A Random Finite Set Model for Data Clustering

Dinh Phung† and Ba-Ngu Vo‡
†Center for Pattern Recognition and Data Analytics, Deakin University, Australia
‡Department of Electrical and Computer Engineering, Curtin University, Australia
Email: *dinh.phung@deakin.edu.au, ba-ngu.vo@curtin.edu.au*

***Abstract*—The goal of data clustering is to partition data points into groups to optimize a given objective function. While most existing clustering algorithms treat each data point as vector, in many applications each datum is not a vector but a point pattern or a set of points. Moreover, many existing clustering methods require the user to specify the number of clusters, which is not available in advance. This paper proposes a new class of models for data clustering that addresses set-valued data as well as unknown number of clusters, using a Dirichlet Process mixture of Poisson random finite sets. We also develop an efficient Markov Chain Monte Carlo posterior inference technique that can learn the number of clusters and mixture parameters automatically from the data. Numerical studies are presented to demonstrate the salient features of this new model, in particular its capacity to discover extremely unbalanced clusters in data.**

I. Introduction

Stochastic geometry is an established area of study with a long history that dates back to the famous problem of Buffon's needle [35]. Stochastic geometric models, including deformable templates and random finite sets have long been used by statisticians to develop techniques for object recognition in static images [3]. Random finite set (RFS) theory (or more generally point process theory) is the study of random point patterns with applications spanning numerous disciplines from agriculture/forestry and epidemiology/public health [36], [25] to communications [2], target tracking [22], [40], computer vision [3], and robotics [26]. The common theme in these applications is the set-valued observation and/or set-valued parameters.

While RFS theory is suitable for inferencing problems involving unknown and random number of parameters, its use has been largely overlooked in *the problem of learning from data*. One of the most popular tasks in learning from data is data clustering where the goal is to partition the data points into groups to optimize a given objective function, such as the distance between data points within a group as in the K-means algorithm. Many clustering methods require the number of clusters to be known apriori, but this is not the case in practice. Nearly all existing clustering algorithms treat each data point as a vector. However, in many applications each data point is a set of vectors (rather than a vector of fixed dimension). For example, in image analysis, the information content of an image is summarized and stored as a set of features. Another example is text modelling, where the 'bag-of-words' representation treats a document as a finite set of words, since the order of appearance of the words is neglected. Other examples include geo-spatial data, epidemiological data etc. In general, a sparse data point in which the order of the non-zero elements is not important can be represented as a set-valued data point.

Acknowledgement: The work of the 2nd author is supported by the Australian Research Council under the Future Fellowship FT0991854.

In this paper we propose a new class of model for data clustering that addresses set-valued data as well as unknown number of clusters based on Poisson RFS. The proposed model is a Dirichlet process mixture of Poisson RFS and is termed the *Dirichlet Poisson RFS Mixture Model* (DP-RFS). In particular, we derive a family of conjugate priors for Poisson RFS likelihoods, and use this result to develop an infinite mixture of Poisson RFS likelihoods with Dirichlet process prior on the mixture weights. We then present an efficient Markov Chain Monte Carlo method to perform posterior inference, from which the number clusters and mixture parameters are automatically learned from the data. More specifically, we exploit the conjugacy of the prior on the parameters of the Poisson RFS likelihood to integrate over these parameters and derive an efficient collapsed Gibbs sampler that converges faster than a standard full Gibbs sampler. A numerical study is presented to demonstrate the capability of the proposed DP-RFS model to learn in scenarios with extremely unbalanced clusters where existing methods typically fail.

II. Background

A. Finite Bayesian mixture models

The most common probabilistic approach to clustering is mixture modelling where the clustering process is treated as a density estimation problem. Mixture models assume in advance the existence of K latent subpopulations in the data and specifies a likelihood of observing each data point x as a mixture:

$$p\left(x \mid \pi_{1:K}, \phi_{1:K}\right) = \sum_{k=1}^{K} \pi_k f\left(x \mid \phi_k\right) \tag{1}$$

where π_k is the probability that x belongs to the k-th subpopulation and $\sum_{k=1}^{K} \pi_k = 1$. This is the parametric and frequentist approach to mixture modeling. The EM algorithm is typically employed to estimate the parameters $\pi_{1:K}$ and $\phi_{1:K}$ from the data. Gaussian mixture models (GMM), for instance, is commonly used in signal processing and target tracking. In this case, each mixture-specific parameter ϕ_k consists of $(\boldsymbol{\mu}_k, \Sigma_k)$ which specifies the mean and covariance matrix for each mixture.

Under a Bayesian setting [12], [32] the parameters $\pi_{1:K}$ and $\phi_{1:K}$ are further endowed with suitable prior distributions. Typically a symmetric Dirichlet distribution $\mathrm{Dir}(\cdot \mid \eta)$ is used as the prior of $\pi_{1:K}$, while the prior distribution for $\phi_{1:K}$ is model-specific depending on the form of the likelihood function f which admits a conjugate prior h. A Bayesian mixture model specifies the generative likelihood for x as:

$$p(x \mid \eta, h) = \int \int \sum_{k=1}^{K} \pi_k f(x \mid \phi_k) P(d\pi_{1:K}) P(d\phi_{1:K})$$

Under this formalism, inference amounts to deriving the joint posterior distribution for $\pi_{1:K}$ and $\phi_{1:K}$, which is often intractable. Markov Chain Monte Carlo methods, such as Gibbs sampling, are common approaches for the inference task [12], [4].

Suppose there are data points $D = \{x_1, ..., x_N\}$. A latent indicator variable z_i is introduced for each data point x_i to specify its mixture component where $z_i \in \{1, ..., K\}$ and $\Pr(z_i = k) = \pi_k$. Conditioning on this latent variable, the distribution for x_i simplifies to:

$$p(x_i \mid z_i = k, \phi_{1:K}, \pi_{1:K}) = f(x_i \mid \phi_k) \tag{2}$$

Full Gibbs sampling for posterior inference becomes straightforward by iteratively sampling the conditional distributions among the latent variables $\pi_{1:K}$, z_i and ϕ_k, i.e.,

$$p(z_i \mid z_{-i}, x_{1:n}, \pi_{1:K}, \phi_{1:K}) \propto p(x_i \mid z_i) = f(x_i \mid \phi_{z_i}) \tag{3}$$

$$p(\pi_{1:K} \mid z_{1:n}, x_{1:n}, \phi_{1:K}) \propto p(z_{1:n} \mid \pi_{1:K}) p(\pi_{1:K}) \tag{4}$$

