Bayesian analysis for mixtures of discrete distributions with a non-parametric component

Baba A Bukar, Hongsheng Dai*, Yoshiko Hayashi,
Veronica Vinciotti, Andrew Harrison, Berthold Lausen

Abstract

Bayesian finite mixture modelling is a flexible parametric modelling approach for classification and density fitting. Many application areas require distinguishing a *signal* from a *noise* component. In practice, it is often difficult to justify a specific distribution for the *signal* component, therefore the *signal* distribution is usually further modelled via a mixture of distributions. However, modelling the *signal* as a mixture of distributions is computationally challenging due to the difficulties in justifying the exact number of components to be used and due to the label switching problem. This paper proposes the use of a non-parametric distribution to model the *signal* component. We consider the case of discrete data and show how this new methodology leads to more accurate parameter estimation and smaller classification error. Moreover, it does not incur the label switching problem. We show an application of the method to data generated by ChIP-sequencing experiments.

Keywords: Bayesian; Gibbs sampler; Label switching; Mixture model

1 Introduction and motivation

Finite mixture modelling can be used to describe data obtained from different populations. The density of a typical mixture distribution can be written as

*Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK; hdaia@essex.ac.uk

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x; \boldsymbol{\theta}_k), \quad \sum_k \pi_k = 1 \tag{1}$$

where K is the number of components, π_k is the weight of component k and $f_k(x; \boldsymbol{\theta}_k)$ is the density of component k, with parameters $\boldsymbol{\theta}_k$. By relaxing distributional assumptions, a mixture model provides a convenient semi-parametric framework for modelling distributions of unknown shape. For example, it is used for model-based density estimation, since any distribution can be approximated by a mixture of elementary components.

In the last two decades, many new methodologies have been proposed for the Bayesian analysis of finite mixture models, such as Diebolt and Robert (1994), West (1997), Richardson and Green (1997), Stephens (2000a), McLachlan and Peel (2004) and Nobile et al. (2007). Although the existing literature has shown that finite mixture models can be inferred in a simple and effective way in a Bayesian estimation framework, persistent challenges still exist in the diagnostic of Markov Chain Monte Carlo (MCMC) convergence due to the following aspects.

The first aspect is the label switching problem, which is caused by the multimodality of the likelihood function. Many methods exist on how to tackle the label switching problem, for example, there are methods that impose identifiability constraints (Diebolt and Robert, 1994; Richardson and Green, 1997; McLachlan and Peel, 2004) and others that are based on relabelling algorithms (Celeux, 1998; Stephens, 2000b; Rodriguez and Walker, 2014; Celeux et al., 2000). For a review and comparison of these methods see, for example, Jasra et al. (2005) and Sperrin et al. (2010). One problem common to the existing methods for dealing with the label switching problem is that they usually require heavy computational costs, which make them unsuitable for large data sets and models with a large number of components. Another drawback of these methods is that they focus on mixture models where all components have the same type of distributions and focus on dealing with the invariance of the likelihood with respect to the permutation of the component labels. When the mixture components have different types of distributions, such as a mixture of Poisson and Negative

binomial distributions, label switching problems will still occur, since the likelihood function may still have multi-modes, but the existing methods for dealing with this problem may not be suitable anymore.

The second aspect is the identification of the number of components, K. Many authors have devised different methodologies for estimating the number of components in a Bayesian finite mixture models, for example reversible jump MCMC (Richardson and Green, 1997) and Birth and Death MCMC (Stephens, 2000a; Nobile et al., 2007). Another approach to deal with the unknown number of components is to use a mixture of Dirichlet processes (Antoniak, 1974; Escobar and West, 1995), which allows an infinite number of components.

The challenges mentioned above limit the applicability of mixture models in the areas involving large data sets and a large number of components. This motivates our study, as we discuss in detail in the following subsection.

1.1 Motivation of the study

In practice, we are often only interested in classifying the observations into two classes. For example, in the analysis of ChIP-Sequencing (ChIP-Seq) data, we are interested in whether a region of the genome is bound by the protein in question or not (Bao et al., 2014). For such ChIP-Seq (discrete) data, although there are only two possible classes, it is inappropriate to use a mixture of two known parametric distributions (e.g. Poisson or Negative Binomial distributions). This is because such data sets usually have long tails and the tails may show multi-modal patterns.

In this paper, we use for illustration the ChIP-Seq data generated by Ramos et al. (2010) for identifying the genomic regions bound by the histone acetyltransferases p300. For each region in the genome, the data report the number of bound fragments that align to that region. A higher value means that the corresponding region is most likely to be bound by the protein in question. Table 1 provides the summary statistics for the data set, where we consider only the data for 1000bp windows along chromosome 21 (Bao et al., 2013). Figure 1 shows a histogram of the count data. The left plot

Sample size	min	max	Mean	Variance
33916	0	282	2.24	18.70

Table 1: Summary statistics of the ChIP-seq data of Ramos et al. (2010) for one experiment on the protein p300 on chromosome21.

shows that the data set has a very long tail. If we zoom in the tail of the distribution (right plot), we see possible multi-modal patterns, suggesting that the distribution of the data is likely to consist of several component distributions. The interest however is that of classifying each region into two possibly states: bound or not bound by the protein in question.

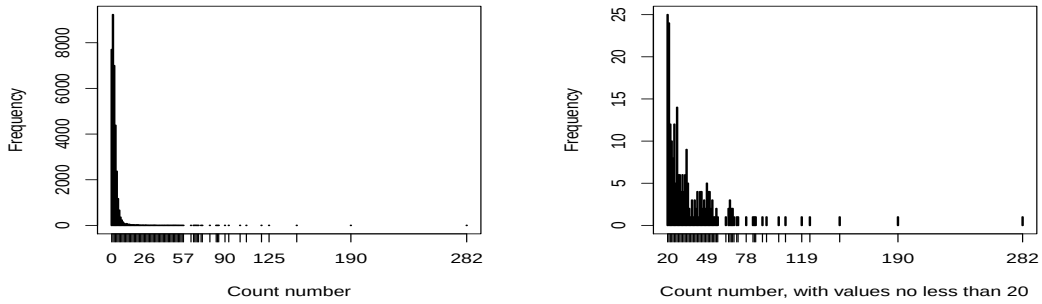

Figure 1: Distribution of ChIP-seq data for one experiment (left), with zoom on the tail (right).

The above situation has been observed also for other ChIP-seq experiments, where a two-component parametric mixture model appears to be too restrictive for the analysis of these data. An alternative approach is to use K components, with $K > 2$. In the context of ChIP-seq data analysis, this is considered by Kuan et al. (2011), who allow the signal distribution to be a mixture of two negative binomial distributions (i.e. $K = 3$). However, it is very challenging to justify what the true value of K is. Although the reversible jump Markov chain Monte Carlo method (Green, 1995) is readily available, the justification of reversible-jump MCMC convergence is non-trivial and it requires heavy computational costs. Another challenge of using K components is that it is non-trivial to determine what the component distributions are. For instance, all components may be chosen as Poisson distributions, or only some components are chosen as Poisson distributions and the others are chosen as Negative Binomial distributions. Finally,

since we are only interested in predicting two classes, using a mixture distribution with K components seems unnecessary. The above arguments and the motivating example have led us to consider a two-component mixture model for discrete observations, with one parametric distribution and one nonparametric distribution.

