Analysis of ChIP-seq data via Bayesian finite mixture models with a non-parametric component

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

Department of Mathematical Sciences, University of Essex, Colchester, UK;
bbalha@essex.ac.uk

Abstract. In large discrete data sets which requires classification into signal and noise components, the distribution of the signal is often very bumpy and does not follow a standard distribution. Therefore the signal distribution is further modelled as a mixture of component distributions.

However, when the signal component is modelled as a mixture of distributions, we are faced with the challenges of justifying the number of components and the label switching problem (caused by multi-modality of the likelihood function). To circumvent these challenges, we propose a non-parametric structure for the signal component. This new method is more efficient in terms of precise estimates and better classifications. We demonstrate the efficacy of the methodology using a ChIP-sequencing data set.

Keywords

BAYESIAN, MIXTURE MODEL, LABEL SWITCHING, CHIP-SEQ

1 Introduction

The observations in a finite mixture model originates independently from a mixture distribution with K components that can be written as

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

where $\pi_k > 0$ with $\sum_k \pi_k = 1$ is the mixing weight of component k and $f_k(x; \boldsymbol{\theta}_k)$ belongs to a given parameterized family $\boldsymbol{\theta}_k$. This model has advantages of relaxing distributional assumptions. It represent subpopulations where the population membership is not known but is inferred from the data (McLachlan and Peel (2004)).

The existing literatures such as Diebolt and Robert (1994) and McLachlan and Peel (2004) have shown that finite mixture models can be inferred in a simple and effective way in a Bayesian estimation framework. Attentions has mostly focused on parametric mixture models, when the component densities are all from the same parametric family having different parameter values for the components. For example, all the distributions could be Poisson with different means or all the distributions could be

Negative Binomial with different parameters (even though, in practice, it is not necessary that all the densities will be of the same kind). This situation causes a persistent challenge in the diagnostic of Markov Chain Monte Carlo (MCMC) convergence due to two aspects.

The first aspect is the label switching problem which results from the multi-modality of the likelihood function. Many methods exist on how to tackle the label switching problem, for example, imposing identifiability constraints (Diebolt and Robert (1994), Richardson and Green (1997), McLachlan and Peel (2004)) and other methods based on relabelling algorithms (Celeux (1998), Stephens (2000b), Celeux et al. (2000), Rodriguez and Walker (2014)). For a review and comparison of these methods see, for example, Jasra et al. (2005) and Sperrin et al. (2010). One limitation to the existing methods for dealing with the label switching problem is that they focus on mixture models where all components having the same type of distributions. Another drawback common to these methods is that they require heavy computational costs, which make them unsuitable for large data sets and models with a large number of components. In practice, mixture components with different types of distributions are sometimes used, such as mixture of Poisson and Negative binomial distributions. In such situations, the likelihood function may still have multi-modes which causes label switching problem. But the existing methods for dealing with this problem may not be applicable in this case.

The other aspect is the justification of the number of components, K. Many authors have devised different stategies for estimating the number of components in 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. This is also computationally non-trivial when a large data set with several components is involved.

These motivates our study, which we discuss in detail in the following subsection.

1.1 Our motivation

In certain application areas, interest may be 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. 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.

For illustration, we use ChIP-seq data generated by Ramos et al. (2010) for the experiment on CREB binding protein (CBP) for identifying the genomic regions bound by the histone acetyltransferases (see Bao et al. (2013) for a description of ChIP-seq technology and this data set). For each region (1000bp) 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. The number of regions in the data set are 33916. The lowest count is zero and the highest count is 214, which means that some regions are tagged but with no protein of interest and a particular region is tagged with 214 counts. The mean and the variance are 2.13 and 8.76 respectively. Figure 1 shows a histogram of the count data. The left plot

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.