$$p(\phi_k \mid z_{1:n}, x_{1:n}, \pi_{1:K}, \phi_{-k}) \propto \prod_{x \in \mathfrak{X}_k} p(x \mid \phi_k) p(\phi_k) \tag{5}$$

where $\mathfrak{X}_k = \{x_i : z_i = k, i = 1, \dots N\}$ is the set of all data points assigned to component k, and z_{-i} denotes the set of all assignment indicators except z_i, i.e., $z_{-i} = (z, \dots, z_{i-1}, z_{i+1}, ..., z_N)$. Due to the conjugacy of Multinomial and Dirichlet distributions the posterior for $\pi_{1:K}$ is again a Dirichlet; and with a conjugate prior, the posterior for $\phi_{1:K}$ will remain in the same form, hence they are straightforward to sample. Collapsed Gibbs inference scheme can also be developed to improve the variance of the estimators by integrating out $\pi_{1:K}$ and $\phi_{1:K}$, leaving out the only following conditional to sample from:

$$p(z_i = k \mid z_{-i}, x_{1:N}) \propto \frac{\int p(x_i \mid \phi_k) p(\phi_k \mid \{x_j : j \neq i, z_j = k\}) d\phi_k}{\eta + n_{-i,k}} \tag{6}$$

where η is the hyperparameter for $\pi_{1:K}$, assumed to be a symmetric Dirichlet distribution, and $n_{-i,k} = \sum_{j=1, j\neq i}^{N} 1_{z_j}(k)$ is the number of assignments to cluster k, excluding position i. The second term involves an integration which can easily be recognized as the predictive likelihood under the posterior distribution for $\phi_{1:K}$. For conjugate prior, this expression can be analytically evaluated. Several results can readily be found in many standard Bayesian text book such as [12].

A key theoretical limitation in the parametric Bayesian mixture model described so far is the assumption that the number of mixtures K in the data is known and *one has to specify it in advance* to apply this model. Recent advances in Bayesian nonparametric modeling (BNP) (e.g., see [13], [18]) provides a principled alternative to overcome these problems by introducing a nonparametric prior distribution on the parameters, which can be derived from Poisson point process or RFS.

B. Poisson RFS

The Poisson RFS, which models "no interaction" or "complete spatial randomness" in spatial point patterns, is arguably one of the best known and most tractable of point processes [35], [9], [38], [25], [20]. The Poisson RFS itself arises in forestry [36], geology [28], biology [24], particle physics [24], communication networks [2], [14], [15] and signal processing [22], [34], [7]. The role of the Poisson RFS in point process theory, in most respects, is analogous to that of the normal distribution in random vectors [8].

We briefly summarize the concept of Poisson RFS since this is needed to address the problem of unknown number of clusters and set-valued data. An RFS X on a state space $\mathcal{X}$ is random variable taking values in $\mathcal{F}(\mathcal{X})$, the space of finite subsets of $\mathcal{X}$. RFS theory is a special case of point process theory–the study of random counting measures. An RFS can be regarded as a simple-finite point process, but has a more intuitive geometric interpretation. For detailed treatments, textbooks such as [35], [9], [38], [25].

Let $|X|$ denotes the number of elements in a set X and $\langle f, g \rangle = \int f(x) g(x) dx$. An RFS X on $\mathcal{X}$ is said to be *Poisson* with a given *intensity function* v (defined on $\mathcal{X}$) if [35], [9]:

1) for any $B \subseteq \mathcal{X}$ such that $\langle v, 1_B \rangle < \infty$, the random variable $|X \cap B|$ is Poisson distributed with mean $\langle v, 1_B \rangle$,
2) for any disjoint $B_1, ..., B_i \subseteq \mathcal{X}$, the random variables $|X \cap B_1|, ..., |X \cap B_i|$ are independent.

Since $\langle v, 1_B \rangle$ is the expected number of points of X in the region B, the intensity value $v(x)$ can be interpreted as the instantaneous expected number of points per unit hyper-volume at x. Consequently, $v(x)$ is not dimensionless in general. If hyper-volume (on $\mathcal{X}$) is measured in units of $\mathcal{K}$ (e.g. m^d, cm^d, in^d, etc.) then the intensity function v has unit $\mathcal{K}^{-1}$.

The number of points of a Poisson point process X is Poisson distributed with mean $\langle v, 1 \rangle$, and condition on the number of points the elements x of X are independently and identically distributed (i.i.d.) according to the probability density $v(\cdot)/\langle v, 1 \rangle$ [35], [9], [38], [25]. It is implicit that $\langle v, 1 \rangle$ is finite since we only consider simple-finite point processes.

The probability distribution of a Poisson point process X with intensity function v is given by ([25] pp. 15):

$$\Pr(X \in \mathcal{T}) = \sum_{i=0}^{\infty} \frac{e^{-\langle v,1 \rangle}}{i!} \int_{\mathcal{X}^i} 1_{\mathcal{T}}(\{x_1, ..., x_i\}) v^{\{x_1,...,x_i\}} d(x_1, ..., x_i) \tag{7}$$

for any (measurable) subset $\mathcal{T}$ of $\mathcal{F}(\mathcal{X})$, where $\mathcal{X}^i$ denotes an i-fold Cartesian product of $\mathcal{X}$, with the convention $\mathcal{X}^0 =$

$\{\emptyset\}$, the integral over $\mathcal{X}^0$ is $1_{\mathcal{T}}(\emptyset)$ and $v^X = \prod_{x\in X} v(x)$. A Poisson point process is completely characterized by its intensity function (or more generally the intensity measure).