There are some existing non-Bayesian methods based on EM-type algorithms, which can deal with a two component mixture model with one parametric component and one non-parametric component. However those methods cannot be applied to our study. For example, Song et al. (2010) proposed a mixture model for sequential clustering of observations. It requires that the component with known location parameter. The classification algorithm relies on this center parameter. However, in our study the location parameter for the noise component is unknown. Xiang et al. (2014) extended the method and does not require the location parameter to be known, but the asymptotic results were not available for the full model as the identifiability problem was not justified there. We therefore focus on the Bayesian approach in this paper, where large sample properties for the estimates are not our concern since simulation from the posterior distribution is generally the main task in Bayesian analysis. The challenge of Bayesian analysis for mixture models is the label switching problem and the determination of the number of components K.

1.2 The contribution and structure of the paper

The advantage of this new method is that it bypasses the challenges involved in the K-component mixture models, such as the label switching problem and the determination of the unknown parameter K. The new method can still distinguish whether an observation is signal or noise, which is the main research interest in the studies that we consider, and it can do so with higher accuracy than a mixture of two parametric distributions, since it is expected to fit the data better.

The rest of this paper is organised as follows. Section 2 is devoted to developing the mixture models and estimation methodologies. Section 3 gives detailed simulation studies, which show that our method is more reliable than existing methods in terms

of better parameter estimation and smaller classification error. The data analysis is provided in Section 4 and a discussion is given in Section 5.

2 The new methodology

Suppose that discrete observations $x_1, \cdots, x_n$ are sampled from a mixture of distributions with two components, where one component is the noise distribution and the other component is a signal distribution. We simply use the following density to model the data,

$$f(x) = \pi_1 f_1(x; \boldsymbol{\theta}_1) + \pi_2 f_2(x; \boldsymbol{\theta}_2) \tag{2}$$

where f_1 is the parametric distribution for the noise, f_2 is the signal distribution and π_1 and π_2 are the corresponding mixture proportions, respectively.

Let z_i, $(i = 1, .., n)$ be an indicator or latent variable associated with each observation x_i, i.e. $z_i = k$ $(k = 1, 2)$ means that the observation x_i is from component k. The complete likelihood function for $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ given the full data is

$$l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{x}, \boldsymbol{z}) \propto \prod_{i=1}^{n} \left\{ \left[\pi_1 f_1(x_i; \boldsymbol{\theta}_1)\right]^{I[z_i=1]} \left[\pi_2 f_2(x_i; \boldsymbol{\theta}_2)\right]^{I[z_i=2]} \right\}. \tag{3}$$

The noise distribution f_1 is usually simpler to determine. For example in ChIP-Seq studies, a Poisson distribution is a natural choice for the noise since a genomic region not bound by the protein in question but tagged is a rare event. In cases where small window sizes are considered for the regions, zero-inflated Poisson distributions have been found to fit the noise distribution very well as they account for large number of zeros (Bao et al., 2014). In contrast to this, the signal distribution can present complicated patterns. As explained in Section 1, it may be difficult to find a suitable parametric distribution model for f_2. On the other hand, if f_2 is further modelled by a mixture distribution, it may not be easy to deal with the label switching problem, to determine the number of mixture components and to determine the component distri-

butions. Since we are only interested in distinguishing the signal and the noise, it is not necessary to identify how many components the signal distribution is formed of and what these component distributions are. We therefore consider to use a nonparametric model for the second component.

As the data are discrete, we can denote with $x_{(1)}, \cdots, x_{(L)}$ the L distinct values of the observations $x_1, \cdots, x_n$. Define

$$f_2^*(x_{(j)}) = p_j, \ \sum_{j=1}^{L} p_j = 1 \tag{4}$$

where p_js $(j = 1, \cdots, L)$ are the unknown parameters. p_j can be interpreted as the probability of $x = x_{(j)}$ given that x is drawn from the signal component. This can be viewed as a nonparametric distribution. Under this model, the distribution of x is given by

$$f(x) = \pi_1 f_1(x; \boldsymbol{\theta}_1) + \pi_2 \sum_{j=1}^{L} f_2^*(x) I[x = x_{(j)}]. \tag{5}$$

Based on the distribution (4), we have the following likelihood function given (x_i, z_i) $(i = 1, \cdots, n)$,

$$\begin{aligned} l(\boldsymbol{\theta}_1, \boldsymbol{p}, \boldsymbol{\pi} | \boldsymbol{x}, \boldsymbol{z}) &\propto \prod_{i=1}^{n} \left\{ [\pi_1 f_1(x_i; \boldsymbol{\theta}_1)]^{I[z_i=1]} \left[\pi_2 \sum_{j=1}^{L} p_j I[x_i = x_{(j)}] \right]^{I[z_i=2]} \right\} \\ &= \pi_1^{n_1} \pi_2^{n_2} \prod_{i=1}^{n} [f_1(x_i; \boldsymbol{\theta}_1)]^{I[z_i=1]} \cdot \prod_{j=1}^{L} p_j^{\sum_{i=1}^{n} I[z_i=2, x_i=x_{(j)}]} \end{aligned}$$

where $n_k = \sum_i I[z_i = k]$, $k = 1, 2$.

If we choose uniform priors for $\boldsymbol{\pi}$ and $\boldsymbol{p}$ and denote the prior for $\boldsymbol{\theta}_1$ as $g_0(\boldsymbol{\theta}_1)$, we have that $\boldsymbol{\pi}$, $\boldsymbol{p}$ and $\boldsymbol{\theta}_1$ are independent under the posterior distributions. In particular, the posterior distribution of $\boldsymbol{\pi}$ is given by the Beta distribution

$$g(\boldsymbol{\pi} | \boldsymbol{x}, \boldsymbol{z}) \ \propto \ \pi_1^{n_1} \pi_2^{n_2} := \ \text{Beta}\ (\boldsymbol{\pi}; n_1 + 1, n_2 + 1), \tag{6}$$

the posterior of $\boldsymbol{p}$ by the Dirichlet distribution

$$\begin{aligned} g(\boldsymbol{p}|\boldsymbol{x},\boldsymbol{z}) &\propto \prod_{j=1}^{L} p_j^{\sum_{i=1}^{n} I[z_i=2, x_i=x_{(j)}]} \\ &:= \text{Dirichlet } (\boldsymbol{p}; 1+\sum_{i=1}^{n} I[z_i=2, x_i=x_{(j)}]), \end{aligned} \tag{7}$$

and the posterior for $\boldsymbol{\theta}_1$ by

$$g(\boldsymbol{\theta}_1|\boldsymbol{x},\boldsymbol{z}) \propto \prod_{i=1}^{n} [f_1(x_i;\boldsymbol{\theta}_1)]^{I[z_i=1]} g_0(\boldsymbol{\theta}_1). \tag{8}$$

