Multidimensional Dirichlet Process-based Non-Parametric Signal Classification for Autonomous Self-Learning Cognitive Radios

Mario Bkassiny, *Student Member, IEEE*, Sudharman K. Jayaweera, *Senior Member, IEEE* and Yang Li, *Student Member, IEEE*

***Abstract*—In this paper, we propose a Bayesian non-parametric signal classification approach for spectrum sensing in cognitive radios (CR's). The proposed classification approach is based on the Dirichlet process mixture model (DPMM) that allows inferring the number and types of signals from their spectral and cyclic properties. The proposed algorithm is completely autonomous and does not require any prior knowledge of the existing signals or the number of distinct signal classes. We assume that the cluster parameters are drawn from a mixture model, where each mixture component parameterizes a specific observation model, including both Gaussian and non-Gaussian models. By using the Gibbs sampling, we estimate the observation model and cluster parameters that best fit the observed data. Given N data points, under certain regularity conditions, we derive an upper bound for the mean-squared error (MSE) in estimating the clusters means. A Bayesian prediction method is also developed to estimate the probability distribution of the data points. The proposed algorithm is applied to detect and classify WiFi and Bluetooth signals in the ISM band. Simulation results validate the proposed classification approach and show its robustness against channel impairments such as Rayleigh channel fading.**

***Index Terms*—Chinese restaurant process, cognitive radio, cyclostationary detection, Dirichlet process mixture model, Gibbs sampling, nonparametric Bayesian statistics, unsupervised learning.**

I. Introduction

In recent years, the concept of cognitive radios (CR's) has been proposed for dynamic spectrum access (DSA). This application was motivated by an FCC report published in 2002 which claimed that a large portion of the spectrum bands is not being utilized most of the time [1]. Thus, many research studies have been focused on using CR's to improve the spectrum utilization by considering CR's as secondary users that try to access the primary channels whenever possible.

However, as it is noted by Mitola [2], the aim of CR's is to improve the quality of information (QoI) of wireless users. DSA is a possible means to achieve this goal. However, it is not the goal itself [2]. Along this line, the authors in [3], [4] proposed a CR architecture that was referred to as the *Radiobot* aimed at achieving: 1) self-learning, 2) self-configuration and 3) spectrum awareness [3], [4].

In order to be aware of its environment, a CR should be able to sense and classify the observed signals. Several feature detection and signal classification methods have been proposed in the literature. For example, [5] proposed a cyclostationarity-based feature detection and a hidden Markov model (HMM)-based signal classification for CR's. However, this technique requires prior training with ideal feature vectors for each signal type, which may not be possible if the CR is operating in an unknown environment without any prior knowledge of the existing signal types. Other classification methods have also been proposed based on neural networks [6] and support vector machines [7], but they also required training data to initialize the classifiers' parameters. On the other hand, feature classification can be performed based on parametric classification approaches such as the Gaussian mixture model (GMM) or K-means algorithm that do not require training data. However, these techniques assume a fixed number of classes, which may not be known in an alien RF environment in which the number of active wireless systems is unknown *a priori*. As an alternative, the authors in [8] proposed to use the X-means algorithm [9] for unsupervised signal classification when the number of clusters is unknown. This approach is based on the K-means algorithms but approximates the number of clusters X by maximizing either the Bayesian information criterion (BIC) or the Akaike information criterion (AIC) [9]. However, similarly to the K-means algorithm, the X-means algorithm assumes spherical Gaussian data, which does not offer enough flexibility when dealing with observations having an arbitrary noise distribution [9]. Moreover, the K-means algorithm can only converge to a local minimum of the distortion measure and its performance heavily depends on the choice of initial center points [9].

To resolve these drawbacks, we resort to non-parametric classification approaches. In particular, the Dirichlet process mixture model (DPMM) that assumes no prior knowledge of the number of clusters [10]. Note that, the DPMM-based classifier is considered to be a Bayesian non-parametric method in the sense of allowing the structure of the model (i.e. number of clusters) to grow with the complexity of the data [10]–[14]. However, the individual observations of the DPMM can still be drawn from parametric distributions. The DPMM-based classifier can infer the number of clusters (or mixture components) from the data itself, making it a suitable candidate for unsupervised and autonomous classifiers. This

M. Bkassiny is with the Department of Electrical and Computer Engineering, State University of New York at Oswego, Oswego, NY, USA, Email: mario.bkassiny@oswego.edu.

S. K. Jayaweera and Y. Li are with the Communication and Information Sciences Laboratory (CISL), Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, USA, Email: {jayaweera, yangli}@ece.unm.edu.

approach has been previously applied for galaxy clustering [15], speaker diarization [16], speaker adaptation [17], image segmentation [18] and compressive sensing [19]. In this paper, we propose the DPMM classification approach to infer the number and types of wireless systems that are sensed by a CR in an unknown environment. The non-parametric nature of the DPMM allows for an arbitrary number of clusters and helps the CR to learn and act autonomously in any RF environment.

Note that, most of the existing DPMM classifiers assume Gaussian observation models, which may not accurately represent complex observations encountered in wireless systems [10], [17]–[23]. In this paper, hence, we extend the DPMM framework to both Gaussian and non-Gaussian observation models by allowing the cluster parameters to be drawn from a mixture model where each mixture component is used to parameterize a particular observation model, including both Gaussian and non-Gaussian distributions. By applying the Gibbs sampling, we determine the observation model that best fits each cluster, while estimating the corresponding parameters. To the best of our knowledge, this is the first DPMM that assumes such a framework, thus offering flexibility in handling arbitrary observation models, as opposed to both K-means and X-means algorithms which assume spherical Gaussian observations [9].

In this paper, we develop the DPMM classifier, as described above, and derive the posterior distribution of the clusters' parameters. The Gibbs sampling is used to sample from the posterior distribution and to update the DPMM hyper-parameters. A Bayesian prediction method is developed to predict the distribution of future feature points, thus allowing the CR to form an RF mapping of the sensed spectrum. We derive an upper bound for the mean-squared error (MSE) in estimating the clusters means, and show that under certain regularity conditions, this upper bound is proportional to $\log N/N$, where N is the number of observed feature points. The proposed algorithm is applied to a cyclostationarity-based feature detection method that was proposed in [21], [22] to provide observations data to the DPMM classifier. The simulation results show that the proposed DPMM-based classifier is able to accurately identify/classify different RF transmissions, in particular, in the ISM band. By comparing our proposed DPMM-based classification algorithm to both K-means and X-means algorithms of [8] and [9], we show, through simulations, that the DPMM-based algorithm achieves superior performance. The efficiency of the DPMM-based classifier stems from the flexibility of the non-parametric DPMM framework and the accuracy of the adopted cyclostationarity-based feature extraction method [21], [22]. Furthermore, since all the DPMM hyper-parameters are updated based on the posterior distributions, the DPMM can accurately approximate the observed model.

The remainder of this paper is organized as follows: Section II gives a description of the DPMM. In Section III, we describe the Bayesian classification method and in Section IV we derive the predictive distribution of the observed feature points. The convergence of the algorithm is discussed in Section V and we derive the MSE of the cluster means in Section VI. Simulation results are presented in Section VII and we conclude the paper in Section VIII. Note that, throughout this paper, we use bold characters to refer to vector and matrix quantities.

II. The Dirichlet Process

A Dirichlet process $DP(\alpha_0, G_0)$ is defined to be the distribution of a random probability measure G over a measurable space $(\Theta, \mathcal{B})$, such that, for any finite measurable partition $(A_1, \cdots, A_r)$ of Θ, the random vector $(G(A_1), \cdots, G(A_r))$ is distributed as a finite dimensional Dirichlet distribution with parameters $(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_r))$ such that:

$$(G(A_1), \cdots, G(A_r)) \sim Dir(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_r)) \ , \tag{1}$$

where $\alpha_0 > 0$. A vector $(X_1, \cdots, X_n) \sim Dir(a_1, \cdots, a_n)$ is said to be distributed according to a Dirichlet distribution with parameters $(a_1, \cdots, a_n)$ if:

$$f\left(x_1, \cdots, x_n | a_1, \cdots, a_n\right) = \frac{\Gamma\left(\sum_{i=1}^{n} a_i\right)}{\prod_{i=1}^{n} \Gamma(a_i)} \prod_{i=1}^{n} x_i^{a_i - 1} \ , \tag{2}$$

subject to $\sum_{i=1}^{n} x_i = 1$, with $x_i > 0$, $a_i > 0$, for all $i = 1, \cdots, n$.

We denote $G \sim DP(\alpha_0, G_0)$ to represent the probability measure G that is drawn from the Dirichlet process $DP(\alpha_0, G_0)$. In other words, G is a *random probability measure* whose distribution is given by the Dirichlet process $DP(\alpha_0, G_0)$ [10]. That is, the realizations G of a Dirichlet process are random *probability distributions*, in contrast with random *variables* or random *processes* that are usually assumed in probabilistic models.

