

Linear Mixing Random Measures Based Mixture Models

CHENG LUO and YANG XIANG*

College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China

(Received on September 03, 2016, revised on October 16, 2016)

Abstract: When observations are organized into groups where commonalities exist amongst them, the traditional clustering models cannot discover shared clusters among groups. In this scenario, the dependent normalized random measures based clustering is a perfect choice. The most interesting property of the proposed LMRM based clustering is that the clusters are assumed to be shared across groups. Hence the problem can be solved immediately. We derive appropriate exchangeable probability partition function, and subsequently also deduce its inference algorithm given any mixture model likelihood. We provide all necessary derivation and solution to this framework. For demonstration, we used mixture of Gaussians likelihood in combination with a dependent structure constructed by linear combinations of completely random measures. Our experiments show superiority performance when using this framework, where the inferred values including both the mixing weights and the number of clusters both respond appropriately to the number of completely random measure used.

Keywords: Clustering models, Mixture of Gaussians, Normalized dependent random measures, Bayesian Non-parameters, Mixed random measures.

1. Introduction

Clustering is one of the most important applications in machine learning and data mining. However, there are two problems in traditional clustering algorithms. The first one is that the number of clusters should be set in advance. Hence experimenters have to try different settings manually, which is surely exhausting. The second one is that when observations are divided into groups, the traditional clustering algorithms can only build models for each of them, and hence different groups use totally different clusters. For example, in clinical data analysis, we divide patients into two groups: treatment group and control group. All the patients can be seen in the same cluster before they take pills or placebos. But after the experiment, the treatment and control groups may have distinct clusters. Hence some clusters are shared by both groups while some clusters are owned by distinct groups. However, the traditional clustering algorithms like K-means, fuzzy C-means or hierarchical clustering cannot deal with this situation.

In order to tackle the first problem, the Infinite Gaussian Mixture Model (IGMM) was proposed [1]. This model is based on the Dirichlet process, one of the most famous random measures. In IGMM, it assumes the observation $x_i \sim F(\theta_i)$, where F is a known likelihood function, and $\theta_i (i = 1, 2, 3 \dots)$ is the parameter of x_i . When θ_i has only finite values, the traditional finite clustering algorithms are formed. However, in IGMM, θ_i is sampled from a Dirichlet Process (DP). Hence there are infinite values that θ_i can take, and thus it is an infinite extension of the traditional mixture models. Because DP is a normalized Gamma process [2], which is in the family of the Completely Random Measures (CRM), one can easily substitute DP with other Normalized Random Measures (NRM) and get different infinite mixture models.

When the observed datasets are divided into groups, we need to use more complex models. In this scenario, we should define a vector of dependent NRMs. The pioneering

*Corresponding author's email: shxiangyang@tongji.edu.cn

work is to consider the dependent random measures [3]. However, it is a rather limiting model. They only considered the dependent stick-breaking process. To extend that model, researchers adopted the dependent Levy measures. For instance, Griffin [5] proposed the Correlated Normalized Random Measures with Independent Increments (CNRMI) by introducing a binary matrix W and constructing NRMs $\tilde{\mu}$ through $\tilde{\mu} = W\mu$, where μ is a vector of independent NRMs.

There are two problems with using CNRMI. Firstly, they use a binary matrix W rather than a real matrix, which clearly limits the ability of the model. Secondly, they expand the random measures as infinite discrete distributions. That means, every NRM $\tilde{\mu}_i$ was written in the form of $\tilde{\mu}_i = \sum_{j=1}^{\infty} l_{i,j} \delta_{x_j}$, where $l_{i,j}$ and x_j are the parameters to be inferred. Hence, they have to truncate the infinite serials at some level K , and approximate $\tilde{\mu}_i$ by $\tilde{\mu}_i \approx \sum_{j=1}^K l_{i,j} \delta_{x_j}$. In this paper, we substitute the binary matrix W with a real matrix, and call the constructed model the Linear Mixing Random Measures (LMRM). The second problem is that we construct an inference algorithm using the Exchangeable Partition Probability Function (EPPF) [5][7], and hence avoid the truncation of $\tilde{\mu}_i$.

