Unsupervised Event Detection with an Infinite Poisson Mixture Model

Vinod Hegde
Insight Centre for Data Analytics
NUI Galway, Ireland
vinod.hegde@insight-centre.org

Milovan Krnjajić
School of Mathematics and Statistics
NUI Galway, Ireland
milovan.krnjajic@nuigalway.ie

Alexei Pozdnoukhov
CEE Systems and Transportation
University of California Berkeley, USA
alexeip@berkeley.edu

***Abstract*—Large amount of time series data generated by sensors and Web users is great source of contextual information. Detecting outliers with unusually high values in time series data is crucial for inferring about any events in the real world. In this work, we describe an infinite Poisson mixture model to detect events by identifying outliers in time series of count data. This unsupervised technique estimates the probability densities of count data which have an unknown Poisson mixture while it simultaneously detects outliers in the data. The advantage of our model is that outliers are mapped to mixture components discovered by infinite mixture model and thus inference can be drawn on the different 'types' of outliers and their proportions in the data. This lets us identify and categorize events based on magnitude of outlier data. We have analysed the performance of our model against a well known event detection technique based on Markov modulated Poisson process (MMPP) using synthetic and real world data. Results show that our approach to detecting events is more appropriate in analysing periodic count data as compared to the MMPP baseline. The experiments demonstrate that the presented model provides robust, detailed, and interpretable results for the analysis of outliers to detect events.**

I. Introduction

In recent years, there has been a surge in the real world data generated by sensors on mobile phones as well as sensors embedded into various physical infrastructures. Data from GPS sensors, accelerometers on mobile phones capture fine-grained activities of individual users. On the other hand, sensors such as RFID tags, traffic polling systems, motion detectors for buildings etc. generate data, based on activities of population. These huge amounts of real world data have been extensively used by researchers to understand specific patterns of human activity. Particularly, analysing unusual trends in behaviour exhibited by users in the context of hourly, daily and weekly periodic variations is crucial to gain useful insights about real world situations. For example, a popular cultural event held in a city would be reflected in the amount of vehicles entering the toll gate of a city. Similarly, a minor increase in the number of vehicles entering a city can be due to a less famous event. These events in the real world can be mapped to unusually high count values in the corresponding time series data of user visits. Detecting such events in an unsupervised manner demands detecting outliers in time series of periodic count data. Also, it is crucial that outliers are classified into various categories based on their values. This lets us identify and categorize events based on the magnitude of increased counts due to them.

Modelling univariate data and detecting outliers has been extensively studied in the field of statistics and machine learning. In [1], the need for careful analysis of time series data when such data is generated by user activities is shown and a model based on non-homogeneous Markov Modulated Poisson Processes (MMPP) has been used for analysis. Similar models based on different versions of MMPP have been used in [2], [3], [4]. The motivation behind MMPP based outlier detection models is the fact that univariate time series data has no clear boundary to differentiate *abnormal* high count value outliers from the *normal* periodic counts. Any MMPP based outlier detection model has the assumption that most of the periodic count data can be modelled by a Poisson distribution with fixed rate parameter. This assumption about the nature of human generated periodic count data does not always hold good as they exhibit multimodality with unknown number of modes. We demonstrate this in Section IV with two datasets. Detecting outliers in multimodal data which is represented with a probabilistic mixture model has been widely studied in the field of statistics. [5] discusses the advantages of detecting outliers with finite mixture model. Specifically, the authors represent data with finite mixtures of Gaussian distributions. These finite parametric models require prior information about the data and hence are less flexible in accurately modelling the data.

On the other hand, nonparametric Bayesian models have high accuracy as they flexibly determine the *right* number of mixture components and mixture weights while estimating the probability density of data. So, they have been used effectively in various research work [6], [7], [8]. To the best of our knowledge, in spite of the flexibility of nonparametric Bayesian models, there has been no study on utilising them for detecting outliers in count data. Motivated by these observations, we have developed an infinite Poisson mixture model to capture the generative mechanism for count data and to identify outliers based on mixture components that are inferred. We have developed an appropriate version of Gibbs sampler, a Markov Chain Monte Carlo (MCMC) algorithm to perform the probabilistic inference. The research contribution made in this paper and its advantages can be summarized as follows.

- We develop an infinite Poisson mixture model for estimating outliers in a count dataset based on the underlying mixture components. We also develop a Gibbs sampler algorithm for parameter learning and inference.
- The proposed model has some major advantages over MMPP based baseline technique and other traditional outlier detection techniques. It is fully unsupervised

and performs accurately even in cases where data has varying degrees of multimodality. It identifies outliers in a dataset and also categorizes them into a previously unknown number of outlier categories. Since outlier and non-outlier data are represented by probability densities, sophisticated probabilistic queries can be made. Finally, any prior information about outliers in a dataset can be incorporated into the model by appropriately choosing prior distribution parameters.

- We apply our outlier detection technique on a real world dataset to detect events in an unsupervised manner. We show that our technique has better accuracy and provides robust analysis.

The rest of the paper is organized as follows. Related work is discussed in Section II. In Section III, we discuss our probabilistic model for detecting outliers in any count dataset. The model detects outliers of unusually high count in a dataset by discovering mixture components of an unknown Poisson mixture density. In Section IV, we analyse the performance of the proposed model against a state-of-the-art version of MMPP-based outlier detection technique using synthetic and real world data. In the last section, we state our conclusions and discuss the future work.