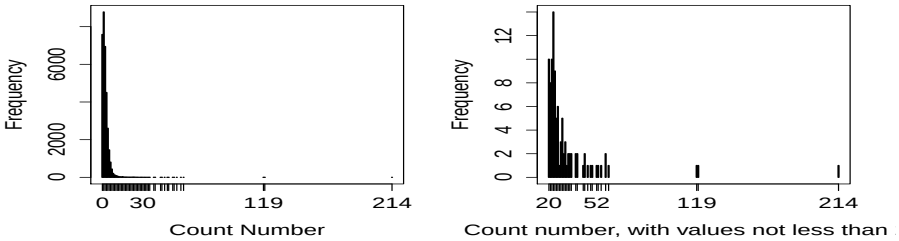

Fig. 1. Distribution of ChIP-seq data for one experiment (left), with zoom on the tail (right).

This situation has been observed also for other ChIP-seq analysis, 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 the true value of K. 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 very difficult to determine exactly the component distributions. 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. As such, using a mixture distribution with K components seems unnecessary. This motivate us to consider a two-component mixture model for discrete observations, with one parametric distribution and one nonparametric distribution.

The non-parametric distribution achieve several advantages. 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. It does not need to justify a particular parametric distribution for the signal. In the context of ChIP-seq data, our method detects the enriched regions in the genome with higher accuracy than the mixture of parametric distributions.

2 The model and the posterior distribution

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 mixture density in (1) to model the data, 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.

Associated to each observation x_i is a latent variable z_i i.e. $z_i = k$ ($k = 1, 2$) which represent the component from which the observation x_i originates. The complete likelihood function for $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ given the full data is

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

The noise distribution f_1 is usually simpler to determine. For example in ChIP-seq studies (for 1000bp where the proportion of zeros is not very large), 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 contrast to this, the signal distribution can present complicated patterns. We therefore consider using 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

$$f_2^*(x_{(j)}) = p_j, \ \sum_{j=1}^{L} p_j = 1\,; \tag{3}$$

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

$$f(x) = \pi_1 f_1(x; \boldsymbol{\theta}_1) + \pi_2 \sum_{j=1}^{L} f_2^*(x) I[x = x_{(j)}]. \tag{4}$$

Based on the distribution (3), 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.

Based on this, Gibbs sampler can be use to draw realisations from the posterior distribution and carry out a Bayesian Monte Carlo analysis. This leads to the following algorithm:

Algorithm 1: The Gibbs sampler

Initialization: select, $z^{(0)}, \boldsymbol{\pi}^{(0)}, \boldsymbol{p}^{(0)}$ and $\boldsymbol{\theta}_1^{(0)}$;
Set $m = 1$;
repeat
  for $i = 1$ *to* n **do**
    Update z_i with probability in

$$P(z_i = 1) \propto \pi_1 f_1(x_i; \boldsymbol{\theta}_1);$$

$$P(z_i = 2) \propto \pi_2 \sum_{j=1}^{L} p_j I[x_i = x_{(j)}];$$

  end
  Update $\boldsymbol{\theta}_1$ from the posterior in

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

  Update $\boldsymbol{\pi}$ from the posterior in

$$g(\boldsymbol{\pi}|\boldsymbol{x}, \boldsymbol{z}) \propto \pi_1^{n_1} \pi_2^{n_2};$$

  Update $\boldsymbol{p}$ from the posterior in

$$g(\boldsymbol{p}|\boldsymbol{x}, \boldsymbol{z}) \propto \prod_{j=1}^{L} p_j^{\sum_{i=1}^{n} I[z_i=2, x_i=x_{(j)}]};$$

  $m = m + 1$
until *Enough MCMC steps have been simulated*;

3 Simulation study

In the simulation study, we consider a mixture distribution with five-components, where the noise component is a Poisson distribution and the signal components are Negative Binomial distributions. We sample 500 observations. Our intention is to compare our proposed method with fully parametric mixture model in terms of estimation and classification. The true model for the simulation is given by

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

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.

In the First scenario, we choose the set of true parameters (Set 1) as $\lambda = 2$, $\pi_1 = 0.6$, $\pi_2 = \cdots = \pi_5 = 0.1$, $\boldsymbol{r} = (15, 13, 10, 8)$ and $\boldsymbol{v} = (0.9, 0.7, 0.6, 0.5)$. This choice of $\boldsymbol{r}$ and $\boldsymbol{v}$ for the NB components gives the corresponding component means as $(1.68, 5.57, 6.67, 8.00)$. Such a choice implies that the means of Poisson component and

all the other NB components are not too far apart. From Table 1 we can see that our method has clear posterior estimates, which approximate the true parameter value. The trace plot confirms that our method does not suffer from the label switching problem (see Figure 2). In fact label switching does not occur in our methodology.