After that, we present an example of how to apply LMRM in real applications. Since LMRM represents a family of dependent NRMs, we need to specify a concrete form of it. In this paper, we choose the Gamma processes as the basic CRMs. Concretely, we choose $\mu = (\mu_1, \dots, \mu_d)$, where μ_1 is Gamma process. We give the details of how to derive the EPPF of this specific LMRM, and apply it to synthetic and real datasets. The experiments show our model is able to find the shared clusters amongst groups as well as the group specific clusters. Interesting readers can also substitute the Gamma process with other CRMs, such as the σ -stable process [4], the generalized Gamma process [10], the inverse Gaussian process [10] or other CRMs under our framework.

The contributions of this paper are summarized as follows:

1. We extent the CNRMI to LMRM and give the EPPF of the LMRM. In CNRMI, the weight matrix W is a binary matrix, while in LMRM, the weight matrix W is a real matrix.
2. We build the LMRM based clustering model, and show how to infer the parameters. We assume the parameters of the clusters are sampled from the LMRM. Hence every group is an infinite mixture model. When the parameter of the clusters are given, the observations are generated by their associated clusters' parameters and a given likelihood function.
3. We show how to apply the model to real applications when the underlying CRMs are Gamma processes.

The rest of the paper is organized as follows: Section 2 derives the EPPF of the LMRM. Section 3 constructs the LMRM based infinite mixture model and shows how to derive the posterior of the parameters. In section 4 we set the underlying CRMs to be Gamma processes and derive its EPPF in detail. The computational examples are given in section 5 and the conclusion and future work are given in the last section.

2. The EPPF of the LMRM

2.1. Basic notations and properties

We first give the formal definition of CRM and NRM, and the corresponding EPPF associated with the dependent NRMs. Then we deduce the EPPF of the vector NRMs.

Let $(X, B(X))$ be a measurable space and μ is called a CRM if $\mu(A_1), \dots, \mu(A_n)$ are independent random variables for any disjoint sets $A_1, \dots, A_n \in B(X)$. The associated NRM is defined by $\mu/\mu(X)$, providing $\mu(X) < \infty$. Let (μ_1, \dots, μ_d) be a vector of CRMs defined on the measurable space $(X, B(X))$. They are called dependent CRMs if $\mu_1(A), \dots, \mu_d(A)$ are dependent for $A \in B(X)$. The vector of dependent NRMs is defined by $(\tilde{\mu}_1, \dots, \tilde{\mu}_d) = (\mu_1/\mu_1(X), \dots, \mu_d/\mu_d(X))$.

By the Levy-Khintchine representation, the dependent CRMs is determined by

$$\mathbb{E} \left[e^{-\sum_{i=1}^d \mu_i(f)} \right] = \exp \left\{ - \int \left(1 - e^{-\sum_{i=1}^d s_i f_i(x)} \right) \nu(ds_1, \dots, ds_d) H(dx) \right\}$$

where f_i is a measurable function almost surely for $i = 1, \dots, d$.

Now we introduce the definition of the EPPF. Let $X = \{X^{(1)}, \dots, X^{(d)}\}$ be a set of observations and $X^{(i)} = \{X_{i,1}, \dots, X_{i,n_i}\}$ denote the observations in the i -th group and $X_{i,j}$ is a X -valued random variable for $i = 1, \dots, d$ and $j = 1, \dots, n_i$. The data X is said to be distributed as the dependent NRMs $(\tilde{\mu}_1, \dots, \tilde{\mu}_d)$ if $X_{i,j} | (\tilde{\mu}_1, \dots, \tilde{\mu}_d) \sim \tilde{\mu}_i$ for all i and j . Pointed out by de Finetti [8], the probability of X can be written in the following way. Let $C_{i,j} \in B(X)$ be a measurable set, the probability of the sample X is

$$\Pr(\{X_{i,j} \in C_{i,j}, i = 1, \dots, d, j = 1, \dots, n_i\}) = \int \prod_{i=1}^d \prod_{j=1}^{n_i} \tilde{\mu}_i(C_{i,j}) \Phi(d\tilde{\mu}_1, \dots, d\tilde{\mu}_d)$$

where Φ is the probability of $(\tilde{\mu}_1, \dots, \tilde{\mu}_d)$.

