2017

Approximate Bayesian computation in nonparametric Bayesian models

Erina Paul

ABSTRACT

APPROXIMATE BAYESIAN COMPUTATION IN NONPARAMETRIC BAYESIAN MODELS

Erina Paul, Ph.D.
Department of Mathematical Sciences
Northern Illinois University, 2017
Sanjib Basu, Director

Many scientific problems require statistical inference in complex models. Bayesian nonparametric models provide a flexible modeling and inference framework for such problems. There is a substantial literature on Bayesian computational methods for nonparametric models, however, in non-conjugate complex models, they can be difficult or computationally expensive to implement. Approximate Bayesian Computation (ABC) provides a computational framework for inference in difficult and intractable Bayesian models. The idea behind ABC is to provide an approximate posterior inference without evaluating the likelihood function based on samples drawn from the sampling distribution. We develop a methodology for statistical inference in nonparametric Bayesian models based on ABC. We utilize the conditionally independent model structure to address the difficult problem of summary statistic choice in ABC. The developed method is generalized to complex nonlinear Bayesian nonparametric models, including generalized linear mixed models and survival models for recurrent data. We further generalize this approach to Bayesian nonparametric models involving the Pitman-Yor process. The approach is further extended to Bayesian nonparametric models involving the stable distributions which are often intractable due to lack

of closed-form expressions. Throughout this dissertation, we illustrate the proposed methods in simulated and real datasets and we compare their performances with preexisting methods.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

MAY 2017

APPROXIMATE BAYESIAN COMPUTATION IN NONPARAMETRIC BAYESIAN MODELS

BY

ERINA PAUL

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE
DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICAL SCIENCES

Dissertation Director:
Sanjib Basu

ACKNOWLEDGEMENTS

First and foremost I would like to express my immense gratitude to my advisor, Dr. Sanjib Basu, for giving me the opportunity to pursue this topic and enlightened me with his valuable guidance, motivation, knowledge, timely suggestions and for being extremely patient with my numerous doubts which helped me to complete my research.

I would like to thank my committee members, Dr. Alan Polansky for being such an excellent teacher, Dr. Michelle Xia for being so generous, supportive and an excellent guide in consulting service, Dr. Duchwan Ryu for his unconditional help, and Dr. Ilya Krishtal for being in my committee.

I am also grateful to Dr. Hamid Bellout, Dr. Zhuan Ye, and Dr. Jeffry Thunder for their help in my PhD requirements. I would also like to thank Shelly and Shannon for their help.

I am feeling obliged in taking the opportunity to express my heartfelt gratitude to all my teachers, especially, Tathagata Banerjee, Gourangadeb Chatterjee, Sourav De, Jugal Kanti Roy; to my friends, Amanda Working, Amartya Chakrabarty, Arnab Maity, Doyel Ghosh, Paramhansa Pramanik, Riddhisha Mukherjee, Saptarshi Chatterjee, Shreya Chakrabarty, and Yiqing Wang for their effective help and moral support.

I am grateful to my parents-in-laws, Nirmalya Kumar Mandal and Kabita Mandal for their support.

I can not thank enough my parents, Pradip Kumar Paul and Malabika Paul for giving me the support and the amazing chances over the years. Finally, I thank my sister, Madorina Paul and my husband, Anjan Mandal for their unconditional love, moral support, and encouragement throughout this entire journey.

DEDICATION

Dedicated to my parents, Pradip Kumar Paul and Malabika Paul.

Contents

List of Tables

List of Figures

Figure Page

CHAPTER 1

INTRODUCTION

Probability approaches are classified into two groups, objective and subjective interpretation. The objective interpretation leads to classical inference whereas the subjective one focuses on Bayesian inference which is based on the personal belief. According to Bayesian perspective, the unknown parameters, θ and the n-dimensional data, $y = (y_1, y_2, \ldots, y_n)$ have a probability distribution. The distribution of θ comes from the model that arises from the past experiences in handling similar data. So in a Bayesian setup, our purpose is to draw inference on θ from the data at hand. Then we are interested in exploring what are the possible values of θ and the extent of uncertainty associated with θ. Let $L(\theta|y)$ be the likelihood function and $\pi(\theta)$ be the prior probability distribution of θ. The inference on θ depends on the calculation of the posterior distribution $\pi(\theta|y) = \frac{\pi(\theta)L(\theta|y)}{p(y)}$, where $p(y)$ is the marginal distribution of y and can be defined as