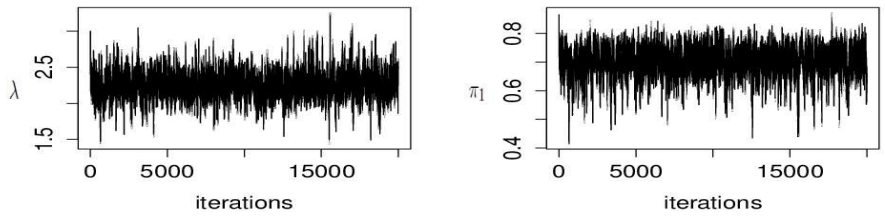

Fig. 2. MCMC trace plots for λ, π_1 for our new model for the true parameters in Table 1

However, for the Poisson component and other NB components, the above situation causes some identifiability problems when traditional Gibbs sampling method is used (see Figure 3). The MCMC trace plots in Figure 3 for π_1 and λ clearly show the occurrence of the label switching problem. This issue severely distorts the posterior estimates, see Table 1. For example the posterior mean for λ is 2.4371 (the true value is 2) and the posterior mean for π_1 is 0.2952 (the true value is 0.6). On the contrary, if we use the proposed method, the estimates for λ and π_1 are 2.2514 and 0.6987, 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. Instead we compare the misclassification rate (the ratio of the number of wrongly classified observations over the total number of observations) for the two methods. This can be easily obtained as the Bayesian approach provides the simulated $\boldsymbol{z}$ from the full posterior. From the last column of Table 1 we can see that our method has smaller misclassification rate than the parametric mixture model.

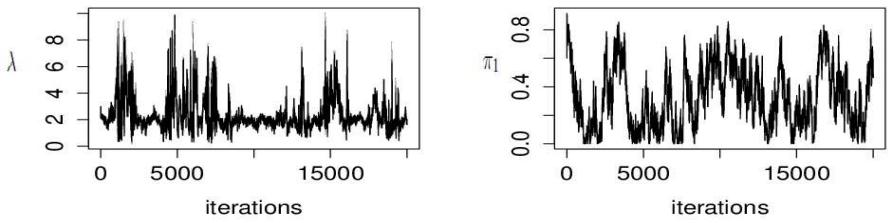

Fig. 3. MCMC trace plots for λ, π_1 for a mixture of a Poisson and four NB distributions for the true parameters in Table 1

In the second set of the simulation, the choice of the true parameters are $\lambda = 7$, $\pi_1 = 0.6$, $\pi_2 = \cdots = \pi_5 = 0.1$, $\boldsymbol{r} = (15, 20, 40, 30)$ and $\boldsymbol{v} = (0.4, 0.3, 0.3, 0.2)$. This choice of $\boldsymbol{r}$ and $\boldsymbol{v}$ for the NB components gives the corresponding component means

Table 1. Parameter Set 1. (i) the new method; (ii) true mixture model of five components.

Model	True value										Posterior mean		Error rate
	λ	π_1	r_1	r_2	r_3	r_4	v_1	v_2	v_3	v_4	λ	π_1	e
(i)	2	0.6	15	13	10	8	0.9	0.7	0.6	0.5	2.2514	0.6987	0.31
											(1.8881,2.6680)	(0.5680,0.7885)	
(ii)	2	0.6	15	13	10	8	0.9	0.7	0.6	0.5	2.4371	0.2952	0.46
											(1.0576,4.9958)	(0.0249,0.7433)	

as $(22.5, 46.7, 93, 120)$. This gives very different component means with the Poisson component having the smallest mean. This situation is similar to the real ChIP-seq data, in terms of long tail and the noise component has the smallest mean value. From Table 2 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. This is because our method does not incur the label switching problem. Contrarily, the larger bias and variation in the estimates in the existing methods is due to the label switching problem, see Figure 4. Still, the new method performs better in terms of misclassification rate.