Probability densities of random finite sets considered in this work are defined with respect to the reference measure μ given by

$$\mu(\mathcal{T}) = \sum_{i=0}^{\infty} \frac{1}{i!\mathcal{K}^i} \int_{\mathcal{X}^i} 1_{\mathcal{T}}(\{x_1,...,x_i\}) d(x_1,...,x_i) \tag{8}$$

for any (measurable) subset $\mathcal{T}$ of $\mathcal{F}(\mathcal{X})$. The measure μ is analogous to the Lebesque measure on $\mathcal{X}$ (indeed it is the unnormalized distribution of a Poisson point process with unit intensity $v = 1/\mathcal{K}$ when the state space $\mathcal{X}$ is bounded). Moreover, it was shown in [40] that for this choice of reference measure, the integral of a function $f : \mathcal{F}(\mathcal{X}) \to \mathbb{R}$, given by

$$\int f(X)\mu(dX) = \sum_{i=0}^{\infty} \frac{1}{i!\mathcal{K}^i} \int_{\mathcal{X}^i} f(\{x_1,...,x_i\}) d(x_1,...,x_i), \tag{9}$$

is equivalent to Mahler's set integral [22]. Note that the reference measure μ, and the integrand f are all dimensionless. Probability densities for Poisson RFS take the form:

$$f(X) = \mathcal{K}^{|X|} e^{-\langle u,1\rangle} u^X. \tag{10}$$

Note that for any (measurable) subset $\mathcal{T}$ of $\mathcal{F}(\mathcal{X})$

$$\begin{aligned}
&\int_{\mathcal{T}} f(X)\mu(dX) \\
&= \int 1_{\mathcal{T}}(X) f(X)\mu(dX) \\
&= \sum_{i=0}^{\infty} \frac{e^{-\langle u,1\rangle}}{i!} \int_{\mathcal{X}^i} 1_{\mathcal{T}}(\{x_1,...,x_i\}) u^{\{x_1,...,x_i\}} d(x_1,...,x_i).
\end{aligned}$$

Thus, comparing with (7), f is indeed a probability density (with respect to μ) of a Poisson RFSs with intensity function u.

C. Infinite mixtures models with Dirichlet process

Recent advances in Bayesian nonparametric modeling (BNP) (e.g., see [13], [18]) addresses the unknown number of clusters by introducing a nonparametric prior distribution on the parameters. One way to motivate the Bayesian nonparametric setting is to reconsider the mixture likelihood in Eq (1). Let $\pi_{1:K} \sim \text{Dir}(\cdot|\eta), \phi_k \overset{\text{iid}}{\sim} h$, $k = 1,\ldots,K$ where $\text{Dir}(\cdot|\eta)$ is the symmetric Dirichlet distribution defined before in section II-A, and construct an atomic measure:

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \tag{11}$$

where δ_{ϕ_k} denotes the Dirac measure concentrated at ϕ_k. Note that for a region A on the parameter space, $G(A) = \sum_{k=1}^{K} \pi_k \mathbf{1}_A(\phi_k)$. The conditional distribution for x given G is

$$\begin{aligned}
p(x \mid G) &= \int f(x \mid \phi) G(d\phi) = \int f(x \mid \phi) \sum_{k=1}^{K} \pi_k \delta_{\phi_k}(d\phi) \\
&= \sum_{k=1}^{K} \pi_k \int f(x \mid \phi) \delta_{\phi_k}(d\phi) = \sum_{k=1}^{K} \pi_k f(x \mid \phi_k)
\end{aligned}$$

which identically recovers the likelihood form in Eq (1). Hence, the generative likelihood for the data point x can be equivalently expressed as: $x \sim f(\cdot \mid \phi)$ where $\phi \sim G$. Under this random measure formalism, inference amounts to deriving the posterior distribution for G.

To model an unknown number of clusters, let Ξ be a Poisson RFS on $\Omega \times \mathbb{R}^+$, with intensity function $v(\phi, w) = \eta h(\phi) w^{-1} e^{-w}$, where $\eta > 0$, and h is a probability density on Ω. Then the random measure

$$G = \frac{1}{\bar{w}} \sum_{(\phi,w)\in\Xi} w\delta_{\phi} \tag{12}$$

where $\bar{w} = \sum_{(\phi,w)\in\Xi} w$, is distributed according to the *Dirichlet process* [11], [19], [21], i.e.[1] $G \sim \text{DP}(\eta, h)$. The RFS Ξ captures the unknown number of clusters as well as the parameters of the clusters. This suggests an elegant and tractable[2] prior for G is the Dirichlet proces.

Briefly, a Dirichlet process $\text{DP}(\eta, h)$ is a distribution over random probability measures on the parameter space Ω and is specified by two parameters: $\eta > 0$ is the *concentration* parameter, and h is the base distribution [11]. The terms 'Dirichlet' and 'base distribution' come from the fact that for any finite partition of the parameter space Ω, the random vector obtained by applying G on this partition is distributed according to a Dirichlet distribution parametrized by ηh. More concisely, we say G is distributed according to a Dirichlet process, written as $G \sim \text{DP}(\eta, h)$ if for any *arbitrary* partition $(A_1,\ldots,A_m)$ of the space Ω, $(G(A_1),\ldots,G(A_m)) \sim \text{Dir}(\eta\langle h, 1_{A_1}\rangle,\ldots,\eta\langle h, 1_{A_m}\rangle)$. The Dirichlet process possesses an extremely attractive conjugate property, also known as the *Polya urn characterization [6]*: let $\phi_1,\ldots,\phi_m$ be i.i.d. samples drawn from G, then

$$p(\phi_m = \phi \mid \phi_1,\ldots,\phi_{m-1}) \tag{13}$$

$$= \frac{\eta h(\phi)}{m-1+\eta} + \frac{1}{m-1+\eta} \sum_{i=1}^{m-1} 1_{\phi_i}(\phi) \tag{14}$$

Using G as a nonparametric prior distribution, the data generative process for an infinite mixture models can be summarized as follows:

$$G \sim \text{DP}(\eta, h) \tag{15}$$

$$\phi_i \sim G \tag{16}$$

$$x_i \sim f(\cdot \mid \phi_i) \tag{17}$$

The recent book [18] provides an excellent account on the theory and applications of the Dirichlet Process.

Alternatively, the nonparametric measure G can be viewed as a limiting form of the parametric measure G in Eq (11)

[1] We note that commonly the Dirichlet process is expressed with a measure instead of its density, i.e., we could otherwise write $G \sim \text{Dir}(\eta, H)$ where H is a base measure whose density is h. However, the use of the density does not compromise the correctness in this paper, hence we equivalently use the notation $G \sim \text{Dir}(\eta, h)$ when the density h is the direct object of interest such as the commonly used likelihood Gaussian in signal processing.