A. Construction of the Dirichlet Process

Since the discrete probability distribution G is drawn from a Dirichlet process, its construction requires special approaches to determine its random parameters. Teh [10] describes several ways of constructing G. A first method is a direct approach that constructs the random probability distribution G based on the *stick-breaking* method. However, this method is impractical since it involves the evaluation of an infinite sum [10]. Interested readers may refer to [10] for a detailed discussion about this method.

A second more practical approach does not define G explicitly. Instead, it characterizes the distribution of the drawings θ of G, given a certain realization G of $DP(\alpha_0, G_0)$. This method constructs G by using the Chinese Restaurant Process (CRP) [10]. The CRP metaphor considers a restaurant with an unbounded number of tables. A customer entering the restaurant is denoted by θ_i, whereas the distinct tables at which the customers sit are denoted by ϕ_k. The i-th customer sits at the table indexed by ϕ_k, with a probability proportional to the number of customers m_k already seated there, and sits at a new table with a probability proportional to α_0 [10]. This construction provides a practical approach to sample the values θ from a distribution G that is drawn from a certain Dirichlet process.

Formally, we let $\theta_1, \theta_2, \cdots$ to be independent identically distributed (i.i.d.) random variables distributed according to G. These random variables are conditionally independent, given

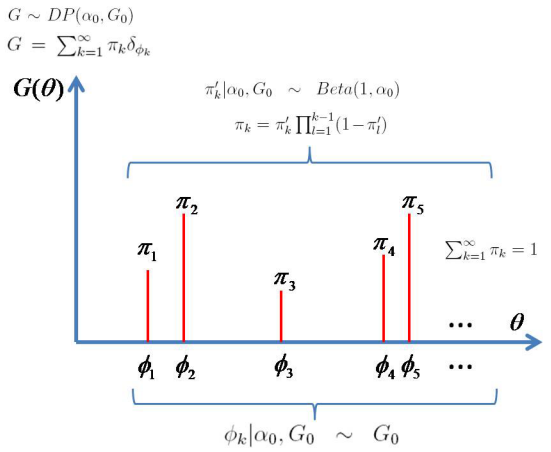


Fig. 1. The Dirichlet process.

G. However, if G is integrated out, $\theta_1, \theta_2, \cdots$ are no more conditionally independent and can be characterized as:

$$\theta_i|\theta_1, \cdots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha_0}\delta_{\phi_k} + \frac{\alpha_0}{i-1+\alpha_0}G_0 \ , \tag{3}$$

where $\{\phi_k\}_{k=1}^K$ are the K distinct values of θ_i's and m_k is the number of θ_i's that are equal to ϕ_k. Note that, this conditional distribution is not necessarily discrete since G_0 might be a continuous distribution (in contrast with G which is discrete with probability 1). The θ_i's that are drawn from G exhibit a clustering behavior since a certain value of θ_i is most likely to re-occur with a nonnegative probability (due to the point mass functions in the conditional distribution). Moreover, the number of distinct θ_i values can be infinite, in general, since there is a nonnegative probability that the new θ_i value is distinct than the previous $\theta_1, \cdots, \theta_{i-1}$. This conforms with the definition of G as a probability mass function over an infinite discrete set. Since, given G, θ_i's are distributed according to G, we denote $\theta_i|G \sim G$.

B. Dirichlet Process Mixture Model (DPMM)

As we have discussed earlier, in Bayesian classification models, a non-parametric classifier assumes that the number of classes (or clusters) can grow with the complexity of the data, as opposed to parametric classifiers which assume that the number of clusters is fixed and known, a priori [10]–[14]. The Dirichlet process makes a perfect candidate for non-parametric classification problems through the Dirichlet process mixture model (DPMM). The DPMM assumes random observations $\mathbf{y}_i$'s that are drawn from a certain mixture model in which the mixture components are identified with the random variables θ_i's drawn from a distribution G from a certain Dirichlet process. This implies that the mixture model may consist of infinitely many mixture components since G is a discrete probability distribution with an infinite support set. Hence, the DPMM endows a non-parametric prior on the parameters of the mixture model [10]. Thus, a DPMM can be defined as follows:

$$\begin{cases} G & \sim & DP(\alpha_0, G_0) \\ \theta_i|G & \sim & G \\ \mathbf{y}_i|\theta_i & \sim & f_{\theta_i}(\mathbf{y}_i) \end{cases} . \tag{4}$$

III. Data clustering based on the DPMM and the Gibbs sampling

Consider a sequence of observations $\mathbf{y}_{1:N} \triangleq \{\mathbf{y}_i\}_{i=1}^N$, where $\mathbf{y}_i \triangleq [y_{i,1}, \cdots, y_{i,d}]^T \in \mathbb{R}^d$, and assume that these observations are drawn from a mixture model. If we do not know the number of mixture components, it is reasonable to assume a non-parametric model, such as the DPMM which allows the number of mixture components to increase with the complexity of the data. Thus, let us assume that the mixture components θ_i are drawn from a $G \sim DP(\alpha_0, G_0)$, for $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$, where ϕ_k are the unique values of θ_i and π_k their corresponding probabilities.

The problem is to estimate the mixture component $\hat{\theta}_i$ for each observation $\mathbf{y}_i$, for all $i \in \{1, \cdots, N\}$. In particular, we are interested in finding maximum a posteriori probability (MAP) estimates of θ_i ($i = 1, \cdots, N$), given the observations $\mathbf{y}_{1:N}$. However, it is hard to find analytical MAP estimates of θ_i's since the joint distribution of $(\theta_1, \cdots, \theta_N)$, given $\mathbf{y}_{1:N}$, is unknown. As an alternative, we may use Monte Carlo methods to compute the MAP estimates by sampling from the posterior distribution of θ_i's, given $\mathbf{y}_{1:N}$ [24], [25]. In particular, in situations that we have the conditional distribution of each θ_i, given the other parameters $\{\theta_j\}_{j\neq i}$, as in (3), we can construct a Markov chain Monte Carlo (MCMC) algorithm based on Gibbs sampling to draw samples from the joint posterior distribution of $(\theta_1, \cdots, \theta_N)$ [26]. The Gibbs sampling algorithm starts with arbitrary estimates of θ_i's and draws samples from the conditional distribution of each parameter θ_i, given the other parameters $\{\theta_j\}_{j\neq i}$, where $\{\theta_j\}_{j\neq i}$ take the values of their most recent estimates [26]. It can be shown that these samples converge in probability to the actual posterior distribution of $(\theta_1, \cdots, \theta_N)$, thus leading to an efficient method for estimating θ_i's [15].

By assuming a DPMM framework, the posterior distribution of $\theta_i|\{\theta_j\}_{j\neq i}, \mathbf{y}_{1:N}$ can be computed as in (5), where $f(\mathbf{y}_i) = \int_\theta f_\theta(\mathbf{y}_i) G_0(\theta) d\theta$ is the marginal distribution of $\mathbf{y}_i$, assuming a prior $G_0(\theta)$, and $f_\theta(\mathbf{y}_i) \triangleq f(\mathbf{y}_i|\theta_i = \theta)$, for all θ's, where θ_i stands for the parameter of observation $\mathbf{y}_i$ [23]. In other words, the assumption of an underlying DPMM for the cluster parameters θ_i's implies that θ_i is equal to θ_j with probability q_j, or it is a new value drawn according to the conditional distributions $f(\theta_i|y_i)$ with probability q_0. Note that, the required posterior distribution $f(\theta_i|\mathbf{y}_i)$ can easily be obtained if θ_i has a conjugate prior for the likelihood $f_{\theta_i}(\mathbf{y}_i)$[1]. In this case, $G_0(\theta_i)$ and $f(\theta_i|\mathbf{y}_i)$ will belong to the same family of distributions. In particular, if both the prior distribution $G_0(\theta_i)$ and the likelihood function $f_{\theta_i}(\mathbf{y}_i)$ are

[1]If the posterior distribution $p(\theta|x)$ is in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a *conjugate prior for the likelihood*. All the members of the exponential family have conjugate priors. In particular, the normal, gamma, exponential, Wishart and inverse-Wishart distributions have conjugate priors [27].

$$\theta_i|\{\theta_j\}_{j\neq i},\mathbf{y}_{1:N}\begin{cases} = \theta_j & \text{with prob.} \quad q_j = \frac{f_{\theta_j}(\mathbf{y}_i)}{\alpha_0 f(\mathbf{y}_i)+\sum_{j=1,j\neq i}^{N} f_{\theta_j}(\mathbf{y}_i)} \\ \sim f(\theta_i|\mathbf{y}_i) & \text{with prob.} \quad q_0 = \frac{\alpha_0 f(\mathbf{y}_i)}{\alpha_0 f(\mathbf{y}_i)+\sum_{j=1,j\neq i}^{N} f_{\theta_j}(\mathbf{y}_i)} \end{cases} . \tag{5}$$

Gaussian, then the posterior distribution $f(\theta_i|\mathbf{y}_i)$ will also be Gaussian. Thus, most of the literature on DPMM problems assumes conjugate priors [10], [18], [23]. In the following, we first present the Gibbs sampling algorithm for the multivariate Gaussian case and then generalize the model to a mixture of Gaussian and non-Gaussian observations.

