A Latent Variable Bayesian Approach to Spatial Clustering with Background Noise

Koray Kayabol*†

First version: November 30, 2011. Revised last version: February 14, 2012.

Abstract

We propose a finite mixture model for clustering of the spatial data patterns. The model is based on the spatial distances between the data locations in such a way that both the distances of the points to the cluster centers and the distances of a given point to its neighbors within a defined window are involved in the model. Nevertheless, we take into consideration the background noise as well in the model. We resort to Classification Expectation-Maximization (CEM) algorithm for both estimating the parameters and clustering the data points. We test the algorithm on some simulated data sets with different background noise levels and apply it to a real earthquake data recorded in Kashmir in 2005.

Keywords: spatial clustering, finite mixture model, earthquake data analysis, background noise, Dirichlet compound multinomial, Bayesian inference, expectation-maximization.

1 Introduction

We use a constrained Finite Mixture Model (FMM) to cluster spatial data points. These points may indicate the locations of some natural events occurred in a region and be concentrated around some centers. As an example, the earthquakes taken places around the main and the following strong shocks may exhibit such a spatial pattern. There are some approaches based on spatial point process [1] and nearest neighbor search [2] for spatial clustering. In these approaches, the data points are clustered regarding to the majority vote of their neighbors given a local region. A drawback of the majority voting is that it may lead the domination of the major cluster in number. In the other hand, model-based clustering methods assign a probability density function to each cluster and perform the clustering regarding to the weights of the points calculated using the cluster densities [3], [4]. FMMs are preferred for model-based clustering. Our clustering approach combines the model and the nearest neighbor based clustering approaches using a constraint FMM. We use a FMM to model the clusters and a latent variable model to introduce the local interactions. Our mixture model contains three key variables, namely location, cluster label and mixture proportion. For the locations, we assume that the points are distributed around the cluster centers as a Gaussian. Mixtures of Gaussians (MoG) based clustering methods are used for spatial point processes [5] [6], [7]. For each data point, we define a latent cluster label to be a categorical random variable which is a special version of the multinomial random variable where each data point belongs to only one cluster [8]. Extension to the classical MoG model, we take the local interaction of each point into consideration by defining a spatially

*Koray Kayabol carried out this work during the tenure of an ERCIM "Alain Bensoussan" Postdoctoral Fellowship Programme.

†K. Kayabol is with the PNA2 CWI, Science Park 123, 1098 XE, Amsterdam, Netherlands, (e-mail: koray.kayabol@cwi.nl).

varying latent model for mixture proportions based on Dirichlet density. There may be found some studies which include the spatial correlations into mixture models, i.e. [9]. Using the model, we aim to improve the clustering performance especially around the cluster borders. By defining appropriate likelihood and prior for the latent mixture proportion and integrating out it from the model, we obtain a Dirichlet Compound Multinomial (DCM) distribution which is also known as Polya distribution and is proposed to model the diffusion of a contagious disease over a population [10]. The DCM based mixture models find varying application areas, in document clustering [11], text retrieval [12] and image segmentation [13], [14].

Apart from the data points concentrated around the cluster centers, there might be some data points located far away from the clusters. We assume these data points to be the background noise or the outliers. The intuitive trend for background noise modelling in spatial clustering is to use a Poisson point process [5] [6], [7]. In this study, we use the noise model proposed in [7] and allocate one of the components in the mixture model for the background noise.

We formulate the problem in a Bayesian framework [15]. The Expectation-Maximization (EM) algorithm [16], [17] is the mostly used method for solving the FMM based clustering problems. We can interpret our Bayesian formulation as a constrained EM method [18]. Our aim is to maximize the FMM w.r.t its parameters subject to a spatially constrained latent variables. Furthermore, we use a computationally less expensive version of EM algorithm, namely Classification EM (CEM) [19], both for parameter estimation and for clustering, using the advantage of categorical random variables. In classification step, CEM uses the Winner-Take-All principle to allocate each data point to the related cluster according to the posterior probability of latent cluster label. After the classification step of CEM, we estimate the parameters of the cluster densities using only the members of the related clusters.

We test the algorithm both on the simulated and the real data. For real data, we use a seismic activity pattern observed in Kashmir in 2005. The seismic signals are analyzed for varying purposes such as surveillance, analysis and prediction of the geological hazards and disasters. These analyzes are generally based on the three variables namely, space, time and magnitude [20],[21]. In this study we focus on the clustering of the spatial earthquake pattern occurred after a main shock. There are some studies on clustering the spatial earthquake data, i.e. [22] uses the mixture of Poisson processes, [23] resorts to Fisher Discriminant Analysis (FDA) and [24] proposes a Dirichlet Process Mixture (DPM) model.

We organize the paper as follows. In Section 2 and 3, the DCM mixture model and CEM algorithm are given. The simulation results are shown in Section 4. Section 5 presents the conclusion and future work.

