Basturk, Nalan; Hoogerheide, Lennart; van Dijk, Herman K.

Working Paper

Bayes estimates of multimodal density features using DNA and Economic Data

Tinbergen Institute Discussion Paper, No. TI 2021-017/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Basturk, Nalan; Hoogerheide, Lennart; van Dijk, Herman K. (2021) : Bayes estimates of multimodal density features using DNA and Economic Data, Tinbergen Institute Discussion Paper, No. TI 2021-017/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/237750

TI 2021-017/III
Tinbergen Institute Discussion Paper

Bayes estimates of multimodal density features using DNA and Economic Data

Nalan Basturk[1]
Lennart Hoogerheide[2]
Herman K. van Dijk[3]

[1] Maastricht University
[2] Vrije Universiteit Amsterdam
[3] Erasmus University Rotterdam

Bayes Estimates of Multimodal Density Features using DNA and Economic Data*

immediate

January 26, 2021

Abstract

In several scientific fields, like bioinformatics, financial and macro-economics, important theoretical and practical issues exist that involve multimodal data distributions. We propose a Bayesian approach using mixtures distributions to approximate accurately such data distributions. Shape and other features of the mixture approximations are estimated including their uncertainty. For discrete data, we introduce a novel mixture of shifted Poisson distributions with an unknown number of components, which overcomes the equidispersion restriction in the standard Poisson which accomodates a wide range of shapes such as multimodality and long tails. Our simulation-based Bayesian inference treats the density features as random variables and highest credibility regions around features are easily obtained. For discrete data we develop an adapted version of the Reversible Jump Markov Chain Monte Carlo (RJMCMC) method, which allows for an unknown number of components instead of the more restrictive approach of choosing a particular number of mixture components using information criteria. Using simulated data, we show that our approach works successfully for three issues that one encounters during the estimation of mixtures: label switching; mixture complexity and prior information and mode membership versus component membership. The proposed method is applied to three empirical data sets: The count data method yields a novel perspective of the data on DNA tandem repeats in Schaap et al. (2013); the bimodal distribution of payment details of clients obtaining a loan from a financial institution in Spain

*The present paper should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. The authors thank Dennis Fok, Sylvia Frühwirth-Schnatter, Jim Griffin, Richard Paap, Jeroen Rombouts, Mark Steel for constructive discussions and helpful comments on an earlier version of this paper.

in 1990 gives insight into the repayment ability of individual clients; and the distribution of the modes of real GDP growth data from the PennWorld Tables and their evolution over time explores possible world-wide economic convergence as well as group convergence between the US and European countries. The results of our descriptive analysis may be used as input for forecasting and policy analysis.

JEL codes: C11, C14, C63.

Keywords: Multimodality, mixtures, Markov Chain Monte Carlo, Bayesian Inference.

1 Introduction

In several scientific fields there exist important theoretical and practical issues that involve multimodal distributions. In bioinformatics it is known, for instance, that more than half of the human genome consists of repetitive DNA. Tandemly repeated DNA sequences comprise a substantial proportion thereof. In different populations one can identify commonalities and differences of so-called Macro-Satellite Repeats (MSRs) as a particular group of repetitive DNA sequences with evidence of multimodal size distributions. Observed deviations in the configurations of MSRs may indicate diseases, see Bruce et al. (2009) and Balog et al. (2012) among several others.

In finance, the probability that a client of a bank is able to repay a loan may depend on the uncertainty that exists with respect to the number of defaulted payments. Analysis of this issue may involve bi- or multimodal data characteristics of the distribution due to the heterogeneity of the population of bank clients, see e.g. Dionne et al. (1996).

Thirdly, in the field of international economic growth, there exists the fundamental question whether or not convergence takes place between a set of countries when the observed distribution of Gross Domestic Product (GDP) growth is multimodal and changing over time, see for instance Quah (1996) and Paap and Van Dijk (1998) for evidence of multi-modality in growth after World War II.

Thus, in several empirical applications one deals with data from distributions that have multimodal densities. Standard exploratory data analysis and inference may be misleading if such a property of the data is not taken into account. Despite its importance for empirical

applications, assessing shape and other features of a multimodal probability density function is not straightforward. One aspect is the distinction between density estimation itself, aiming for a good fit to the data, and the estimation of such features as the relevant number of modes, the classification issue of belonging to a particular mixture component and accurate estimation of tail probabilities. These issues give rise to separate estimation problems (Good and Gaskins, 1980).

Existing methods to analyze multimodal densities are mostly limited to continuous distributions, see e.g. Silverman's test and its extensions (Silverman, 1981; Fischer et al., 1994), the DIP test (Hartigan and Hartigan, 1985; Hartigan, 1985), and the 'smooth bootstrap' test (Efron and Tibshirani, 1994) for 'bump hunting' or counting the number of modes in a probability density. Furthermore, even when the number of modes are estimated, most existing methods do not reveal uncertainty around these values (Fischer et al., 1994; Hall and Ooi, 2004), with a few exceptions based on nonparametric density estimates (Müller and Sawitzki, 1991; Minnotte, 1997; Chaudhuri and Marron, 1999). We refer to McLachlan and Peel (2004) for an extensive summary of frequentist and Bayesian methods for determining the number of modes.

This paper proposes a Bayesian approach to estimate shape and other features of a multimodal density and the uncertainty around these values. The method we propose is applicable for continuous and discrete data distributions. For continuous multimodal data, we show that estimates based on mixtures of normal densities with an unknown number of components provide a straightforward method to evaluate density features. For discrete such data, we introduce a novel mixture of shifted Poisson distributions with an unknown number of components, which overcomes the equidispersion restriction in the standard Poisson. This implies that our shifted Poisson mixture accommodates a wide range of shapes such as multimodality and long tails.

Density estimates of the mixture distribution are obtained using simulation-based Bayesian inference with density features treated as random variables and highest credibility regions around features are automatically obtained without any extra computational effort. For discrete data we develop an adapted version of the Reversible Jump Markov Chain Monte Carlo (RJMCMC) of Green (1995). Our method allows for an unknown number of components instead of the more restrictive approach of choosing a particular number of mixture components using information criteria as done in Schaap et al. (2013).

Using simulated continuous and discrete data, we focus on three issues that one encounters during the estimation of mixtures. Label switching between mixture components due to weak or no identification of individual components is tackled through parameter restrictions defined by flexible priors. This worked well in all cases considered. On the issue of mixture complexity and prior information we show conclusive evidence that our methodology is robust to alternative parameter specifications when the number of prior components is known. For the case of an unknown number of components we show that in our experiments the number of models with the highest posterior probability correspond to the true number of modes for each data set that we consider. As third issue we report results on mode membership versus component membership, making use of the K-means algorithm as in (Fruhwirth-Schnatter, 2006, p. 97 and Fruwirth-Schnatter(2011).

The proposed methods are applied to three empirical data sets. First, the count data method yields a novel perspective on the data of DNA tandem repeats in Schaap et al. (2013). We show that posterior distributions of the number of modes and quantiles are concentrated around their true value in all our diverse DNA examples. We compare results of the proposed method with two well-known frequentist tests, the Silverman test (Silverman, 1981) and the DIP test (Hartigan and Hartigan, 1985) and show that our approach performs better than both of these frequentist tests. Note that this also holds for the simulated data.

Next, payment details of clients obtaining a loan from a financial institution in Spain in 1990 are studied. Given that the bimodal data distribution has a fat tail, this implies that a standard Poisson distribution (or mixtures) may not be sufficient to approximate that data feature. Our more accurate descriptive analysis provided by a mixture of shifted Poisson distributions and the estimated modes may be used for the institutions policy of granting loans. For instance, according to our descriptive analysis the proposed method can already be used to categorise clients according to their defaulted instalment behaviour based on the estimated modes of the distribution. In further research with extensive micro data on explanatory variables, differences between individuals can even be better categorised.

As a final application, we consider the modes of cross-country GDP per capita distribution using data collected from 170 countries from Penn World Tables for the period 1960-2009. We apply a mixture of normal densities to approximate the distribution of the data and to analyse

the number of modes over time. These modes can be used to analyse the highly debated topic of convergence or divergence in economic growth between countries. Furthermore, it is of interest to see the evolution of these modes over time. Using the Penn World Tables for the period 1960-2009, we show that the changing number of modes does not necessarily indicate convergence, but instead, a middle income category seems to be emerging over time. We also analyse convergence between US and European countries. At the beginning of the sample, the results indicate 2 or 3 modes for GDP growth, while the number of modes is 1 for the period 1990-2009. Apparently, group behaviour with respect to convergence is more prevalent than overall convergence in the world. The results of our descriptive analysis may be used later as input for a causal analysis and policy measures.