We also have that the posterior probability of z_i given $\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{p}$ and $\boldsymbol{\theta}_1$ is

$$\begin{aligned} P(z_i=1|\boldsymbol{x},\boldsymbol{\pi},\boldsymbol{p},\boldsymbol{\theta}_1) &\propto \pi_1 f_1(x_i;\boldsymbol{\theta}_1) \\ P(z_i=2|\boldsymbol{x},\boldsymbol{\pi},\boldsymbol{p},\boldsymbol{\theta}_1) &\propto \pi_2 \sum_{j=1}^{L} p_j I[x_i=x_{(j)}]. \end{aligned} \tag{9}$$

Based on all the above posterior distributions, we can use the Gibbs sampler to draw realisations from the posterior distribution and carry out a Bayesian Monte Carlo analysis. To implement the Gibbs sampler, we need to update the unknown parameters and the latent variable $\boldsymbol{z}$ by sampling from the conditional posterior distributions in (6), (7), (8) and (9). This leads to the algorithm:

Initialization, select, $\boldsymbol{z}^{(0)}, \boldsymbol{\pi}^{(0)}$, $\boldsymbol{p}^{(0)}$ and $\boldsymbol{\theta}_1^{(0)}$;
Set $m = 1$;
repeat
  for $i = 1$ to n **do**
    Update z_i with probability (9)
  end
  Update $\boldsymbol{\theta}_1$ from the posterior in (8) ;
  Update $\boldsymbol{\pi}$ from the posterior in (6);
  Update $\boldsymbol{p}$ from the posterior in (7);
  $m = m + 1$
until enough MCMC steps have been simulated;

Algorithm 1: The Gibbs sampler.

2.1 The interpretation of the model

The second component in (5) can be viewed as a nonparametric component, since we allocate a probability to each $x_{(j)}$. The probabilities p_j can be viewed as the empirical probabilities estimated via a sampling approach. It is easy to interpret the idea of this nonparametric component in the following way. If the Poisson component has $\lambda = 5$, say, then the probability that an observation with value 30 comes from the Poisson (noise) component will be very small (about e^{-13}). If the empirical distribution (the signal distribution) tells that $P(X = 30) \approx 0.00001$, then we should indeed classify the observation 30 into the signal component, provided that the component proportion values (π_1 and π_2) are in a similar scale. That means regardless of what the signal distribution is, we can always classify observations into either signal or noise.

The posterior predictive distribution of the new model also has a reasonable interpretation, which is actually linked with the Dirichlet process distribution. If we assume that the latent variable $\boldsymbol{z}$ is known, then the posterior predictive distribution is given by

$$\begin{aligned} & f_{pre}(y|\boldsymbol{x}, \boldsymbol{z}) \\ = & \int_{\boldsymbol{\theta}_1, \boldsymbol{p}, \boldsymbol{\pi}} \left(\pi_1 f_1(y; \boldsymbol{\theta}_1) + \pi_2 \sum_{j=1}^{L} p_j I[y = x_{(j)}] \right) \frac{l(\boldsymbol{\theta}_1, \boldsymbol{p}, \boldsymbol{\pi}|\boldsymbol{x}, \boldsymbol{z}) g_0(\boldsymbol{\theta}_1)}{c} d\boldsymbol{\theta}_1 d\boldsymbol{p} d\boldsymbol{\pi} \end{aligned}$$

where c, depending on $\boldsymbol{x}, \boldsymbol{z}$, is the normalising constant for the full posterior distribution. We can further write the posterior predictive distribution as

$$\begin{aligned} & f_{pre}(y|\boldsymbol{x}, \boldsymbol{z}) = \\ \propto & \mathbf{E}(\pi_1) \int_{\boldsymbol{\theta}_1} f_1(y; \boldsymbol{\theta}_1) \left(\prod_{i=1}^{n} [f(x_i; \boldsymbol{\theta}_1)]^{I[z_i=1]} \right) g_0(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1 \\ & + \mathbf{E}(\pi_2) \int_{\boldsymbol{p}} \left(\sum_{j=1}^{L} p_j I[y = x_{(j)}] \right) \prod_{j=1}^{L} p_j^{\sum_{i=1}^{n} I[z_i=2, x_i=x_{(j)}]} d\boldsymbol{p} \\ := & \mathbf{E}(\pi_1) \cdot \mathbf{E}_1(f_1(y; \boldsymbol{\theta}_1)) + \mathbf{E}(\pi_2) \cdot \sum_{j=1}^{L} I[y = x_{(j)}] \mathbf{E}_2(p_j) \end{aligned}$$

where $\mathbf{E}(\pi_1)$ and $\mathbf{E}(\pi_2)$ are the posterior expectation of $\boldsymbol{\pi}$, $\mathbf{E}_1(f_1(y;\boldsymbol{\theta}_1))$ is a posterior expectation conditional on all observations allocated to the first component ($z_i = 1$) and $\mathbf{E}_2(p_j)$ is the posterior expectation for $\boldsymbol{p}$ conditional on all observations allocated to the second component ($z_i = 2$).

Based on (7) we know that

$$\mathbf{E}_2(p_j) = \frac{1 + \sum_i I[z_i = 2, x_i = x_{(j)}]}{L + \sum_i I[z_i = 2]}$$

and based on (6) we know that

$$\mathbf{E}(\pi_k) = \frac{n_k + 1}{n + 2}, \quad k = 1, 2.$$

Then with simple calculations we further have

$$f_{pre}(y|\boldsymbol{x}, \boldsymbol{z}) \propto \frac{n_1 + 1}{n + 2} \cdot \mathbf{E}_1(f_1(y;\boldsymbol{\theta}_1)) + \frac{n_2 + 1}{n + 2} \cdot \frac{1 + \sum_i I[z_i = 2, x_i = y]}{L + n_2} \quad (10)$$

which has a very close connection with the posterior predictive distribution for Dirichlet process distributions in Ferguson (1973).

Suppose that a random sample $X_1, \cdots, X_n$ is from a probability space $(\mathbf{R}, \mathcal{B})$ with a random probability measure $\mathbf{P}$, which is a Dirichlet process with a base measure parameter α. Then Ferguson (1973) showed that the conditional distribution of $\mathbf{P}$ given $X_1, \cdots, X_n$ is still a Dirichlet process with parameter $\alpha + \sum_i \delta_{X_i}$, where δ_u denotes the measure giving mass one to the point u. Based on this result, Ferguson (1973) derived the posterior predictive distribution for a new variable Y from $\mathbf{P}$, as

$$\mathbf{P}_{pre}(\cdot|X_1, \cdots, X_n) = \frac{\alpha(\mathbf{R})}{\alpha(\mathbf{R}) + n} \cdot \frac{\alpha(\cdot)}{\alpha(\mathbf{R})} + \left(1 - \frac{\alpha(\mathbf{R})}{\alpha(\mathbf{R}) + n}\right) \frac{\alpha(\cdot) + \sum_i \delta_{X_i}(\cdot)}{\alpha(\mathbf{R}) + n} \quad (11)$$

which is a mixture of the prior belief α and the empirical distribution. Comparing (10) and (11) we can see that the posterior predictive distribution of our model is a mixture of the parametric predictive distribution $\mathbf{E}_1(f_1(y;\boldsymbol{\theta}_1))$ conditional on all

observations allocated in the first component, and the empirical distribution conditional on all observations allocated in the second component.