2 Dirichlet Compound Multinomial Mixture Model with Noise

We denote each data as a vector $\mathbf{x}_n = (x_1, x_2) \in \mathbb{R}^2$ where $n \in \mathcal{R} = \{1, 2, \ldots, N\}$. The components of $\mathbf{x}_n$ represent the longitude and the latitude respectively. If we denote $\boldsymbol{\mu}_k$ to be the cluster centers, we can define the cluster density $p(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)$ as a Gaussian as

$$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{2\pi|\Sigma_k|^{\frac{1}{2}}} \exp\left\{(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right\} \tag{1}$$

where $\Sigma_k \in \mathbb{R}^2 \times \mathbb{R}^2$ is the covariance matrix of the kth cluster. We denote $\theta_k = \{\boldsymbol{\mu}_k, \Sigma_k\}$ to be the parameters of the cluster densities for $k = 1, \ldots, K$.

Every data point has a latent cluster label. Denoting by K the number of clusters, we encode the cluster label as a K dimensional categorical random vector $\mathbf{z}_n$ whose elements $z_{n,k}$, $k \in \mathcal{C} = \{0, 1, \ldots, K\}$ have the following properties: 1) $z_{n,k} \in \{0, 1\}$ and 2) $\sum_{k=0}^{K} z_{n,k} = 1$. This binary random vector indicates the cluster label of the related pixel. We allocate the 0th cluster for background noise

with parameter α. We assume the elements of $\mathbf{z}_n$ to be distributed a priori as a multinomial density with parameters $\boldsymbol{\pi}_n = [\pi_{n,0}, \ldots, \pi_{n,K}]$. We denote the prior of $\mathbf{z}_n$ as $p(\mathbf{z}_n|\boldsymbol{\pi}_n) = \text{Mult}(\mathbf{z}_n|\boldsymbol{\pi}_n)$. The parameters $\pi_{n,k}$ represents the mixture proportions and ensure that $\sum_{k=0}^{K} \pi_{n,k} = 1$. We may write the probability of $\mathbf{x}_n$ to be the marginalization of the joint probability density $p(\mathbf{x}_n, \mathbf{z}_n|\Theta, \boldsymbol{\pi}_n) = p(\mathbf{x}_n|\mathbf{z}_n, \Theta)p(\mathbf{z}_n|\boldsymbol{\pi}_n)$, [8], over $\mathbf{z}_n$ as

$$\begin{aligned} p(\mathbf{x}_n|\Theta) &= \sum_{\mathbf{z}_n} p(\mathbf{x}_n|\mathbf{z}_n, \Theta)p(\mathbf{z}_n|\boldsymbol{\pi}_n) \\ &= \sum_{\mathbf{z}_n} \alpha^{z_{n,0}} \pi_{n,0}^{z_{n,0}} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\theta_k)^{z_{n,k}} \pi_{n,k}^{z_{n,k}} \\ &= \sum_{\mathbf{z}_n} \alpha^{z_{n,0}} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\theta_k)^{z_{n,k}} \prod_{k=0}^{K} \pi_{n,k}^{z_{n,k}} \end{aligned} \tag{2}$$

where $\Theta = \{\alpha, \theta_1, \ldots, \theta_K\}$ is the set of the parameters. We assume the parameter α to be fixed. By taking into consideration that $\mathbf{z}_n$ is a categorical random vector distributed as a multinomial, and assuming that $\boldsymbol{\pi}_n$ is spatially invariant, (2) is reduced to classical MoG model with a noise term [5], [6], [7] as follow

$$p(\mathbf{x}_n|\Theta) = \alpha\pi_0 + \sum_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\theta_k)\pi_k \tag{3}$$

In our study, we do not use the classical MoG model in (3) but use the spatially varying mixture model in (2) to include the spatial local statistics. We use a fully Bayesian approach to include the local statistics to the clustering process. In fully Bayesian approaches, all the parameters might be estimated from the same data. We give the details of the model and the algorithm in the following section.

3 Bayesian Estimation and Clustering with Classification EM Algorithm

In this section, we formulate the clustering problem under fully Bayesian framework. Since the spatial data is comprised of the spatial coordinates, any spatial interaction model has to depend on the data. We use the following setting to introduce the spatial interaction to the model. Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be the set of all data points. If we define a local region bounded with a circle located at a data point $\mathbf{x}_n$ with radius r as $\mathcal{D}_n = \{m \in \mathcal{R} : ||\mathbf{x}_m - \mathbf{x}_n)|| \leq r\}$, we can denote $\mathcal{X}_{\partial n} = \{\mathbf{x}_m : m \in \mathcal{D}_n\}$ to be the neighbor set of $\mathbf{x}_n$ and $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$ to be the set of all local interactions in the model. We do not include the parameter r into our Bayesian model and release its value to be determined by the user.