The remainder of this paper is as follows: Section 2 presents the normal and Poisson mixture model and the RJMCMC algorithms for both models. Section 3 presents simulated data illustrations. Section 4 presents the results from applying these models to the three different datasets. Section 5 concludes. We emphasize that a computer package – The R-package MultiMode: Efficient and Robust Simulated Multimodal Densities – accompanies the present paper. Additional results on the use of the algorithm and set-up and results of the simulation experiments are shown in an appendix containing supplementary material.

2 Bayesian estimation of Normal and Shifted Poisson Mixture Distributions using RJMCMC

2.1 Mixture of normal densities for continuous data

A mixture of normal densities can be used to approximate an empirical multimodal distribution of continuous data. For a mixture of J normal densities, each mixture component $j \in \{1, \ldots, J\}$ has three parameters: mean μ_j, variance σ_j^2 and probability of mixture component π_j with the restriction $\pi_j \geq 0, \forall j$ and $\sum_{j=1}^{J} \pi_j = 1$. The total number of parameters for the mixture of normal densities is thus $3 \times J$, where the number of free parameters is $3 \times J - 1$ due to the restriction $\sum_{j=1}^{J} \pi_j = 1$. The approximation properties of a mixture of normal densities are well-studied for the continuous data case in the literature, see Frühwirth-Schnatter (2006) for general background and more specific references cited there. It is well-known that with a sufficiently

large number of normal mixture components any empirical distribution of continuous data and its features like skewness, fat tails and/or multimodality can be accurately approximated. However, a nontrivial problem is to determine the proper number of components in empirical situations. Regular MCMC methods cannot be used for this purpose since the number of model parameters changes with the number of mixture components. For this reason, we opt for the RJMCMC algorithm of Green (1995) which allows 'jumping' between parameter subspaces of different dimension.

A major idea of RJMCMC is to equate the number of parameters between different models, in this case for different J, and then to use standard MCMC tools. The 'jump' between mixtures with a different numbers of components occurs in two ways. The first possible move, 'split', indicates that the algorithm starts with a mixture model with J components and jumps to $J+1$ components. The second possible move, 'combine', indicates that the algorithm starts with a mixture model with J components and jumps to $J-1$ components. These 'proposed' jumps are accepted with a probability derived from the posterior probabilities of the models. For the split move, an algorithm for obtaining means, variances and probabilities of the *new* components has to be specified. For the combine move, an algorithm for obtaining mean, variance and probability of *reduced* mixture component has to be specified. We employ the split and combine moves proposed in Richardson and Green (1997) and, for convenience, summarise this methodology in the supplementary material. Given posterior draws of model parameters, inference on the number and location of the modes are obtained using the algorithm in Appendix B.

2.2 Mixture of shifted Poisson distributions for count data

We specify a mixture of shifted Poisson distributions that is intended to describe accurately multimodal count data. Let y_i for $i = 1, \ldots, n$ be independent realizations from a mixture of J shifted Poisson distributions:

$$y_i - \kappa_j \sim \text{Poisson}(\lambda_j) \text{ if } z_{ij} = 1 \text{ for } i = 1, \ldots, n; j = 1, \ldots, J, \tag{1}$$

where $z_{ij} = 1$ if y_i belongs to cluster j, and 0 otherwise and the latent variable distribution is defined as $\Pr[z_{ij} = 1] = \pi_j$, for $i = 1, \ldots, n; j = 1, \ldots, J$, with $\pi_j \geq 0$ for $j = 1, \ldots, J$ and

$\sum_{j=1}^{J} \pi_j = 1$, and where the shift parameter κ_j is a non-negative integer.

In (1), unlike for the case of a regular Poisson, an equidispersion restriction is not present even when the number of mixture components is 1. Intuitively, the shift parameter κ_j identifies the amount of dispersion between the mean and variance for each component in the mixture. The parameters of a shifted Poisson specification have a more direct interpretation in term of moments than a negative binomial distribution since the mean and variance in the latter are governed by two parameters jointly, whereas the shifted Poisson has variance equal to the parameter λ_j. Furthermore, in (1), the shift parameter κ_j allows for underdispersion, whereas the mixing of multiple components allows for overdispersion. A mixture of Negative Binomial distributions would, however, not allow for underdispersion, only overdispersion. For background we refer to Frühwirth-Schnatter (2006).

The augmented likelihood of model (1) is:

$$\ell\left(y, z|\theta\right) = \begin{cases} \prod_{i=1}^{n} \prod_{j=1}^{J} \left[\exp(-\lambda_j) \frac{\lambda_j^{y_i - \kappa_j}}{(y_i - \kappa_j)!}\right]^{z_{ij}} \pi_j^{z_{ij}}, & y_i = \kappa_j, \kappa_j + 1, \ldots, \ \forall i, j \text{ with } z_{ij} = 1 \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where $y = \{y_1, \ldots, y_n\}$, $z_i = \{z_{i1}, \ldots, z_{iJ}\}$, $z = \{z_1, \ldots, z_n\}$ and $\theta = \{\lambda, \kappa, \pi\}$, $\lambda = \{\lambda_1, \ldots, \lambda_J\}$, $\kappa = \{\kappa_1, \ldots, \kappa_J\}$, $\pi = \{\pi_1, \ldots, \pi_J\}$.

For a known number of components, J, uninformative but proper priors can be assigned to parameters. We make use of uniform priors defined on bounded regions for λ_j and κ_j and a symmetric Dirichlet prior for the weight parameters π_j:

$$\lambda_j \sim \text{unif}(\lambda_{\min}, \lambda_{\max}) \tag{3}$$

$$\kappa_j \sim \text{unif}(\kappa_{\min}, \kappa_{\max}) \tag{4}$$

$$(\pi_1, \ldots, \pi_J) \sim \text{Dirichlet}(\alpha, \ldots, \alpha) \tag{5}$$

$$[\lambda_{\min}, \lambda_{\max}] = [\kappa_{\min}, \kappa_{\max}] = [0, \max(y_i | y_i = 1, \ldots, n)] \tag{6}$$

Details of the posterior sampler, i.e. Gibbs sampling and RJMCMC steps, for the model in (1) under the priors in (3)–(6) are given in Appendix A. For simulated data illustrations

and empirical applications, we define the prior probability of 0.5 for split and combine moves, implicating an equal prior probability for models with a different number of mixture components, J.

Remark. An alternative to RJMCMC is to use continuous time samplers. In this method, the jump process is replaced by an additional parameter on the number of components which has a Poisson prior distribution. Parameters of new components are drawn from the prior distribution. A basic reference for this method is Stephens (2000). We note that in Cappé et al. (2002), it is shown that RJMCMC and the continuous time sampler are theoretically equivalent in terms of convergence. In the present paper we have good experience with RJMCMC and leave comparison with an alternative method as a topic for further research.

3 Mixture estimation issues illustrated

Using simulated data, we show the results of several experiments dealing with three issues that one encounters during the estimation of mixture distributions. The issues refer to label switching, mixture complexity and prior information and mode membership versus component membership.

Label switching. Label switching between mixture components refers to the feature that the posterior distribution may be invariant to switching between components. One can make use of restrictions on means, variances or probabilities of each component that are defined in a flexible way through the prior, see e.g. Malsiner Walli, Frühwirth-Schnatter and Grün (2014). We note that in the case of the shifted Poisson, one of the following label switching constraints can be imposed: $\kappa_l < \kappa_j,$ for $l < j$, (based on the shift parameter), $\kappa_l + \lambda_l < \kappa_j + \lambda_k,$ for $l < j$, (based on the mean), or $\pi_l < \pi_j,$ for $l < j$, (based on the component probabilities). We emphasise that these constraints are not needed to analyse such features as accuracy of the estimated/predicted distribution of the data and the number of modes of multimodal distributions. But it is a relevant issue in connection with the classification of mixture components. We summarize this issue when we discuss the experiments relating to mode versus component membership.

Mixture complexity. An important methodological and practical issue is the relation between prior parameters and mixture complexity, *i.e.* a posterior density with a large number of components. The mixture approximation used in this paper aims to obtain an accurate estimate of the shape and several features of the posterior density. We emphasise that in applications with very weak or flat proper priors and a relatively small number of observations in particular components, over fitting or lack of sparsity may lead to finding a spurious number of modes, and the method may also be sensitive to outliers. This issue is also referred to as determining 'genuine multimodality' in situations where the danger is that minor modes are captured by a mixture model, see e.g., Grün and Leisch (2009); Frühwirth-Schnatter (2011).

For RJMCMC methods, the issue of over fitting is analysed in several papers. The behaviour of parameters in models with spurious modes can be linked to the specification of uniform priors for Dirichlet parameters, see Nobile (2004). For symmetric Dirichlet priors, as in equation (5), overfitting is likely to occur with (near) empty mixture components or with mixture components which are 'identical' in parameter values. The former is shown to be the case when the Dirichlet parameter α is small with $\alpha < d/2$ where d denotes the number of parameters in each mixture component. For symmetric Dirichlet priors with a relatively high parameter, $\alpha > d/2$, identical clusters are likely, see Rousseau and Mengersen (2011). We follow Frühwirth-Schnatter (2011) and take $\alpha = 4$ in the reminder of this paper.