Therefore, under our modelling framework and given all observations x_i with its classification indicator z_i, a new observation can be viewed as from a random probability measure $\mathbf{P}$, which is a Dirichlet process with a base measure parameter proportional to $\mathbf{E}_1(f_1(y;\boldsymbol{\theta}_1))$. Ferguson (1973) uses the base measure parameter α as the prior information and the predictive distribution converges to the empirical distribution as the sample size $n \to \infty$. In our study, the base measure parameter can be viewed as the information for the first component. The latent variable z_i determines the proportion of samples in each component and it can be sampled via the proposed Gibbs sampler algorithm. If we fix all $z_i = 2$, our model degenerates to Ferguson's Bayesian nonparametric analysis, which cannot deal with classification since the target distribution is estimated via a nonparametric distribution.

3 Simulation studies

3.1 Scenario 1

To verify the validity of our methodology, we simulate a data set of $n = 500$ observations from a mixture of a Poisson distribution and a Negative Binomial distribution. The true model is

$$f(x) = \pi_1 \text{Poi}(x;\lambda) + \pi_2 \text{NB}(x;r,v), \tag{12}$$

where λ is the mean of the Poisson distribution, r is the nonnegative dispersion parameter and v is the probability parameter for the Negative Binomial distribution. We choose different values of the true parameters in order to study the performance of our proposed method under different situations. We consider three cases, (a) the means of the two components are far apart, (b) the means of the two components are very close and (c) the means of the two components are neither too close nor too far apart. We choose $\pi_1 = 0.8$, i.e. having a larger proportion for the noise component, to reflect our real ChIP-seq data. We also consider the case where the signal and noise have the

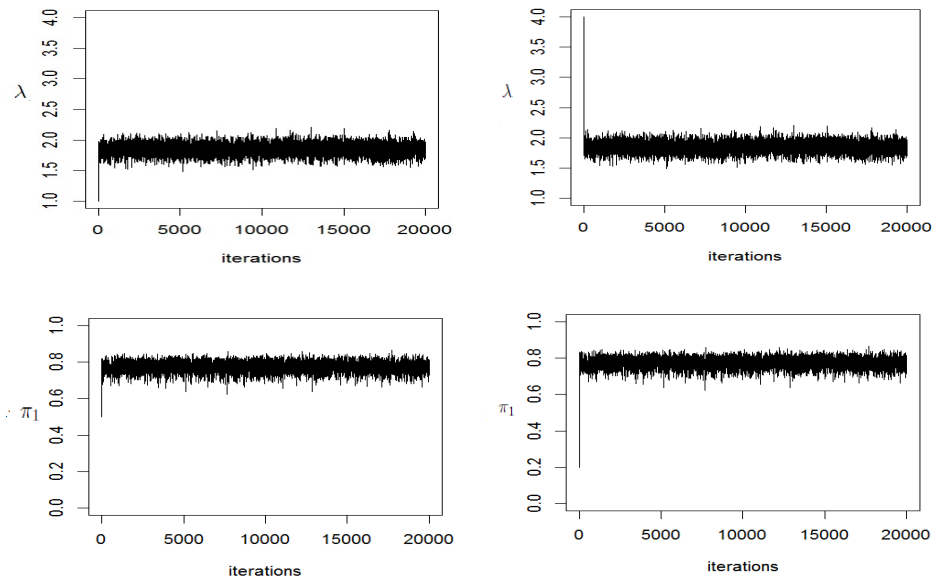

Figure 2: Trace plots for π_1 and for λ, with different starting values. The true parameter values are $\pi_1 = 0.8$, $\lambda = 2$, $r = 15$ and $\nu = 0.4$.

same component weights, $\pi_1 = \pi_2 = 0.5$.

The simulation studies are based on 20,000 iterations with 10,000 burn-in iterations, repeated for 100 times. Different starting values for the Gibbs sampler are chosen to justify the convergence of the Markov chains. From Figure 2 we can see that 20,000 steps are enough to guarantee the convergence for the Markov chains. We also choose different prior distributions to study the sensitivity of our model to the prior used. The results provided in the supplementary material demonstrate that the method is robust to different priors.

Table 2 shows the posterior means of the parameters of the proposed model under a number of different cases. We can see that the estimates are very good when the two components are clearly separated (Set 1 case). But for Set 2 and Set 3, there is some bias in the estimate of π_1. This is because for Set 2 and Set 3, the two component means are very close and many observations from the signal (having larger mean values than the noise) are treated as a sample from the noise component, leading to an inflated estimate of π_1. This kind of bias occurs in all analyses based on mixture models when the component densities are very close, i.e. the two components are not easily identifiable. For the purpose of comparison, we calculate the misclassification

	True value			True $\pi_1 = 0.8$			True $\pi_1 = 0.5$		
	λ	r	v	$E(\lambda)$	$E(\pi_1)$	Error	$E(\lambda)$	$E(\pi_1)$	Error
Set 1[a]	6	10	0.3	5.9271	0.7547	0.08	6.3299	0.4207	0.13
				(5.6054,6.2460)	(0.6895,0.8092)		(5.7362,6.9172)	(0.3362,0.4922)	
	10	20	0.2	9.6081	0.7489	0.05	9.8108	0.4061	0.08
				(8.9084,10.1492)	(0.6723,0.8114)		(9.0310,10.3848)	(0.3322,0.4665)	
	2	15	0.4	1.8425	0.7722	0.05	1.9963	0.4171	0.08
				(1.6765,2.0060)	(0.7136,0.8202)		(1.7091,2.2801)	(0.3355,0.4828)	
Set 2[b]	2	5	0.6	2.0375	0.9300	0.20	2.3493	0.8285	0.44
				(1.8412,2.2088)	(0.8344,0.9823)		(2.0519,2.6615)	(0.7175,0.9115)	
	4	2	0.4	3.6699	0.8019	0.31	3.2550	0.7131	0.47
				(3.1780,4.0766)	(0.6320,0.9201)		(2.5142,3.9011)	(0.5582,0.8262)	
	6	5	0.5	5.6248	0.8854	0.24	5.3592	0.7631	0.48
				(5.2943,5.9334)	(0.7899,0.9552)		(4.6386,5.9127)	(0.6155,0.8681)	
Set 3[c]	1	7	0.6	1.0527	0.8276	0.14	1.2148	0.5316	0.24
				(0.8823,1.2257)	(0.7379,0.8961)		(0.8289,1.9641)	(0.3769,0.6317)	
	2.5	6	0.5	2.7537	0.8969	0.18	3.0378	0.7282	0.36
				(2.5479,2.9584)	(0.8171,0.9479)		(2.6840,3.4131)	(0.6246,0.8061)	
	3	5	0.4	3.2014	0.8828	0.16	3.7587	0.7137	0.36
				(2.9778,3.4199)	(0.8151,0.9313)		(3.3226,4.2288)	(0.6093,0.7937)	

Table 2: Simulation results (posterior means, classification error and 95% credible intervals) where the true model is (12).