We may write the marginal likelihood of the parameter set Θ given the data $\mathcal{X}$ and spatial interaction model $\mathcal{D}$ as

$$p(\mathcal{X}, \mathcal{D}|\Theta) = p(\mathcal{D}) \sum_{\mathcal{Z}} p(\mathcal{X}|\Theta, \mathcal{Z})p(\mathcal{Z}|\mathcal{D}) \tag{4}$$

where $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ is the set of the cluster labels. The maximum likelihood estimation of Θ from (4) can be found iteratively using the EM algorithm. We can write the EM objective function to be maximized as

$$Q_{EM}(\Theta|\Theta^{t-1}) = C + \sum_{\mathcal{Z}} \log\{p(\mathcal{X}|\Theta, \mathcal{Z})\}p(\mathcal{Z}|\mathcal{X}, \mathcal{D}, \Theta^{t-1}) \tag{5}$$

We can formulate the constraint clustering problem by maximizing $Q_{EM}(\Theta|\Theta^{t-1})$ subject to

$$p(\mathcal{Z}|\mathcal{X},\mathcal{D},\Theta^{t-1}) = \int p(\Pi,\mathcal{Z}|\mathcal{X},\mathcal{D},\Theta^{t-1})d\Pi \tag{6}$$

where $\Pi = \{\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_N\}$ is the set of mixture proportions. Since our aim is not only to fit the parameters to the objective function but also to cluster the data, we resort to Classification EM algorithm [19] using the advantage of working with categorical random variables. The CEM algorithm incorporates a classification step between the E-step and the M-step which performs a Maximum-a-Posteriori (MAP) estimation. In the following three sections, we give the details of the CEM algorithm.

3.1 E-step

In order to calculate the posterior $p(\mathcal{Z}|\mathcal{X},\mathcal{D},\Theta^{t-1})$ introduced in (6), we factorize the integrand using the pseudo likelihood approximation [25]. This approximation leads to a kind of binary mean field approximation on $\mathcal{Z}$. We may write

$$\begin{aligned} p(\Pi,\mathcal{Z}|\mathcal{X},\mathcal{D},\Theta^{t-1}) &\approx \prod_{n=1}^{N} p(\boldsymbol{\pi}_n,\mathbf{z}_n|\mathcal{Z}_{-n},\mathbf{x}_n,\mathcal{D}_n,\Theta^{t-1}) \\ &= \prod_{n=1}^{N} p(\boldsymbol{\pi}_n,\mathbf{z}_n|\mathcal{Z}_{\partial n},\mathbf{x}_n,\Theta^{t-1}) \\ &= \prod_{n=1}^{N} \frac{p(\mathbf{x}_n|\mathbf{z}_n,\Theta^{t-1})p(\mathcal{Z}_{\partial n}|\boldsymbol{\pi}_n)p(\mathbf{z}_n|\boldsymbol{\pi}_n)p(\boldsymbol{\pi}_n)}{p(\mathbf{x}_n,\mathcal{Z}_{\partial n}|\Theta^{t-1})} \end{aligned} \tag{7}$$

where $\mathcal{Z}_{-n}$ represents the set from which the $\mathbf{z}_n$ is extracted and $\mathcal{Z}_{\partial n} = \{\mathbf{z}_m : m \in \mathcal{D}_n\}$ is the cluster labels around $\mathbf{z}_n$. To obtain the last expression, we also assume that the data points are i.i.d. distributed as

$$p(\mathcal{X}|\mathcal{Z},\Theta^{t-1}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n,\Theta^{t-1}) \tag{8}$$

We may integrate out $\boldsymbol{\pi}_n$ from (6) by considering (7) as

$$p(\mathbf{z}_n|\mathcal{Z}_{\partial n},\mathbf{x}_n,\Theta^{t-1}) \propto p(\mathbf{x}_n|\mathbf{z}_n,\Theta^{t-1}) \int p(\mathbf{z}_n|\boldsymbol{\pi}_n)p(\mathcal{Z}_{\partial n}|\boldsymbol{\pi}_n)p(\boldsymbol{\pi}_n)d\boldsymbol{\pi}_n \tag{9}$$

The first term in (9) is the exact likelihood of the cluster label $\mathbf{z}_n$ given the estimated parameters from the previous step as introduced in (2) implicitly as

$$p(\mathbf{x}_n|\mathbf{z}_n,\Theta^{t-1}) = \alpha^{z_{n,0}} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\theta_k)^{z_{n,k}}. \tag{10}$$

The first term inside the integral is a multinomial, $\text{Mult}(\mathbf{z}_n|\boldsymbol{\pi}_n)$, as introduced in Section 2. Assuming that $\boldsymbol{\pi}_n$ is distributed as i.i.d. multinomial inside the local region $\mathcal{D}_n$, we may write the second term inside the integral in (9) as

$$p(\mathcal{Z}_{\partial n}|\boldsymbol{\pi}_n) \propto \prod_{m\in\mathcal{D}_n} \prod_{k=0}^{K} \pi_{n,k}^{z_{m,k}} \tag{11}$$

