator,” *IEEE Transactions on Antennas Propagation*, vol. 40, pp. 13–19, 1992.
 - ⁶ L. C. Potter, D. M. Chiang, R. Carriere, and M. J. Gerry, “A gtd-based parametric model for radar scattering,” *IEEE Transactions on Antennas Propagation*, vol. 43, pp. 1058–1067, 1995.
 - ⁷ M. McClure, R. C. Qiu, and L. Carin, “On the superresolution identification of observables from swept-frequency scattering data,” *IEEE Transactions on Antennas Propagation*, vol. 45, pp. 631–641, 1997.
 - ⁸ C. E. Baum, *The singularity expansion method*. New York: Springer-Verlag, 1976.
 - ⁹ D. Kralj, L. Mei, T. T. Hsu, and L. Carin, “Short-pulse propagation in a hollow waveguide: analysis, optoelectronic measurement and signal processing,” *IEEE Transactions on Microwave Theory Technology*, vol. 43, pp. 2144–2150, 1995.
 - ¹⁰ T. S. Ferguson, “A Bayesian analysis of some nonparametric problems,” *Annals of Statistics*, vol. 1, no. 2.
 - ¹¹ M. D. Escobar and M. West, “Bayesian density estimation and inference using mixtures,” *Journal of the American Statistical Association*, vol. 90, pp. 577–588, 1995.
 - ¹² S. N. MacEachern and P. Müller, “Estimating mixture of Dirichlet process models,” *Journal of Computational and Graphical Statistics*, vol. 7, pp. 223–238, 1998.
 - ¹³ Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical Dirichlet processes,” *Journal of the American Statistical Association*, 2006, to be appear.
 - ¹⁴ ———, “Hierarchical Dirichlet processes,” U.C. Berkeley Statistics, Tech. Rep. 653, 2004.
 - ¹⁵ P. Runkle, L. Carin, L. Couchman, T. Yoder, and J. Bucaro, “Multiaspect target identifi-

- cation with wave-based matched pursuits and continuous hidden Markov models,” *IEEE Transactions on Pattern Anal. Machine Intell.*, vol. 21, pp. 1371–1378, 1999.
- ¹⁶ J. Sethuraman, “A constructive definition of Dirichlet priors,” *Statistica Sinica*, vol. 4, pp. 639–650, 1994.
- ¹⁷ D. Blackwell and J. B. MacQueen, “Ferguson distributions via Polya urn schemes,” *Annals of Statistics*, vol. 1, pp. 353–355, 1973.
- ¹⁸ C. Rasmussen, “The infinite gaussian mixture model,” in *Neural Information Processing Systems*, vol. 12, 2000.
- ¹⁹ L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, pp. 275–285, 1989.
- ²⁰ N. Dasgupta, P. Runkle, L. Carin, L. Couchman, T. Yoder, J. Buraco, and G. Dobeck, “Class-based target identification with multi-aspect scattering data,” *IEEE Journal of Oceanic Engineering*, vol. 28, pp. 271–282, 2003.
- ²¹ A. Stolcke and S. Omohundro, “Hidden Markov model induction by Bayesian model merging” *Advances in Neural Information Processing Systems*, vol. 5, pp. 11–18, 1993.
- ²² M. J. Beal, Z. Ghahramani, and C. Rasmussen, “The infinite hidden Markov model,” in *Neural Information Processing Systems*, 2002.
- ²³ J. Pitman, “Random discrete distributions invariant under size-biased permutation,” *Advances in Applied Probability*, vol. 28, pp. 525–539, 1996.
- ²⁴ D. J. C. MacKay, “Ensemble learning for hidden Markov models,” University of Cambridge, Tech. Rep., 1997.
- ²⁵ C. Antoniak, “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems,” *Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.
- ²⁶ G. E. Hinton and T. J. Sejnowski, *Learning and relearning in boltzmann machines*. Cambridge, MA: MIT Press, 1986.
- ²⁷ R. M. Gray, “Vector quantization,” *IEEE ASSP Magazine*, pp. 4–29, 1984.