[2] By 'tractable' we mean that the posterior is also a Dirichlet process.

when $K \to \infty$ and the weights $\pi_{1:K}$ are drawn from a symmetric Dirichlet Dir $\left(\frac{\eta}{K}, \ldots, \frac{\eta}{K}\right)$ [37]:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \tag{18}$$

The representation for G in Eq (18) is known as the *stick-breaking* representation, where $\phi_k \overset{\text{iid}}{\sim} h, k = 1, \ldots, \infty$ and $\pi_{1:\infty}$ are the weights constructed through a 'stick-breaking' process [33]. Imagine we are given a stick of length 1, if we infinitely break this stick into small pieces and assigned each piece to π_k, then clearly, $\sum_{k=1}^{\infty} \pi_k = 1$. Since the support of a Beta distribution is between 0 and 1, one may repeatedly sample a value from a Beta distribution and use this proportion as a principled way to break the stick. Formally, we construct the infinite dimensional vector $\pi_{1:\infty}$ as follows:

$$\begin{aligned} v_k \overset{\text{iid}}{\sim} &\text{Beta}\left(1, \eta\right), k = 1, \ldots, \infty \\ \pi_k = &v_k \prod_{s<k} \left(1 - v_s\right) \end{aligned}$$

It can be shown that with probability one $\sum_{k=1}^{\infty} \pi_k = 1$, and we denote this process as $\pi_{1:\infty} \sim \text{GEM}\left(\eta\right)$ (e.g., see [18], [29] for details).

III. Dirichlet Poisson RFS Mixture Models

A. Bayesian inference with Poisson RFS

In the previous section we see how Poisson-RFS are used to derive tractable priors, in this section we use Poisson-RFS to develop a tractable data model. Central to Bayesian analysis is the characterization of the posterior distribution and the predictive density that expresses the likelihood of a new data point upon the update of the posterior distribution.

We start by introducing some necessary notations. Let $f\left(\cdot \mid \Psi\right)$ be a parametric distribution. Occasionally, we use the parameter Ψ to index the distribution f_Ψ. For example, f is a Gaussian distribution, then $\Psi = (\boldsymbol{\mu}, \Sigma)$ specifies the mean and covariance matrix. Unless otherwise stated, we further use $h\left(\cdot \mid \gamma\right)$ to denote the conjugate prior for f in the sense that the posterior distribution $p\left(\Psi \mid x, \gamma\right) \propto f\left(x_i \mid \Psi\right) h\left(\Psi \mid \gamma\right)$ also has the same form as h (with a new parameters γ'). For example if f is a Gaussian with unknown mean and fixed variance, then h is a Gaussian, or if f is Poisson, then h is Gamma (e.g., see [12]).

As described previously, an RFS is a *random point pattern*. What distinguishes a RFS from a classic random vector-valued random variable is that the number of points, or elements, is random; and the points themselves are random and unordered, or simply, an RFS is a *finite-set-valued random variable* [39]. An RFS X can be fully parametrized by a discrete probability distribution to specify the cardinality of X and a family of joint distributions to describe the distribution of values of the points.

To facilitate our exposition in the sequel *we express a Poisson RFS X explicitly as an RFS whose cardinality distribution follows a Poisson distribution with the rate λ and elements x of X are independently and identically distributed (i.i.d) according to a probability distribution f_Ψ and write $X \sim$ PoissonRFS$\left(\lambda, f_\Psi\right)$.*

A Poisson RFS X can be sampled as follows: $X = \emptyset, n \sim \text{Poisson}\left(\lambda\right)$, then for $i = 1, \ldots, n$ we set $X = X \bigcup \{x_i\}$ where $x_i \overset{\text{iid}}{\sim} f_\Psi$ and Poisson $\left(\lambda\right)$ is a standard Poisson distribution with mean rate λ. Assume unit volume $\mathcal{K} = 1$, we express Eq (10) for Poisson-RFS likelihood density as:

$$p(X \mid \lambda, f_\Psi) = e^{-\lambda} \lambda^{|X|} f_\Psi^X$$

And when we wish to express the elements of X explicitly as $X = \{x_1, \ldots, x_n\}$, this likelihood density becomes [23]:

$$p\left(X = \{x_1, \ldots, x_n\} \mid \lambda, f_\Psi\right) = e^{-\lambda} \lambda^n \prod_{i=1}^{n} f_\Psi\left(x_i\right) \tag{19}$$

By convention, when X is an empty set, the RHS reduces to $e^{-\lambda}$. We note that X is parametrized by two parameters λ and Ψ; Let us write them jointly as $\theta = (\lambda, \Psi)$. Bayesian inference for Poisson-RFS requires the specification of the prior distribution over θ. *Furthermore we wish to develop a conjugate prior so that the posterior has the same form as the prior distribution*. The following proposition summaries our result.

Proposition 1. *Let $X \sim$ PoissonRFS $\left(\lambda, f_\Psi\right)$, and $h(\cdot \mid \gamma)$ be a conjugate prior of f_Ψ. Then the distribution given by*

$$p\left(\lambda, \Psi \mid \alpha, \beta, \gamma\right) = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} \lambda^{\alpha-1} e^{-\beta\lambda} h\left(\Psi \mid \gamma\right) \tag{20}$$

is the conjugate prior for X, in the sense that the posterior distribution $p\left(\lambda, \Psi \mid X, \alpha, \beta, \gamma\right)$ has the same form as (20).

Proof: To prove this, we note that the Gamma distribution is a conjugate prior for a Poisson distribution and H is conjugate to F, hence our first guess is that this conjugate structure will carry on for a Poisson-RFS. And, it turns out that this intuition is indeed correct as described below.

To see why, let $\lambda \sim \text{Gamma}\left(\alpha, \beta\right)$ so that $p\left(\lambda \mid \alpha, \beta\right) \propto \lambda^{\alpha-1} e^{-\beta\lambda}$ and using Baye's rule, the ***posterior*** distribution takes the form

$$\begin{aligned} p\left(\lambda, \Psi \mid X, \alpha, \beta, \gamma\right) &\propto p\left(X \mid \lambda, \Psi\right) p\left(\lambda, \Psi \mid \alpha, \beta, \gamma\right) \\ &\propto e^{-\lambda} \lambda^{|X|} f_\Psi^X \lambda^{\alpha-1} e^{-\beta\lambda} h\left(\Psi \mid \gamma\right) \\ &\propto \left[\lambda^{|X|+\alpha-1} e^{-(\beta+1)\lambda}\right] \left[f_\Psi^X h\left(\Psi \mid \gamma\right)\right] \end{aligned}$$

It is clear that this has the same form as the prior distribution in Eq (20) since the last term will results in h-like distribution due to conjugacy of h and f_Ψ. Given an observed X, the rate λ now follows Gamma $\left(|X| + \alpha, \beta + 1\right)$ and Ψ follows $h(\cdot \mid \gamma')$ where γ' is the posterior parameter resulting from $f_\Psi^X h\left(\Psi \mid \gamma\right)$ due the conjugacy of h and f_Ψ and have values depending on specification of h and f_Ψ. ■

By induction, the posterior distribution after observing N set-valued observation $\{X_1, \ldots, X_N\}$ is

$$\begin{aligned} &p\left(\lambda, \Psi \mid X_1, \ldots, X_n, \alpha, \beta, \gamma\right) \\ &\propto \left[\lambda^{\sum_{i=1}^{N} |X_i| + \alpha - 1} e^{-(\beta+N)\lambda}\right] \left[\prod_{i=1}^{N} f_\Psi^{X_i} h\left(\Psi \mid \gamma\right)\right] \end{aligned}$$

The posterior for λ is now Gamma $\left(\alpha_N, \beta_N\right)$ with $\alpha_N = \alpha + \sum_{i=1}^{N} |X_i|, \beta_N = \beta + N$; whereas Ψ follows $h(\cdot \mid \gamma_N)$

where γ_N is posterior parameter obtained from evaluating $\prod_{i=1}^{N} f_{\Psi}^{X_i} h(\Psi \mid \gamma)$.