II. Related Work

Bayesian parametric framework for outlier detection has been studied extensively in statistics and machine learning [9], [10], [11], [12]. Particularly, a non-homogeneous Markov Modulated Poisson Process (MMPP) based outlier detection techniques have been developed for situations where events need to be detected based on data generated by periodic human activities. In this approach, periodic count data is modelled by Poisson process regulated by a Markov chain with a fixed number of states [1]. Effectiveness of the approach has been studied on vehicular traffic data and user presence data to detect events. In [2], number of tweets[1] generated by users have been modelled using MMPP to quantify abnormal volume of tweets. Modelling and predicting voice packet data with a version of MMPP has been studied in [3]. [4] studies identifying network intrusion using MMPP. These works have analysed how variations of MMPP model can be used to detect outliers and thus unusual bursts in human activity which correspond to events. Traditional models such as MMPP based techniques have assumed fixed number of Poisson mixture components in time series data for estimating outliers. This assumption can lead to either overfitting or underfitting of data depending on number of mixture components used in the model.

On the other hand, [5] discusses the advantages of detecting outliers with finite mixture model. Specifically, the authors represent data using a finite mixture of Gaussian distributions as follows and show that outliers can be detected.

$$f(x) = (1-\epsilon) f_n(x) + \epsilon f_0(x) \tag{1}$$

$$f_n(x) = \sum_{i=1}^{K-M} \pi_i \phi(x|\mu_i, \sigma_i^2) \tag{2}$$

$$f_0(x) = \sum_{j=K-M+1}^{K} \pi_j \phi(x|\mu_j, \sigma_j^2) \tag{3}$$

$$\tag{4}$$

Here, ϵ is a small positive fraction which represents the proportion of outliers. $f_n(x)$ is the probability density of data that are non-outliers and $f_0(x)$ is probability density of outlier data. However, the number of mixture components representing non-outlier and outlier data are fixed in this model using K. This is a limited assumption as number of mixture components in a dataset can not be known a priori. So, outlier detection solutions based on previously described finite mixture models face the problem of either underfitting, if few mixture components are assumed or overfitting otherwise. This potentially leads to less accuracy in detecting outliers. Interestingly, the work in [13] considers real world data and shows the need to use an infinite mixture model to fit count data more accurately compared to a parametric model. Similar results have been presented in [6], [7] and numerous other works. These observations motivate us to use an infinite mixture model for outlier detection in any univariate dataset.

III. Generative model for outlier detection in Poisson Mixture data

In this section, we discuss our generative model that can detect outliers by nonparametrically fitting a count dataset and deriving the modes that generate anomalies of atypically high values. The model is based on the standard results of Dirichlet Process mixture model [14], [15] for nonparametric density estimation of a given dataset. In Figure 1, we show the generative model in plate notation. Note that we use the term mixture component and cluster interchangeably in the following discussion. The random variables in this nonparametric Bayesian model are defined as follows.

$$\begin{array}{ll} p|a,b \sim Beta(a,b) & e_i|p \sim Bernoulli(p) \\ \pi|\alpha \sim GEM(1,\alpha) & z_i|\pi \sim Discrete(\pi) \\ \theta_k|\lambda \sim G_0(\lambda) & \theta_l^{'}|\lambda^{'} \sim G_0(\lambda^{'}) \end{array}$$

$$x_i|e_i, z_i, \{\theta_k\}_{k=1}^{\infty}, \{\theta_l^{'}\}_{l=1}^{\infty} \sim e_i F(\theta_{z_i}^{'}) + (1-e_i) F(\theta_{z_i}) \tag{5}$$

Here, we consider N items $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ which are discrete count data from an unknown Poisson mixture. p is a random variable which has a Beta prior distribution with a, b as hyperparameters. p represents the probability that a count value in the dataset is an outlier with unusually high value. e_i is a Bernoulli random variable which would be 1 when the data x_i is an outlier. π is a sample from a stick breaking process [16] denoted by GEM with the parameter α. It represents the