[a]Component means are far apart
[b]Component means are close
[c]Components means are neither too close nor too far apart

rate (the ratio of the number of wrongly classified observations over the total number of observations). We can see that Set 1 has much smaller misclassification rate than other sets (Sets 2 and 3).

From Figure 2 we can see that label switching does not occur. In fact we did not find any label switching in the trace plots based on all simulations in Table 2. Note that if we use a mixture model with a Poisson component and a NB component (the true underlying model) to analyse the data, the label switching problem still exists although the two components are different. This is shown in Figure 3, which is the simulation results for a mixture with two components: one Poisson component with a small mean value 2 and a negative binomial component with a larger mean value around 22.5. We can see from the trace plots that in this case the MCMC chain manages to estimate λ and π_1 close to their true values, 2 and 0.8 respectively. However, the algorithm

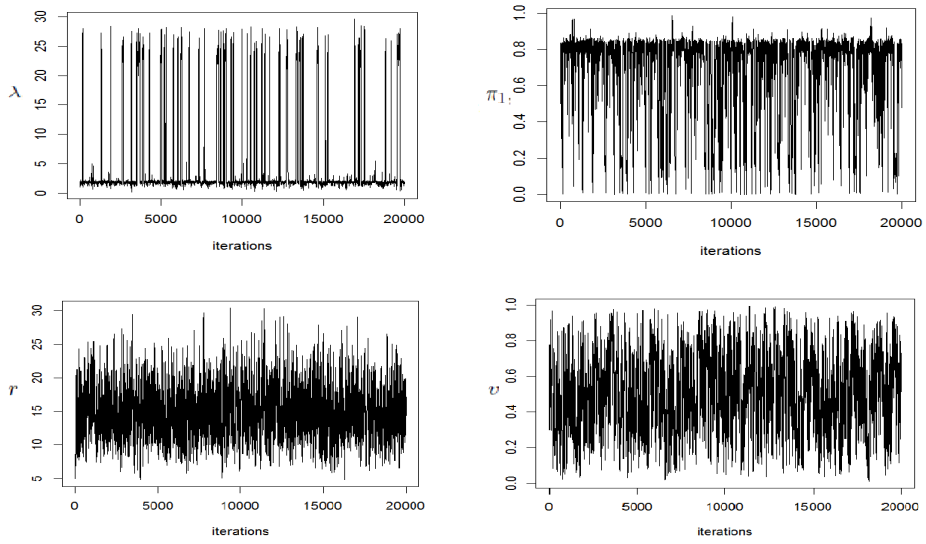

Figure 3: MCMC trace plots for λ, π_1, r and v by using a mixture of Poisson and NB distributions; the true model is (12) with $\lambda = 2$, $\pi_1 = 0.8$, $r = 15$ and $v = 0.4$.

sometimes returns estimates of λ around 20 and very small estimates of π_1, meaning that the Poisson distribution is used to model observations with large values but the NB distribution is used to model observations with small values. Such a label switching is due to the multi-modality property of the likelihood and makes it impossible to draw conclusions from the MCMC chains without some form of relabelling. However, all existing methods cannot deal with such label switching problems since they require the component distributions to be of the same type. Here we cannot simply relabel a Poisson parameter say $\lambda = 20$ to the pair of NB parameters (r, v). For simplicity of presentation we did not provide any results (such as posterior means and credible intervals) based on a mixture of Poisson and NB distribution here, since those results are severely biased.

3.2 Scenario 2

We now consider a more general mixture distribution with five-components, where the noise component is a Poisson distribution and the signal components are Negative Binomial distributions. The sample size is also chosen as $n = 500$. The aim here is to show that our methodology outperforms the fully parametric mixture model, under

general mixture distributions, in terms of estimation and classification. The true model for this simulation is given by

$$f(x) = \pi_1 \text{Poi}(x; \lambda) + \sum_{k=2}^{5} \pi_k \text{NB}(x; r_k, v_k). \tag{13}$$

We chose different values for the parameters λ, r_k and v_k in order to compare our method with existing methods under different settings.

First we choose the set of true parameters (Set 1) as $\lambda = 1$, $\pi_1 = 0.6$, $\pi_2 = \cdots = \pi_5 = 0.1$, $\boldsymbol{r} = (3, 5, 8, 10)$ and $\boldsymbol{v} = (0.3, 0.5, 0.7, 0.8)$. This choice of $\boldsymbol{r}$ and $\boldsymbol{v}$ for the NB components gives the corresponding component means as $(7, 5, 3.43, 2.5)$. Such a choice means that the Poisson component has the smallest mean, but the means of all components are not too far away. This would cause some identifiability problems for the Poisson component and other NB components if we use traditional Gibbs sampling methods. Indeed this is confirmed by Figure 4 where the MCMC trace plots for π_1 and λ clearly show the occurrence of the label switching problem. This issue severely distorts the posterior estimates, see Table 3. For example the posterior mean for λ is 1.9227 (the true value is 1) and the posterior mean for π_1 is 0.3101 (the true value is 0.6). On the contrary, if we use the new method, the estimates for λ and π_1 are 1.352 and 0.749, respectively, which are closer to the true values. For simplicity we did not provide the estimates for $\boldsymbol{r}$ and $\boldsymbol{v}$ since the main aim here is classification and under the new model $\boldsymbol{r}$ and $\boldsymbol{v}$ are not involved.

To justify the classification performance of the new method, one may use the posterior probability distribution for $\boldsymbol{z}$ as the classification criteria. The posterior probability of $z_i = 1$ is given by,

$$g_i = P(z_i = 1 | \boldsymbol{x}, \boldsymbol{\theta}) := \frac{\pi_1 f_1(x_i; \boldsymbol{\theta}_1)}{\pi_1 f_1(x_i; \boldsymbol{\theta}_1) + \pi_2 \sum_{j=1}^{L} p_j I[x_i = x_{(j)}]}. \tag{14}$$

If g_i is less than a threshold, say ρ, the value x_i will be classified into class 2. Based on this idea, false discovery rate (FDR) is commonly used to justify the performance of a classifier and was for example used by (Bao et al., 2013) in the context of mixture

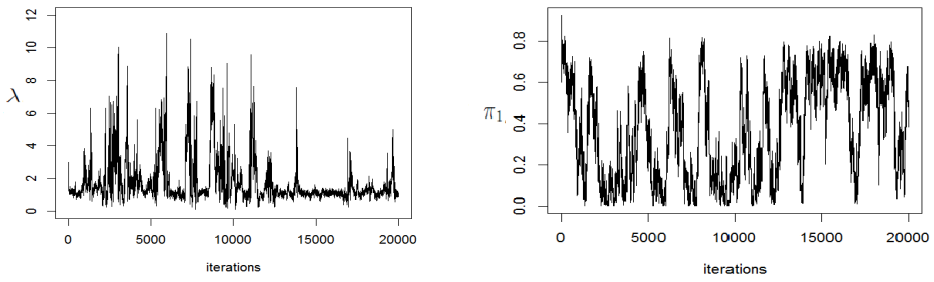

Figure 4: MCMC trace plots for λ, π_1 by using the true model, a mixture of a Poisson and four NB distributions; the true parameter values are $\lambda = 1$, $r_1 = 3, r_2 = 5, r_3 = 8, r_4 = 10$ and $v_1 = 0.3, v_2 = 0.5, v_3 = 0.7, v_4 = 0.8$, $\pi_1 = 0.6$, $\pi_2 = \cdots = \pi_5 = 0.1$.

models. It is defined as

$$\begin{aligned} FDR &= \frac{\#\{\text{false positive discovery}\}}{\#\{\text{declared positive}\}} \\ &= \frac{\#\{\text{false positive discovery}\}}{\sum_i I[g_i < \rho]}. \end{aligned}$$

We fixed the FDR at level 0.01 and find the threshold ρ and further calculate the false non-discovery rate (FNDR) based on the existing method and our new proposed method. The FNDR is defined as

$$FNDR = \frac{\#\{\text{false negative discovery}\}}{\#\{\text{declared negative}\}}$$

The FNDR values are shown on the last column of Table 3. The new method has smaller FNDR.