As in a standard Bayesian analysis problem, given the observed data $D = \{X_1, \ldots, X_N\}$ it is important to be able to specify the *predictive* likelihood of an unseen observation X for a prediction task. For our mixture model developed in sequel, we use this likelihood in the Gibbs sampler to assess the likelihood of data points being assigned to cluster components. It turns out that this predictive density is also tractable for our Bayesian Poisson-RFS case. With a small effort of manipulation, this can be shown to be:

$$\begin{aligned} & p(X \mid X_{1:N}, \alpha, \beta, \gamma) \\ &= \int \int p(X \mid \lambda, \Psi)\, p(\lambda, \Psi \mid X_{1:N})\, d\lambda d\Psi \\ &= \left[\frac{\alpha_N \beta_N^{\alpha_N}}{(\beta_N + 1)^2}\right] \left[\int f_{\Psi}^{X} h(\Psi \mid \gamma_N)\, d\Psi\right] \end{aligned} \tag{21}$$

Again, depending on the specific forms for h and f_Ψ, the last term can be evaluated analytically (see [4] for several examples).

B. The Dirichlet Poisson RFS mixture model

The intuition for our proposed Dirichlet Poisson RFS Mixture Model (DP-RFS) is that *each mixture component is now a Poisson-RFS*, hence the model's support is now the space of finite sets. Therefore, we model set-valued data as random quantities and estimate a mixture density with an infinite number mixture components over these data. Since the data likelihood is a mixture of Poisson RFS densities, each mixture component is parameterised by the tuple $\phi_k = (\lambda_k, \Psi_k)$. To do so, let G follows a Dirichlet process whose base distribution is a conjugate prior specified in Eq (20). Using G as a nonparametric prior distribution, the data generative process for our model for N set-valued observations $\{X_1, \ldots, X_N\}$ can be summarized as follows:

$$G \sim \text{DP}(\eta, h') \tag{22}$$

$$\phi_i = (\lambda_i, \Psi_i) \sim G, \tag{23}$$

$$X_i \sim \text{PoissonRFS}(\lambda_i, f_{\Psi_i}) \tag{24}$$

where

$$h'(\lambda, \Psi \mid \alpha, \beta, \gamma) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} h(\Psi \mid \gamma)$$

taken as the conjugate prior developed in Eq (20). Our Dirichlet Poisson RFS mixture model then specifies an infinite mixture over a set-valued observation X as:

$$p(X \mid \pi_{1:\infty}, \phi_{1:\infty}) = \sum_{k=1}^{\infty} \pi_k \left[e^{-\lambda_k} \lambda_k^{|X|} f_{\Psi_k}^{X}\right] \tag{25}$$

C. Markov Chain Monte Carlo Inference

Given only the data $\{X_1, X_2, \ldots, X_N\}$, the concentration parameter η and the parameters α, β, γ for the base distribution h', our task is to infer a posterior distribution for $\pi_{1:\infty}$ and $\phi_{1:\infty}$. This is an intractable Bayesian inference problem and an MCMC inference scheme is needed. A full Gibbs inference similar to the scheme described in section II-A (cf. Eq 3–5) can be developed. For faster convergence, we describe in this section a collapsed Gibbs inference. We introduce the latent cluster indicators z_i to explicitly indicate the mixture component to which the data point X_i being assigned to and sample them directly, whereas $\pi_{1:\infty}$ and $\phi_{1:\infty}$ will be integrated out.

Using the stick-breaking represention for the Dirichlet process the data generative process can be now equivalently expressed as:

$$\begin{aligned} & (\lambda_k, \Psi_k) \overset{\text{iid}}{\sim} h'(\cdot \mid \alpha, \beta, \gamma) \text{ for } k = 1, 2, \ldots \\ & \quad \pi_k = v_k \prod_{s<k} (1 - v_s) \text{ where } v_k \overset{\text{iid}}{\sim} \text{Beta}(1, \eta) \\ & \text{For } i = 1, \ldots, N \\ & \quad z_i \sim \text{Discrete}(\pi_{1:\infty}) \\ & \quad X_i \sim \text{PoissonRFS}(\lambda_{z_i}, \Psi_{z_i}) \\ & \text{End} \end{aligned}$$

where the extra notation Discrete $(\cdot)$ denote a discrete distribution whose support is the set of positive integers.

Our aim is to perform posterior inference on the $p(z_{1:N} \mid X_{1:N}, \Phi)$, where $\Phi = \{\alpha, \beta, \gamma, \eta\}$ is the set of so-called *hyper-parameters*.

This inference can be carried out under a Gibbs sampling scheme using the Polya urn characterization of the Dirichlet process [6], otherwise also known as the Chinese restaurant process [30]. The structure of our inference scheme follows the work [27] for generic Gibbs inference for Dirichlet Process Mixture model. Central to this Gibbs inference scheme is the conditional distribution $p(z_i \mid z_{-i}, X_{1:N}, \Phi)$ from which one iteratively scans through each z_i and sample it. This conditional distribution can be expressed as follows using Bayes' rule and recall that the notation z_{-i} denotes the set of all assignment indicators except z_i, and likewise for X_{-i}:

$$\begin{aligned} & p(z_i = k \mid z_{-i}, X_{1:N}, \Phi) \\ &= p(z_i = k \mid z_{-i}, X_i, X_{-i}, \Phi) \\ &\propto p(X_i \mid z_i = k, z_{-i}, X_{-i}, \Phi)\, p(z_i = k \mid z_{-i}, X_{-i}, \Phi) \\ &\propto p(X_i \mid z_i = k, z_{-i}, X_{-i}, \Phi)\, p(z_i = k \mid z_{-i}, \Phi) \end{aligned} \tag{26}$$