We assume the sample X has K distinct values $\{Y_1, \dots, Y_K\}$, and for each k there are $q_{i,k}$ variables in group i having the value Y_k . Let C_1, \dots, C_K be disjoint measurable sets, the probability of the sample can be rewritten as

$$\Pr(X, \{Y_k \in C_k, k = 1, \dots, K\}) = \int \prod_{k=1}^K \prod_{i=1}^d \tilde{\mu}_i(C_k)^{q_{i,k}} \Phi(d\tilde{\mu}_1, \dots, d\tilde{\mu}_d)$$

Let $C_k = C_{k,\epsilon} = \{y: d(Y_k, y) < \epsilon\}$ and $\epsilon \downarrow 0$, the EPPF of the dependent NRMs is defined by

$$\Pi(K, \{n_i, q_{i,k}, i = 1, \dots, d, k = 1, \dots, K\}) = \int \mathbb{E} \left[\prod_{k=1}^K \prod_{i=1}^d \tilde{\mu}_i(dx)^{q_{i,k}} \right] \quad (1)$$

where K is the number of partitions, n_i is the number of observations in each group, and $q_{i,k}$ is the number of observations in group i and partition k . Here we substitute \Pr with Π to denote the special probability of the EPPF. From (1) we can see that the probability of an EPPF is determined by the number of distinct values K , and the counting of Y_k in each group, which is denoted by $q_{i,k}$.

2.2. The EPPF of the dependent NRMs

The EPPF of the NRMs (the expression of Equation (1)) is determined by the following proposition. The auxiliary variables u_1, \dots, u_d are introduced in the following way. We first substitute $\tilde{\mu}_i$ with μ_i/T_i , where $T_i = \mu_i(\mathbf{X})$. Then we remove the denominator by introducing an auxiliary variable u_i

$$T_i^{-n_i} = \frac{1}{\Gamma(n_i)} \int u_i^{n_i-1} e^{-T_i u_i} du_i$$

The proof of the following proposition is similar with that in [6], which we omit it here to save space.

Proposition 1. Let K, n_1, \dots, n_d be the positive integers such that $K \leq \sum_{i=1}^d n_i$, the EPPF of the dependent NRMs is

$$\begin{aligned} \Pi(K, \{n_i, q_{i,k}, i = 1, \dots, d, k = 1, \dots, K\}) \\ = \int \left(\prod_{i=1}^d \frac{u_i^{n_i-1}}{\Gamma(n_i)} \right) e^{-\psi(u_1, \dots, u_d)} \left(\prod_{k=1}^K \tau_{q_k}(u_1, \dots, u_d) du_1, \dots, du_d \right) \end{aligned}$$

where $\mathbf{q}_k = (q_{1,k}, \dots, q_{d,k})$ and

$$\psi(u_1, \dots, u_d) = \int \left(1 - e^{-\sum_{i=1}^d u_i s_i} \right) \nu(ds_1, \dots, ds_d) \quad (2)$$

$$\tau_{q_k}(u_1, \dots, u_d) = \int e^{-\sum_{i=1}^d s_i u_i} \prod_{i=1}^d s_i^{q_{i,k}} \nu(ds_1, \dots, ds_d) \quad (3)$$

2.3. The EPPF of the LMRM

The LMRM is indeed a special case of the dependent NRMs. By the general form shown in Proposition 1, the EPPF of the LMRM follows once we know the form of $\psi(u_1, \dots, u_d)$ and $\tau_{q_k}(u_1, \dots, u_d)$. Since they are determined by the Levy measure $\nu(ds_1, \dots, ds_d)$, we only need to compute it given the underlying CRMs and the matrix \mathbf{W} .