If we assign a noninformative prior for $\boldsymbol{\pi}_n$ as $p(\boldsymbol{\pi}_n) = \prod_{k=1}^{K} \pi_{n,k}^{-1}$, we can reorganize $p(\mathcal{Z}_{\partial n}|\boldsymbol{\pi}_n)p(\boldsymbol{\pi}_n)$ as a Dirichlet density as

$$\mathrm{Dir}(\boldsymbol{\pi}_n|\mathbf{v}_n) = \frac{\Gamma(\sum_{k=1}^{K} v_{n,k})}{\prod_{k=1}^{K}\Gamma(v_{n,k})} \prod_{k=1}^{K} \pi_{n,k}^{v_{n,k}-1} \tag{12}$$

where

$$v_{n,k} = 1 + \sum_{m \in \mathcal{D}_n} z_{m,k}^{t-1}. \tag{13}$$

and $\Gamma(\cdot)$ represents the Gamma function. If we perform the integration in (9) by considering (12),

$$\int \mathrm{Mult}(\mathbf{z}_n|\boldsymbol{\pi}_n)\mathrm{Dir}(\boldsymbol{\pi}_n|\mathbf{v}_n)d\boldsymbol{\pi}_n = \frac{\Gamma\left(\sum_{k=1}^{K} v_{n,k}\right)}{\Gamma\left(\sum_{k=1}^{K} v_{n,k} + z_{n,k}\right)} \prod_{k=1}^{K} \frac{\Gamma(v_{n,k} + z_{n,k})}{\Gamma(v_{n,k})} \tag{14}$$

we obtain the DCM or Polya distribution. Furthermore, using the identity $\Gamma(x+1) = x\Gamma(x)$, we obtain a simpler non-parametric version of the DCM density as

$$p(\mathbf{z}_n|\mathbf{v}_n) = \prod_{k=1}^{K} \left(\frac{v_{n,k}}{\sum_{k=1}^{K} v_{n,k}} \right)^{z_{n,k}} \tag{15}$$

Now, we are able to write the posterior in (6) as follows

$$p(z_{n,0}|\mathbf{x}_n, \mathbf{v}_n, \Theta^{t-1}) \propto \left(\frac{\alpha v_{n,k}}{\sum_{k=1}^{K} v_{n,k}} \right)^{z_{n,0}}, \quad k = 0 \tag{16}$$

$$p(z_{n,k}|\mathbf{x}_n, \mathbf{v}_n, \Theta^{t-1}) \propto \left(\mathcal{N}(\mathbf{x}_n|\theta_k^{t-1}) \frac{v_{n,k}}{\sum_{k=1}^{K} v_{n,k}} \right)^{z_{n,k}}, \quad k \neq 0. \tag{17}$$

3.2 C-step

In the C-step, we perform the clustering by assigning a cluster label to each data point such as for all $n = 1, \ldots, N$, classify the nth pixel into class j as $z_{n,j} = 1$ by choosing j which maximizes the posterior $p(z_{n,k}|\mathbf{x}_n, \mathbf{v}_n, \Theta^{t-1})$ over $k = 0, 1, \ldots, K$ as

$$j = \arg\max_k p(z_{n,k}|\mathbf{x}_n, \mathbf{v}_n, \Theta^{t-1}) \tag{18}$$

3.3 M-step

After C-step, we can partition the data points domain $\mathcal{R}$ into K non-overlapping groups such that $\mathcal{R} = \bigcup_{k=1}^{K} \mathcal{R}_k$ and $\mathcal{R}_k \bigcap \mathcal{R}_l = 0$, $k \neq l$. We can write the classification log-likelihood function by modifying the EM objective function in (5) as

$$Q_{CEM}(\Theta|\Theta^{t-1}) = \sum_{k=1}^{K} \sum_{m(k) \in \mathcal{R}_k} \log \mathcal{N}(\mathbf{x}_{m(k)}|\theta_k) \tag{19}$$

To maximize this function, we alternate among the variables α, μ_k and Σ_k. The CEM functions of the parameters are written as follows

$$Q_{CEM}(\boldsymbol{\mu}_k|\Theta^{t-1}) = \sum_{n \in \mathcal{R}_k} -\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{20}$$

Table 1: RMSE values of the clustered simulated data with 50% of noise for different initial cluster centers obtained by perturbing the ground-truth centers by adding 5, 10, 15, 20 and 25, and related numbers of iterations.

	5	10	15	20	25
RMSE( )	0.8316	0.8964	0.8688	1.5202	3.1609
RMSE(Σ)	0.5140	0.6667	0.6342	1.2565	1.9516
# of iter.s	3	4	4	6	8

$$Q_{CEM}(\Sigma_k|\Theta^{t-1}) = \frac{N_k}{2}\log|\Sigma_k| + \sum_{n\in\mathcal{R}_k} -\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{21}$$

The solutions to (20) and (21) can be easily found as

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n\in\mathcal{R}_k}\mathbf{x}_n \tag{22}$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n\in\mathcal{R}_k}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{23}$$

4 Simulation Results

In this section, we present the clustering results of the proposed model on simulated and real data.