A. DPMM-based clustering with a Gaussian observation model

A Gibbs sampling algorithm for estimating the parameters θ_i of a DPMM was proposed in [23], which showed that the outcomes of the developed algorithm converge, in probability, to those of the posterior distribution of $(\theta_1, \cdots, \theta_N)$, given $\mathbf{y}_{1:N}$. However, [23] assumed that the prior distribution $G_0(\theta_i)$ can be chosen as a uniform distribution, presuming prior knowledge of the range of the observations, which, in general, may not be available. In addition, it also assumed that the observations $\mathbf{y}_{1:N}$ are distributed according to a standard Gaussian distribution, given the parameters θ_i's. This assumption was relaxed in [15] in which a Bayesian method was proposed to estimate both mean and variance of the Gaussian observation model from the observations $\mathbf{y}_{1:N}$.

In this section, we follow an approach similar to [15] in developing a multi-dimensional Bayesian non-parametric estimator for DPMM's. In the next section, we generalize this method to non-Gaussian observation models.

Let us assume a sequence of observations $\mathbf{y}_{1:N}$ from a DPMM that are normally distributed given the mixture component parameters $\theta_{1:N} \triangleq \{\theta_i\}_{i=1}^{N}$. We may thus denote $\mathbf{y}_i|\theta_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{V}_i)$, where $\theta_i = (\boldsymbol{\mu}_i, \mathbf{V}_i)$ for $i \in \{1, \cdots, N\}$. The prior distribution $G_0(\theta_i)$ can be modeled as the normal/inverse-Wishart conjugate prior such that $\mathbf{V}_i^{-1} \sim W(\mathbf{S}/2, s/2)$[2] and $\boldsymbol{\mu}_i|\mathbf{V}_i \sim \mathcal{N}(\mathbf{m}, \tau\mathbf{V}_i)$, for some mean $\mathbf{m}$ and scale factor $\tau > 0$. Note that, this is the most commonly used conjugate prior distribution for the mean and the covariance matrix of a multivariate Gaussian observation model[3]. Furthermore, a large value of τ implies a large dispersion among the cluster means, whereas parameter $\mathbf{m}$ is a prior estimate of these means [15].

On the other hand, the parameter s reflects the confidence in the value of the covariance matrix $\mathbf{V}_i$. That is, a large value of s corresponds to the case where $\mathbf{V}_i$ is believed to be approximately equal to its prior estimate $\mathbf{S}$. However, a small value of s corresponds to the case where little knowledge is available about $\mathbf{V}_i$ [15].

The posterior distribution $f(\theta_i|\mathbf{y}_i)$ is a bivariate normal/inverse-Wishart distribution whose components are [15]:

$$\begin{aligned} \mathbf{V}_i^{-1} &\sim W\left(\frac{\mathbf{S}_i}{2}, \frac{1+s}{2}\right) , \\ \boldsymbol{\mu}_i|\mathbf{V}_i &\sim \mathcal{N}(\mathbf{x}_i, X\mathbf{V}_i) , \end{aligned}$$

where $\mathbf{S}_i = \mathbf{S} + \frac{(\mathbf{y}_i-\mathbf{m})(\mathbf{y}_i-\mathbf{m})^T}{1+\tau}$, $X = \frac{\tau}{1+\tau}$ and $\mathbf{x}_i = \frac{\mathbf{m}+\tau\mathbf{y}_i}{1+\tau}$. The corresponding weights q_0 and q_j in (5) can shown to be [15]:

$$q_0 \propto \frac{\alpha_0 c(s)}{|\mathbf{M}|^{1/2}} \left(1 + \frac{(\mathbf{y}_i - \mathbf{m})^T\mathbf{M}^{-1}(\mathbf{y}_i - \mathbf{m})}{s}\right)^{-(1+s)/2}$$

and

$$q_j \propto \frac{1}{\sqrt{2|\mathbf{V}_j|}} e^{\frac{-(\mathbf{y}_j-\boldsymbol{\mu}_j)^T\mathbf{V}_j^{-1}(\mathbf{y}_j-\boldsymbol{\mu}_j)}{2}} ,$$

for $j \in \{1, \cdots, N\}$, $j \neq i$ and subject to $\sum_{j=1,j\neq i}^{N} q_j = 1$, with $\mathbf{M} = \frac{1+\tau}{s}\mathbf{S}$ and $c(s) = \Gamma(\frac{1+s}{2})\Gamma(\frac{s}{2})s^{-1/2}$.

We may use the above posterior marginal distribution to perform Gibbs sampling. The resulting number of distinct values of $\theta_{1:N}$ (denoted by $\{\phi_k\}_{k=1}^{K}$) is then an estimate of the number of components (or clusters) in the mixture model. Algorithm 1 summarizes this DPMM classification procedure based on the Gibbs sampling. Upon convergence, the observations $\mathbf{y}_i$'s that share identical values of θ_i's are assumed to belong to the same cluster.

Algorithm 1 Clustering algorithm.

```
Initialize θ̂_i = y_i, ∀i ∈ {1, ··· , N}.
while Convergence condition not satisfied do
  for i = shuffle {1, ··· , N} do
    Use Gibbs sampling to obtain θ̂_i from the posterior
    distribution in (5).
  end for
end while
```

B. DPMM-based clustering with a mixture prior for θ_i

Most of the existing DPMM-based classification problems assume that the observations $\mathbf{y}_{1:N}$ are normally distributed, given the cluster parameters θ_i's [10], [18], [23]. In this paper, however, we relax this condition to allow $\mathbf{y}_i|\theta_i$ to be non-Gaussian distributed. In modifying the likelihood $f_{\theta_i}(\mathbf{y}_i)$, however, we also need to adapt the prior distribution of θ_i accordingly so that it is a conjugate prior for the assumed likelihood. This is necessary since if we were to loose the conjugate property of the prior, a closed-form expression for the posterior distribution of θ_i, as in (5), may not be possible. For example, the Gaussian prior is conjugate for the Gaussian likelihood. However, if we were to use a different likelihood function, such as the log-normal distribution, the Gaussian prior is no more conjugate for this particular likelihood. In

[2]The Wishart distribution $W(\mathbf{V}, n)$ is characterized by a positive definite scale matrix V and n denoting the degrees of freedom.

[3]Note that, families of conjugate priors are not unique. In particular, the set of all probability distributions is always a conjugate prior.

this case, a possible conjugate prior would be the Gamma distribution [28]. Thus, modifying the likelihood $f_{\theta_i}(\mathbf{y}_i)$ should be done in conjunction with adapting the prior distribution of θ_i, accordingly.

Hence, we allow the likelihood function $f_{\theta_i}(\mathbf{y}_i)$ to belong to one of the L different distributions (e.g. Gaussian, or Gamma or log-normal, etc.). The parameter θ_i denotes the distribution parameter and we let $Z_i \in \{1, \cdots, L\}$ to denote the distribution index, which specifies the type of the distribution $f_{\theta_i}(\mathbf{y}_i)$. Clearly, θ_i can be modeled as a *mixture model* of L components where each component is a random parameter drawn from a certain set $\mathcal{S}_l$, for $l = 1, \cdots, L$. The set $\mathcal{S}_l$ contains all possible parameters of the l-th distribution model. By following a Bayesian approach, we can estimate the parameters θ_i's, given the observations $\mathbf{y}_i$'s, by using (5).

We denote a discrete prior distribution for Z_i such that $P\{Z_i = l\} \triangleq \kappa_l$, for $l = 1, \cdots, L$. Given a certain observation model $Z_i = l$ for the observation $\mathbf{y}_i$, we denote the conditional prior distribution of θ_i as $\theta_i|\{Z_i = l\} \sim G_0^{(l)}(\theta_i)$, where $\theta_i \in \mathcal{S}_l$.

We define $f_\theta^{(l)}(\mathbf{y}_i) \triangleq f(\mathbf{y}_i|\theta_i = \theta, Z_i = l)$, for all $\theta \in \mathcal{S}_l$, to be the likelihood function of the observation $\mathbf{y}_i$, given that $Z_i = l$. Thus, we can write $\mathbf{y}_i|\{\theta_i, Z_i\} \sim f_{\theta_i}^{(1)}(\mathbf{y}_i)\mathcal{I}_{\{Z_i=1\}} + \cdots + f_{\theta_i}^{(L)}(\mathbf{y}_i)\mathcal{I}_{\{Z_i=L\}}$, where the indicator function $\mathcal{I}_A$ is defined as $\mathcal{I}_A = 1$ if the event A is true, and 0 otherwise. Note that, the distribution of $\mathbf{y}_i|\{\theta_i, Z_i\}$ is defined for $\theta_i \in \mathcal{S}_{Z_i}$ such that θ_i is a valid parameter for the Z_i-th distribution model.