[1]https://twitter.com/

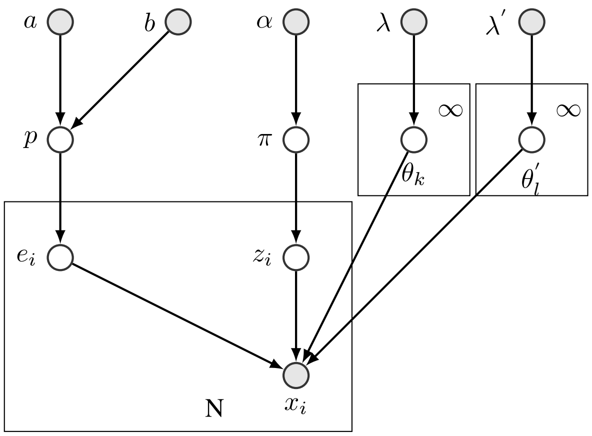


Fig. 1: Generative model for detecting the outliers in periodic count dataset

discrete distribution over the cluster indices of the parameter vectors that exist for the mixture component of the dataset. z_i represents the index of cluster chosen for the data x_i. $\theta_l^{'}$ represents a sample from the base distribution with parameters $\lambda^{'}$ and θ_k is a sample from the distribution parametrized by λ. $\theta_l^{'}$, θ_k represent the Poisson rate parameter for outlier and non-outlier clusters respectively. Since we need to model an unknown Poisson mixture data, we have used Gamma distribution as the base distribution G_0 which is the conjugate prior distribution for Poisson distribution. Thus, $\lambda^{'}$ and λ are hyper parameter vectors defined as $\lambda^{'} = (shape^{'}, rate^{'})$ and $\lambda = (shape, rate)$ for $\theta_l^{'}, \theta_k$ respectively. The data x_i have the Poisson likelihood given by the last statement in the Equation (3). x_i would be generated by a Poisson distribution with rate parameter $\theta_l^{'}$ if it is an outlier data and by a Poisson distribution with rate parameter θ_k otherwise. The data $\mathbf{x}$ is modelled with two infinite mixtures as we need to model an unknown number of non-outlier and outlier clusters present in the data. In Figure 1, we show the Bayesian network for the generative model described. Here, the hyper parameters whose values are known are marked dark and random variables with unknown distribution are marked white. The parameter learning and inference for this network involves obtaining the joint distribution of the random variables. We use one of the widely adopted Gibbs sampling, a Markov Chain Monte Carlo (MCMC) algorithm which requires the full conditional distributions for every random variable involved in order to obtain samples from the joint posterior distribution. We now show the conditional distributions derived for each of the random variables.

$$\begin{aligned} &Pr(e_i = 1|x_i, \mathbf{e}_{-i}, a, b, z_i, \{\theta_k\}_{k=1}^{\infty}, \{\theta_l^{'}\}_{l=1}^{\infty}) \\ &= Pr(e_1 = 1|\mathbf{e}_{-i}, z_i, \{\theta_k\}_{k=1}^{\infty}, \{\theta_l^{'}\}_{l=1}^{\infty}) \\ &\quad Pr(x_i|e_i = 1, z_i, \{\theta_k\}_{k=1}^{\infty}, \{\theta_l^{'}\}_{l=1}^{\infty}) \\ &= Pr(e_i = 1|\mathbf{e}_{-i}, a, b) F(x_i|\theta_{z_i}^{'}) \end{aligned} \tag{6}$$

Here $\mathbf{e}_{-i}$ denotes the set with all e_j, $j \neq i$. We use similar notation for $\mathbf{x}_{-i}$, $\mathbf{z}_{-i}$. $\mathbf{x}_k$ denotes the set of data items belonging to k^{th} cluster. $\mathbf{e}$ denotes the set with all e_j. We can derive the conditional distribution of $\mathbf{e}$ as follows.

$$\begin{aligned} &Pr(p|a, b) = p^{a-1}(1-p)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &Pr(e_1, \ldots, e_n|p) = p^{\sum e_i}(1-p)^{n-\sum e_i} \\ &Pr(e_1, \ldots, e_n|a, b) \\ &\quad = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int p^{(\sum e_i + a - 1)}(1-p)^{(n-\sum e_i + b - 1)} dp \\ &\quad = \frac{\Gamma(a+b)\Gamma(\sum e_i + a)\Gamma(n - \sum e_i + b)}{\Gamma(a)\Gamma(b)\Gamma(n+a+b)} \end{aligned} \tag{7}$$

$$\begin{aligned} Pr(e_i = 1|\mathbf{e}_{-i}, a, b) &= \frac{Pr(\mathbf{e}|a, b)}{Pr(\mathbf{e}_{-i}|a, b)} \\ &= \frac{\Gamma(a + \sum \mathbf{e}_i)\Gamma(n - 1 + a + b)}{\Gamma(a + \sum \mathbf{e}_{-i})\Gamma(n + a + b)} \\ &= \frac{a + \sum \mathbf{e}_{-i}}{a + b + n - 1} \end{aligned} \tag{8}$$

In a similar fashion to Equation (8) we can derive that

$$Pr(e_i = 0|\mathbf{e}_{-i}, a, b) = \frac{b + n - \sum \mathbf{e}_{-i}}{a + b + n - 1} \tag{9}$$

$$\begin{aligned} &Pr(e_i = 0|x_i, \mathbf{e}_{-i}, a, b, z_i, \{\theta_k\}_{k=1}^{\infty}, \{\theta_l^{'}\}_{l=1}^{\infty}) \\ &= Pr(e_i = 0|\mathbf{e}_{-i}, a, b) F(x_i|\theta_{z_i}) \end{aligned} \tag{10}$$

similar to Equation (6). We use Chinese Restaurant Process [17] as the basis for deriving the clustering property of the dataset. The probability that a data item falls into one of the active cluster j with parameter ϕ is derived as follows. Note that ϕ is equal to an existing value of θ_k if the cluster j is marked as non-outlier cluster and $\theta_k^{'}$ otherwise.

$$\begin{aligned} &Pr(z_i = j|\mathbf{x}, \alpha, e_i = 1, \mathbf{z}_{-i}, \{\theta_k\}_{k=1}^{\infty}, \{\theta_l^{'}\}_{l=1}^{\infty}) \\ &= Pr(z_i = j|x_i, \alpha, \mathbf{z}_{-i}, \phi) \\ &= Pr(z_i = j|\alpha, \mathbf{z}_{-i}, \phi) Pr(x_i|z_i = k, \alpha, \mathbf{z}_{-i}, \phi) \\ &= Pr(z_i = j|\alpha, \mathbf{z}_{-i}) Pr(x_i|\phi) \\ &= \frac{n_{j,-i}}{n + \alpha - 1} F(x_i|\phi) \end{aligned} \tag{11}$$