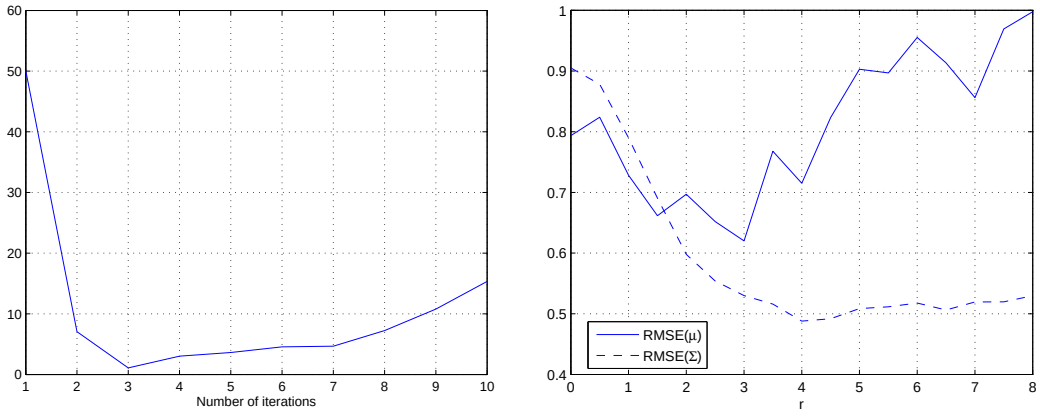
4.1 Simulated data

In order to test our algorithm, we simulate a spatial point pattern which has two Gaussian clusters and uniformly distributed background noise or clutter. The point pattern similar to one used in [7] with a difference that the cluster centers are much closer to each others. The Gaussian clusters are centered at $(70, 70)$ and $(120, 120)$ and have standard deviations in the horizontal and vertical directions of $(10, 20)$ and $(18, 10)$. We simulate 250 data points for each cluster and test the algorithms for different noise ratios changing from %10 to %70.

We first test the dependence of the DCM algorithm on initial values. Table 1 lists the different the Root Mean Squared Errors (RMSE) of the estimated mean and covariance parameters obtained with different initial cluster centers presence of 50% of noise. We estimate the mean error by averaging errors found by running the algorithms 200 times with different random noise realizations. The error in the estimation of $\boldsymbol{\mu}$ is found by taking the difference between the estimated $\hat{\boldsymbol{\mu}}$ and real $\boldsymbol{\mu}^*$ means. For the estimated covariance matrix, we use an error such as $||(\mathbf{I}_2 - \hat{\Sigma}^{-1}\Sigma^*)||_2$ where $\mathbf{I}_2$ is the 2×2 identity matrix and $||\mathbf{J}||_2$ is the square root of the maximum eigenvalue of $\mathbf{J}^T\mathbf{J}$. In this experiment, we initialize the ground-truth cluster centers by adding 5, 10, 15, 20 and 25 both in vertical and horizontal directions. The covariance matrices are estimated from the data using the initial means. The initial values up to 15 give quite good results. With the initial values after 20, the clustering performance is deteriorated. As seen from Table 1, the necessary number of iterations and the errors are increasing proportional to the distance between the initial and the ground-truth cluster centers.

To show the convergence of the algorithm, we may use the Total RMSE, TRMSE = RMSE($\boldsymbol{\mu}$) + RMSE($\boldsymbol{\mu}$). Fig. shows the three plots. Regarding to TRMSE, the algorithm converges its optimum value after 4 iterations.

We have performed some experiments by changing r from 0 to 8 to determine the radius r and understand the sensitivity of the clustering to r. Fig. 1(b) illustrates the plot of the RMSEs of the



(a) (b)

Figure 1: (a) TRMSE plot versus number of iterations. (b) RMSE plots of the clustered simulated 50% noisy data versus neighborhood radius r.

Table 2: RMSE values of the clustered simulated noisy data with MoG and DCM mixture model.