We focus on the fit of Poisson mixtures for discrete data and on such features as the determination of the correct number of modes for the case of bi- and tri-modal distributions.[1] The simulation study we consider is similar to Woo and Sriram (2007) and Umashanger and Sriram (2009), with the extension of using a shifted Poisson distribution in part of the simulations.

For the simulation setup, the following choices have been made, see Table 1. Standard and shifted Poisson distributions are selected with 2 and 3 components and with model weights close to equal and very unequal. Sample sizes are taken as small (100) and large (1000). The true parameter values of the simulated data set are denoted by $\theta^{(j)} = (p_j, \lambda_j, \kappa_j) = (p_j, \lambda_j, 0)$ for $j \in \{1, \ldots, J\}$ for a J-component mixture of standard Poisson distributions, and by $\theta_s^{(j)} = (p_j, \lambda_j, \kappa_j)$

[1] These properties are well-studied in the literature for normal mixtures, see Frühwirth-Schnatter (2006). In the supplementary material we report simulation studies showing that a mixture of normal densities has good approximation properties for a large number of components in simulated large data sets but that the estimated shape and features of the posterior density are sensitive for the choice of this density in case of small samples.

Table 1: Parameters of simulated data from mixtures of standard and shifted Poisson distributions

	Parameters definitions for simulated datasets			
	J	$\theta_{(s)}^{(1)}$	$\theta_{(s)}^{(2)}$	$\theta_{(s)}^{(3)}$
Mixture of standard Poisson distributions				
dataset 1	2	$(0.50, 1, 0)$	$(0.50, 9, 0)$	–
dataset 2	2	$(0.80, 1, 0)$	$(0.20, 9, 0)$	–
dataset 3	3	$(0.45, 1, 0)$	$(0.45, 5, 0)$	$(0.10, 10, 0)$
Mixture of shifted Poisson distributions				
dataset 4	2	$(0.50, 1, 1)$	$(0.50, 9, 2)$	–
dataset 5	2	$(0.80, 1, 1)$	$(0.20, 9, 2)$	–
dataset 6	3	$(0.45, 1, 1)$	$(0.45, 5, 2)$	$(0.10, 10, 3)$

Note: The table presents true parameters for each simulated dataset from mixture of standard and shifted Poisson distributions. $J \in \{1, \ldots, 2\}$ denotes the number of mixture components for each dataset. Parameters of each component are defined as $\theta^{(J)} = (p_j, \lambda_j, 0)$ for $j \in \{1, \ldots, J\}$ for mixtures of standard Poisson distributions and $\theta_s^{(J)} = (p_j, \lambda_j, \kappa_j)$ for $j \in \{1, \ldots, J\}$ for mixtures of shifted Poisson distributions.

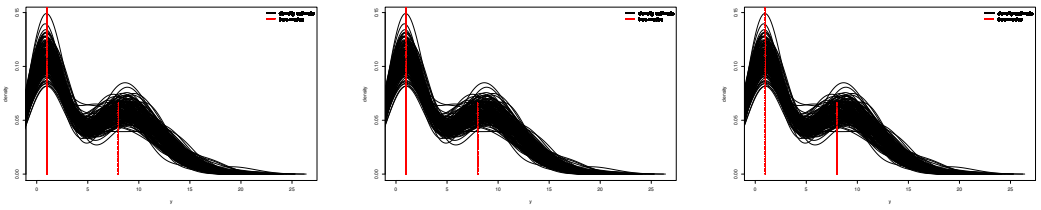
for a J-component mixture of shifted Poisson distributions. For the parameter specifications, we make use of the equivalence property between standard and shifted Poisson distributions with the value of shift parameter chosen as $\kappa = 0$. More background on the different experiments is provided in the supplementary material appendix.

In Figure 1 kernel density estimates for simulated data and true modes of the distribution (in vertical lines) are shown for each simulation setup. For all simulation setups, the smaller dataset with $n = 100$ has more uncertainty in density estimates compared to those with $n = 1000$. However, the obtained density properties, such as the mean, mode or quantiles, are shown with reasonable accuracy even for small samples.

Known number of mixture components We first consider a known number of mixture components for each simulated data set and report estimated density features that are particularly relevant for discrete data simulations. To the best of our knowledge, theoretical approximation properties do not exist in the literature for this case.

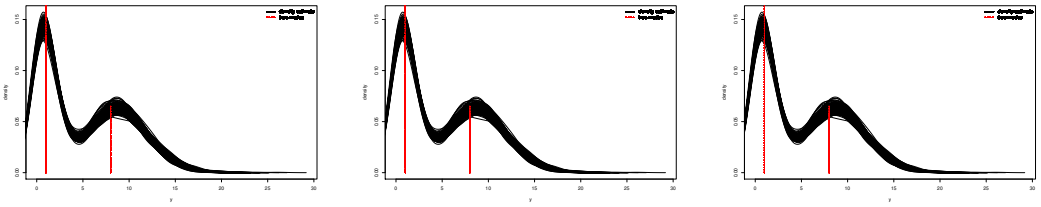
We estimate the model parameters together with the number of modes for each simulated data set using the model and the posterior sampler in Appendix A. The Dirichlet prior parameter $\alpha = 4$ and posterior results are based on 10000 total number of draws, and 5000 draws are disregarded as burn-in draws.

Figure 1: True and estimated probability density functions for simulation experiments. In each plot, the simulation study is replicated 150 times. Solid lines are kernel density estimates for each simulation replication. Vertical lines are the theoretical modes of the true pdf.



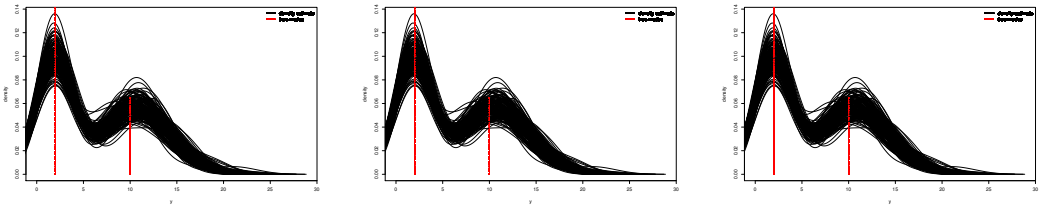
dataset 1 dataset 2 dataset 3

Mixture of standard Poisson densities, 100 observations



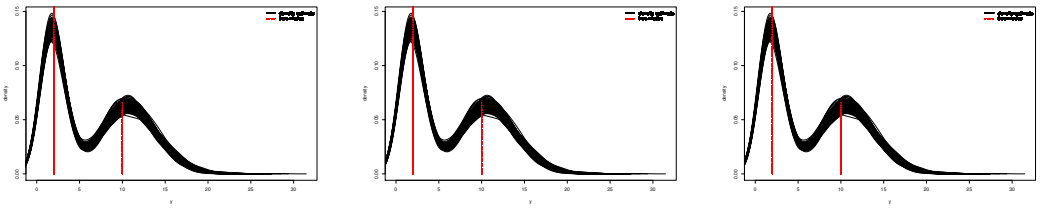
dataset 1 dataset 2 dataset 3

Mixture of standard Poisson densities, 1000 observations



dataset 4 dataset 5 dataset 6

Mixture of shifted Poisson densities, 100 observations



dataset 4 dataset 5 dataset 6

Mixture of shifted Poisson densities, 1000 observations

Posterior results for the number of modes, together with the true number of modes for each simulated data are reported in Table 2. In this table, a high probability value implies that the correct number of modes is found to be very likely according to the estimates. For most simulated data, the corresponding probability is very high (above 0.9). We conclude that the methodology is robust to alternative parameter specifications for the case where the number of mixture components is known. So, a preliminary conclusion is that prior knowledge on the number of components appears very helpful in order to obtain accurate results on density fit and features.

Table 2: Posterior modes for simulated discrete data with known J mixture components

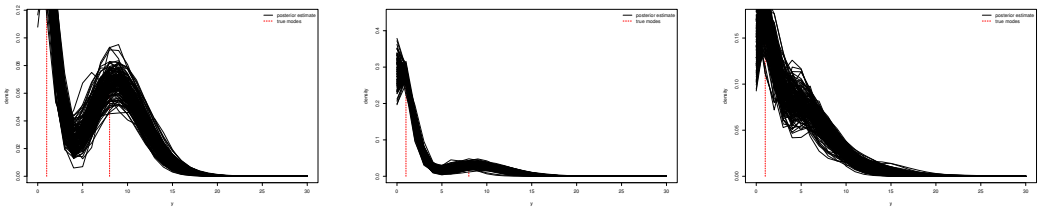
	number of components	number of modes	probability
Mixture of standard Poisson distributions			
data 1	2	2	1.00
data 2	2	2	1.00
data 3	3	1	0.95
Mixture of shifted Poisson distributions			
data 4	2	2	1.00
data 5	2	2	1.00
data 6	3	2	0.94

Note: The table reports the number of mixture components (column 1), true number of modes (column 2) and the posterior probability of true number of modes (column 3) for each simulated dataset. Posterior results are based on 10000 posterior draws (5000 burnaman bakalim, -in draws).