Model	True value										Posterior mean		FNDR
	λ	π_1	r_1	r_2	r_3	r_4	v_1	v_2	v_3	v_4	$E(\lambda)$	$E(\pi_1)$	
(i)	1	0.6	3	5	8	10	0.3	0.5	0.7	0.8	1.3516	0.7488	0.078
											(1.0855,1.6115)	(0.6385,0.8260)	
(ii)	1	0.6	3	5	8	10	0.3	0.5	0.7	0.8	1.9227	0.3101	0.406
											(0.6301,3.6701)	(0.0368,0.7504)	

Table 3: Parameter Set 1. (i) the new method; (ii) existing mixture model, the true mixture model of five components is used. FDR is controlled at level 0.01.

Note that, for the results in Table 3, we run the Gibbs sampler for 20,000 steps with 10,000 steps as burn-in iterations. For both methods, we choose a Gamma$(2, 1)$ prior

distribution for λ and a uniform prior distribution for $\boldsymbol{\pi}$. For the new method we choose uniform priors for $\boldsymbol{p}$, whereas for the Poisson-NB mixture we choose a Gamma$(20, 1)$ prior for the elements of $\boldsymbol{r}$ and a uniform distribution for the elements in $\boldsymbol{v}$. Furthermore, for the parametric mixture, we use a Metropolis-Within-Gibbs sampler to simulate from the posterior distributions, given the difficulty in simulating the parameters $\boldsymbol{r}$ and $\boldsymbol{v}$ for NB distributions.

In a second simulation, we choose the set of true parameters (Set 2) as $\lambda = 5$, $\pi_1 = 0.6$, $\pi_2 = \cdots = \pi_5 = 0.1$, $\boldsymbol{r} = (5, 7, 10, 14)$ and $\boldsymbol{v} = (0.4, 0.6, 0.8, 0.9)$. This choice of $\boldsymbol{r}$ and $\boldsymbol{v}$ for the NB components gives the corresponding component means as $(7.5, 4.67, 2.5, 1.56)$. Such a choice still gives very close means for each component, but now the Poisson component does not have the smallest mean. The posterior estimates based on the traditional parametric mixture model is still very poor and our method returns a smaller FNDR (see Table 4).

Model	True value										Posterior mean		FNDR
	λ	π_1	r_1	r_2	r_3	r_4	v_1	v_2	v_3	v_4	$E(\lambda)$	$E(\pi_1)$	
(i)	5	0.6	5	7	10	14	0.4	0.6	0.8	0.9	4.5530	0.8010	0.24
											(4.1507,4.9451)	(0.6720,0.8930)	
(ii)	5	0.6	5	7	10	14	0.4	0.6	0.8	0.9	4.1706	0.4278	0.52
											(0.6884,5.7709)	(0.0082,0.8059)	

Table 4: Parameter Set 2. (i) the new method; (ii) existing mixture model, the true mixture model of five components is used. FDR is controlled at level 0.01.

In the final simulation, we choose the set of true parameters (Set 3) as $\lambda = 6$, $\pi_1 = 0.6$, $\pi_2 = \cdots = \pi_5 = 0.1$, $\boldsymbol{r} = (8, 12, 30, 40)$ and $\boldsymbol{v} = (0.3, 0.3, 0.4, 0.3)$. This choice of $\boldsymbol{r}$ and $\boldsymbol{v}$ for the NB components gives the corresponding component means as $(18.7, 28, 45, 93.3)$. Such a choice will give very different component means with the Poisson component having the smallest mean. This situation is similar to the real ChIP-seq data, in the sense that there is a long tail and the noise component has the smallest mean value. From Table 5 we can see that our method gives posterior mean estimates for λ and π_1 with smaller bias and shorter credible intervals than the parametric mixture approach. Once again, the larger bias and variation in the

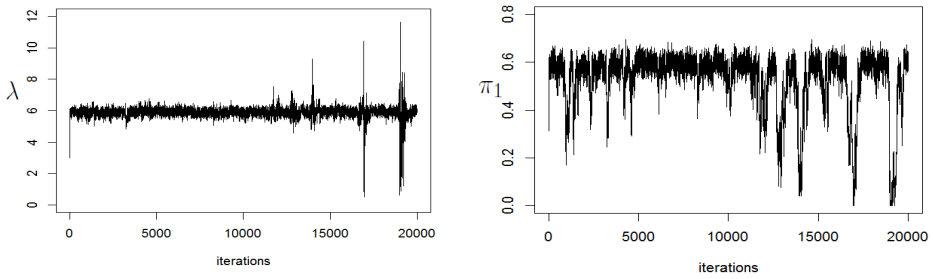

Figure 5: MCMC trace plots for λ, π_1 by using the true model, a mixture of a Poisson and four NB distributions; the true parameter values are $\lambda = 6$, $r_1 = 8, r_2 = 12, r_3 = 30, r_4 = 40$ and $v_1 = 0.3, v_2 = 0.3, v_3 = 0.4, v_4 = 0.3$, $\pi_1 = 0.6$, $\pi_2 = \cdots = \pi_5 = 0.1$.

estimates given by the existing methods is due to the label switching problem, see Figure 5. However, for this scenario, since the signal and noise are far apart, the new method did not gain an advantage in terms of FNDR, when controlling the FDR at level 0.01.

Model	True value										Posterior mean		FNDR
	λ	π_1	r_1	r_2	r_3	r_4	v_1	v_2	v_3	v_4	$E(\lambda)$	$E(\pi_1)$	
(i)	6	0.6	8	12	30	40	0.3	0.3	0.4	0.3	5.7662	0.5799	0.10
											(5.4113,6.1190)	(0.5227,0.6325)	
(ii)	6	0.6	8	12	30	40	0.3	0.3	0.4	0.3	5.8718	0.5372	0.03
											(5.0773,6.5490)	(0.0548,0.6376)	

Table 5: Parameter Set 3. (i) the new method; (ii) existing mixture model, the true mixture model of five components is used. FDR is controlled at level 0.01.