	MoG			DCM		
Noise ratio %	RMSE( )	RMSE(Σ)	$1/\alpha$	RMSE( )	RMSE(Σ)	$1/\alpha$
10	0.9852	0.3283	20.0	0.9699	0.3447	75.0
20	1.2636	0.3793	10.0	1.0645	0.3893	30.0
30	1.2694	0.4605	6.0	1.0005	0.4262	10.0
40	0.9011	0.5824	5.0	0.6961	0.4795	8.0
50	0.7070	0.6778	4.6	0.7331	0.4885	4.4
60	0.9654	0.9352	3.8	1.0063	0.5404	3.2
70	4.5738	1.1133	3.6	1.0320	0.6179	2.0

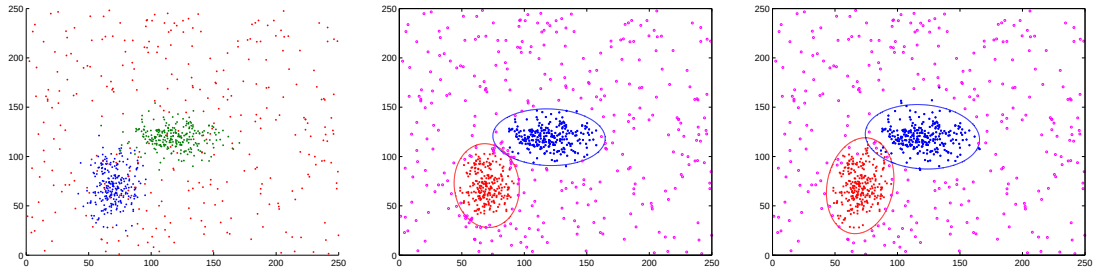
estimated mean and covariance parameters presence of 50% noise versus r. From the plot, we can see that the best value of r can be found between 2 and 4.

Table 2 lists the RMSE of the estimated mean and covariance parameters obtained by using MoG and DCM models. For initialization, we perturb the ground-truth mean values by adding 1. We fix the radius $r = 4$ and the number of iterations to 4 for all experiments regarding to Fig. 1(b) and Table 1, respectively.

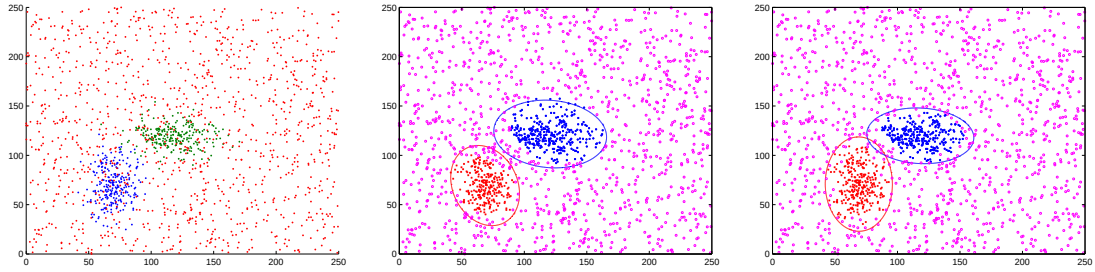
From Table 2, we can see that the performance of the mixture model with Bernoulli background noise model increases in case of the noise higher than %20. The DCM model is better in over all and especially in the estimation of the covariance matrix. Fig 2 shows the clustering results in case of two different noise realizations. Since the mean parameter is related with location and the covariance is with the shape and the orientation of the clusters, we can reach such an interpretation of the results that the DCM model enables to estimate the shape and the orientation of the cluster better than the MoG model in case of noisy background.

4.2 Earthquake data

We use a real earthquake data which consist of the locations of a sequence of earthquakes happened between October 8, 2005 and November 7, 2005 in the Kashmir area in Pakistan. The earthquakes in the data set have the magnitudes higher than 4.5 and are originated at a depth of less than 70 km.



(a) Noise ratio 40% (b) MoG (c) DCM, $r = 3.2$



(d) Noise ratio 70% (e) MoG (f) DCM $r = 2$

Figure 2: Clustering results with MoG and DCM mixtures in case of the presence of %40 and %70 of background noises. The ellipses represent the contours (or isolines) of the fitted 2D Gausians evaluated at probability 0.5.

Table 3: MNCL values of the clustered earthquake data with MoG and DCM mixture models and their noisy versions.