Unknown number of mixture components We next perform a more extensive simulation study with an unknown number of components J for discrete and continuous data. Discrete data sets are simulated from mixtures of standard Poisson distributions and from mixtures of shifted ones. For each distribution type, different parameter settings are considered with $J \in \{1, 2, 3, 4\}$ mixture components. Each sample consists of $n = 100$ observations and the simulation study is repeated 150 times.

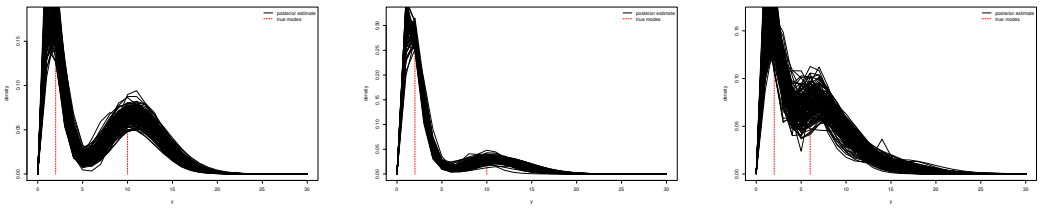
Table 3 contains estimation results for data simulated from standard and shifted mixtures of Poisson distributions. The true number of modes, the estimated mode with the highest posterior probability and the corresponding probability are reported for each data set. The number of modes with the highest posterior probability correspond to the true number of modes

Figure 2: RJMCMC estimates and true modes for simulated data. In each plot, simulation study is replicated 150 times. Vertical lines are the theoretical modes of the true pdf.



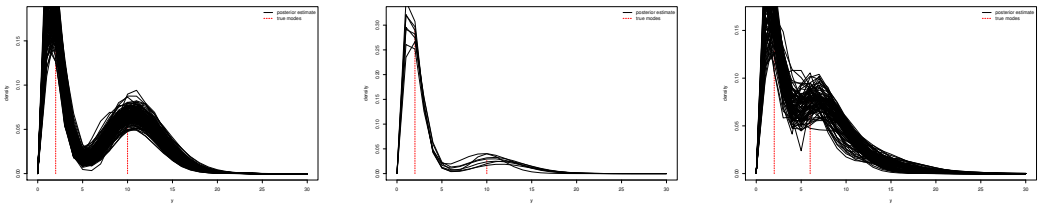
dataset 1 dataset 2 dataset 3

Mixture of standard Poisson distributions, 100 observations



dataset 4 dataset 5 dataset 6

Mixture of shifted Poisson distributions, 100 observations



dataset 4 dataset 5 dataset 6

Mixture of shifted Poisson distributions, 1000 observations

Table 3: True and estimated number of modes from simulated discrete data

	true value	value (max. pr.)	post. prob.	std.dev.	p_{S}	p_{DIP}	p_{HY}
Mixture of standard Poisson distributions							
data 1	2	2.00	0.98	0.12	0.01 (0.02)	0.00 (0.00)	
data 2	2	2.00	0.99	0.07	0.07 (0.04)	0.00 (0.00)	
data 3	2	2.00	0.74	0.46	0.15 (0.11)	0.00 (0.00)	
Mixture of shifted Poisson distributions							
data 4	2	2	0.97	0.05	0.15 (0.08)	0.07 (0.14)	0.02 (0.03)
data 5	2	2	0.99	0.02	0.34 (0.19)	0.69 (0.26)	0.10 (0.13)
data 6	2	2	0.75	0.18	0.60 (0.21)	0.58 (0.32)	0.34 (0.24)

Estimation results are based on 10000 draws (5000 burn-in draws) and averages from 10 simulation replications reported. Number of mixture components J is estimated together with the rest of the model parameters. p_{S} and p_{DIP} denote average p-values from Silverman and DIP tests for 150 simulation replications, with standard deviations of p-values in parentheses.

in each data set we consider. Furthermore, the posterior probability associated with the true number of modes is higher than 0.5 in all simulated data sets. In a few cases, reported in the additional material, this posterior probability is not close to 1, although the number of modes with maximum posterior probability does correspond to the true number of modes. We refer to the additional material appendix for more details.

Mode membership versus component membership. Component membership, i.e. the classification of belonging to a particular mixture component, can be done in a simple practical manner using the K-means algorithm as in (Frühwirth-Schnatter, 2006, pp. 97) and Frühwirth-Schnatter (2011). We explain how we execute this procedure in the supplementary material. We apply the k-means algorithm to posterior draws from the RJMCMC algorithm applied to simulated data from a mixture of Poisson distributions. We consider two simulated data sets. In the first data set, parameters of the Poisson distribution are chosen to be clearly different between the mixture components and as a result the generated distribution of the data is clearly bimodal with two distinct modes. In the second simulated data set, parameters of the Poisson distribution do not lead to such clear set of modes.

Re-labeled draws for simulated data with distinct modes: We show that in the first simulated data set with $n = 100$ observations with clearly distinct modes, the k-means algorithm is useful in *a posteriori* relabeling of parameter draws. A histogram of the data, posterior draws from RJMCMC and posterior draws relabeled using the k-means algorithm are shown in Figure 3.

We summarize the results as follows.

The top panel in Figure 3 shows the histogram of the simulated data, the true density and the density of each mixture component for $J = 2$ components. With a relatively small number of draws, the data histogram is very similar to the true density, and two clear modes can be easily observed in the data.

The bottom left panel in Figure 3 shows the subset of draws where the number of mixture components is $J^\star = 2$ and the number of mixture components with maximum posterior probability according to the RJMCMC method. Around 70% of posterior draws lead to 2 mixture components, hence only 30% of draws are 'lost' at this step of the k-means algorithm.

The bottom right panel in Figure 3 shows the re-labeled draws after application of the k-means algorithm. These draws correspond to a subset of draws where $J^\star = 2$, and k-means clustering of each parameter in the draws are permutations of $\{1, 2\}$. Around 64% of the initial number of posterior draws satisfy this condition.

The bottom right panel in Figure 3 also shows that k-means algorithm is successful in this case: Clustering of draws are clear, for example compared to the 'unlabelled' posterior draws on the bottom left panel of Figure 3.

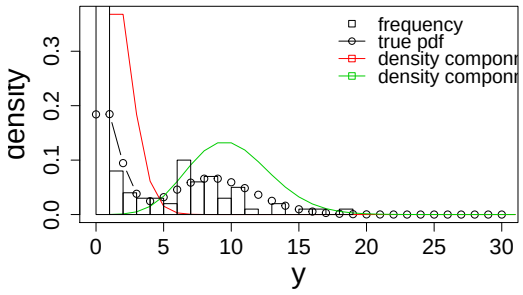
Thus, the k-means algorithm is successful in this case: The effective number of draws from the algorithm is close to the actual number of draws, ie. $\bar{M} \approx \tilde{M} \approx M$.

Re-labeled draws for simulated data without distinct modes: We show that in the second simulated data set with $n = 100$ observations, the k-means algorithm has the problem that only a small subset of draws are left after application of the algorithm. Furthermore, the label switching problem does not seem to be completely removed. The histogram of the data, posterior draws from RJMCMC and posterior draws labeled using the k-means algorithm are shown in Figure 4. The results are as follows.

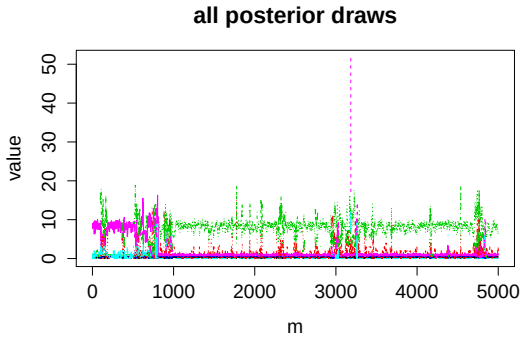
The top panel in Figure 4 shows the histogram of the simulated data, the true density and the density of each mixture component for $J = 4$ components. With a relatively small number of draws, the data histogram is very different from the the true density. The actual modes of the distribution are hardly visual in the figure.