Let $\bar{\mu}_1, \dots, \bar{\mu}_R$ be independent CRMs and $\mathbf{W} = (w_{i,r})$ be the weight matrix, where $i = 1, \dots, d$ and $r = 1, \dots, R$. The i -th element of the LMRM $\tilde{\mu}_i$ is constructed by $\tilde{\mu}_i = \mu_i / \mu_i(\mathbf{X})$, where $\mu_i = \sum_{r=1}^R w_{i,r} \bar{\mu}_r$. Then we focus on the derivation of the Levy measure $\nu(ds_1, \dots, ds_d)$ of $(\tilde{\mu}_1, \dots, \tilde{\mu}_d)$. Assume the Levy measure of $\bar{\mu}_1, \dots, \bar{\mu}_R$ is $\nu^*(ds)$, then we have the Laplace functional

$$\mathbb{E} \left[e^{-\sum_{i=1}^d \tilde{\mu}_i(f_i)} \right] = \exp \left\{ \sum_{r=1}^R \left(1 - e^{-\sum_{i=1}^d s_i f_i(x)} \right) \delta_{sW_{1,r}, \dots, sW_{d,r}}(s_1, \dots, s_d) \nu^*(ds) H(dx) \right\}$$

This gives us the Levy measure of $(\tilde{\mu}_1, \dots, \tilde{\mu}_d)$

$$\nu(ds_1, \dots, ds_d) = \sum_{r=1}^R \int \delta_{sW_{1,r}, \dots, sW_{d,r}}(ds_1, \dots, ds_d) \nu^*(ds) \quad (4)$$

Insert the expression of $\nu(ds_1, \dots, ds_d)$ in Equation (4) back into $\psi(u_1, \dots, u_d)$ and $\tau_{q_k}(u_1, \dots, u_d)$ shown in Proposition 1 gives the EPPF of the LMRM.

3. The LMRM based clustering model

3.1. Model description

In this section, we describe the LMRM based clustering model and its inference algorithm. We assume the observations X are divided into d groups and some of the clusters are shared across groups. Denote $X_i = \{X_{i,1}, \dots, X_{i,n_i}\}$ the i -th group and $\theta_{i,j}$ the parameter of $X_{i,j}$, where $j = 1, \dots, n_i$. We do not assume directly that the observation $X_{i,j}$ is generated by the LMRM. Instead, we assume its parameter $\theta_{i,j}$ is sampled from the LMRM, and $X_{i,j}$ is sampled from $f(X_{i,j} | \theta_{i,j})$, where f is the link function. Since the LMRM is determined by its underlying CRMs's Levy measure and their number R , we parametrize the LMRM by $\text{LMRM}(\nu^*, H, R)$. We assume the observations are generated by the following model

$$\begin{aligned} (\tilde{\mu}_1, \dots, \tilde{\mu}_d) &\sim \text{LMRM}(\nu^*, H, R) \\ X_{i,j} | \theta_{i,j} &\sim f(\cdot | \theta_{i,j}) \end{aligned}$$

$$\theta_{i,j} | (\tilde{\mu}_1, \dots, \tilde{\mu}_d) \sim \tilde{\mu}_i(\cdot)$$

Remark. By the property of the LMRM, the mixing measure $\tilde{\mu}_i$ is a discrete distribution almost surely, and some of the atoms are shared (See [22] for the explanation). Hence the probability of $\theta_{i,j} = \theta_{k,l}$ is greater than 0. When, $\theta_{i,j} = \theta_{k,l}$ the observations $X_{i,j}$ and $X_{k,l}$ are in the same cluster.

3.2. Inference

Suppose we have observed n data points, and there are n_i data points in group i . Assume there are K clusters, and for cluster k there are $q_{i,k}$ ($q_{i,k} \geq 0, 1 \leq k \leq K$) observations in group i attached to it. Now we compute the probability of a new observation X_{i,n_i+1} to be assigned to cluster k or a new cluster. Denote $c_{i,j}$ the indicator of $X_{i,j}$, and $c_{i,j} = k$ means $c_{i,j}$ is assigned to cluster k while $c_{i,j} = \text{new}$ means it is assigned to a new cluster. By virtual of Proposition 1, we have

$$\Pr(c_{i,n_i+1} = k, u_1, \dots, u_d) \propto \frac{\tau_{q_k + \delta_i}(u_1, \dots, u_d)}{\tau_{q_k}(u_1, \dots, u_d)} f(x_{i,j} | \theta_k) \quad (5)$$

and

$$\Pr(c_{i,n_i+1} = \text{new}, u_1, \dots, u_d) \propto \psi(u_1, \dots, u_d) \int f(x_{i,j} | \theta_k) H(d\theta) \quad (6)$$

where δ_i is a binay vector of length d with the only non-zero value in the i -th element.