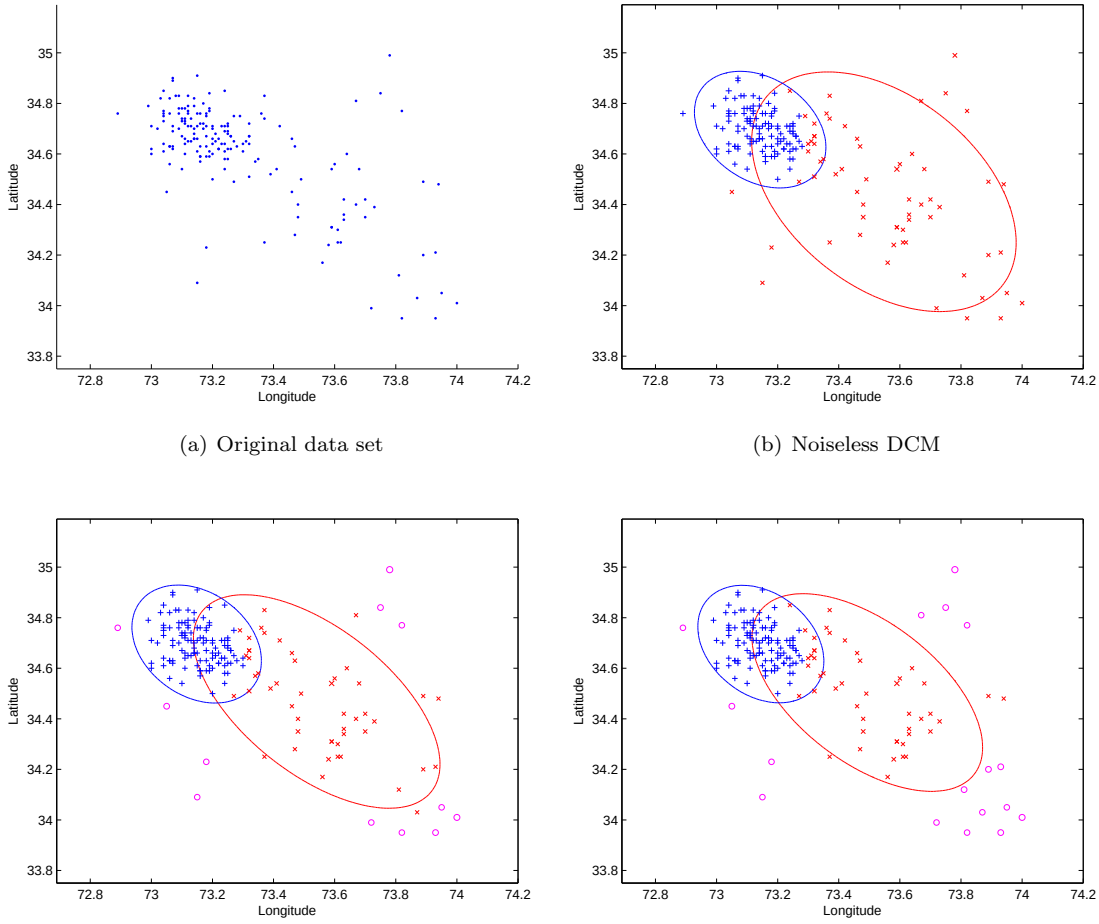
MoG	DCM	N-MoG	N-DCM
1.9038	1.9286	1.8441	1.7515

Fig. 3(a) shows the pattern of 176 spatial locations in corresponding longitudes and latitudes.

We test the noisy and noiseless mixture models on the entire data set. Since in real data case, we do not have any ground-truth to calculate RMSE, we calculate the Mean Negative Completed Likelihood (MNCL) excluding the estimated outliers as

$$MNCL = -2\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\log\{p(\mathbf{x}_n|\hat{\theta}_k)^{\hat{z}_{n,k}}p(\hat{z}_{n,k}|\hat{\mathbf{v}}_n)\} \tag{24}$$

Table 3 presents the average MNCL values calculated using clustering results of MoG and DCM mixture models. Fig. 3 shows the original data set and the clustering results of the noiseless DCM and the noisy MoG and DCM models. Noisy mixture models fit the data more tightly especially the noisy DCM model. The results depend on the parameter α. In this study, we initialize it by 0.1. We give the elements of the estimated covariance matrices by FDA [23], MoG and DCM models in Table 4. The FDA method provides some clusters with large variances because it use the entire data without considering background noise. The DCM with background noise model fits the data more tightly than the others.



(a) Original data set

(b) Noiseless DCM

(c) Noisy MoG

(d) Noisy DCM

Figure 3: (a) The original data set and (b) the clustering results of (b) the noiseless DCM, (c) the noisy MoG and (d) the noisy DCM models. The ellipses represent the contours (or isolines) of the fitted 2D Gausians evaluated at probability 0.5.

Table 4: Estimated cluster covariance matrices with MoG and DCM mixture models and FDA in [23].

		$\sigma_{1,1}$	$\sigma_{2,2}$	$\sigma_{1,2}$
clust 1	MoG	0.0370	0.0338	-0.0208
	DCM	0.0293	0.0273	-0.0158
	FDA [23]	0.0394	0.0813	-0.0188
clust 2	MoG	0.0068	0.0056	-0.0018
	DCM	0.0067	0.0053	-0.0018
	FDA [23]	0.0117	0.0105	-0.0039

5 Conclusion

The proposed spatial varying mixture model exhibits good performance in the estimation of the shape and the orientation of the clusters. Especially, in the presence of background noise the success of the proposed method is more significant compared to classical Gaussian mixture model with background noise component. The algorithm is sensitive to the parameter of the background noise. The user should define it regarding to an a prior knowledge or by observing the clustering results. For the earthquake data, the noise level should be determined by an expert in the area by considering the expected number of background events regarding to previous earthquake data analysis. The earthquake cluster model may be extended by including magnitude and time information. The proposed approach can be also used for other spatial clustering applications, i.e. minefield detection, astronomical point source clustering.