Note that in the last term X_{-i} has been removed due to the fact that z_i is conditionally independent of X_{-i} given z_{-i} in the absence of X_i. Due to the Polya urn characterization of the Dirichlet process as described in Eq (13) the second term $p(z_i = k \mid z_{-i}, \Phi)$ can be written as:

$$p(z_i = k \mid z_{-i}, \Phi) = \begin{cases} \frac{n_{-i,k}}{n-1+\eta} & \text{if } k \text{ exists} \\ \frac{\eta}{n-1+\eta} & \text{if } k \text{ is new} \end{cases}$$

where we recall that $n_{-i,k} = \sum_{j=1, j\neq i}^{N} 1_{z_j}(k)$. This is also known the *Chinese Restaurant Process* in combinatorial stochastic process [30]. This expression illustrates the *clustering property* induced by the mode: a future data observation is more likely to return to an existing cluster with a probability

Algorithm 1 Collapsed Gibbs inference for the proposed Dirichlet Poisson RFS Mixture Models.

Input

- Set-valued observations $X_1, \ldots, X_N$
- Concentration parameter η and prior parameters α, β, γ
- Number of Gibbs samples L.

Collapse Gibbs inference

1) Initialize a random number of mixtures K (say 1)
2) Initialize randomly $z_1^{(0)}, \ldots, z_N^{(0)}$ so that $1 \leq z_i^{(0)} \leq K$
3) For $l = 1$ to L

   For $i = 1$ to n sample z_i from

$$p(z_i^{(l)} = k \mid z_{-i}^{(l-1)}, X_{1:N}, \gamma) \propto \begin{cases} n_{-i,k} f_k\left(X_i; X_{-i}\right) & \text{if } k \leq K \\ \eta f\left(X_i\right) & \text{if } k = K+1 \end{cases}$$

   If $z_i^{(l)} = K+1$, set $K \leftarrow K+1$
4) Remove any empty mixture component and decrease K accordingly.

Output:

- The number of mixture components learned K.
- L Gibbs samples $\left\{z_1^{(l)}, \ldots, z_N^{(l)}\right\}_{l=1}^{L}$ for the cluster indicators.

proportional to its popularity $n_{-i,k}$, but it is also flexible enough to pick on a new value if needed as data grows beyond the complexity that current model can explain. Furthermore, the number of clusters grow at $O\left(n \log \gamma\right)$ under the Dirichlet process prior [11], [1].

The first term $p\left(X_i \mid z_i = k, z_{-i}, X_{-i}, \Phi\right)$ in Eq (26) can be recognized as a form of predictive likelihood with respect to the mixture component k, where the predictive likelihood for unseen data point under Bayesian inference for Poisson RFS has been developed previously in section III-A (cf. Eq 21)

$$p\left(X_i \mid z_i = k, z_{-i}, X_{-i}, \Phi\right) = \int\int p\left(X_i \mid \lambda_k, \Psi_k\right) p\left(\lambda_k, \Psi_k \mid \{X_j : z_j = k, j \neq i\}\right) d\lambda_k d\Psi_k$$

and we shall denote this likelihood as $f_k\left(X_i; X_{-i}\right)$. Gibbs sampling then simply involves iteratively sampling $z_1, \ldots, z_N$ as summarized in Algorithm 1.

Note in this algorithm that when z_i takes on a new cluster, i.e., $z_i = K+1$ the predictive likelihood $f\left(X_i\right)$ is simply an integration over the prior distribution without observation any data point in this newly mixture component yet, i.e.,

$$f\left(X_i\right) = \int\int p\left(X_i \mid \lambda, \Psi\right) p\left(\lambda, \Psi \mid \alpha, \beta, \gamma\right) d\lambda d\Psi$$

In practice, we discard some initial Gibbs samples, a strategy commonly known as burn-in period in MCMC literature. In our experiment, to provide robustness we also sample the concentration parameter η according the procedure described in [10]; however it is not essential to understanding the Gibbs inference routine here, hence its description will be skipped.

IV. Numerical Results

This secion demonstrates the key properties of the proposed model via two numerical studies. We focus on one typical phenomenon in data modelling known as data clustering with extremely unbalanced datasets – an open challenging problem in data clustering analysis [17], [41]. We construct *five* Gaussians arranged in a star-shape: sitting at the center is a large-variance Gaussian specified with Poisson rate of 100, which *dominates* the generation of data; four other Gaussians scattered over the four corners and are specified with an extremely low Poisson rate of 0.5. Hence, as seen in Figure 1, the data looks as if it is generated solely by the dominant Gaussian and consequently this scenario presents a very challenging case to model the other four 'outlier' clusters. This is also known as an imbalanced data problem in related field of unsupervised learning and data mining and is frequently encountered in novelty and abnormality detection problem [17], [41], [5], [16].

Our baseline comparison is the state-of-the-art infinite Gaussian mixture model (iGMM) [31] which is a Bayesian nonparametric version of the classic Gaussian mixture models. This model can also bypass the model selection problem to automatically discover the number of clusters from the data. Input to iGMM is vector-valued data, hence we take the union of set-valued observations as the data for iGMM. We ensure that the initializations for our DP-RFS model and iGMM are as similar as possible and ran 500 Gibbs iterations after a small burn-in period. We keep track of the mode of the number of clusters as we progress and use the last result as our estimated result (equivalent to a MAP estimation with Gibbs sample).

Figure 1 presents the results of the simulation. The top figure shows the estimated number of clusters K varies with Gibbs iteration. We initialize $K = 1$ for both iGMM and our model. Note that iGMM tends to under estimate the number of clusters due to dominant cluster; our DP-RFS model, on the other hand, tends to over estimate the number of clusters at first, but gradually approaches the true number of cluster. This is partially explained by the use of Poisson RFS likelihood in the model, which provides the flexibility in creating spurious and skewed clusters to explain the data.

At termination, iGMM yields three clusters as seen in the bottom-left of the figure; and completely missed the four outlier clusters. The two Gaussians with diagonal direction appears to be affected and confused by the outlier clusters. Our DP-RFS model discovers 6 clusters, however one has an infinite variance and hence eliminated leaving five clusters plotted in the bottom-right of Figure 1. Our proposed technique has correctly identified the dominant cluster and all other four outlier clusters. Further, it estimates the Poisson rate for the dominant cluster to be 77.32 and the other four are $0.34, 0.35, 0.38$ and 0.37, which are quite close to the groundtruth.