Under the above formulation, the posterior distribution of the parameter θ_i, given the observation $\mathbf{y}_i$, is defined over the set $\mathcal{S} \triangleq \bigcup_{l=1}^{L} \mathcal{S}_l$ such that:

$$f(\theta_i|\mathbf{y}_i) = \sum_{l=1}^{L} \hat{\kappa}_{l,i} f(\theta_i|\mathbf{y}_i, Z_i = l) \ , \tag{6}$$

where

$$\begin{aligned} \hat{\kappa}_{l,i} &\triangleq P\{Z_i = l|\mathbf{y}_i\} \\ &= \frac{\kappa_l f(\mathbf{y}_i|Z_i = l)}{\sum_{l'=1}^{L} \kappa_{l'} f(\mathbf{y}_i|Z_i = l')} \\ &= \frac{\kappa_l \int_{\theta \in \mathcal{S}_l} f_\theta^{(l)}(\mathbf{y}_i) G_0^{(l)}(\theta) d\theta}{\sum_{l'=1}^{L} \kappa_{l'} \int_{\theta \in \mathcal{S}_{l'}} f_\theta^{(l')}(\mathbf{y}_i) G_0^{(l')}(\theta) d\theta} \ , \end{aligned} \tag{7}$$

and $f(\theta_i|\mathbf{y}_i, Z_i = l) = 0$ if $\theta_i \notin \mathcal{S}_l$. In general, if a closed-form expression can not be obtained for (7), $\hat{\kappa}_{l,i}$ can be evaluated numerically.

The expression in (6) implies that θ_i can be sampled from the posterior distribution $f(\theta_i|\mathbf{y}_i, Z_i = l)$ with a probability $\hat{\kappa}_{l,i}$, for $l = 1, \cdots, L$. In other words, given an observation $\mathbf{y}_i$, the distribution index Z_i is first sampled from the discrete set $\{1, \cdots, L\}$, with corresponding probabilities $\{\hat{\kappa}_{l,i}\}_{l=1}^{L}$. Given the sampled value of Z_i, θ_i can be sampled from $\mathcal{S}_{Z_i}$ using the posterior distribution $f(\theta_i|\mathbf{y}_i, Z_i)$. Furthermore, if $f(\theta_i|\mathbf{y}_i, Z_i = l)$ and $G_0^{(l)}(\theta_i)$ are conjugate for the likelihood $f_{\theta_i}^{(l)}(\mathbf{y}_i)$, $\forall l \in \{1, \cdots, L\}$, then the posterior in (6) can be expressed in closed-form. If not, the posterior may not be derived in closed-form. However, the approach can still be used with numerical methods.

The marginal distribution of the observation $\mathbf{y}_i$ can be computed as:

$$f(\mathbf{y}_i) = \sum_{l=1}^{L} \kappa_l \int_{\theta \in \mathcal{S}_l} f_\theta^{(l)}(\mathbf{y}_i) G_0^{(l)}(\theta) d\theta \ . \tag{8}$$

By substituting (6) and (8) in (5), we obtain the posterior distribution of $\theta_i|\{\theta_j\}_{j \neq i}, \mathbf{y}_{1:N}$.

An Example (Clustering with a mixture of Gamma, log-normal and Gaussian observation models):

For example, let us assume that $\mathbf{y}_i = [y_{i,1}, \cdots, y_{i,d}]^T \in \mathbb{R}^d$ and $L = 3$, so that each $\mathbf{y}_i|\theta_i$ is a mixture of Gaussian, Gamma and log-normal distributions. For analytical tractability, the likelihood functions of the observations $\mathbf{y}_i$'s are selected so that the prior and posterior distributions of θ_i are conjugate. We also assume that the elements of $\mathbf{y}_i$'s are independent in the case of non-Gaussian observation models.

First, as in Section III-A, we may define $\mathcal{S}_1 \triangleq \mathbb{R}^d \times \mathbb{R}^{d \times d}$ to be the set of possible parameters of the Gaussian likelihood function corresponding to $\theta_i|\{Z_i = 1\} \triangleq (\boldsymbol{\mu}_i, \mathbf{V}_i)$. In this case, the likelihood $f_{\theta_i}^{(1)}(\mathbf{y}_i)$, the posterior $f(\theta_i|\mathbf{y}_i, Z_i = 1)$, the marginal $\int_{\theta \in \mathcal{S}_1} f_\theta^{(1)}(\mathbf{y}_i) G_0^{(1)}(\theta) d\theta$ and the prior $G_0^{(1)}(\theta_i)$ can be computed as described in Section III-A.

Next, we define $\mathcal{S}_2 \triangleq \mathbb{R}^d$, such that $\theta_i|\{Z_i = 2\} \triangleq \mathbf{a}$, where $\mathbf{a} = [a_1, \cdots, a_d]^T$ are the shape parameters of a Gamma distributed likelihood function (assuming fixed rate parameters $\{b_k\}_{k=1}^{d}$) such that:

$$f_\theta^{(2)}(\mathbf{y}_i) = \prod_{k=1}^{d} \frac{b_k^{a_k}}{\Gamma(a_k)} y_{i,k}^{a_k - 1} e^{-b_k y_{i,k}} \ , \tag{9}$$

where we have let $\theta = \mathbf{a}$, i.e. $y_{i,k}|\{\theta_i = \theta, Z_i = 2\} \sim Ga(a_k, b_k)$ and are i.i.d. Note that, (9) denotes the likelihood of observation $\mathbf{y}_i$ joining a cluster with parameter θ. In this case, to preserve the conjugate property, the prior distribution of $\mathbf{a}$ is assumed to be equal to:

$$G_0^{(2)}(\theta_i) = G_0^{(2)}(\mathbf{a}) = \prod_{k=1}^{d} \frac{1}{J(a_0, b_0, b_k, c_0)} \cdot \frac{a_0^{a_k - 1} b_k^{c_0 a_k}}{\Gamma(a_k)^{b_0}} \ , \tag{10}$$

where a_0, b_0 and c_0 are the corresponding hyper-parameters and $J(a_0, b_0, b_k, c_0) \triangleq \int_0^\infty \frac{a_0^{x-1} b_k^{c_0 x}}{\Gamma(x)^{b_0}} dx$ is the normalization term. The posterior distribution of θ_i can be obtained as in [28] and can be shown to be equal to:

$$\begin{aligned} &f(\theta_i|\mathbf{y}_i, Z_i = 2) = f(\mathbf{a}|\mathbf{y}_i) = \\ &= \prod_{k=1}^{d} \frac{1}{J(a_0 y_{i,k}, b_0 + 1, b_k, c_0 + 1)} \cdot \frac{(a_0 y_{i,k})^{a_k - 1} b_k^{(c_0+1)a_k}}{\Gamma(a_k)^{b_0+1}} \ . \end{aligned}$$

The marginal distribution of $\mathbf{y}$ can thus be computed as:

$$\begin{aligned} &f(\mathbf{y}_i|Z_i = 2) = \\ &= \prod_{k=1}^{d} \int_0^\infty \frac{b_k^z}{\Gamma(z)} y_{i,k}^{z-1} e^{-b_k y_{i,k}} \frac{a_0^{z-1} b_k^{c_0 z}}{\Gamma(z)^{b_0}} \left[\int_0^\infty \frac{a_0^{t-1} b_k^{c_0 t}}{\Gamma(t)^{b_0}} dt \right]^{-1} dz \ . \end{aligned}$$

Note that, in practice, the above marginal distribution of $\mathbf{y}$ can be estimated using numerical methods since it has to be only evaluated for a particular value of $\mathbf{y}_i$.

Finally, we define $\mathcal{S}_3 \triangleq \mathbb{R}^d$ such that $\theta_i|\{Z_i=3\} \triangleq \boldsymbol{\rho}$, where $\boldsymbol{\rho}=[\rho_1,\cdots,\rho_d]^T$ are the log-scale parameters of a log-normal likelihood function (assuming fixed shape parameters $\{\xi_k\}_{k=1}^d$) such that:

$$f_\theta^{(3)}(\mathbf{y}_i)=\prod_{k=1}^{d}\frac{1}{y_{i,k}\sqrt{2\pi\xi_k^2}}e^{-\frac{\left(\ln y_{i,k}-\rho_k\right)^2}{2\xi_k^2}}\ , \qquad (11)$$

where we let $\theta=\boldsymbol{\rho}$, i.e. $y_{i,k}|\{\theta_i=\theta,Z_i=3\}\sim\ln\mathcal{N}(\rho_k,\xi_k^2)$ and are i.i.d. The prior distribution of $\boldsymbol{\rho}$ is assumed to be equal to:

$$G_0^{(3)}(\theta_i)=G_0^{(3)}(\boldsymbol{\rho})=\prod_{k=1}^{d}\frac{1}{\sqrt{2\pi\xi_{0,k}^2}}e^{-\frac{\left(\rho_k-\rho_{0,k}\right)^2}{2\xi_{0,k}^2}}\ , \qquad (12)$$

i.e. $\rho_k\sim\mathcal{N}\left(\rho_{0,k},\xi_{0,k}^2\right)$, where $\rho_{0,k}$ and $\xi_{0,k}$ $(k=1,\cdots,d)$ are the corresponding hyper-parameters. The posterior distribution of θ_i is equal to [28]:

$$f(\theta_i|\mathbf{y}_i,Z_i=3)=f(\boldsymbol{\rho}|\mathbf{y}_i)=\prod_{k=1}^{d}\frac{1}{\sqrt{2\pi\psi_k}}e^{-\frac{(\rho_k-\nu_k)^2}{2\psi_k}}\ , \qquad (13)$$

i.e. $\rho_k|y_{i,k}\sim\mathcal{N}(\nu_k,\psi_k)$, where $\nu_k=\frac{\xi_{0,k}^2\rho_{0,k}^2+\xi_k^2y_{i,k}}{\xi_{0,k}^2+\xi_k^2}$ and $\psi_k=\xi_{0,k}^2+\xi_k^2$. The marginal distribution of $\mathbf{y}$ can thus be computed as:

$$\begin{aligned}&f(\mathbf{y}_i|Z_i=3)=\\ &=\prod_{k=1}^{d}\frac{1}{2\pi y_{i,k}\sqrt{\xi_k^2\xi_{0,k}^2}}\int_{-\infty}^{\infty}e^{-\frac{\left(\ln y_{i,k}-\rho\right)^2}{2\xi_k^2}}e^{-\frac{\left(\rho-\rho_{0,k}\right)^2}{2\xi_{0,k}^2}}d\rho\ .\end{aligned}$$

which can again be estimated numerically.