In addition to the observations, we need also to sample the values of the auxiliary variables u_1, \dots, u_d . From Proposition 1, we can see that the density of u_i is propositional to

$$u_i^{n_i-1} \psi(u_1, \dots, u_d) \prod_{k=1}^K \tau_{q_k}(u_1, \dots, u_d)$$

Lastly, the parameter θ_k is updated with density

$$\left(\prod_{c_{i,j}=k} f(x_{i,j} | \theta) \right) H(d\theta)$$

4. The Gamma Process Based LMRM

In the previous section, we present the LMRM based clustering model and how the parameters are inferred. However, in real applications we need to specify the form of \mathbf{v}^* . In this paper, we choose \mathbf{v}^* to be the Levy measure for Gamma process

$$\nu^*(ds) = \alpha s^{-1} e^{-s}$$

Insert this formula into (2) and (3) to get the following equalities

$$\tau_{q_k}(u_1, \dots, u_d) = \int e^{-\sum_{i=1}^d s_i u_i} \prod_{i=1}^d s_i^{q_{i,k}} \nu(ds_1, \dots, ds_d)$$

$$\psi(u_1, \dots, u_d) = \sum_{r=1}^R \frac{\prod_{i=1}^d w_{i,r}^{q_{i,k}}}{(h_r + 1)^{t_k}}$$

where $t_k = \sum_{i=1}^d q_{i,k}$ and $h_r = \sum_{i=1}^d w_{i,r} u_i$. Substitute these terms into (5). (6) shows the probability of $c_{i,j} = k$ is proportional to

$$t_k f(x_{i,j} | \theta_k) \left(\sum_{r=1}^R \frac{w_{i,r}^{q_{i,k+1}} \prod_{l=1, l \neq i}^d w_{l,r}^{q_{l,k}}}{(h_r + 1)^{t_k}} \right) \left(\sum_{r=1}^R \frac{\prod_{l=1}^d w_{l,r}^{q_{l,k}}}{(h_r + 1)^{t_k}} \right)^{-1}$$

And the probability of $c_{i,j} = \text{new}$ is proportional to

$$\alpha \int f(x_{i,j} | \theta) H(d\theta) \sum_{r=1}^R \frac{w_{i,r}}{w_{i,r} u_i + 1}$$

5. Illustration Examples

5.1. Synthetic data

We define $\mu_1 = \frac{1}{2}N(-10,1) + \frac{1}{2}N(-5,1)$, $\mu_2 = \frac{1}{2}N(0,1) + \frac{1}{2}N(5,1)$, and

$\mu_3 = \frac{1}{2}N(10,1) + \frac{1}{2}N(15,1)$, where $N(a, b)$ denotes the Normal distribution with mean

a and standard deviation b . We construct the two groups $\tilde{\mu}_1 = 0.3\mu_1 + 0.01\mu_2 + 0.69\mu_3$ and $\tilde{\mu}_2 = 0.3\mu_1 + 0.69\mu_2 + 0.01\mu_3$. It can be seen that the first cluster μ_1 is shared by the two groups while μ_2 is held only by the first group $\tilde{\mu}_1$ and μ_3 is held only by the second group $\tilde{\mu}_2$. We generate 1000 data points for each of the group and run the LMRM based clustering and want to show that the structure of the data set can be discovered.

In our experiments, we set R to be 2, 3 and 4 and observe the values of the mixing weights $w_{i,r}$, and the activated number of clusters. We set the base measure to be $N(0, 2.6)$, where 2.6 is the approximate standard deviation for the sample. We test the performance of the model by running Gibbs sampling [9] for 10000 iterations and discard the first 2000 iterations as burn in to show the performance of the model. The average

values of the mixing weights and the number of clusters are shown in Table 1.