When the data belongs to an outlier cluster with $e_i = 1$, the probability that z_i gets a new value of cluster index $K+1$ where already K clusters exist is given by

$$\begin{aligned} &Pr(z_i = K+1|\mathbf{x}, \alpha, e_i = 1, \mathbf{z}_{-i}, \lambda, \lambda^{'}) \\ &= Pr(z_i = K+1|x_i, \alpha, \mathbf{z}_{-i}, \lambda^{'}) \\ &= Pr(z_i = K+1|\alpha, \mathbf{z}_{-i}, \lambda^{'}) Pr(x_i|z_i = K+1, \alpha, \mathbf{z}_{-i}, \lambda^{'}) \\ &= Pr(z_i = K+1|\alpha, \mathbf{z}_{-i}) Pr(x_i|\lambda^{'}) \\ &= \frac{\alpha}{n + \alpha - 1} \int F(x_i|\theta^{'}) G_0(\theta^{'}|\lambda^{'}) d\theta^{'} \end{aligned} \tag{12}$$

The conditional distributions for the parameter vectors of the data can be determined as follows.

$$\begin{aligned} &Pr(\theta_l^{'}|\mathbf{x}, \mathbf{e}, \mathbf{z}, \lambda^{'}) = Pr(\theta_l^{'}|\mathbf{x}_l, \lambda^{'}) \\ &\propto G_0(\theta_l^{'}|\lambda^{'}) \mathbf{L}(\mathbf{x}_l|\theta_l^{'}) \end{aligned} \tag{13}$$

If a new cluster is formed for generating outlier data x_i, the conditional distribution for the parameter of that cluster $\theta_l^{'}$ is derived as follows.

$$Pr(\theta_l^{'}|x_i) = \frac{G_0(\theta_l^{'})F(x_i|\theta_l^{'})}{\int G_0(\theta')F(x_i|\theta')} \quad (14)$$

We can obtain the conditional distributions for the cluster indices and parameters when e_i=0 in a similar fashion described in Equations (10) - (14).

The Gibbs sampler is used to derive the posterior samples of random variables in question using the full conditional distributions described above. Specifically, at each MCMC sample, we infer on the number of mixture components or clusters in the data and their weights and parameter values. Additionally, we label each of the inferred cluster as outlier or non-outlier cluster and the data belonging to those clusters as outlier data and non-outlier data respectively. We give the details of the direct Gibbs sampling in Algorithm 1 that we use for the inference. Here, we initialise two clusters where one represents the non-outlier data and the other represents outlier data. The Poisson rate parameters for these two clusters are initialised using the minimum and maximum values of the dataset. All data items are initially assigned to non-outlier cluster and are marked as non outlier data items. The hyper-parameters of Gamma distribution are initialised using the simple statistics of the data, namely the maximum and minimum. A Gamma distribution with shape parameter k and rate parameter 1 has a mean and variance value as k. The restriction on the Gamma prior distribution for drawing the Poisson rate parameters for outlier clusters thus results in the formation of outlier clusters with Poisson rate parameters biased towards the maximum value in the dataset. Similarly, Poisson rate parameters of non-outlier clusters are biased towards the minimum value in the dataset.

At each iteration of the algorithm, a data item is labelled as an outlier or non-outlier. The data item is then assigned to one of existing clusters or assigned to a new cluster with certain probabilities. If data item is assigned to a new cluster, the new cluster is marked as an outlier cluster if the assigned data item has been marked as outlier and as non-outlier cluster otherwise. The cluster parameters are updated for outlier clusters and non-outlier clusters using the samples from corresponding base distributions. After all the iterations, we use all samples to compute the number of clusters in the data using the posterior mode of number of active clusters in samples. The number of clusters is used to select matching samples which we use to compute outlier probability, cluster parameter values, cluster weights. We use the same samples to compute the posterior mode of the cluster outlier labels to identify clusters as outlier and non-outlier clusters. We then assign each data item to a cluster where it has the maximum probability density.

Algorithm 1 Gibbs Sampling for outlier detecting infinite mixture model

function INITIALISEPARAMETERS($\mathbf{x}$)
  $\mathbf{e} \leftarrow 0$
  $theta_1^{'} \leftarrow maximum(\mathbf{x})$
  $theta_1 \leftarrow minimum(\mathbf{x})$
  $(rate^{'}, rate) \leftarrow (1, 1)$
  $(shape^{'}, shape) \leftarrow (max(\mathbf{x}), min(\mathbf{x}) + 1)$
end function
function GIBBSSAMPLER($\mathbf{x}$,$\mathbf{e}$,$\mathbf{z}$,θ)
  IntialiseParameters($\mathbf{x}$)
  At any iteration t with $t > 1$ of the sampling
  For $i = 1, \ldots, n$
  if x_i is the single element in its cluster **then** remove the cluster, it's parameter and decrement C by 1
  end if
  Sample e_i with the following probabilities

  Choose a non-outlier cluster parameter$\theta_{i(t-1)}$ where x_i has the maximum density

$$Pr(e_{it} = 0) \propto \frac{a + \sum \mathbf{e}_{-i}}{a + b + n - 1} F(x_i|\theta_{i(t-1)})$$