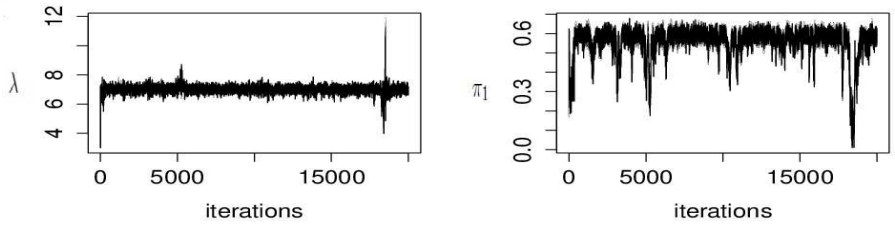

Fig. 4. MCMC trace plots for λ, π_1 for a mixture of a Poisson and four NB distributions for the true parameters in Table 2

Table 2. Parameter Set 2. (i) the new method; (ii) the true mixture model of five components.

Model	True value										Posterior mean		Error rate
	λ	π_1	r_1	r_2	r_3	r_4	v_1	v_2	v_3	v_4	λ	π_1	e
(i)	7	0.6	15	20	40	30	0.4	0.3	0.3	0.2	6.8676	0.5787	0.06
											(6.4998,7.2305)	(0.5226,0.6292)	
(ii)	7	0.6	15	20	40	30	0.4	0.3	0.3	0.2	6.9622	0.5349	0.10
											(6.4599,7.4080)	(0.2279,0.6329)	

For the results, we run the Gibbs sampler for 20,000 steps with 10,000 steps as burn-in iterations over 100 simulations. Furthermore, we use a Metropolis-Within-Gibbs

sampler to simulate from the posterior distributions for the parametric mixtures, given the difficulty in simulating the parameters $\boldsymbol{r}$ and $\boldsymbol{v}$ for NB distributions.

4 Data analysis

4.1 ChIP-seq data

We present the application of the new method to ChIP-seq data. We consider the GBPT301.1000bp data set from the R package `enRich`. Our aim is to detect the regions in the genome that are enriched, so it is a natural two-component mixture model problem with a noise and a signal components. Several methods for the analysis of ChIP-seq data assume a parametric signal distribution mixed with a parametric noise 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. (2013) propose a Poisson/NB for noise and a Poisson/NB for the signal. We consider the signal component as unknown and use nonparametric distribution to model it.

Based on the posterior distribution, the posterior classification probability can be used to predict whether a region is enriched or not.

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

The region i will be classified as an enriched region if $D_i < c$. The threshold value c is determined by controlling the false discovery rate (FDR) at a predefined level (Bao et al. (2013)), say 0.002. The expected false discovery rate corresponding to the threshold value c is given by

$$\widehat{FDR} := \frac{\sum_{i \in \text{enriched region}} (D_i)}{\sum_i I[D_i < c]}.$$

We present the result in Figure 5, which shows a Venn diagram of the detected regions as enriched for GBP experiment of ChIP-seq data for our proposed model, compared with a mixture of two Poisson distributions and a mixture of two NB distributions, at 0.2% 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.

5 Conclusion

We developed mixture model with a parametric and a nonparametric components. We achieved several advantages by using the nonparametric component. Firstly, we neither need to specify the distributions for the signal component nor to consider the number of components. Secondly, the method circumvents the label switching problem. Results on simulated data verify the validity of the approach and show a better performance in terms of estimation and classification. We illustrate the proposed method on ChIP-seq data (GBPT301.1000bp) to detect the enriched regions bound by proteins of interest.

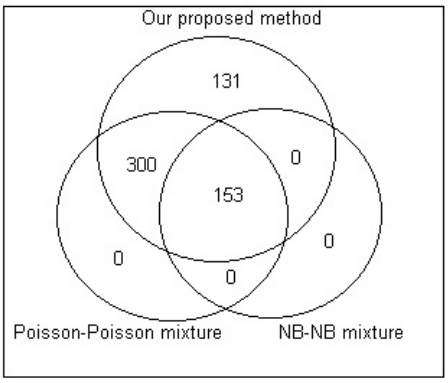

Fig. 5. Number of enriched regions identified by the proposed model, Poisson-Poisson mixture model, NB-NB mixture model on chromosome21 at the 0.2% FDR.