Table 1: Parameters discovered for synthetic dataset

	$W_{1,1}$	$W_{1,2}$	$W_{1,3}$	$W_{1,4}$	$W_{2,1}$	$W_{2,2}$	$W_{2,3}$	$W_{2,4}$	K
$R=2$	2.1560	0.0743	---	---	0.0504	2.0894	---	---	8.07
$R=3$	2.0379	0.0635	1.0038	---	0.0346	1.6804	0.8997	---	6.04
$R=4$	2.3880	1.0882	0.7872	1.0874	0.0504	1.3252	2.3666	11.3251	6.14

It can be seen that when $R = 3$ or 4 (the second and the last row in Table 1), the average number of clusters is approximately 6, and one of the weights is very small in each group as expected. Furthermore, the proportions of the weights in each group are approximately equal to the true proportions (0.01 : 0.35 : 0.69).

However, when we set $R = 2$ (the first row), the model cannot detect the structure of the sample since there are fewer assumed CRMs than the real settings. As a result, the activated number of the clusters is 8 instead of 6. It can be inferred that our algorithm split the shared CRM into two and allocate them to the two groups.

5.1. Clinical data

Besides the synthetic data set, we also test our algorithm in a clinical data set. This data consists of drug information collected on 50 patients. The data set comprises of 6 variables: Subject, Treatment, Age, Gender, Before_exp_BP and After_exp_BP, where Treatment is a binary variable with 1 for treatment and 0 for placebo, and Before_exp_BP and After_exp_BP are for blood pressure before and after experiment respectively. We construct the first group using the blood pressure for patients taking real pills and the second group for placebo. Figure 1 shows the histograms of the two groups. It can be seen that the first group consists of three clusters with centers are approximately 85, 95, 105 while the second group consists of two clusters with centers are approximately 91, 98. It means one of the clusters is shared by the two groups. In fact, most of the patients have similar blood pressures (95 v.s. 98) before experiments except for a few people having extremely high blood pressure (105 to 115). But after the experiments, the patients taking real pills have a lower blood pressure than those having placebos (85 v.s. 91).

Table 2: Parameters discovered for clinical dataset

	$W_{1,1}$	$W_{1,2}$	$W_{1,3}$	$W_{1,4}$	$W_{2,1}$	$W_{2,2}$	$W_{2,3}$	$W_{2,4}$	K
$R=2$	0.9590	0.9300	---	---	6.0546	6.1913	---	---	3.04
$R=3$	1.4742	0.3496	0.6308	---	0.0012	1.1795	0.9358	---	4.02
$R=4$	0.5355	0.8963	1.4557	0.5148	1.1265	3.6203	0.0002	1.0332	4.02

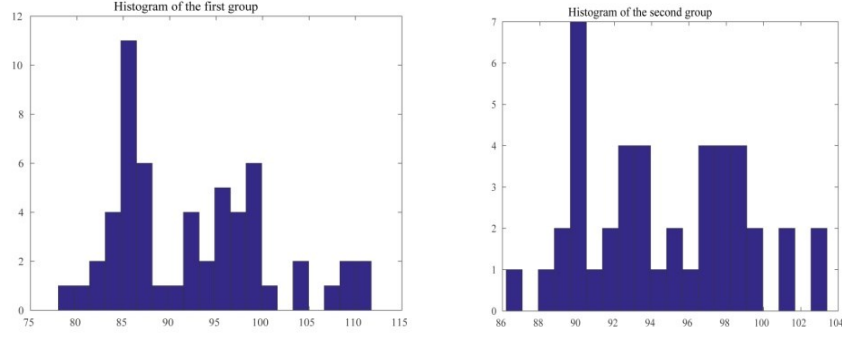


Fig.1: The histogram of the clinical data set

Table 3: Comparing LMRM with the traditional clustering model CRP

	LMRM				CRP				
	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4	c_5
G_1	---	109.6	85.27	98.27	---	110.4	86.35	96.43	---
G_2	92.30	---	---	98.27	90.62	---	---	---	97.71

We assume the observations are distributed as mixtures of Gaussians like the synthetic data. The standard deviation for each cluster is assumed to be 3 and the standard deviation for the base measure is assumed to be 6. We run the Gibbs sampling for 10000 iterations. The average values of the mixing weights are shown in Table 2. It is clear that when we set $R = 3$ or $R = 4$, one of the mixing weight in group 2 ($w_{2,1}$ and $w_{2,2}$) is close to 0, meaning there is one CRM being ignored in the second group. However, when we set $R = 2$, all the mixing weights are relatively large, indicating that this setting cannot discover the structure of the sample.