Once we have the marginal posterior distributions characterized as above, we can apply the Gibbs sampling as in Algorithm 1 to find the best observation model that fits each cluster.

C. Prior and posterior distributions for α_0

In [29], it was shown that the posterior distribution for α_0 can be represented in a simple conditional form, given a certain class of prior distributions for α_0 [14]. In particular, if the prior distribution of α_0 follows the Gamma distribution, such that $\alpha_0\sim Ga(a,b)$ with shape $a>0$ and scale $b>0$[4], then the conditional posterior distribution of α_0 may be expressed as a mixture of two Gamma distributions, where the mixing parameter follows a Beta distribution, such that:

$$\begin{aligned}\alpha_0|x,K\ &\sim\ \pi_x Ga\left(a+K,b-\log\left(x\right)\right)+\\ &+\ \left(1-\pi_x\right)Ga\left(a+K-1,b-\log\left(x\right)\right)\ ,\end{aligned} \qquad (14)$$

where $K>1$ is the number of clusters and $x|\alpha_0,K\sim Beta\left(\alpha_0+1,N\right)$ with $Beta$ denoting the Beta distribution [14], [29]. The mixing parameter π_x is defined such that:

$$\frac{\pi_x}{1-\pi_x}=\frac{a+K-1}{N\left(b-\log\left(x\right)\right)}\ , \qquad (15)$$

[4]It is very hard to estimate a and b from real-world data. However, it is noticed in [14] that small values of a and b lead to nearly similar values of the α probability density, thus resulting in a lack of variability in the distribution of θ_i.

It should be noted that α_0 and K should be sampled at each iteration of the Gibbs sampling and that the prior distribution of K is given by [29]:

$$P(K|\alpha_0,N)=c_N(K)N!\alpha_0^K\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+N)}\ , \qquad (16)$$

where $c_N(K)=P(K|\alpha_0=1,N)$ can be computed using recurrence formulae for Stirling numbers [29]. Note that this prior distribution depends only on the number of data points N and on the concentration parameter α_0.

Moreover, for large N, the number of clusters generated by this model can be approximated as $K=X+1$, where X is a Poisson random variable with mean $\alpha_0\left(\gamma+\log\left(N\right)\right)$ and $\gamma\approx0.5772156649$ being the Euler constant [29]. This approximation is useful if the number of clusters K is much smaller than the number of data points N, when N is large [29]. In wireless applications, we may assume that different wireless systems form different clusters. The data points within each cluster may represent the signals corresponding to that system (cluster). If the signals are detected frequently w.r.t. the operation time of a certain system, a large number of feature points will be observed in a single cluster, which makes the number of feature points N to grow at a much faster rate compared to K, thus justifying the use of above approximation.

On the other hand, in order to compute the posterior distribution of K, given the observed data points, the authors in [15], [29] proposed a Monte Carlo approach. This method was based on counting the number of distinct mixture components at each Gibbs iteration and updating the posterior of K accordingly. Hence, the empirical posterior probability of K can be approximated by the histogram of the number of mixture components that are encountered throughout the Gibbs sampling iterations.

IV. Bayesian Prediction (or Density Estimation) of the Observation Variables

Upon observing and classifying N feature points, a CR may need to predict the occurrence of a particular observation $\mathbf{y}_{N+1}$ in the next time step. The predictive probability distribution of the random observation $\mathbf{Y}_{N+1}$ can help to achieve this goal by using the previously observed features. Such predictive distribution can be useful in decision-making applications, allowing CR's to coordinate their actions with other wireless users by predicting their behavior.

The posterior distribution of $\mathbf{Y}_{N+1}$, given the observations $\mathbf{y}_{1:N}$ and the cluster parameters $\theta_{1:N}$, is denoted by $P\left(\mathbf{Y}_{N+1}|\theta_{1:N},\mathbf{y}_{1:N}\right)$. Since $\{\mathbf{Y}_i\}_{i=1}^N$ are i.i.d., given $\theta_{1:N}$, we have $P\left(\mathbf{Y}_{N+1}|\theta_{1:N},\mathbf{y}_{1:N}\right)=P\left(\mathbf{Y}_{N+1}|\theta_{1:N}\right)$ which may be evaluated as $\int P\left(\mathbf{Y}_{N+1}|\theta_{N+1}\right)dP(\theta_{N+1}|\theta_{1:N})$ [15]. According to [15], the probability distribution of $\mathbf{Y}_{N+1}$, given the components $\theta_{1:N}$, can be computed as:

$$\left(\mathbf{Y}_{N+1}|\theta_{1:N}\right)\sim\frac{\alpha_0}{\alpha_0+N}f(\mathbf{y}_{N+1})+\frac{1}{\alpha_0+N}\sum_{i=1}^{N}f_{\theta_i}\left(\mathbf{y}_{N+1}\right)\ , \qquad (17)$$

where $f(\mathbf{y}_{N+1})$ is the marginal distribution of Y_{N+1} which was defined in (8). We may re-write (17) as:

$$(\mathbf{Y}_{N+1}|\theta_{1:N}) \sim \frac{\alpha_0}{\alpha_0+N} f(\mathbf{y}_{N+1}) + \frac{N}{\alpha_0+N} \sum_{k=1}^{K} \frac{n_k}{N} f_{\theta_k}(\mathbf{y}_{N+1}) \tag{18}$$

where n_k is the number of data points in cluster $k \in \{1, \cdots, K\}$. Note that (18) implies that the observation $\mathbf{Y}_{N+1}$ is drawn from a mixture of a Student t-distribution and an observation mixture model with mixing parameters $\frac{\alpha_0}{\alpha_0+N}$ and $\frac{n}{\alpha_0+N}$, respectively. In wireless applications, it is reasonable to assume that a detected signal may belong to a previously detected system (cluster) with a probability proportional to the number of signals observed from that system. However, since we assume that the number of systems (clusters) is unknown, *a priori*, a signal belonging to a new system may arise with a probability proportional to α_0. Thus, the probability distribution in (17) may be used to predict the occurrence of a certain signal, given past information.

Since past information may consist of only noisy observations $\mathbf{y}_{1:N}$, in the following, we show the predictive distribution of $\mathbf{Y}_{n+1}$, given the past observations $\mathbf{y}_{1:N}$. Thus, we integrate out the cluster parameters $\theta_{1:N}$ from the posterior distribution of $\mathbf{Y}_{N+1}$ since these parameters are not fully observable by the classifier. Hence, the Bayesian prediction, or density estimation, problem can be solved by evaluating the unconditional predictive distribution:

$$P(\mathbf{Y}_{N+1}|\mathbf{y}_{1:N}) = \int P(\mathbf{Y}_{N+1}|\theta_{1:N})\, dP(\theta_{1:N}|\mathbf{y}_{1:N}) \ . \tag{19}$$

The complexity of the above expression stems from the inherent complexity of the posterior $P(\theta_{1:N}|\mathbf{y}_{1:N})$. However, by using the Monte Carlo approach of [15], [23], it is possible to obtain an approximation for this density function, iteratively. For a given $\mathbf{m}$ and τ parameters, the estimated density function is given by [15]:

$$\begin{aligned} P(\mathbf{Y}_{N+1}|\mathbf{y}_{1:N}) &\approx \frac{1}{N_r} \sum_{r=1}^{N_r} P(\mathbf{Y}_{N+1}|\theta_{1:N}(r)) \\ &= \frac{1}{N_r} \sum_{r=1}^{N_r} \left[\frac{\alpha_0(r)}{\alpha_0(r)+N} f(\mathbf{y}_{N+1}) + \right. \\ &\quad \left. + \frac{1}{\alpha_0(r)+N} \sum_{i=1}^{N} f_{\theta_i(r)}(\mathbf{y_{N+1}}) \right] , \end{aligned}$$

where N_r is the number of Gibbs sampling iterations, $\theta_i(r)$ and $\alpha_0(r)$ are the sampled parameters at the r-th iteration. The authors in [15] have shown the convergence of the above estimate to the actual predictive distribution $P(\mathbf{Y}_{N+1}|\mathbf{y}_{1:N})$ for almost all starting values. That is:

$$\lim_{N_r \to \infty} \frac{1}{N_r} \sum_{r=1}^{N_r} P(\mathbf{Y}_{N+1}|\theta_{1:N}(r)) = P(\mathbf{Y}_{N+1}|\mathbf{y}_{1:N}) \ . \tag{20}$$

The above identity shows that the predictive distribution of Y_{N+1} is equivalent to the average likelihood function of Y_{N+1}, averaged over the Gibbs sampling iterations.