  Choose an outlier cluster $\theta_{i(t-1)}^{'}$ where x_i has the maximum density

$$Pr(e_{it} = 1) \propto \frac{b + n - \sum \mathbf{e}_{-i}}{a + b + n - 1} F(x_i|\theta_{i(t-1)}^{'})$$

  Draw a sample for the cluster index with the probabilities as follows

$$Pr(z_{it} = k, k <= C) \propto \frac{n_{k,-i}}{n + \alpha - 1} F(x_i|\theta_{k(t-1)})$$

  if e_{it} is 1 **then**

$$Pr(z_{it} = k, k = C{+}1) \propto \frac{\alpha}{n + \alpha - 1} \int F(x_i|\theta) G_0(\theta|\lambda^{'}) d\theta$$

  else

$$Pr(z_{it} = k, k = C{+}1) \propto \frac{\alpha}{n + \alpha - 1} \int F(x_i|\theta) G_0(\theta|\lambda) d\theta$$

  end if
  if a new cluster is formed **then**

$$C = C + 1$$

  and mark the new cluster as an outlier cluster if e_{it} is 1 or 0 otherwise
  end if
  For all clusters marked as outlier clusters, sample the cluster parameters from the posterior distributions

$$Pr(\theta_{kt}) \propto G_0(\theta_{k(t-1)}|\lambda^{'}) F(\mathbf{x}_{k(t-1)}|\theta_{k(t-1)})$$

  For all clusters marked as non-outlier clusters, sample the cluster parameters from the posterior distributions

$$Pr(\theta_{kt}) \propto G_0(\theta_{k(t-1)}|\lambda) F(\mathbf{x}_{k(t-1)}|\theta_{k(t-1)})$$

end function

IV. EVALUATION

We have evaluated the performance of the presented outlier detection technique using synthetic and real world datasets.

We first compared the effectiveness of our outlier detection technique against an appropriate version of MMPP based technique in identifying outliers and outlier probabilities for synthetic datasets. Then, we evaluated the performance in identifying real world events using the *buildings* event dataset mentioned in [1]. Results show that the proposed technique is more accurate and does not require manual tuning of parameters as opposed to MMPP baseline.

A. Synthetic datasets

We have considered four types of Poisson mixture data varying in the number of mixture components or clusters and dispersion. These have the typical distribution that time series data generated by periodic human activities have, as described in [1], [2]. We have used variance to mean ratio or dispersion index $D = \frac{\sigma^2}{\mu}$ to measure the dispersion in data. Statistical description of the datasets is as follows:

1) dataset 1 - *Data with multiple clusters and large dispersion*: Data from a Poisson mixture of 5 components with rate parameters (10,30,130,320,520) and mixture weights of (0.2,0.2,0.25,0.15,0.2).
2) dataset 2 - *Data with multiple clusters and small dispersion*: Data from a Poisson mixture of 5 components with mixture weights (0.13,0.18,0.2,0.24,0.25) with rate parameters (10,22,35,58,80).
3) dataset 3 - *Data with two clusters and large dispersion*: dataset contains two distinct Poisson mixture components. The mixture weights are (0.75, 0.25) with the rate parameters (30,130).
4) dataset 4 - *Data with two clusters and small dispersion*: dataset has been generated from a mixture of two Poisson distributions with rate parameters(30,60) and mixture weights (0.75, 0.25).

We have run our algorithm and obtained outlier probability for the datasets and outlier labels for each datum. We have used the uninformative priors for Beta distribution as (a=1,b=1). These values ensure that the algorithm has weak and uniform prior information regarding outlier probability. In case there is strong prior information about outlier probability, that information can be incorporated by appropriately altering the values of a and b. Since the number of clusters formed is sensitive to α [18], we have used a weak informative prior value of α=0.01. We experimented with various types of data using the presented version of Gibbs sampler algorithm. The sampler converges within the first few tens of iterations for datasets with large dispersion. In contrast, the sampler converged within first few hundred iterations for datasets with small dispersion. So, we have run the algorithm to generate 2000 samples of random variables for each dataset. We have used a burn-in period of 400 iterations and discarded those samples and obtained i.i.d samples at a lag of 10 iterations to obtain posterior predictive density of data and posterior outlier probability. The posterior predictive density is defined as

$$Pr(\widetilde{x}|\mathbf{x},\alpha) = \sum_{\theta} Pr(\widetilde{x}|\theta)Pr(\theta|\mathbf{x},\alpha) \tag{15}$$

Here, $\widetilde{x}$ is the data element for which density is to be predicted using the observed data $\mathbf{x}$, hyperparameters α and parameters θ. We use the samples of the parameters $\theta_k^{'}$, θ_k and mixture weight samples to compute posterior predictive density of the data. The posterior outlier probability at any iteration t is calculated with

$$F(p) \sim Beta(a + n_t^{'}, b + n_t) \tag{16}$$

Here $n_t^{'}$ is the number of data items assigned to clusters marked as outlier clusters and n_t is the number of data items assigned to clusters marked as non-outlier clusters. At every iteration of the algorithm, we compute the posterior mean of the outlier probability which is the considered the outlier probability of the entire dataset for that iteration. The results of running Algorithm 1 are shown in Figures 2-4. Figure 2 shows the posterior predictive density that is derived for each of the datasets from the samples of Gibbs sampler. Figure 3 shows the number of clusters formed for each dataset over the samples. We show the posterior mean of outlier probability for each of the datasets in Figure 4. Here, the horizontal lines show the cumulative mixture weights of the mixture components sorted by their Poisson rate parameter. The algorithm identifies appropriate number of clusters in the dataset within first few iterations as seen in Figure 3.