To illustrate further clustering behaviors in the existence of imbalanced clusters, we present the results that used the common Mixture of Gaussians (MoG) for clustering tasks. While iGMM [31] and our proposed DP-RFS mixture model can automatically infer the number of clusters K from data, MoG requires us to specify this number in advance. Figure 2 presents the results for $K = 2, 3, 4, 5$ and 6. Again, in addition to the fact that MoG is unable to infer the number of clusters, it suffers a similar effect as observed in iGMM wherein the

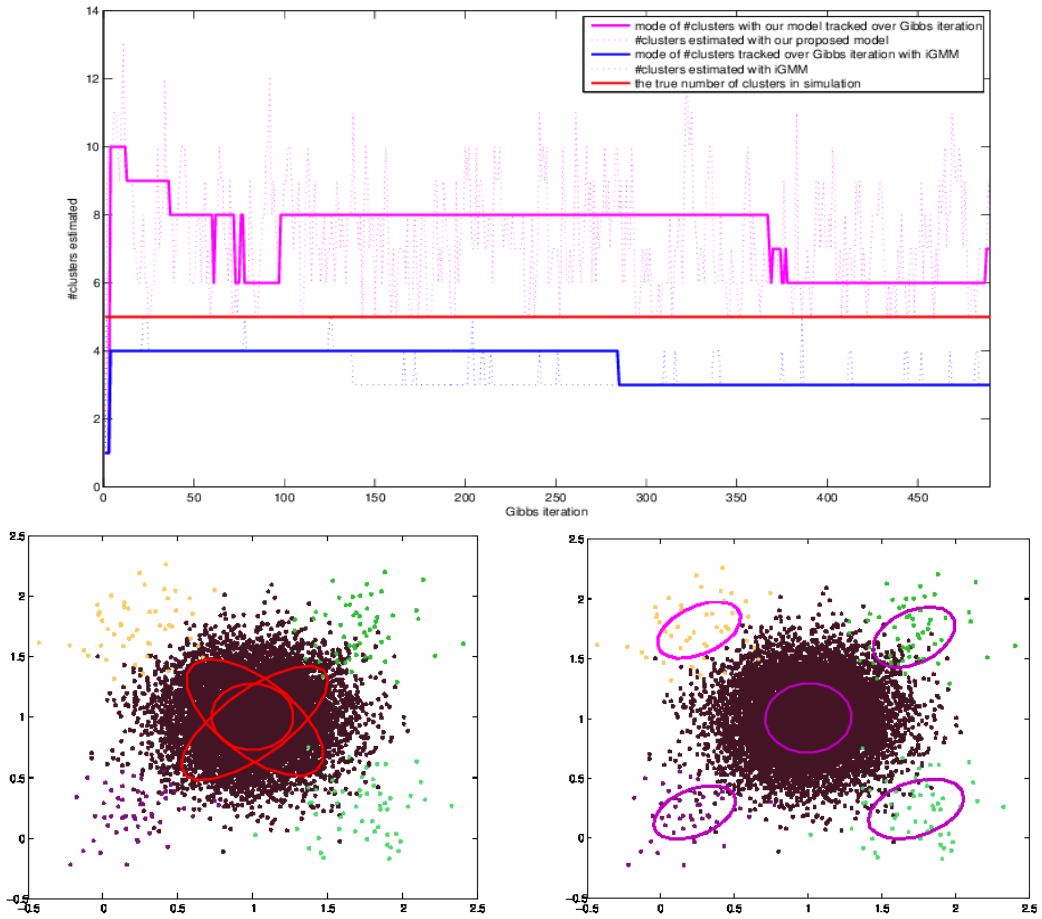

Figure 1. Numerical results. Five clusters created with one dominant clusters at the center (see main text for more details). Top: the estimated number of clusters varies with Gibbs iteration. Bottom-left: result of the current state-of-the-art infinite mixture of Gaussian [31] which misses four outlier clusters completely; Bottom-right: results from our DP-RFS mixture which correctly identifies all five clusters.

existence of the dominant cluster makes it almost impossible to learn the other four outlier clusters.

V. Discussion and Conclusion

In this paper we have shown how Poisson RFS can be used to develop infinite mixture model data clustering. In particular, we developed a conjugate prior for a Poisson-RFS likelihood with all of the properties of a typical Bayesian conjugate setting, including its conjugate posterior distribution and predictive density. Using this result, we constructed an infinite mixture of Poisson-RFS using the recently developed Dirichlet process theory for Bayesian nonparametric mixture models. This results in a new class of statistical models to both signal processing and machine learning: *it is an infinite mixture over set-valued data observations* and we term this model the Dirichlet Poisson Random Finite Set mixture model (DP-RFS). As set-valued observations arises naturally in everyday analysis tasks, we anticipate that this line of modelling will accommodate a wide range of applications. The numerical study presented in this paper has demonstrated the capacity of the proposed DP-RFS model to tackle the open challenge of modelling and clustering imbalanced data. Lastly, beyond Poisson-RFS, our framework opens the door to more general RFS models for data clustering.

References

[1] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[2] F. Baccelli and B. Blaszczyszyn. *Stochastic Geometry and Wireless Networks: Volume 1: Theory Foundation and Trends in Networking*, volume 1. Now Publishers Inc, 2010.

[3] A. Baddeley and M. Lieshout. Stochastic geometry models in high-level vision. *Journal of Applied Statistics*, 20(5-6):231–256, 1993.

[4] J. M. Bernardo and A. F. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[5] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[6] D. Blackwell and J. MacQueen. Ferguson distributions via Pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[7] F. Caron, P. Del Moral, A. Doucet, M. Pace, et al. On the conditional distributions of spatial point processes. *Advances in Applied Probability*, 43(2):301–307, 2011.

[8] D. R. Cox and V. Isham. *Point processes*. Chapman & Hall: Monographs on Applied Probability and Statistics, 1980.

[9] D. Daley and D. Vere-Jones. *An introduction to the theory of point processes*. Springer-Verlag, 1988.

[10] M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[11] T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[12] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2003.

[13] J. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics*. Springer Verlag, 2003.