The bottom left panel in Figure 4 shows the subset of draws where the number of mixture

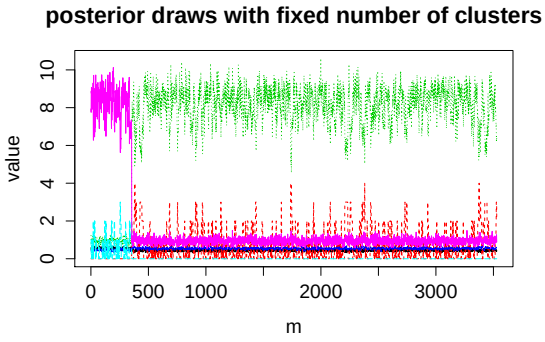
Figure 3: Histogram of simulated data with distinct modes, posterior draws from RJMCMC and re-labelled posterior draws from RJMCMC



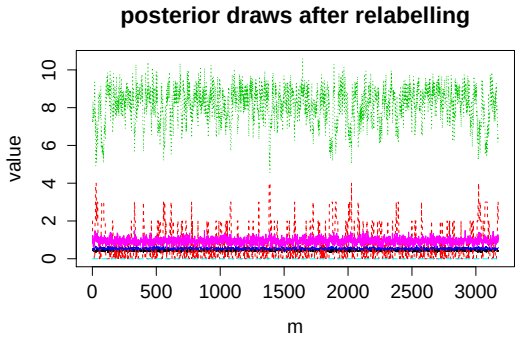
histogram of simulated data



$M = 5000$ posterior draws

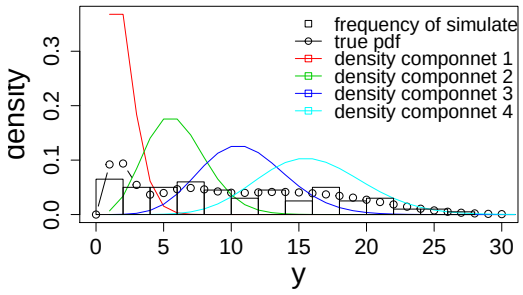


$\bar{M} = 3559$ posterior draws

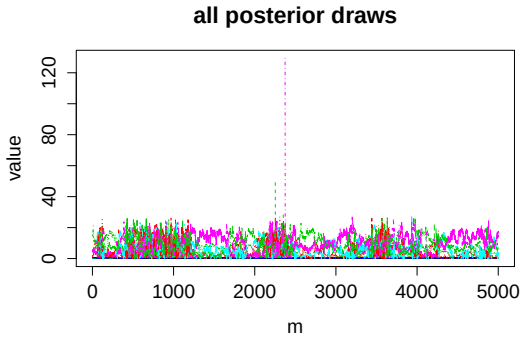


$\tilde{M} = 3175$ posterior draws

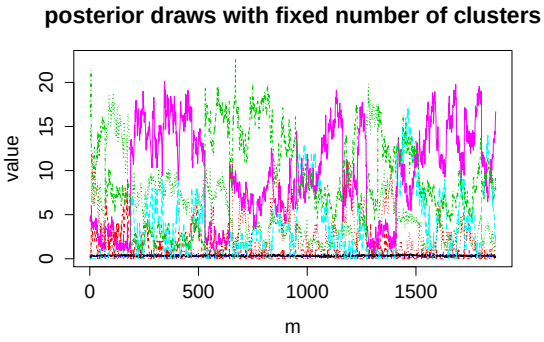
Figure 4: Histogram of simulated data with 'non-distinct' modes, posterior draws from RJMCMC and re-labelled posterior draws from RJMCMC



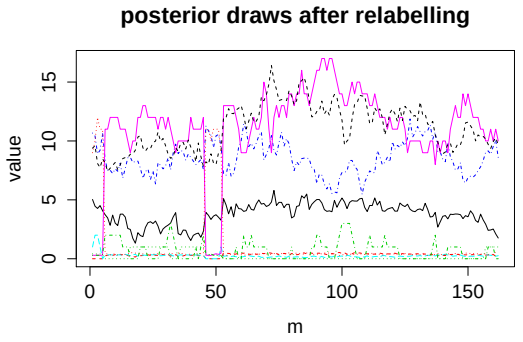
histogram of simulated data



$M = 5000$ posterior draws



$\bar{M} = 2368$ posterior draws



$\tilde{M} = 307$ posterior draws

components is $J^\star = 2$ which is the number of mixture components with maximum posterior probability in RJMCMC. Less than 50% of posterior draws lead to 4 mixture components. More than half of the posterior draws are 'disregarded' in this case, and label switching seems to be an important problem.

The bottom right panel in Figure 4 shows the re-labeled draws after the k-means algorithm. These draws correspond to a subset of draws where $J^\star = 4$, and k-means clustering of each parameter in the draws are permutations of $\{1, 2\}$. Only 307 draws satisfy these conditions. Furthermore, label switching problem is not eliminated completely particularly in the solid pink line in the figure. The k-means algorithm is not successful in this case: The effective number of draws from the algorithm are very small (for different random seeds, none of the draws satisfy k-means conditions). The bottom right panel in Figure 4 also shows that k-means algorithm is not successful in this case: Clustering of draws still show label switching around draw 50. Next, we investigate how well the simulation results hold in a variety of empirical bi- and multi-modal data distributions.

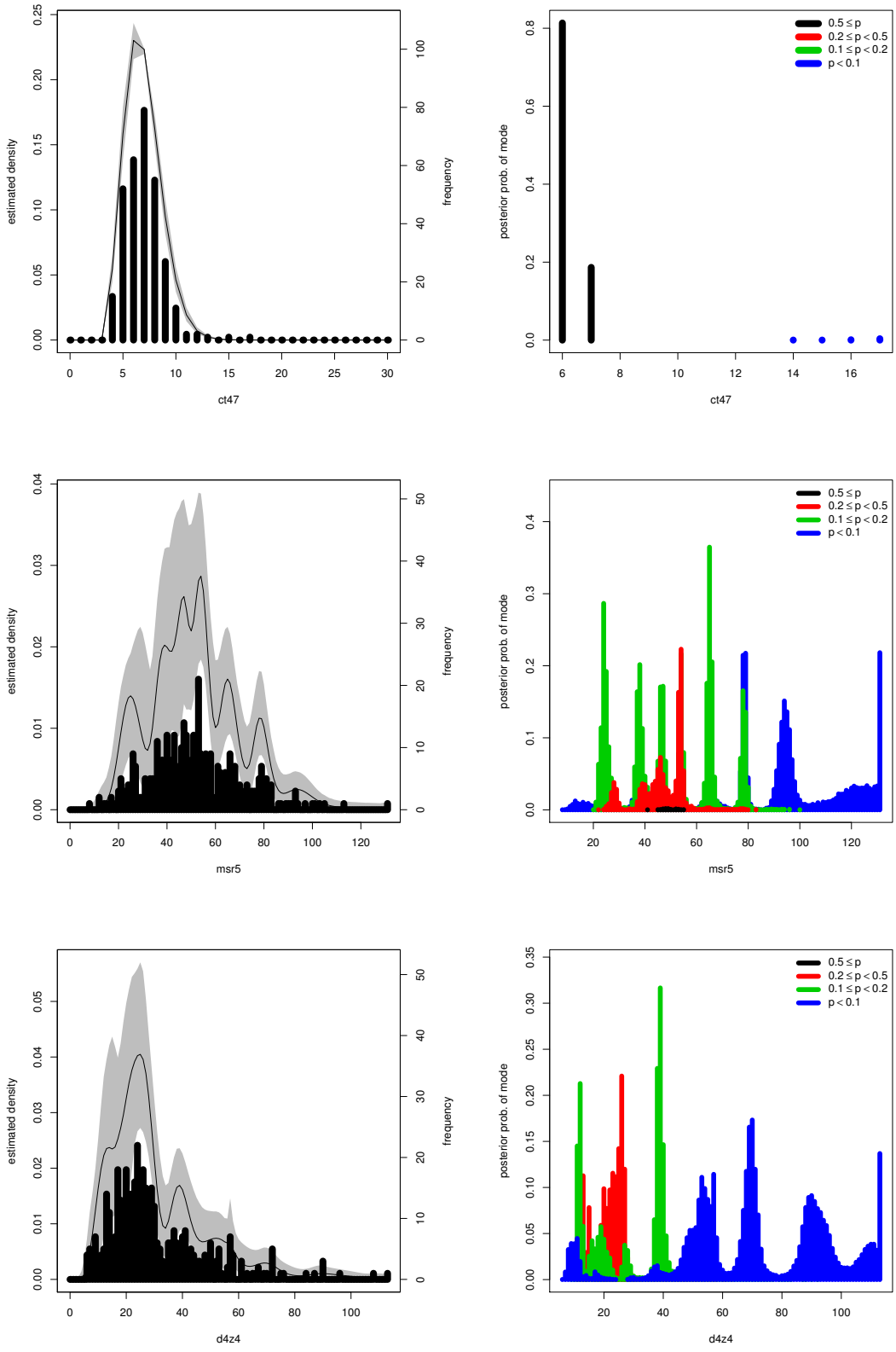
4 Three empirical applications

4.1 DNA tandem repeats data

In this subsection, we make use of a model that consists of a mixture of shifted Poisson distributions to estimate posterior features using counts of DNA tandem repeat data for the case of three specific DNA sequences denoted by CT47, MSR5 and D4Z4. These data are obtained from 270 unrelated human DNA samples from Asian, African and Caucasian origin, see Schaap et al. (2013). It is of substantial interest to analyse the number and location of modes in the data, since differences in these values may next be linked to e.g. genetic diseases.