In order to compare the performance of our algorithm, we have used MMPP model discussed in [1] which identifies real world events based on outliers. This model considers that any count value in a periodic count dataset can be described as

$$N(t) = N_0(t) + N_E(t) \tag{17}$$

where $N_0(t)$ represents counts due to periodic activity. $N_E(t)$ represents any additional counts that are due to an event and hence represent an outlier.

The counts observed due to periodic activities are assumed to follow a Poisson distribution with rate parameter $\lambda(t)$ dependent on time. In order to account for the effect of day and hour on the periodic counts, the time dependent rate parameter is calculated as $\lambda(t) = \lambda_0 \delta_{d(t)} \phi_{d(t),h(t)}$. Here $\delta_{d(t)}$ controls the day effect and $\phi_{d(t),h(t)}$ controls the hour effect for a day.

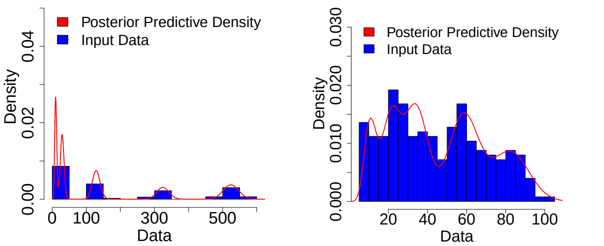
The presence of an event results in increased counts in addition to periodic counts which is captured in $z(t)$ and $N_E(t)$ defined as

$$N_E(t) = \begin{cases} 0, & \text{if } z(t) = 0 \\ P(N, \gamma(t)), & \text{if } z(t) = 1 \end{cases} \tag{18}$$

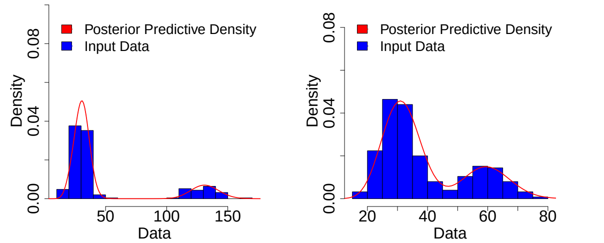
Here $z(t)$ is controlled by Markovian transition probability matrix defined as

$$\begin{pmatrix} 1 - z_0 & z_1 \\ z_0 & 1 - z_1 \end{pmatrix} \tag{19}$$

and z_0 and z_1 have prior distributions and their values determine the number of events detected by the system as outliers with increased count values. The model also contains other appropriate prior distributions and hyperparameters needed for the Bayesian inference. We have obtained MCMC samples

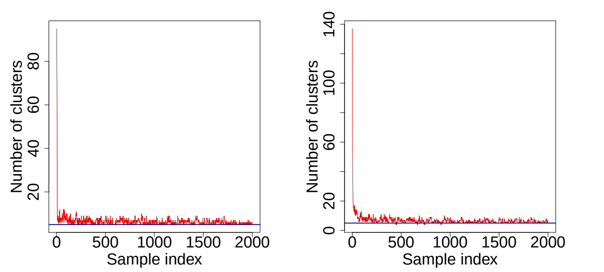


(a) Posterior predictive density for dataset 1 (b) Posterior predictive density for dataset 2

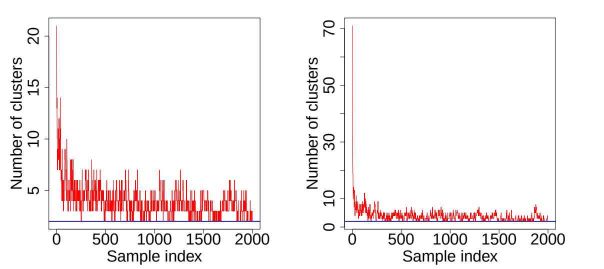


(c) Posterior predictive density for dataset 3 (d) Posterior predictive density for dataset 4

Fig. 2: Posterior predictive density against data



(a) Samples of number of clusters for dataset 1 (b) Samples of number of clusters for dataset 2

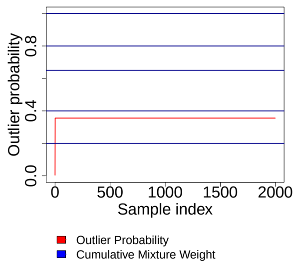


(c) Samples of number of clusters for dataset 3 (d) Samples of number of clusters for dataset 4

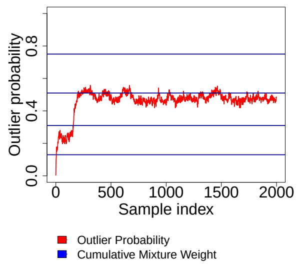
Fig. 3: Samples of number of clusters obtained during MCMC sampling