4 Data analysis

4.1 ChIP-seq data

As described in Section 1.1, we now show the applicability of the new method to ChIP-seq data. In ChIP-seq technology, the DNA is sheared into smaller fragments, typically 200 - 1000 base pairs (bp) long beforehand, this facilitates throughput sequencing. The dataset considered in this analysis is p300T301.1000bp dataset from the R package `enRich`, which is size-selected into 1000 base pairs (See Bao et al. (2013) for a description of the ChIP-seq technology and this particular dataset). The aim of the analysis is to detect the regions in the genome bound by the histone acetyltransferases p300, so it is a natural two-mixture problem with a background and a signal

component. Several methods for the analysis of ChIP-seq data assume a parametric signal distribution mixed with a parametric background distribution. For example, Kuan et al. (2011) propose a mixture of Negative Binomial distributions; Qin et al. (2010) adopt a generalized Poisson distribution for the signal and Bao et al. (2014) propose a zero-inflated Poisson/NB for noise and a NB for the signal. This paper considers a non-parametric model for the signal distribution, so that the variability in the signal can properly be accounted for.

Based on the posterior distribution, the posterior classification probability in (14) can be computed to predict a region is enriched or not. The region i will be classified as an enriched region if $g_i < \rho$. The threshold value ρ is determined by controlling the false discovery rate at a predefined level (Bao et al., 2014) say 0.001. The expected false discovery rate corresponding to the threshold value ρ is given by

$$0.001 = \widehat{FDR} \;=\; \frac{\sum_{i \in \text{enriched region}} (g_i)}{\sum_i I[g_i < \rho]}.$$

Figure 6 shows a Venn diagram of the regions detected as enriched by p300 using the model proposed in this paper, compared with a mixture of two Poisson distributions and a mixture of two NB distributions, at 0.1% false discovery rate. For the Poisson and NB mixtures we use the implementation in the `enRich` R package. Our method detects more enriched regions than the existing methods at the same false discovery rate. We use ChromHMM (Ernst and Manolis, 2010) to validate the enriched regions identified by the methods. Figure 7 shows the results based on ChromHMM with 3 chromatin states. The top plots give the emission probabilities for the different analyses, that is the probability of the observed enrichment given each of the three possible states. These plots show that two of the three states explain most of the enrichment pattern in the identified lists. The bottom plot give the relative fold enrichment for several annotations. These plots show how these two states are mostly enriched with TSSs, active and weak promoters, and weak enhancers. Furthermore, the plots show how the second state, which is mainly identified by our method, reflects a larger degree of

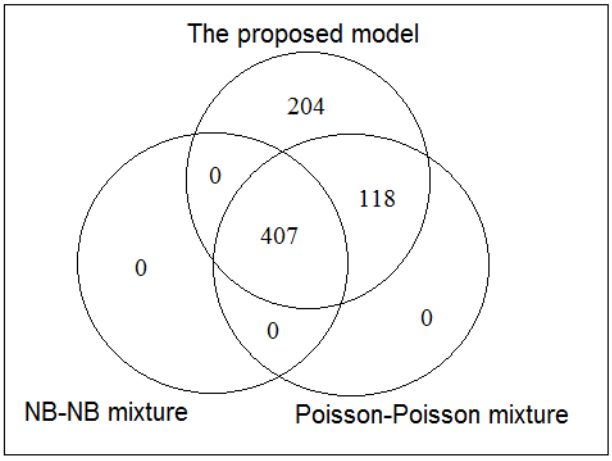

Figure 6: Number of enriched regions identified by the proposed model, Poisson-Poisson mixture model and NB-NB mixture model on chromosome21 at the 0.1% FDR.

enrichment of active and weak promoters. Therefore, we can conclude that by using the proposed method, more regions are found at the same FDR, and that these regions are generally of the same quality as those found by the existing methods.

5 Discussion

This paper developed a mixture model with a parametric component and a nonparametric component for modelling noise and signal, respectively. We showed several advantages to using a nonparametric component. Firstly, we neither need to specify the distributions for the signal component nor to consider how many components there are. Secondly, the method does not incur the label switching problem. Results on simulated data verify the validity of the approach and show a better performance of the method compared to fully parametric finite mixture distributions under general cases.

In our analysis the second mixture component is modelled as a nonparametric distribution, which actually involves L unknown parameters, the probabilities for distinct observations values in the enriched region. Therefore, if L is very large, the computational cost could be heavy. For the data set analyzed in this paper, the value of L is not too large, in the scale of 100, therefore the method is efficient. But if L is up