The estimated numbers of modes are compared to results using the standard Silverman and DIP tests. The null hypotheses in both frequentist tests is a single mode, and the alternative hypothesis is at least two modes. In Figure 5 the estimated data distributions for three sequences are presented together with the posterior modes classified according to the 'severeness of modes', that is, calculated by the posterior probabilities of each mode. Sequence CT47 clearly has a single mode according to the estimate, while MSR5 and D4Z4 appear to have multiple modes according

Figure 5: DNA sequences CT47, MSR5 and D4Z4: estimated distributions of the data and posterior mode probabilities



Estimated density, 95% interval, posterior probabilities of modes

Table 4: DNA sequences CT47, MSR5 and D4Z4: Estimated number of modes using the Bayesian procedure and p-values of Silverman and DIP tests

data set	post. prob. for # of modes					post. std. dev.	Silverman test p-value	DIP test p-value
# of modes	1	2	3	4	5			
CT47	1.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00
# of modes	5	6	7	8	9			
MSR5	0.01	0.11	0.34	0.44	0.09	0.97	0.00	0.21
# of modes	5	6	7	8	9			
D4Z4	0.04	0.23	0.40	0.26	0.06	0.98	0.05	0.18

Estimation results are based on 10000 draws (5000 burn-in draws).

to the figures. In Table 4 the estimated numbers of modes are shown in detail, together with the p-values from the standard Silverman and DIP tests. Considering the first sequence, CT47, the results of our proposed method and those of the standard tests are very different since both frequentist tests do reject strongly the null hypothesis of a single mode while our Bayesian procedure indicates a single mode with substantial credibility. For the sequence MSR5, the Silverman test is in line with the proposed method, indicating a rejection of the null hypothesis of a single mode but the DIP test, on the other hand, does not reject the null even at a 10% level. For the sequence D4Z4 both Silverman and DIP test do not reject the null of a single mode at the 10% level, while our Bayesian procedure indicates the presence of several modes albeit with a certain degree of uncertainty. There is no clear pattern in Silverman and DIP tests in terms of how conservative these are in testing for the number of modes, but the two tests are clearly not appropriate for assessing the number of modes in this count data example. We note that the results of the Silverman and DIP tests may also suffer due to the fact that these tests assume a continuous underlying data distribution and the DNA data is a typical count data example.[2]

4.2 Defaulted payment instalments

For the second empirical application, we apply the mixture of shifted Poisson distributions to the case of count data on payment details of clients obtaining a loan from a financial institution in Spain in 1990. These data, the number of defaulted payment instalments, consists of 4329

[2]The minor differences between the estimates we report and those in Schaap et al. (2013) may be due to the fact that we estimate the number of mixtures together with the rest of the model parameters and do not remove outliers from the estimated density.

Table 5: Default data: Posterior distributions of modes and quantiles

post. prob. for # of modes					post. std. dev.
1	2	3	4	5	
0.00	0.20	0.78	0.02	0.00	0.43

Posterior distribution of quantiles

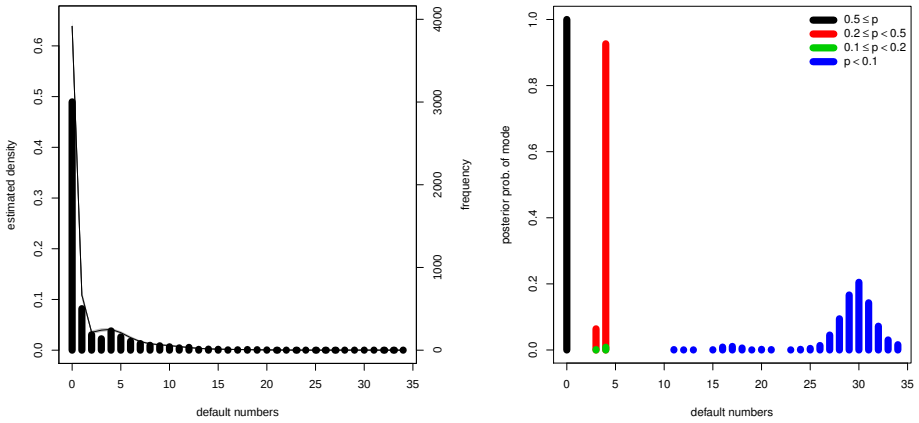
	0.05	0.1	0.5	0.9	0.95
mean	0.00	0.00	0.00	5.70	8.21
std. dev.	0.00	0.00	0.00	0.46	0.41

Estimation results are based on 10000 draws (5000 burn-in draws).

observations from 0 to 34 defaulted instalments and have been analysed in Dionne et al. (1996), Woo and Sriram (2007) and Karlis and Xekalaki (2001).

The estimated density using a mixture of shifted Poisson distributions and the posterior probabilities of the modes are shown in Figure 6. In Table 5 the posterior probabilities of the number of modes and estimated quantiles are presented.

Figure 6: Default data: Estimated distribution of the data and posterior probability of modes.



The figure shows the data histogram, estimated density (95% interval) and posterior probability of modes

We start to observe that these data is a typical example of 'zero inflated count data'. A standard Poisson distribution or even mixtures of Poissons may fail to approximate this data density given the long tail. The proposed mixture of *shifted* Poisson distribution, on the other hand, leads to accurate density estimates of these data as shown in Figure 6. As stated, given

that the data distribution has a fat tail, this implies that a standard Poisson distribution (or mixtures) may not be sufficient to approximate that data feature.

This accurate 'descriptive' analysis provided by a mixture of shifted Poisson distributions and the estimated modes may be used for the institution's policy of granting loans. According to our descriptive analysis the proposed method can already be used to categorise clients according to their defaulted instalment behaviour based on the estimated modes of the distribution. In further research with extensive micro data on explanatory variables, differences between individuals can even be better categorised.

4.3 Economic growth in many countries

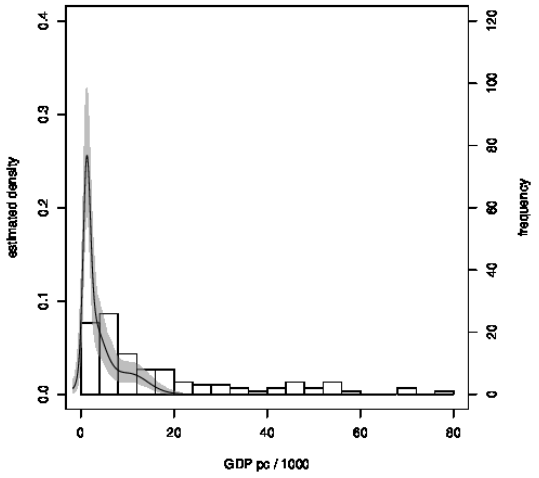
As a final application, we consider the modes of cross-country GDP per capita distribution. These modes can be used to indicate a highly debated topic, convergence or divergence in economic growth between countries, see Paap and Van Dijk (1998), Başţürk et al. (2010) among several others. Furthermore, it is of interest to see the full distribution of GDP per capita to analyse the evolution of these modes.

The data for this application are the average GDP per capita over 10 year intervals, collected from 170 countries, from Penn World Tables. We apply a mixture of normal densities to approximate the data density and to analyse the number of modes over time.

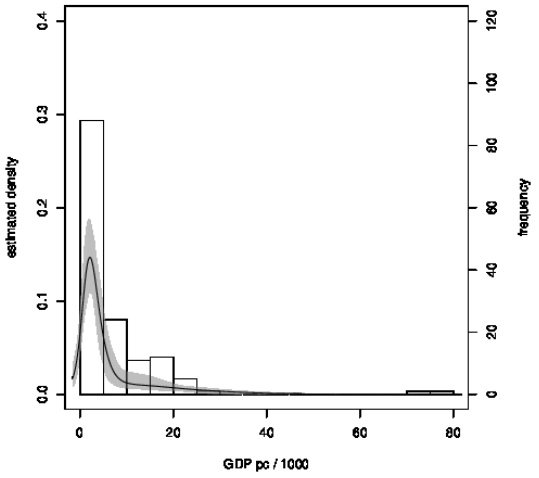
Posterior probabilities of the number of modes for these data are given in Table 6 for different time periods. These results indicate 1–4 modes for cross country GDP per capita data. Furthermore, the number of modes seem to decrease over time and this decreasing number of modes may indicate GDP convergence. The estimated number of modes is naturally linked to the countries included in the analysis.