V. Convergence of Algorithm 1

The convergence of Algorithm 1 has been proven in [15], [23] based on the MCMC approach. The convergence result can be stated as follows.

Let $Q_I(\theta_{1:N}(0), A)$ be the probability that, with an initial value $\theta_{1:N}(0)$ and after one iteration, Algorithm 1 produces a sample value that is contained in the measurable set A, i.e. $Q_I(\theta_{1:N}(0), A) = P\{\theta_{1:N}(1) \in A|\theta_{1:N}(0)\}$. $Q_I(.,.)$ is called the transition kernel of the Markov chain. Similarly, let $Q_I^s(\theta_{1:N}(0), A) = P\{\theta_{1:N}(s) \in A|\theta_{1:N}(0), s\}$. Let's denote by $P(\theta_{1:N}|\mathbf{y}_{1:N})$ the posterior distribution of $\theta_{1:N}$.

Theorem 1 of [15] states that, for almost all starting values of $\theta_{1:N}(0)$, the probability measure Q_I^s (defined over the measurable space $\Omega \supset A$) converges in total variation norm to the posterior distribution as s goes to infinity. That is, for almost all $\theta_{1:N}(0)$, $\lim_{s\to\infty} \|Q_I^s(\theta_{1:N}(0), .) - P(\theta_{1:N}|\mathbf{y}_{1:N})\| = 0$. Of course, this convergence *in probability* is a weaker type of convergence, compared to the *almost sure* convergence for which $P\{\lim_{r\to\infty} \|\theta_{1:N}(r) - \theta_{1:N}\| > \delta\} = 0$, for some $\delta > 0$. In other words, Theorem 1 does not state that $\theta_i(r) \to \theta_i$ for all $i \in \{1, \cdots, N\}$. However, it ensures that the Gibbs sampling outcomes $\theta_{1:N}(r)$ will be distributed according to the actual posterior distribution of $\theta_{1:N}|\mathbf{y}_{1:N}$, for large r. This result is particularly important to justify the use of the Gibbs sampling outcomes in constructing the posterior distribution of $\theta_{1:N}|\mathbf{y}_{1:N}$ and finding an estimation of $\theta_{1:N}$.

VI. Mean-Squared Error (MSE) Analysis of the Estimated Cluster Means

In this section, we derive the mean-squared error (MSE) of the estimated cluster means and, under certain regularity conditions, we establish an asymptotic upper bound on the MSE. Denote by $\hat{\boldsymbol{\mu}}_k$ and $\boldsymbol{\mu}_k$ to be, respectively, the estimated and actual mean vectors of cluster $k \in \{1, \cdots, K\}$.

By assuming that the DPMM-based classifier results in correct clustering of the observation points (after sufficiently many Gibbs sampling iterations), the MSE of the estimated cluster means $\boldsymbol{\mu}_k$ can be expressed as:

$$MSE_k = tr\left(\frac{1}{n_k}\mathbf{V}_k\right) = \frac{1}{n_k} tr(\mathbf{V}_k) \ , \tag{21}$$

where $\mathbf{V}_k$ is the covariance matrix of the observations in cluster k, and n_k is the number of data points belonging to cluster k.

In a DPMM with N data points and with K clusters, the average MSE becomes:

$$MSE = \mathbb{E}\left\{ \frac{1}{K} \sum_{k=1}^{K} MSE_k | N \right\} \ , \tag{22}$$

where the prior distribution of K is as given in (16). For large N, K can be approximated with a Poisson random variable such that [29]:

$$P\{K = k|\alpha_0, N\} = \frac{e^{-\alpha_0(\gamma+\log N)} [\alpha_0(\gamma + \log N)]^k}{k!} \ , \text{ for } k = 0, 1, \cdots . \tag{23}$$

Thus, we have:

$$MSE = \mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^{K} MSE_k|N\right\} \quad (24)$$

$$= \mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}tr\left(\mathbf{V}_k\right)|N, n_k \neq 0\right\} . \quad (25)$$

Due to the complexity of the distribution of $\frac{1}{n_k}$, it is hard to obtain a closed form for the above MSE expression. However, if the observations are equally partitioned among the clusters (i.e. $n_k = \frac{N}{K}$), we have:

$$\begin{aligned} MSE &= \mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}tr\left(\mathbf{V}_k\right)|N, n_k \neq 0\right\} \\ &= \mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{K}{N}tr\left(\mathbf{V}_k\right)|N\right\} \\ &\leq \frac{1}{N}\mathbb{E}\left\{\sum_{k=1}^{K}V_{max}|N\right\} \\ &= \frac{1}{N}V_{max}\mathbb{E}\left\{K|N\right\} \\ &= \frac{1}{N}V_{max}\left(\gamma+\log N\right)\mathbb{E}\left\{\alpha_0\right\} \\ &= \frac{ab}{N}V_{max}\left(\gamma+\log N\right) \\ &= \overline{MSE} , \end{aligned} \quad (26)$$

where $V^{max} = \max_{k=1,\cdots,K} tr(\mathbf{V}_k)$ and $\alpha_0 \sim Ga(a,b)$. Thus, under the above assumed conditions and for large N, an upper bound for MSE of the cluster mean estimates can be taken to be proportional to:

$$\overline{MSE} \propto \frac{\log N}{N} .$$

This result shows that the MSE of the cluster mean estimates decreases with N. However, the convergence of the Gibbs sampling algorithm becomes slower as N increases. Thus, a tradeoff should be made between the estimation accuracy and the convergence speed when selecting a particular data set of size N for clustering.

The above asymptotic bound is valid for large values of N, which can be justified in spectrum sensing applications when the sensing periods are very short, as in [21]. In this case, we consider a time window that includes a large number of sensing intervals as the processing period. Feature points are extracted after each sensing interval, thus leading to a large number of feature points N during this time window. These N feature points are then used in DPMM classification, justifying the use of large N in the above result. In addition, if the RF activities remain constant during the time window, feature points will be observed from the same clusters over successive sensing intervals. Then, we may assume that the total number of feature points will be equally partitioned among all the clusters.

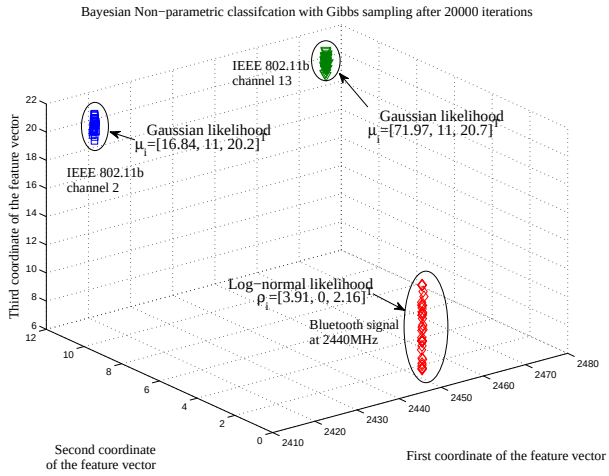


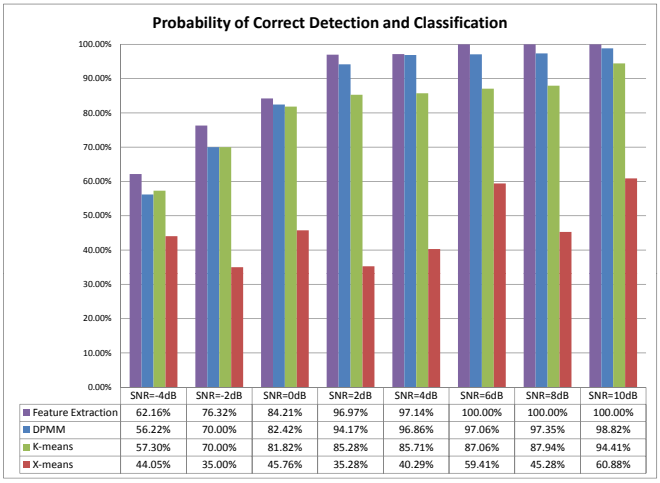
Fig. 2. Signal Classification of 2 WiFi and a Bluetooth signal. The feature point is denoted by (f_c, α, B), where f_c is the carrier frequency, α is the cyclic frequency component corresponding to the symbol rate and B is the estimated bandwidth. Energy detection is applied for $30\mu s$ at an SNR of 5 dB with Rayleigh fading (fast fading). The probability of correct classification is 100% after 20000 Gibbs sampling iterations.