List of Figures

FIG. 1	(a) Graphical representation of a DP mixture model. Circles denote random variables and arrows denote the dependence. Rounded squares represent fixed hyperparameters and the rectangle denotes replications. The bright circle is the observed data. (b) Graphical representation of a hierarchical Dirichlet process (HDP) mixture model.	31
FIG. 2	Graphical representation of an HMM.	32
FIG. 3	The infinite hidden Markov model interpreted as an HDP. For observation o_{jt} , s_{jt-1} defines the mixture model to be used and s_{jt} selects the mixture component according to infinite dimension weight vector $w_{s_{jt-1}}$	33
FIG. 4	Comparison of grouped data structure under the HDP and the iHMM. . . .	34
FIG. 5	Evolution trajectories of represented states M versus iterations of Gibbs sampling. Each curve represents a distinct initialization of M with $M = [2, 4, 6, 10, 20, 50, 100]$ respectively. In all cases the iHMM posterior for the number of states is peaked at four.	35
FIG. 6	Hinton diagrams for HMM model parameters estimation using the iHMM and a traditional HMM with EM method. The first column (most left) shows the true model used for generating the data. The second through fourth columns correspond the iHMM with initial number of states $M = [2, 4, 10]$, respectively. The last two columns are examples of EM algorithm with a bad initialization as well as a good one. The first row (top) are transition matrices, the second row are observation matrices and the third row are initial state probabilities. Here iHMM parameters are inferred from the last posterior sample of the Gibbs sweeps and EM results are given when the likelihood changing in consecutive iterations below the level of 10^{-4}	36
FIG. 7	Trajectory of state evolution in mixture HMMs problem. Starting from $M = 2$.	37

FIG. 8	Top row: True transition matrices of the two HMMs. Bottom left: iHMM result of transition matrix for the HMM mixture. Two states are shared between the HMMs, resulting in a total of 7 states. Bottom right: iHMM result of initial state probability.	38
FIG. 9	Observation models learned from the iHMM. Black contours are the true models while red contours represent the models derived from the iHMM. Each contour is plotted by 2σ . Blue dots are data drawn from HMM 1 and green dots are data drawn from HMM 2.	39
FIG. 10	Five elastic shell targets, with all units in centimeters. The scattered fields are observed as a function of angle at a fixed distance from the target center, in a plane bisecting the target axis. The nominal external sizes and shapes of the targets are the same, with the acoustic scattering distinguished by the internal structure.	40
FIG. 11	Scattering data of five elastic targets as shown in Fig. 10.	41
FIG. 12	Trajectories of number of underlying states on Shell data shown in Fig 11. Each column is associated with the studying case with different initial number of states. The results for each target are shown in one row.	42
FIG. 13	Posterior distribution on the number of states for Target 5 (initialization with $M=2$).	43
FIG. 14	Scattered fields of six false targets. The six false targets are a small missile-like object, a plastic container, a 55-gallon drum, two rocks and a log respectively.	44
FIG. 15	Detection results for different sets of targets using the SHMM and the class-based iHMM. The results are presented with the respective shell target held-out of the training set.	45
FIG. 16	Posterior distribution on model structure (the number of states) for case 1 in Fig. 15	46

FIG. 17	Hidden state distribution across all five shell targets. Note that many targets within the same class share states	47
FIG. 18	Relationship between physical angles and HMM states for shell targets. Some angles share the same signal statistics (state) while some are distinct from target to target.	48

FIG. 1.

FIG. 2.

FIG. 3.

	HDP (CRF view)	iHMM
Group	Restaurant j (fixed)	s_{t-1} (random)
Data	Customer x_{ji}	o_t
Mixture Component	Dish $\theta_{jt}^* = \theta_k^*$, $k = 1 \sim \infty$, $\theta_k \sim H$	$b_{s_t}(\cdot)$, $s_t = 1 \sim \infty$
DP weights	Popularity w_j	$A(s_{t-1}, :)$
Likelihood	$F(x_{ji} \theta_{jt}^* = \theta_k^*)$	$b_{s_t}(o_t)$

FIG. 4.

FIG. 5.

FIG. 6.

FIG. 7.

Transition matrix of HMM 1

Transition matrix of HMM 2

iHMM result for the mixture model

Initial state probability
derived from iHMM result

FIG. 8.

FIG. 9.

FIG. 10.

FIG. 11.

FIG. 12.

FIG. 13.

FIG. 14.

FIG. 15.

FIG. 16.

FIG. 17.

FIG. 18.