For the GDP convergence analysis, we next consider the estimated distributions of the data over different periods. Mean estimates and 95% intervals are shown in Figure 7. Estimates at the beginning of the sample period have more high peaks and almost no probability mass in the mid-point of the data range. At the end of the sample, despite the decreasing number of modes, the probability mass at the mid-point increases while the peaks at the tails of the distributions are less pronounced. We therefore conclude that the changing number of modes do not necessarily indicate 'convergence', but instead, a 'middle income' category seems to be

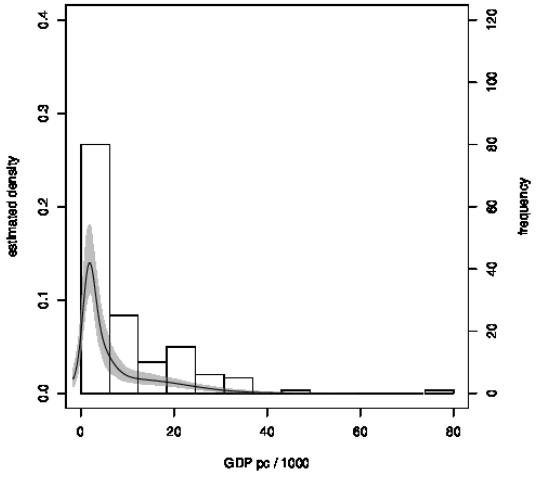
Figure 7: GDP data: Mean data densities and 95% intervals



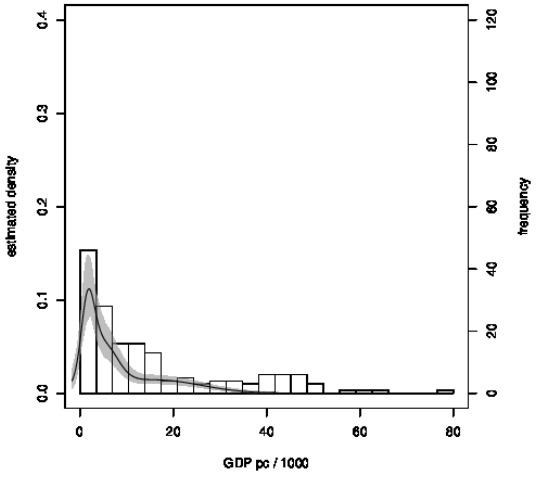
1960–1969



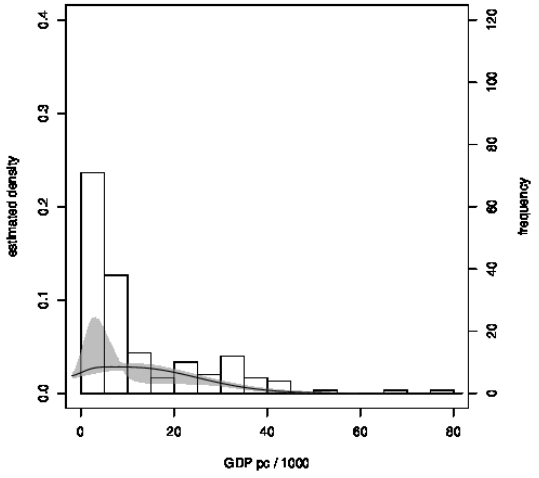
1970–1979



1980–1989



1990–1999



2000–2009

Table 6: GDP data: posterior probabilities of number of modes

	post. pr.					post. std. dev.
	1 modes	2 modes	3 modes	4 modes	5 modes	
1960–1969	0.34	0.49	0.15	0.02	0.00	0.75
1970–1979	0.72	0.27	0.01	0.00	0.00	0.48
1980–1989	0.72	0.27	0.00	0.00	0.00	0.46
1990–1999	0.44	0.46	0.10	0.00	0.00	0.65
2000–2009	0.96	0.03	0.00	0.00	0.00	0.21

Estimation results are based on 10000 draws (5000 burn-in draws).

emerging over time according to these results.

Convergence analysis for US and European countries The convergence analysis we had so far has a large number of countries. It may be argued that it is unreasonable to find convergence between all developing and developed countries which are quite heterogenous.

We apply the continuous data model with a mixture of normal densities to a subset of countries which are expected to be more homogeneous. Table 7 presents results for countries in US and Europe. At the beginning of the sample, the results indicate 2 or 3 modes for GDP growth, while the number of modes is 1 at for the period 1990-2009. Apparently, club behaviour with respect to convergence is more prevalent than overall convergence in the world.

Table 7: Posterior results for number of mixture components for the US and countries in Europe

	post. pr.					post. std. dev.	90% HPDI
	1 mode	2 modes	3 modes	4 modes	5 modes		
1960-1969	0.08	0.45	0.40	0.07	0.00	0.76	[1.00, 4.00]
1970-1979	0.77	0.22	0.01	0.00	0.00	0.44	[1.00, 2.00]
1980-1989	0.88	0.12	0.00	0.00	0.00	0.34	[1.00, 2.00]
1990-1999	1.00	0.00	0.00	0.00	0.00	0.03	[1.00, 1.00]
2000-2009	1.00	0.00	0.00	0.00	0.00	0.00	[1.00, 1.00]

5 Conclusions and future work

We presented a Bayesian approach for detecting the number of distinct modes in continuous data using mixtures of normal distributions and for discrete data we introduced a novel model with mixtures of shifted Poisson distributions. The methodology is illustrated with different

simulated data and compared to using standard tests for the number of modes in the data. Three different data sets with different properties ranging from DNA data via financial loan data unto international growth data of real Gross Domestic Product are analysed using the proposed methodology. Results show that our methodology leads to robust probabilistic conclusions about determining modes and their estimated uncertainty. The approach works better than several frequentist tests.

In future research, we plan to compare the proposed method with other tests to detect multimodality and to estimate quantiles of non-standard distributions. The method can also be extended to multivariate (e.g. panel) data. Furthermore, robustness of results with respect to the specification of the RJMCMC algorithm will be analysed. We finally note that an accompanying R package MultiMode will be available shortly.

References

Balog J, Miller D, Sanchez-Curtailles E, Carbo-Marques J, Block G, Potman M, De Knijff P, Lemmers RJ, Tapscott SJ, Van Der Maarel SM. 2012. Epigenetic regulation of the x-chromosomal macrosatellite repeat encoding for the cancer/testis gene ct47. *European Journal of Human Genetics* **20**: 185.

Baştürk N, Paap R, Van Dijk D. 2010. Financial development and convergence clubs. Technical Report 2010–52, Econometric Institute, Erasmus University Rotterdam.

Bruce H, Sachs N, Rudnicki D, Lin S, Willour V, Cowell J, Conroy J, McQuaid D, Rossi M, Gaile D, Nowak N, Holmes S, Sklar P, Ross C, Delisi L, Margolis R. 2009. Long tandem repeats as a form of genomic copy number variation: structure and length polymorphism of a chromosome 5p repeat in control and schizophrenia populations. *Psychiatric Genetics* **19**: 64.

Cappé O, Robert CP, Rydén T, Enz TR. 2002. Reversible jump mcmc converging to birth-and-death mcmc and more general continuous time samplers.

Chaudhuri P, Marron JS. 1999. Sizer for exploration of structures in curves. *Journal of the American Statistical Association* **94**: 807–823.

Dionne G, Artís M, Guillén M. 1996. Count data models for a credit scoring system. *Journal of Empirical Finance* **3**: 303–325.

Efron B, Tibshirani R. 1994. *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC.

Fischer NI, Mammen E, Marron JS. 1994. Testing for multimodality. *Computational Statistics & Data Analysis* **18**: 499–512.

Frühwirth-Schnatter S. 2006. *Finite mixture and Markov switching models: Modeling and applications to random processes*. Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.

Frühwirth-Schnatter S. 2011. Dealing with label switching under model uncertainty. In

Mengersen K, Robert CP, Titterington D (eds.) *Mixture estimation and applications*, chapter 10. Chichester: Wiley, 193–218.

Good I, Gaskins R. 1980. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association* **75**: 42–56.

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.

Grün B, Leisch F. 2009. Dealing with label switching in mixture models under genuine multimodality. *Journal of Multivariate Analysis* **100**: 851–861. ISSN 0047-259X.

Hall P, Ooi H. 2004. Attributing a probability to the shape of a probability density. *Annals of Statistics* **32**: 2098–2123.

Hartigan JA, Hartigan PM. 1985. The DIP test of unimodality. *The Annals of Statistics* **13**: 70–84. ISSN 00905364.

Hartigan PM. 1985. Computation of the DIP statistic to test for unimodality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **34**: 320–325. ISSN 00359254.