VII. Simulation results: Signal Classification in the ISM band

In this section, we apply above developed non-parametric signal classification algorithm based on DPMM to the problem of RF mapping. In particular, to start with, we consider 2 IEEE 802.11.b WiFi signals (channels 2 and 13) transmitting at 2.417 and 2.472GHz, respectively. We also consider a Bluetooth signal transmitting at 2.45 GHz during the sensing process. The SNR at the receiver is 5 dB and each sensing window is $30\mu s$. We assume a fast-fading Rayleigh channel with normalized fading coefficients h such that $\mathbb{E}\{h^2\} = 1$.

After each $30\mu s$ sensing time, feature points (f_c, α, B) are extracted from the sensed signal, where f_c denotes the carrier frequency (down-converted to zero-IF), α is the cyclic frequency component corresponding to the symbol rate and B is the estimated signal bandwidth. The carrier frequencies f_c and cyclic frequencies α are obtained by applying the energy and cyclostationary detection algorithms in [22] and the signal bandwidth is estimated from the smoothed PSD of the received signal. In this setup, each WiFi signal has a bandwidth of 22 MHz and the Bluetooth signal has a bandwidth of 1 MHz. Furthermore, the Bluetooth signal has a symbol rate of 1 Mbaud and the WiFi has a chiprate of 11 Mchips/s that is manifested in the α component of the feature points.

We perform 50 repetitions of the sensing process (over a total sensing time of $50 \times 30\mu s$) and obtain the feature points. We then apply our proposed DPMM-based feature classification algorithm to classify the observed feature points. The feature points that are marked with the same marker shape in Fig. 2 are assigned to the same cluster. We show in Fig. 2 the results of the DPMM classification in a 3D feature space where the two WiFi signals are estimated to have Gaussian observation models while the Bluetooth signal is assigned a log-normal model. The classification accuracy, denoting the



	SNR=-4dB	SNR=-2dB	SNR=0dB	SNR=2dB	SNR=4dB	SNR=6dB	SNR=8dB	SNR=10dB
Feature Extraction	62.16%	76.32%	84.21%	96.97%	97.14%	100.00%	100.00%	100.00%
DPMM	56.22%	70.00%	82.42%	94.17%	96.86%	97.06%	97.35%	98.82%
K-means	57.30%	70.00%	81.82%	85.28%	85.71%	87.06%	87.94%	94.41%
X-means	44.05%	35.00%	45.76%	35.28%	40.29%	59.41%	45.28%	60.88%

Fig. 3. The classification accuracy of the DPMM, X-means and K-means ($K = 4$) algorithms using 4 different wireless signals at different SNR's. The feature extraction accuracy is also computed to show its impact on the classification performance.

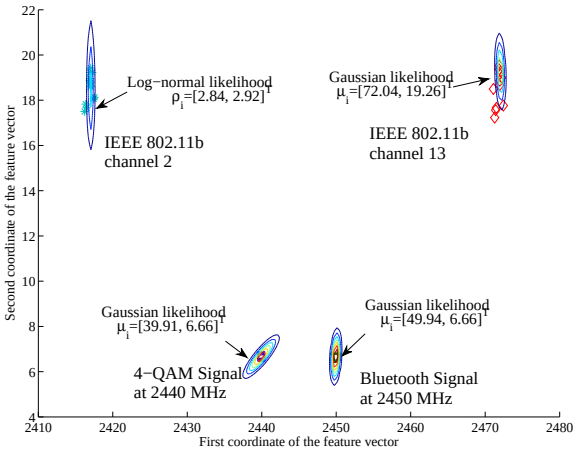


Fig. 4. Signal Classification of 2 WiFi and a Bluetooth signal. The feature point is denoted by (f_c, B), where f_c is the carrier frequency and B is the estimated bandwidth of the signal. Energy detection is applied for $30\mu s$ at an SNR of 5 dB with Rayleigh fading (fast fading). The probability of correct classification is 100% after 5000 Gibbs sampling iterations.

percentage of feature points classified into correct clusters, is estimated as 100% in this setup.

In the next set of simulations, we compare performance of the proposed DPMM-based classification algorithm to that of the approach proposed in [8] based on the K-means and X-means algorithms [9]. In the simulation setup, we consider an additional 4-QAM digital signal transmitting at 2440 MHz. For simplicity, we limit the feature vectors to be 2-D data (f_c, B). We analyze the performance of the DPMM, K-means and X-means classification algorithms at different SNR's in terms of the classification accuracy. The classification accuracy is defined as the proportion of feature vectors that are correctly detected and classified. Obviously, this quantity depends on both feature extraction and signal classification performance. Hence, we also compute the feature extraction accuracy as the proportion of correctly detected feature vectors. This will show the impact of a particular feature extraction algorithm on the overall classification accuracy.

At each SNR, we compute the feature extraction accuracy as well as the classification performance of each of the three above algorithms. The classification accuracy is averaged over 10 independent runs and is shown in Fig. 3. This figure shows that the classification accuracy of each of the three classifiers is upper-bounded by the feature extraction accuracy which is considered as the bottleneck for the classification performance. Obviously, this is the case since we define the correctly classified features as a subset of the correctly detected features. As can be seen from Fig. 3, by increasing the SNR, the feature extraction accuracy improves, as well as both DPMM and K-means ($K = 4$) accuracies. Although the K-means might have similar (yet lower) performance, compared to the DPMM, it requires additional prior information about the number of clusters to achieve good results. Even with such information, however, the K-means does not always reach the performance level of the DPMM due to its underlying Gaussian spherical assumption that is not able to match complex observations, as in spectrum sensing applications. On the other hand, the DPMM can achieve better performance, compared to the K-means, yet without any prior knowledge about the number of clusters. This is due to its better ability to match the observation model and infer hidden information about the data by following a Bayesian approach. The X-means algorithm, however, suffers from poor performance, even at high SNR, since it is not able to estimate accurately the number of clusters because of its presumption of Gaussian spherical observation model. Furthermore, as can be shown in Fig. 3, a high SNR does not necessarily improve the classification performance of the X-means, as long as the number of clusters is not estimated correctly. The DPMM classifier, however, can avoid this problem by having a better estimation of the number of clusters.

In Fig. 4, we plot the predictive probability distribution of future feature points. For simplicity of representation, we again consider a 2D feature space with feature points (f_c, B) and represent the probability density function of the predictive distribution in contour lines. The result shows four main clusters corresponding to the WiFi, Bluetooth and QAM signals where the feature points corresponding to channel 2 of the WiFi system is estimated to have a log-normal distribution while the other feature points are estimated to have Gaussian distributions. The obtained distribution forms an RF mapping of the RF environment and can help CR's to adapt their actions by using this information (beyond the scope of this paper).

Finally, we verify the analytical MSE expression of Section VI in the case of $K = 2$. In particular, we consider two WiFi signals (channels 2 and 13) and we compute the corresponding MSE resulting from the DPMM-based classification. We consider a scalar feature point f_c consisting of the measured center frequencies. Since the number of systems is assumed to be fixed during the classification process, we have $\mathbb{E}\{K|N\} = 2$ in (26), thus resulting in an analytical upper bound equal to $\frac{KV^{max}}{N}$. The mean and variance of the

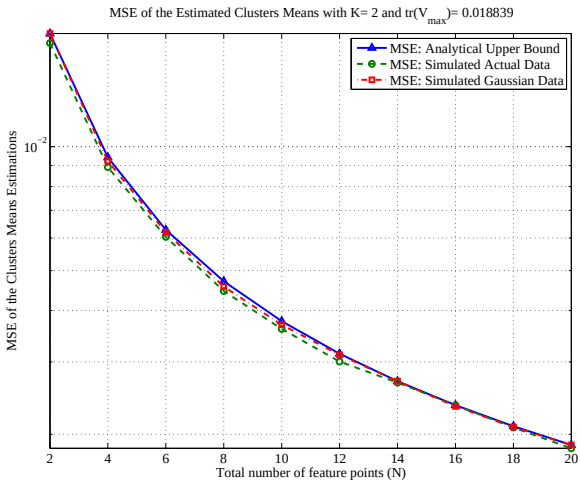


Fig. 5. Analytical upper bound on the MSE of clusters means estimation for both simulated Gaussian data and simulated actual data with $K = 2$ clusters.

feature points f_c's corresponding to the two WiFi signals are, respectively, $2416.94MHz$ and $0.0188MHz^2$, for channel 2, and $2471.94MHz$ and $0.0185MHz^2$, for channel 13. Note that, the means of the extracted features deviate slightly from the transmit carrier frequencies (i.e. $2417MHz$ and $2472MHz$) due to the limited spectral resolution in discrete spectral estimation. We also generate i.i.d. data observations from a GMM with 2 components and whose means and variances are identical to the extracted feature points. We compute the MSE of cluster means for both simulated actual data and simulated Gaussian data, with respect to the number of feature points N. We compare the corresponding MSE's to the above analytical upper bound, as shown in Fig. 5. The result shows that the MSE's in both simulated actual data and simulated Gaussian data are very close to the analytical upper bound, thus justifying the use of this upper bound with different data models.