through Gibbs sampling of the random variables described above. The posterior mode of $z(t)$ in the samples for any count value determines whether that count value is an outlier and likely represents an event. The posterior mean computed with $z(t)$ for all the data can be used to detect outlier probability at each sample. We use this outlier probability to analyse the effectiveness of above described technique. In Table I, we summarize the cumulative mixture weights and outlier probabilities inferred by both algorithms to analyse their effectiveness. In order to analyse the performance of the algorithm in detecting outlier probabilities, consider dataset 3 which has two distinct clusters. Mixture weights for the clusters are (0.75,0.25) with Poisson rate parameters (30,130) respectively. Data items belonging to the second cluster represent outliers with high count values. In Figure 4c, it is shown that the algorithm appropriately identifies the mixture weight of the second cluster as proportion of outliers. In dataset 2, there is very low dispersion in spite of generating the data from 5 distinct clusters. We can see that the algorithm identifies the combined weights of last three clusters as outlier probability. In Figure 4, the four plots show that the algorithm identifies sum of the mixture weights of high value rate parameter clusters, as outlier probability. This approach to finding outliers is more informative as outliers are categorized with different rate parameters and the proportion of each category of outliers can be known. Since we assign each datum to a single cluster where it has the maximum density, all the data items in a given cluster are either marked as outliers or non-outliers. Also, any prior information regarding outlier probability for the dataset can be easily incorporated using suitable Beta prior distribution.

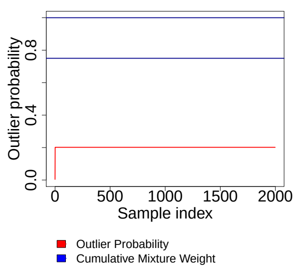
In Figure 5, we show the outlier probabilities derived for synthetic datasets. Here the horizontal lines show the cumulative mixture weights of mixture components sorted by their Poisson rate parameter. The third row of the table shows that both models identify mixture weight of the cluster with large Poisson parameter as outlier probability for dataset 3. We can see these inferences in Figures 4c and 5c. Note that dataset 3 has a huge dispersion with clear bimodality. Since MMPP based model specifically distinguishes between normal and outlier dataset with its parameters, it successfully infers about the proportion of outlier data. However, in Figures 5a, 5b and 5d, we can see that MMPP based algorithm fails to identify the outlier probabilities correctly for datasets 1, 2 and 4. For each of these datasets, it labels some of the data belonging to a cluster as outliers and the remaining in the same cluster as non-outliers. For example, consider the posterior mean of the outlier probabilities identified by the algorithm for dataset 2 is 0.62. This implies that the algorithm identifies some of the data points belonging to cluster with rate parameter 35 as outliers and rest of the data of same cluster as non-outliers. Similarly, the algorithm infers that the outlier probability for dataset 4 is 0.38 which implies that 13% of the data belonging to non-outlier cluster with rate parameter 30 are misclassified as outliers. These results clearly show that MMPP based outlier detection is not effective for datasets which have no strong



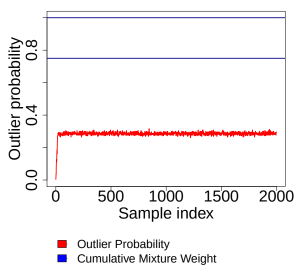
(a) Samples of outlier probability for dataset 1



(b) Samples of outlier probability for dataset 2

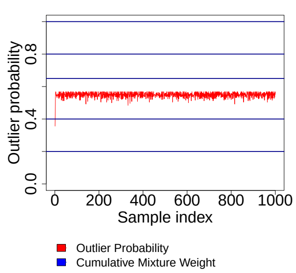


(c) Samples of outlier probability for dataset 3

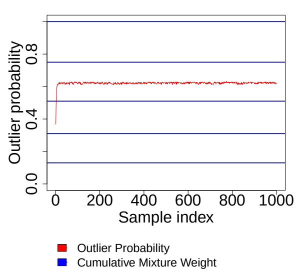


(d) Samples of outlier probability for dataset 4

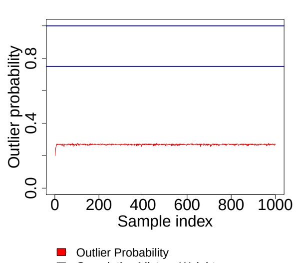
Fig. 4: Samples of outlier probability obtained from the algorithm



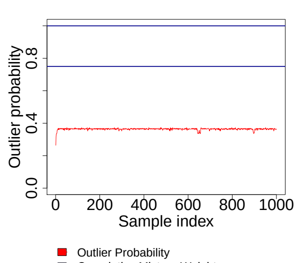
(a) Samples of outlier probability for dataset 1



(b) Samples of outlier probability for dataset 2



(c) Samples of outlier probability for dataset 3



(d) Samples of outlier probability for dataset 4

Fig. 5: Samples of outlier probability obtained using MMPP model

dataset	*Cumulative Mixture Weights*	*Presented Model*	*MMPP model*
1	(0.2,0.4,**0.65**,0.8,1)	**0.35**	0.57
2	(0.13,0.31,**0.51**,0.75,1)	**0.48**	0.62
3	(0.75,**1.0**)	**0.22**	**0.24**
4	(0.75,**1.0**)	**0.28**	0.38

TABLE I: Comparison of outlier probabilities inferred by the models

bimodality.

On the other hand, the proposed model correctly identifies outlier probabilities based on mixture weights of different clusters in datasets as seen in Table I. For example, in the first row, we can see that inferred outlier probability is 0.35 for the first dataset. This means that our model identifies the sum of mixture of weights of last two clusters as outlier probability. So, it infers that last two clusters are outlier clusters and data belonging to them are outliers. This also lets us make probabilistic queries about outliers belonging to outlier clusters. Similar observations can be made for the remaining datasets as shown in the table. We now explain the advantages of our technique in detecting real world events against MMPP based event detection technique.