Karlis D, Xekalaki E. 2001. Robust inference for finite Poisson mixtures. *Journal of Statistical Planning and Inference* **93**: 93–115.

Malsiner Walli G, Frühwirth-Schnatter S, Grün B. 2014. Model-based clustering based on sparse finite gaussian mixtures. *Statistics and Computing* : forthcoming.

McLachlan G, Peel D. 2004. *Finite mixture models*. John Wiley & Sons.

Minnotte MC. 1997. Nonparametric testing of the existence of modes. *The Annals of Statistics* **25**: 1646–1660.

Müller DW, Sawitzki G. 1991. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association* **86**: 738–746.

Nobile A. 2004. On the posterior distribution of the number of components in a finite mixture. *Annals of Statistics* **32**: 2044–2073.

Paap R, Van Dijk HK. 1998. Distribution and mobility of wealth of nations. *European Economic Review* **42**: 1269–1293.

Quah DT. 1996. Empirics for economic growth and convergence. *European Economic Review* **40**: 1353–1375.

Richardson S, Green PJ. 1997. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)* **59**: pp. 731–792. ISSN 00359246.

Rousseau J, Mengersen K. 2011. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**: 689–710.

Schaap M, Lemmers RJLF, Maassen R, van der Vliet PJ, Hoogerheide LF, Van Dijk HK, Baştürk N, de Knijff P, an der Maarel SM. 2013. Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. *BMC Genomics* **14**: 143.

Silverman BW. 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)* **41**: 97–99.

Stephens M. 2000. Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of Statistics* : 40–74.

Umashanger T, Sriram T. 2009. L2E estimation of mixture complexity for count data. *Computational Statistics & Data Analysis* **53**: 4243–4254. ISSN 0167-9473.

Woo MJ, Sriram T. 2007. Robust estimation of mixture complexity for count data. *Computational Statistics & Data Analysis* **51**: 4379 – 4392. ISSN 0167-9473.

Supplementary Material for Bayesian Estimation of Multimodal Density Features applied to DNA and Economic Data

by

Nalan Basturk, Lennart Hoogerheide, and Herman K. van Dijk

APPENDIX A RJMCMC algorithm for a mixture of shifted Poisson distributions

In this section we introduce the RJMCMC algorithm for posterior sampling of the parameters in the model defined as a mixture of shifted Poisson distributions in (1) under the priors in (3)–(5). Given posterior draws of model parameters, inference on the number and location of the modes are obtained using the algorithm in Appendix B.

Given the priors in (3)–(5), Gibbs sampling steps are straightforward. For $j = 1, \ldots, J$, under the condition that $y_i \geq \kappa_j \ \forall i, j$ with $z_{ij} = 1$

$$p\left(\kappa_j | y, z, \theta_{-\kappa_j}\right) \propto \frac{\lambda_j^{\sum_{i|z_{ij}=1} y_i - n_j \kappa_j}}{\prod_{i|z_{ij}=1} (y_i - \kappa_j)!} \tag{A.1}$$

$$p\left(\lambda_j | y, z, \theta_{-\lambda_j}\right) \propto \text{Gamma}_{[\lambda_{\min}, \lambda_{\max}]}\left(\frac{1}{n_j}, 1 + \sum_{i|z_{ij}=1} (y_i - \kappa_j)\right) \tag{A.2}$$

$$p\left(\pi | y, z, \theta_{-\pi}\right) \propto \text{Dirichlet}\left(n_1 - 1, \ldots, n_J - 1\right), \tag{A.3}$$

where $n_j = \sum_{i=1}^{n} z_{ij}$ is the number of observations in component j and κ_j is an integer in $[\max\{\kappa_{\min}, \min_{i|z_{ij}=1}(y_i)\}, \kappa_{\max}]$.

For unknown J, we propose the following RJMCMC procedure: The combine rule to move from clusters j_1 and j_2 to j is defined as:

$$\begin{aligned} \pi_j &= \pi_{j_1} + \pi_{j_2}, \quad \lambda_j = \tfrac{\lambda_{j_1}\pi_{j_1} + \lambda_{j_2}\pi_{j_2}}{\pi_{j_1} + \pi_{j_2}} \\ \kappa_j &= \min(\kappa_{j_1}, \kappa_{j_2}) \\ z_{ij} &= z_{ij_1} + z_{ij_2}, \forall i \text{ with } z_{ij_1} = 1 \text{ or } z_{ij_2} = 1. \end{aligned} \tag{A.4}$$

That is, the two clusters to be combined are chosen based on the proximity of the means,

$\lambda_j + \kappa_j$.

The split move introduces three random variables $u_1 \sim \text{Beta}(2,2)$, $u_2 \sim \text{Beta}(2,2)$, $u_3 \sim \text{Pois}(2)$ such that the move from cluster j to clusters j_1 and j_2 is:

$$\begin{aligned}
\pi_{j_1} &= \pi_j u_1; \quad \pi_{j_2} = \pi_j(1-u_1); \\
\lambda_{j_1} &= \lambda_j u_2; \quad \lambda_{j_2} = \lambda_j \tfrac{1-u_1u_2}{1-u_1}; \\
\kappa_{j_1} &= \kappa_j; \quad \kappa_{j_2} = \kappa_j + u_3 \\
z_{ij_1} &= \begin{cases} 1 & \text{with probability } \pi_{j_1}/(\pi_{j_1}+\pi_{j_2}) \\ 0 & \text{with probability } \pi_{j_2}/(\pi_{j_1}+\pi_{j_2}) \end{cases} \quad \forall i \text{ with } z_{ij}=1 \\
z_{ij_2} &= 1 - z_{ij_1}, \forall i \text{ with } z_{ij} = 1
\end{aligned} \tag{A.5}$$

where the determinant of the Jacobian of this parameter transformation is: $|J| = \pi_j\lambda_j/(1-u_1)$.

APPENDIX B Posterior inference for the number and location of modes

We treat the number of modes as an unknown parameter and for each draw of the number of clusters J, we calculate the number and the location of modes. This is applied for both the mixture of normal and shifted Poisson distributions.

We start as follows. Each draw, $m = 1, \ldots, M$ leads to a value of the posterior density:

$$p(\tilde{y}|\theta^{(m)}) = \sum_{j=1}^{J} p\left(y|\theta^{(m)}\right) \pi_j^{(m)} \tag{B.6}$$

where $\theta^{(m)}$ is the set of model parameters for the normal or Poisson distribution.

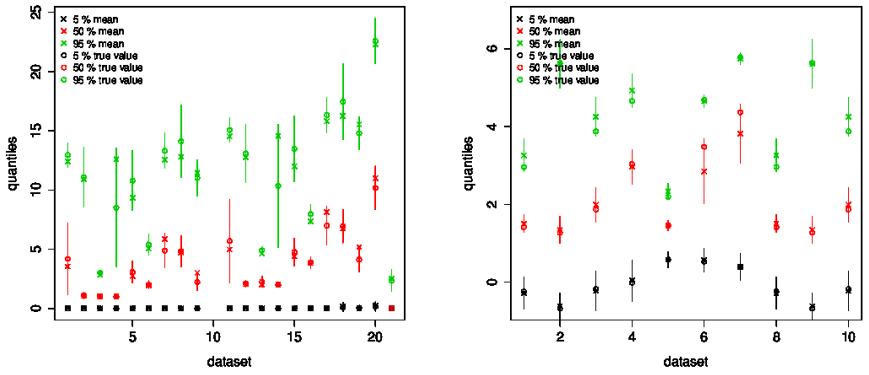
For count data, we calculate the posterior probability of being a mode for integers $y = \{\tilde{y}_1, \ldots, \tilde{y}_L\}$ on the range $[\min(y), \max(y)]$ where the modes $\hat{y}_{1^{(m)}}, \ldots, \hat{y}_{\hat{J}^{(m)}}$ satisfy $p(\tilde{y}_{j^{(m)}}) > p(\tilde{y}_{j^{(m)}} - 1)$ and $p(\tilde{y}_{j^{(m)}}) < p(\tilde{y}_{t^\star})$ where $t^\star = \min_{t;t>j^{(m)}}(p(\tilde{y}_{j^{(m)}}) \neq p(\tilde{y}_t))$, $j = 1, \ldots, \hat{J}$.

Similarly, for continuous data, the modes are calculated using a grid of 1000 points within the data range since usually the number of modes of a mixture of normal distributions with $J > 2$ components can not be analytically computed.

APPENDIX C Estimation of quantiles using simulated data

We report results on the estimation of quantiles of the posterior distributions of simulated data in Table 3. Figure C1 summaries these quantile estimates and true quantiles. All 90% intervals for quantile estimates include the true quantile value, hence data quantiles from mixture of Poisson distributions are accurately estimated from the density estimates. This is important when one uses such results for a risk analysis.

Figure C1: Quantile estimates for simulated discrete data



unknown J known J

Averages from 20 simulation replications reported