VIII. Conclusion

In this paper, we proposed a non-parametric signal classification method to identify/classify active wireless systems in an unknown RF environment. This proposed technique is suitable for autonomous CR's, such as Radiobots of [3] and [4], in performing spectrum sensing and signal classification in alien RF bands. Since our non-parametric technique does not require any prior knowledge of the existing signals in the sensed spectrum, it can ensure autonomous operation of CR's such as Radiobots. The proposed DPMM framework extends to both Gaussian and non-Gaussian observation models and it uses the Gibbs sampling to estimate the appropriate distribution for each cluster. We derived an upper bound for the MSE of the estimate of the cluster means as a function of the number of feature points N. A Bayesian predictive distribution was also derived to construct an RF mapping for the on-going RF activity. Simulation results were presented to compare the performance of the proposed DPMM-based algorithm to those of existing classifiers such as K-means and X-means.

References

[1] FCC, "Report of the spectrum efficiency working group," FCC spectrum policy task force, Tech. Rep., Nov. 2002.

[2] J. Mitola, "Cognitive radio architecture evolution," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 626 –641, Apr. 2009.

[3] S. K. Jayaweera and C. G. Christodoulou, "Radiobots: Architecture, algorithms and realtime reconfigurable antenna designs for autonomous, self-learning future cognitive radios," University of New Mexico, Technical Report EECE-TR-11-0001, Mar. 2011. [Online]. Available: http://repository.unm.edu/handle/1928/12306

[4] S. Jayaweera, Y. Li, M. Bkassiny, C. Christodoulou, and K. Avery, "Radiobots: The autonomous, self-learning future cognitive radios," in *International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS '11)*, Chiangmai, Thailand, Dec. 2011.

[5] K. Kim, I. Akbar, K. Bae, J.-S. Urn, C. Spooner, and J. Reed, "Cyclostationary approaches to signal detection and classification in cognitive radio," in *2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '07)*, Dublin, Ireland, Apr. 2007, pp. 212 –215.

[6] J. Popoola and R. van Olst, "Application of neural network for sensing primary radio signals in a cognitive radio environment," in *AFRICON 2011*, Livingstone, Zambia, Sep. 2011, pp. 1 –6.

[7] M. Ramon, T. Atwood, S. Barbin, and C. Christodoulou, "Signal classification with an svm-fft approach for feature extraction in cognitive radio," in *SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC '09)*, Belem, Brazil, Nov. 2009, pp. 286 –289.

[8] T. Clancy, A. Khawar, and T. Newman, "Robust signal classification using unsupervised learning," *IEEE Transactions on Wireless Communications*, vol. 10, no. 4, pp. 1289 –1299, Apr. 2011.

[9] D. Pelleg and A. Moore, "X-means: extending k-means with efficient estimation of the number of clusters," in *Seventh International Conference on Machine Learning (ICML '00)*, Stanford, CA, June-July 2000.

[10] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006. [Online]. Available: http://www.jstor.org/stable/27639773

[11] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "Bayesian nonparametric methods for learning markov switching processes," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 43 –54, Nov. 2010.

[12] ——, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1569 –1585, Apr. 2011.

[13] N. Bouguila and D. Ziou, "A dirichlet process mixture of dirichlet distributions for classification and prediction," in *IEEE Workshop on Machine Learning for Signal Processing (MLSP '08)*, Oct. 2008, pp. 297 –302.

[14] A. Rabaoui, N. Viandier, J. Marais, and E. Duflos, "On selecting the hyperparameters of the dpm models for the density estimation of observation errors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '11)*, Prague, Czech Republic, May 2011, pp. 4092 –4095.

[15] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1995. [Online]. Available: http://www.jstor.org/stable/2291069

[16] K. Ishiguro, T. Yamada, S. Araki, T. Nakatani, and H. Sawada, "Probabilistic speaker diarization with bag-of-words representations of speaker angle information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 447 –460, Feb. 2012.

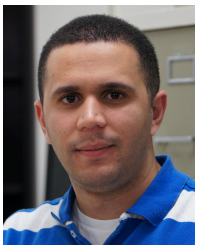
[17] A. Harati Nejad Torbati, J. Picone, and M. Sobel, "Applications of Dirichlet process mixtures to speaker adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '12)*, Kyoto, Japan, Mar. 2012, pp. 4321 –4324.

[18] S. Li, Z. Yanning, M. Miao, and T. Guangjian, "SAR image segmentation method using DP mixture models," in *International Symposium on Computer Science and Computational Technology (ISCSCT '08)*, vol. 2, Shanghai, China, Dec. 2008, pp. 598 –601.

[19] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin, "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6140 –6155, Dec. 2010.

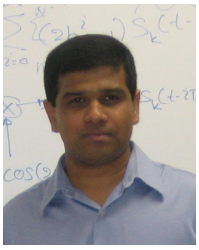
[20] N. Shetty, S. Pollin, and P. Pawelczak, "Identifying spectrum usage by unknown systems using experiments in machine learning," in *IEEE Wireless Communications and Networking Conference (WCNC '09)*, Budapest, Hungary, Apr. 2009, pp. 1 –6.

[21] M. Bkassiny, S. K. Jayaweera, Y. Li, and K. A. Avery, "Wideband spectrum sensing and non-parametric signal classification for autonomous self-learning cognitive radios," *IEEE Transactions on Wireless Communications*, vol. 11, no. 7, pp. 2596 –2605, July 2012.
[22] ——, "Blind cyclostationary feature detection based spectrum sensing for autonomous self-learning cognitive radios," in *IEEE International Conference on Communications (ICC '12)*, Ottawa, Canada, June 2012.
[23] M. D. Escobar, "Estimating normal means with a Dirichlet process prior," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 268–277, Mar. 1994. [Online]. Available: http://www.jstor.org/stable/2291223
[24] M. DeGroot, *Optimal Statistical Decisions*. McGraw-Hill, 1970.
[25] H. W. Sorenson, *Parameter Estimation: Principles and Problems*. Marcel Dekker, 1980.
[26] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721 –741, Nov. 1984.
[27] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Chapman & Hall/CRC, 2003.
[28] D. Fink, "A compendium of conjugate priors," Tech. Rep., 1997.
[29] M. West, "Hyperparameter estimation in Dirichlet process mixture models," Duke University, Tech. Rep., 1992.



Mario Bkassiny (S'06) received the B.E. degree in Electrical Engineering with High Distinction and the M.S. degree in Computer Engineering from the Lebanese American University, Lebanon, in 2008 and 2009, respectively. He received his PhD degree in Electrical Engineering from the University of New Mexico, Albuquerque, NM, USA in 2013. He is currently an Assistant Professor at the Department of Electrical and Computer Engineering at State University of New York (SUNY) at Oswego, Oswego, NY, USA. He worked as a research assistant at the Communication and Information Sciences Laboratory (CISL), Department of Electrical and Computer Engineering at the University of New Mexico from 2009 to 2013.

Dr. Bkassiny served as a technical co-chair of the Workshop on Wideband Cognitive Radio Communication and Networks (WCRCN) at the IEEE Vehicular Technology Conference (VTC-Fall 2013). His current research interests are in cognitive radios, distributed learning and reasoning, signal classification, wideband spectrum sensing, cyclostationary detection and dynamic spectrum leasing (DSL).



Sudharman K. Jayaweera (S'00, M'04, SM'09) was born in Matara, Sri Lanka. He completed his high school education at the Rahula College, Matara, and worked as a science journalist at the Associated Newspapers Ceylon Limited (ANCL) till 1993. Later, he received the B.E. degree in Electrical and Electronic Engineering with First Class Honors from the University of Melbourne, Australia, in 1997 and M.A. and PhD degrees in Electrical Engineering from Princeton University, USA in 2001 and 2003, respectively. He is currently an Associate Professor in Electrical Engineering at the Department of Electrical and Computer Engineering at University of New Mexico, Albuquerque, NM. Dr. Jayaweera held an Air Force Summer Faculty Fellowship at the Air Force Research Laboratory, Space Vehicles Directorate (AFRL/RVSV) from 2009-2011.

Dr. Jayaweera is currently an associate editor of IEEE Transactions on Vehicular Technology. He has also served as a member of the Technical Program Committees of numerous IEEE conferences and was the Tutorial and Workshop Chair of the 2013 Fall IEEE Vehicular Technology Conference. His current research interests include cooperative and cognitive communications, machine learning, information theory of networked-control systems, statistical signal processing and smart-grid control and optimization.

Yang Li received the B.E. degree in Electrical Engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2005 and the M.S. degree in Electrical Engineering from New Mexico Institute of Mining and Technology, Socorro, New Mexico, USA in 2009. He is currently working towards his PhD degree in Electrical Engineering at the Communication and Information Sciences Laboratory (CISL), Department of Electrical and Computer Engineering at the University of New Mexico, Albuquerque, NM, USA. His current research interests are in cognitive radios, spectrum sensing, cooperative communications, and dynamic spectrum access (DSA).