Relatively large window size (1000bp) in the ChIP-seq data motivates the use of traditional mixture models that do not account for Markov dependencies. For a smaller window size (say 200bp) we expect spatial dependencies between the neighboring windows. More elaborate models such as HMMs or Markov random fields should be considered in this case such as the method developed in Bao et al. (2014). The possible extension of this method to account for Markov dependencies is currently under investigation.

The proposed method is only valid for discrete data sets, thus a possible extension might be 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 of z_i in Algorithm 1 will not be valid anymore. This can be explored 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), 1152–1174.*

BAO, Y., VINCIOTTI, V., WIT, E. and 'T HOEN P.A.C. (2014): Joint Modelling of ChIP-seq Data Via a Markov Random Field Model. *Biostatistics, 15, 2, 296-310.*

BAO, Y., VINCIOTTI, V., WIT, E. and 'T HOEN P.A.C. (2013): Accounting for Immunoprecipitation Eficiencies in the Statistical Analysis of ChIP-seq Data. *BMC Bioinformatics, 14, 169.*

CELEUX, G. (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, G., HURN, M. and ROBERT, C.P. (2000): Computational and Inferential Dificulties With Mixture Posterior Distributions. *Journal of American Statististical Association, 95, 957-970.*

DIEBOLT, J. and ROBERT, C.P. (1994): Estimation of Finite Mixture Distributions Through Bayesian Sampling. *Journal of the Royal Statistical Society. Series B, 56, 363-375.*

ESCOBAR, M.D. and WEST, M. (1995): Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association, 90(430), 577-588*.

GREEN P. (1995): Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika, 82(4), 711-732*.

JASRA, A., 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, 50-67*.

KUAN, P.F., CHUNG, D., PAN, G., THOMSON, J.A., STEWART, R. and KELE, S.(2011): A Statistical Framework for the Analysis of ChIP-seq Data. *Journal of the American Statistical Association, 106(495), 891–903*.

MCLACHLAN, G. and PEEL, D. (2004): *Finite Mixture Models*. Wiley.com.

NOBILE, A. and FEARNSIDE, A.T. (2007): Bayesian Finite Mixtures With an Unknown Number of Components: The Allocation sampler. *Statistics and Computing, 17(2), 147-162*.

QIN, Z.S., YU, J., SHEN, J., MAHER, C.A., HU, M., KALYANA-SUNDARAM, S., YU, J. and CHINNAIYAN, A.M. (2010): HPeak: an HMM-Based Algorithm for Defining Read-Enriched Regions in ChIP-seq Data. *BMC bioinformatics, 11(1), 369*.

RAMOS, Y.F.M., HESTAND, M.S., VERLAAN, M., KRABBENDAM, E., ARIYUREK, Y., VAN GALEN M., VAN DAM, H., VAN OMMEN G.B.,DEN DUNNEN J.T., ZANTEMA A. and 'T HOEN P.A.C. (2010): Genome-Wide Assessment of Differential Roles for p300 and CBP in Transcription Regulation *Nucleic Acids Research, 39(16), 5396-5408*.

RICHARDSON, S. and GREEN, P.J. (1997): Bayesian Analysis of Mixtures With an Unknown Number of Components (With Discussion). *Journal of the Royal Statistical Society: series B, 59:(4), 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, 25-45*.

SPERRIN M., JAKI T. and WIT E. (2010): Probabilistic Relabelling Strategies for the Label Switching Problem in Bayesian Mixture Models. *Journal of Statistics and Computing, 20, 357-366*.

STEPHENS, M. (2000a): Bayesian Analysis of Mixture Models With an Unknown Number of Components An Alternative to Reversible Jump Methods. *Annal of Statistician, 28, 40-74*.

STEPHENS, M. (2000b): Dealing With Label Switching in Mixture Models. *Journal of the Royal Statistical Society: Series B, 62:(4), 795–809*.

WEST, M. (1997): Hierarchical Mixture Models in Neurological Transmission Analysis. *Journal of the American Statistical Association, 92(438), 587–606*.