6 Acknowledgement

The author would like to thank Dr. Marie-Colette van Lieshout for her valuable comments and Prof. Alfred Stein for providing the Kashmir earthquake data set.

References

[1] J. Møller, "Markov chain Monte Carlo and spatial point processes" in *Stochastic Geometry: Likelihood and Computation*, editors O.E. Barndoff-Nielsen, W.S. Kendall and M.N.M. van Lieshout, Boca Raton, Florida (U.S.A.): Chapman & Hall/CRC, 1999.

[2] F.P. Preparata and M.I. Shamos, *Computational Geometry - An Introduction*, 1st ed. New York (U.S.A.): Springer, 1985.

[3] D.A. Binder, "Bayesian cluster analysis", *Biometrika*, vol.65, no.1, pp. 31–38, 1978.

[4] C. Fraley and A.E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis", *Computer J.*, vol.41, pp. 578–588, 1998.

[5] J.D. Banfield and A.E. Raftery, "Model-based Gaussian and non-Gaussian clustering", *J. Am. Statistical Assoc.*, vol.93, no.441, pp. 294–302, 1998.

[6] A. Dasgupta and A.E. Raftery, "Features in spatial point processes with clutter via model-based clustering", *Biometrics*, vol.49, no.3, pp. 803–821, 1998.

[7] C. Fraley and A.E. Raftery, "Model-based clustering, discriminant analysis, and density estimation", *J. Am. Statistical Assoc.*, vol.97, no.458, pp. 611–631, 2002.

[8] D. Titterington, A. Smith and A. Makov, *Statistical Analysis of Finite Mixture Disributions*, 3rd ed. Chichester (U.K.): John Wiley & Sons, 1992.

[9] C. Fernandez and P.J. Green, "Modelling spatially correlated data mixtures: a Bayesian approach", *J. R. Statist. Soc. B*, vol.64, no.4, pp. 805–826, 2002.

[10] F. Eggenberger and G. Polya, "Über die statistik verketter vorgänge", *Z. Angew. Math. Mech.*, vol.3, no.4, pp. 279–289, 1923.

[11] C. Elkan, "Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution", in *Int. Conf. Machine Learning, ICML'06*, pp. 289–296, 2006.

[12] Z. Xu and R. Akella, "New probabilistic retrieval model based on Dirichlet compound multinomial distribution", in *ACM Int. Conf. Research and Development in Information Retrieval, SIGIR'08*, pp. 427–434, 2008.

[13] A. Banerjee, P. Burlina and F. Alajaji, "Image segmentation and labeling using the Polya urn model", *IEEE Trans. Image Process.*, vol.8, no.9, pp. 1243–1253, 1999.

[14] C. Nikou, A.C. Likas and N.P. Galatsanos, "A Bayesian framework for image segmentation with spatially varying mixtures", *IEEE Trans. Image Process.*, vol.19, no.9, pp. 2278–2289, 2010.

[15] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, *Bayesian Data Analysis*, 2rd ed. , (U.S.A.): Chapman & Hall/CRC, 2004.

[16] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. R. Statist. Soc. B*, vol.39, pp. 1-22, 1977.

[17] R.A. Redner and H.F. Walker, "Mixture densities, maximum likelihood and the EM algorithm", *SIAM Review*, vol.26, no.2 pp. 195–239, 1984.

[18] R.J. Hathaway, "A constrained EM algorithm for univariate normal mixtures", *J. Statist. Comput. Simul.*, vol.23, no.3, pp. 211–230, 1986.

[19] G. Celeux and G. Govaert, "A classification EM algorithm for clustering and two stochastic versions", *Comput. Statist. Data Anal.*, vol.14, pp. 315–332, 1992.

[20] Y. Ogata, "Space-time point-process models for earthquake occurrences", *Ann. Inst. Statist. Math.*, vol.50, no.2, pp. 379–402, 1998.

[21] Y. Ogata, K. Katsura and M. Tanemura, "Heterogeneous space-time occurrences of earthquakes and its residual analysis", *J. R. Statist. Soc. C*, vol.52, no.4, pp. 499–509, 2003.

[22] T. Pei, C.H. Zhou, M. Yang, J.C. Luo and Q.L. Li, "The algorithm of decomposing superimposed 2-D Poisson processes and its application to the extracting earthquake clustering pattern", *ACTA Seismologica Sinica*, vol.17, no.1, pp. 54–63, 2004.

[23] M.N.M. van Lieshout and A. Stein, "Earthquake modelling at the country level using aggregated spatio-temporal point processes", Tech. Rep. PNA-1102, CWI, Netherlands, 2011.

[24] S. Hernandez and P. Sallis, "Modelling seismic activity using a Bayesian non-parametric method", *Int. J. Geology*, vol.5, no.4, pp. 126–130, 2011.

[25] J. Besag, "Statistical analysis of non-lattice data", *Statistician*, vol.24, no.3, pp. 179–195, 1975.