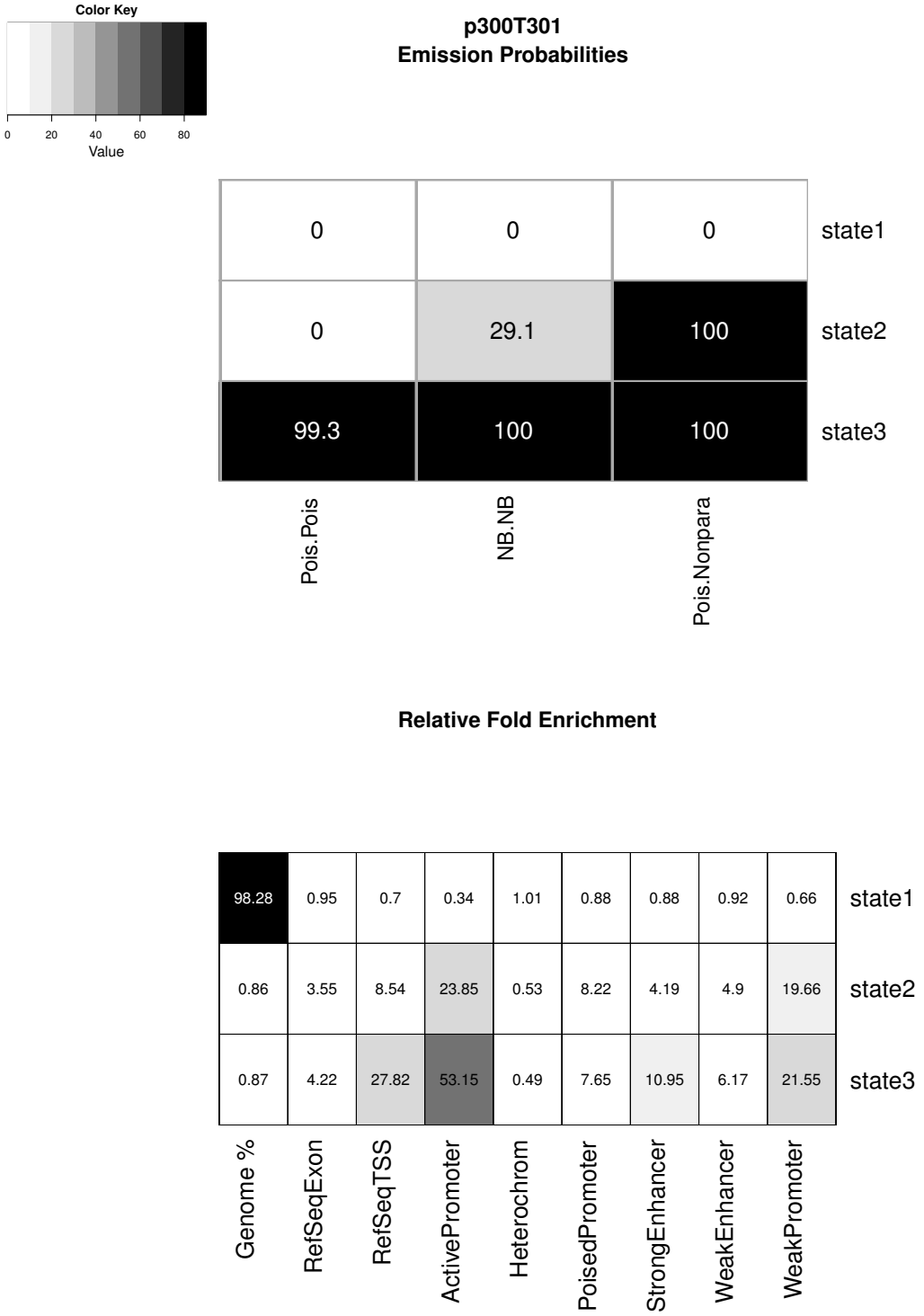

Figure 7: Validation of the enriched bins detected. The top plots show heatmaps of the probabilities (in percentages) that the p300 detected bins are enriched given each identified chromatin-state. The bottom plot shows the relative percentage of the genome represented by each chromatin state (first column) and the relative fold enrichment for several types of annotation (remaining columns).

to several thousands, the method may not be practical. In the context of ChIP-seq data, one solution for this is to consider smaller windows for genome regions. This will automatically reduce the value of L. However, when smaller windows for genome regions are considered, true enriched regions could easily cross two or more adjacent windows. In this case, the spatial dependencies between neighboring windows along the genome should be taken into account. Furthermore, for smaller fragments window size of 200bp, it is generally expected that greater part of the genome is not enriched with excess zeros in the output of ChIP-seq experiment, which form part of the noise component. Therefore, zero-inflated models (e.g. zero-inflated Poisson) or models with greater variance (e.g. negative binomial) are better choices for the noise component combine with more elaborate models with Markov property, such as HMMs or Markov random fields should be considered in this case such as the methods developed in Spyrou et al. (2009) and Bao et al. (2014). We are currently working on an extension of the methodology proposed in this paper to account for Markov dependencies.

The proposed method is only valid for discrete data sets, thus a possible extension is to develop methods able to deal with continuous data sets. In this case, a continuous distribution would be chosen for the noise component $f_1(x)$. However, new methods would need to be developed for the nonparametric component, since the posterior (9) of z_i in Algorithm 1 will not be valid anymore. This is left as a future research work.

References

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

Bao, Yanchun and Vinciotti, Veronica and Wit, Ernst and 't Hoen Peter A.C. (2014), Joint modelling of ChIP-seq data via a Markov random field model, Biostatistics, **15**(2), pp.296–310.

Bao, Yanchun and Vinciotti, Veronica and Wit, Ernst and 't Hoen Peter A.C. (2013),

Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data, BMC Bioinformatics, **14**, p.169.

Celeux, Gilles (1998), Bayesian inference for mixture: The label switching problem, in R. Payne and P.J. Green, eds, COMPSTAT 98, Physica, Heidelberg, pp.227–232.

Celeux, Gilles and Hurn, Merrilee and Robert, Christian P. (2000), Computational and inferential difficulties with mixture posterior distributions, Jounal of the American Statistical Association, **95**(451), pp.957–970.

Diebolt, Jean and Robert, Christian P. (1994), Estimation of finite mixture distributions through Bayesian sampling, Journal of the Royal Statistical Society, Series B, **56**, pp.363–375.

Escobar, Michael D. and West, Mike (1995), Bayesian density estimation and inference using mixtures, Journal of the American Statistical Association, **90**(430), pp.577–588.

Ernst J. and Manolis K. (2010), Discovery and characterization of chromatin states for systematic annotation of the human genome, Nature Biotechnology, **28**(8), pp.817-827.

Ferguson T. S. (1973), A Bayesian analysis of some nonparametric problems, The Annals of Statistics, **1**, pp.209–230.

Green P. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika, **82**(4), pp.711–732.

Jasra, A. and Holmes, C. C. and Stephens, D. A. (2005), Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling, Statistical Science, **20**, pp.50–67.

Kuan, Pei Fen and Chung, Dongjun and Pan, Guangjin and Thomson, James A and Stewart, Ron and Kele, S. (2011), A statistical framework for the analysis of ChIP-seq data, Journal of the American Statistical Association, **106**(495), pp.891-903.

McLachlan, Geoffrey and Peel, David (2004), Finite mixture models, Wiley.com.

Nobile, Agostino and Fearnside, Alastair T. (2007), Bayesian finite mixtures with an unknown number of components: The allocation sampler, Statistics and Computing, **17**(2), pp.147–162.

Qin, Zhaohui S. and Yu, Jianjun and Shen, Jincheng and Maher, Christopher A. and Hu, Ming and Kalyana-Sundaram, Shanker and Yu, Jindan and Chinnaiyan, Arul M. (2010), HPeak: an hmm-based algorithm for defining read-enriched regions in ChIP-seq data, BMC Bioinformatics, **11**(1), p.369.

Ramos, Yolande F. M. and Hestand, Matthew S. and Verlaan, Matty and Krabbendam, Elise and Ariyurek, Yavuz and van Galen Michiel and van Dam, Hans and van Ommen Gert-Jan B. and den Dunnen Johan T. and Zantema Alt and t Hoen Peter A.C. (2010), Genome-wide assessment of differential roles for p300 and CBP in transcription regulation, Nucleic Acids Research, **39**(16), pp.5396–5408.

Richardson, Sylvia and Green, Peter J. (1997), Bayesian analysis of mixtures with an unknown number of components (with discussion), Journal of the Royal Statistical Society, Series B, **59**(4), pp.731–792.

Rodriguez C. E. and Walker S. G. (2014), Label switching in bayesian mixture models: Deterministic relabeling strategies, Journal of Computational and Graphical Statistics, **23**, pp.25–45.

Song S., Nicolae D.L. and Song J. (2010), Estimating the mixing proportion in a semiparametric mixture model, Computational Statistics and Data Analysis, **54**, pp.2276–2283.

Spyrou, Christiana and Stark, Rory and Lynch, Andy G and Tavaré, Simon (2009), BayesPeak: Bayesian analysis of ChIP-seq data, BMC Bioinformatics, **10**(1), p.299.

Sperrin M., Jaki T. and Wit E. (2010), Probabilistic relabelling strategies for the

label switching problem in Bayesian mixture models, Statistics and Computing, **20**, pp.357–366.

Stephens, Matthew (2000a), Bayesian analysis of mixture models with an unknown number of components, an alternative to reversible jump methods, The Annals of Statistics, **28**, pp.40–74.

Stephens, Matthew (2000b), Dealing with label switching in mixture models, Journal of the Royal Statistical Society, Series B, **62**(4), pp.795–809.

West, Mike (1997), Hierarchical mixture models in neurological transmission analysis, Journal of the American Statistical Association, **92**(438), pp.587–606.

Xiang S., Yao W. and Wu J. (2014), Minimum profile hellinger distance estimation for a semiparametric mixture model, The Canadian Journal of Statistics, **42**, pp.246–267.