We also compared our model with traditional Chinese Restaurant Process (CRP), which clusters different groups separately. It can be easily seen from Table 3 that CRP creates redundant clusters. In Table 3, we use c_k to denote the center of cluster k . In fact, it splits the data of the people before taking the experiments into two (c_4 in LMRM and c_4 and c_5 in CRP). Obviously, before the experiments, the blood pressure in both groups should be in one cluster regardless of whether they take real pills or placebos. But our proposed LMRM can avoid this problem.

6. Conclusions

In this paper, we have proposed a framework for modeling mixture models when observations are organized in groups and the prior is a LMRM. We pointed out that when the Levy measure of the LMRM is given, the EPPF can be derived analytically and hence

the inference resemble that of the product of independent Chinese Restaurant Processes with an additional term, parameterized by (u_1, \dots, u_d) to capture their dependency structure. As a special case, we show the details of how to apply the model to real applications when the underlying CRMs are Gamma processes.

In experiments, we showed the performance and the superiority of our model in both synthetic and clinical data. In particular, we systematically evaluate our model under different R , i.e., the number of the assumed CRMs. It can be seen that when R is greater or equal to the ground-truth, the inferred clustering information largely agrees with the true structure of the data samples, and across the board in all groups. We also noted that the clusters will split up or mixed together when R is smaller than of the ground-truth.

Acknowledgements

This work was supported by 973 Program of China [grant numbers 2014CB340401]; International Exchange Program for Graduate Students, Tongji University.

References

- [1]. Tamara Broderick, Michael I Jordan, and Jim Pitman. *Cluster and Feature Modeling from Combinatorial Stochastic Processes*. Statistical Science, 2013; 28(3): 289–312.
- [2]. Changyou Chen, Vinayak Rao, Wray Buntine, and Yee Whye Teh. *Dependent normalized random measures*. Proceedings of The 30th International Conference on Machine Learning. 2013; 28(1): 969–977.
- [3]. S. Favaro, A. Lijoi, and I. Prunster. *On the stick-breaking representation of normalized inverse Gaussian priors*. Biometrika. , 2012; 99(3): 663–674.
- [4]. Stefano Favaro and Yee Whye Teh. *MCMC for Normalized Random Measure Mixture Models*. Statistical Science. 2013; 28(3): 335–359.
- [5]. J. E. Griffin, M. Kolossiatos, and M. F J Steel. *Comparing distributions by using dependent normalized random-measure mixtures*. Journal of the Royal Statistical Society. Series B: Statistical Methodology. 2013; 75(3): 499–529.
- [6]. Fabrizio Leisen, Antonio Lijoi, and Dario Spano. *A vector of dirichlet processes*. Electronic Journal of Statistics. 2013; 7(1): 62–90.
- [7]. Antonio Lijoi and Bernardo Nipoti. *Dependent mixture models: Clustering and borrowing information*. Computational Statistics and Data Analysis. 2014; 71(1): 417–433.
- [8]. Liming Wang and Xiaodong Wang. *Hierarchical Dirichlet process model for gene expression clustering*. EURASIP journal on bioinformatics & systems biology. 2013; 2013(1): 5–10.
- [9]. Yi-an Ma, Tianqi Chen, and Emily B Fox. *A Complete Recipe for Stochastic Gradient MCMC*. In Nips, number Mcmc. 2015; 2015(1): 1–19.
- [10]. Mingyuan Zhou and Carin, Lawrence. *Negative binomial process count and mixture modeling*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015; 37(2): 307–320.

Cheng Luo is a PHD student at Tongji University. His current research interests include statistical machine learning and Bayesian statistics.

Dr. Yang Xiang is a professor and doctoral supervisor at Tongji University. His research interests include computational linguistics and data mining.