B. *Buildings dataset*

Now, we analyse the performance of our algorithm in detecting events using human generated periodic count data. We have used *buildings* dataset mentioned in [1] which consists of count data of people's movements recorded every 30 minutes at the Calit2 institute building in the University of California, Irvine campus for a duration of 3 months. The data was recorded by optical detectors which count the number of people entering and exiting the building. Additionally, there are details of *events* that took place in the building. The time series data is effected by events which are aperiodic activities held in the building. There were 89 hours during which events have taken place in the building. Note that all the hours during which events have occurred might not have increased counts of people moving in the building. Few events in the building were unscheduled or unofficial and hence were not recorded. Some of these events can be directly seen as increased count values which are *outliers* in the time series data. We have used the time series data of hourly counts of people entering the building and real world events in the building to evaluate the two models. Detecting outliers in such data to infer about any real world events is challenging.

We have run our algorithm on count datasets corresponding to each hour of a day and detected outliers. We have used outliers to detect hours during which events were held in the building. We inferred Poisson rate parameters of mixture components for hourly dataset. So, the various *rates* at which people arrive at any given hour and unusual rates among them are known. For example, five different rates at which people arrive in the building at any hour shows the variability in the arrival rates. If two rate parameters represent outlier clusters, then count values belonging to these two clusters potentially

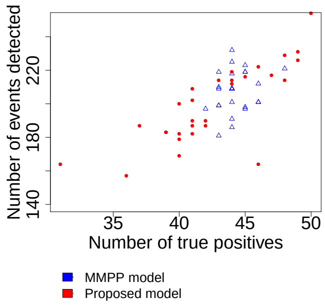


Fig. 6: Performance evaluation for event detection

represent events in the building. In order to analyse the performance of the MMPP based event detection technique, we set up the transition probability matrix over $z(t)$ with 30 sets of random weights to detect the same number of event hours as detected by our model. The performance of both the techniques are shown in Figure 6. Our technique had a true positive rate of 28% whereas the average true positive rate for the MMPP model was 21%. Another advantage of our technique is that the user need not have any prior knowledge about what constitutes an event. In contrast, such information is required in order to detect events using MMPP model as different weights on the transition probability matrix over $z(t)$ detect different number of events.

V. Conclusion

In this work, we have presented a nonparametric Bayesian model to detect outliers in count datasets. We have used this model to describe the generative process behind an unknown Poisson mixture of probability densities and developed an appropriate version of Gibbs sampling algorithm for probabilistic inference. This algorithm identifies outliers based on the mixture components revealed by probability density of the data. We have first analysed the performance of the algorithm on a synthetic dataset varying in the number of mixture components and data dispersion. Further, its performance has been compared against MMPP based outlier detection technique. We have shown that the algorithm effectively identifies outliers and outlier probabilities even in cases where MMPP based technique has limited performance. Then, we have used the people's movement and events data for a university building and shown that our technique is more effective than MMPP based technique in detecting events.

References

[1] A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying poisson processes," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 207–216.

[2] A. Pozdnoukhov and C. Kaiser, "Space-time dynamics of topics in streaming text," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. ACM, 2011, pp. 1–8.

[3] H. Heffes and D. Lucantoni, "A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *Selected Areas in Communications, IEEE Journal on*, vol. 4, no. 6, pp. 856–868, 1986.

[4] S. L. Scott, "A bayesian paradigm for designing intrusion detection systems," *Computational statistics & data analysis*, vol. 45, no. 1, pp. 69–83, 2004.

[5] K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," *Data Mining and Knowledge Discovery*, vol. 8, no. 3, pp. 275–300, 2004.

[6] A. Haghighi and D. Klein, "Unsupervised coreference resolution in a nonparametric bayesian model," in *Annual meeting-Association for Computational Linguistics*, vol. 45, no. 1, 2007, p. 848.

[7] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *Proceedings of the national conference on artificial intelligence*, vol. 21, no. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, p. 381.

[8] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky, "Describing visual scenes using transformed dirichlet processes," *Advances in neural information processing systems*, vol. 18, p. 1297, 2006.

[9] E. Marshall and D. Spiegelhalter, "Identifying outliers in bayesian hierarchical models: a simulation-based approach," *Bayesian Analysis*, vol. 2, no. 2, pp. 409–444, 2007.

[10] J. Wakefield, A. Smith, A. Racine-Poon, and A. Gelfand, "Bayesian analysis of linear and non-linear population models by using the gibbs sampler," *Applied Statistics*, pp. 201–221, 1994.

[11] F. A. Quintana and P. L. Iglesias, "Bayesian clustering and product partition models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 2, pp. 557–574, 2003.

[12] K. Chaloner and R. Brant, "A bayesian approach to outlier detection and residual analysis," *Biometrika*, vol. 75, no. 4, pp. 651–659, 1988.

[13] M. Krnjajić, A. Kottas, and D. Draper, "Parametric and nonparametric bayesian model specification: A case study involving models for count data," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2110–2128, 2008.

[14] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[15] R. M. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[16] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[17] D. Aldous, "Exchangeability and related topics," *École d'Été de Probabilités de Saint-Flour XIII1983*, pp. 1–198, 1985.

[18] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the american statistical association*, vol. 90, no. 430, pp. 577–588, 1995.