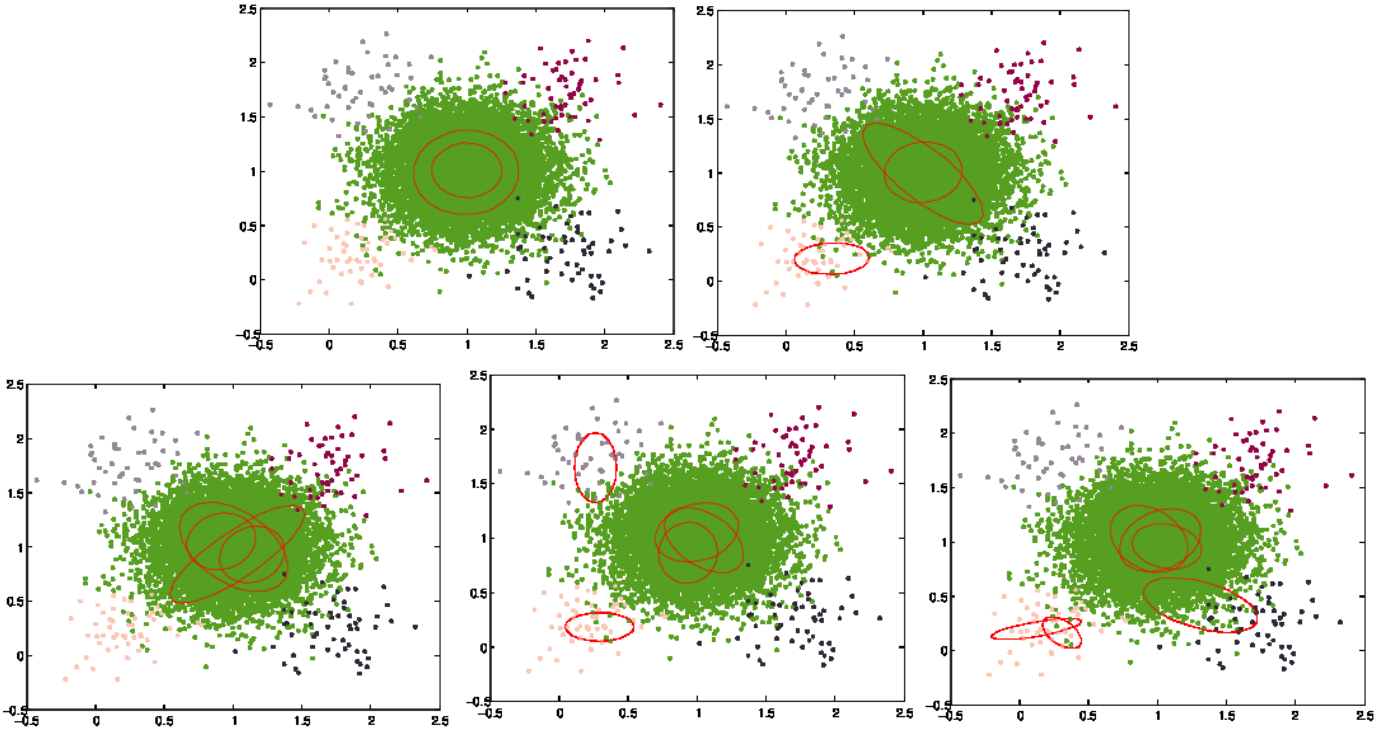

Figure 2. Clustering results using finite Mixture of Gaussian model (MoG). In this case, we need to specify the number of clusters in advances. From left to right and top to bottom, we set the number of clusters to be $2, 3, 4, 5, 6$ respectively.

[14] M. Haenggi. On distances in uniformly random networks. *IEEE Transactions on Information Theory*, 51(10):3584–3586, 2005.

[15] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *Selected Areas in Communications, IEEE Journal on*, 27(7):1029–1046, 2009.

[16] J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[17] H. He and E. A. Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

[18] N. Hjort, C. Holmes, P. Müller, and S. Walker. *Bayesian nonparametrics*. Cambridge University Press, 2010.

[19] M. Jordan. Hierarchical models, nested models and completely random measures. In P. M. D. S. M.-H. Chen, DK Dey and K. Ye, editors, *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*. Springer-Verlag, New York, NY, 2010.

[20] J. Kingman. *Poisson Processes*. Oxford University Press, 1993.

[21] D. Lin, E. Grimson, and J. Fisher. Construction of dependent dirichlet processes based on poisson processes. *Advances in Neural Information Processing Systems*, 2010.

[22] R. Mahler. Multi-target Bayes filtering via first-order multi-target moments. *IEEE Trans. Aerospace & Electronic Systems*, 39(4):1152–1178, 2003.

[23] R. P. Mahler. *Statistical multisource-multitarget information fusion*, volume 685. Artech House Norwood, 2007.

[24] V. Marmarelis and T. Berger. General methodology for nonlinear modeling of neural systems with poisson point-process inputs. *Mathematical biosciences*, 196(1):1–13, 2005.

[25] J. Moller and R. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall CRC, 2004.

[26] J. Mullane, B. Vo, M. Adams, and B.-T. Vo. A random finite set approach to Bayesian SLAM. *IEEE Transactions on Robotics*, 27(2):268–282, 2011.

[27] R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[28] Y. Ogata. Seismicity analysis through point-process modeling: A review. *Pure and Applied Geophysics*, 155(2-4):471–507, 1999.

[29] D. Phung. Bayesian nonparametric modelling of correlated data sources and applications (poster). In *International Conference on Bayesian Nonparametrics*, Amsterdam, The Netherlands, June 10-14 2013.

[30] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

[31] C. E. Rasmussen. The infinite Gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.

[32] C. P. Robert. *Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer-Verlag New York, 2001.

[33] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[34] S. Singh, B.-N. Vo, A. Baddeley, and S. Zuyev. Filters for spatial point processes. *SIAM Journal of Control and Optimization*, 48(4):2275–2295, 2009.

[35] D. Stoyan, D. Kendall, and J. Mecke. *Stochastic Geometry and its Applications*. John Wiley & Sons, 1995.

[36] D. Stoyan and A. Penttinen. Recent applications of point process methods in forestry statistics. *Statistical Science*, 15(1):61–78, 2000.

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

[38] M. van Lieshout. *Markov Point Processes and their Applications*. Imperial College Press, 2000.

[39] B. Vo. *Random finite sets in multi-object filtering*. PhD thesis, School of Electrical, Electronic and Computer Engineering, The University of Western Australia, 2008.

[40] B.-N. Vo, S. Singh, and A. Doucet. Sequential Monte Carlo methods for multi-target filtering with random finite sets. in *IEEE Trans. Aerospace & Electronic Systems*, 41(4):1224–1245, 2005.

[41] S.-J. Yen and Y.-S. Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, 2009.