$$p(y) = \int \pi(\theta)L(\theta|y)d\theta. \tag{1.1}$$

The computation of the marginal distribution (1.1) and the estimators from the posterior distribution may be difficult except for conjugate structure between the likelihood and the prior. It may also happen that the posterior distribution is known but the dimension of θ and y are large. In this situation, the computation is expensive and time-consuming. There are various methods (see Robert (2007), for example) that can expedite the computation of the posterior mean of $h(\theta)$,

$$E(h(\theta)|y) = \int h(\theta)\pi(\theta|y)d\theta = \int m(\theta|y)d\theta, \tag{1.2}$$

where $m(\theta|y) = h(\theta)\pi(\theta|y)$.

1.1 Bayesian Computation

Here, we discuss the various methods used in the context of Bayesian computation. The main purpose of the following methods is to approximate the posterior distribution when it is computationally difficult to handle.

- **Numerical Integration.** This method allows approximating the integral when no closed form expression is available. One way to approximate $m(\theta|y)$ in (1.2) is the Simpson's method (Stigler (1986)). Another approach is the polynomial quadrature (Naylor & Smith (1982)) which approximates the posterior means by weighted sums of orthogonal polynomials and the posterior may be close to normal distribution. There are various methods which are based on orthogonal bases. If the dimension is low, this method can produce accurate approximations to the integral. However, increase in the dimension of θ may affect the accuracy of the numerical integration approach.

- **Monte Carlo Methods.** The idea behind the Monte Carlo method (Metropolis & Ulam (1949), Von Neumann (1951)) is to draw samples from the posterior distribution $\pi(\theta|y)$ and then to estimate the integral using the samples, that is, if $\theta^{(1)}, \ldots, \theta^{(B)}$ are independent and identically distributed samples from $\pi(\theta|y)$,

$$\frac{1}{B}\sum_{b=1}^{B} h(\theta^{(b)})$$

converges almost surely to (1.2) as $B \to \infty$ by the Strong Law of Large Numbers under the appropriate regularity conditions. Another representation of Monte Carlo method through the importance function can be found in Robert & Casella (2004). Sometimes it is difficult

to draw samples from $\pi(\theta|y)$ and it can be approximated by $g(\theta)$, where $g(\cdot)$ is a density from which we can easily generate θ, that is, if $\theta^{(1)},\ldots,\theta^{(B)}$ are independent and identically distributed samples from $g(\cdot)$, under appropriate regularity conditions of the Strong Law of Large Numbers,

$$\frac{\sum_{b=1}^{B} h(\theta^{(b)})w(\theta^{(b)})}{\sum_{b=1}^{B} w(\theta^{(b)})},$$

where $w(\theta^{(b)}) = \frac{\pi(\theta^{(b)})L(\theta^{(b)}|y)}{g(\theta^{(b)})}$, converges almost surely to (1.2) as $B \to \infty$. The choice of an efficient importance function $g(\cdot)$ can be difficult in complex problems.

- **Laplace Approximation.** An alternative to Monte Carlo method is the Laplace approximation, see Tierney & Kadane (1986). Let $T(\theta) = \log\pi(\theta) + \log L(\theta|\mathbf{y})$ and $T^*(\theta) = \log h(\theta) + T(\theta)$. Here, we can approximate the posterior mean of $h(\theta)$ by expanding $T(\theta)$ and $T^*(\theta)$ in the Taylor series about $\hat{\theta}$ and $\hat{\theta}^*$, the posterior mode of $T(\theta)$ and $T^*(\theta)$, respectively. Hence, the approximation of $T(\theta)$ is

$$T(\theta) \approx T(\hat{\theta}) - \frac{(\theta-\hat{\theta})^2}{2\sigma^2},$$

where $\sigma^2 = -[T''(\theta)]^{-1}$. Similarly, we can get the approximation of $T^*(\theta)$ and the posterior mean of $h(\theta)$ in (1.2) is approximated by the normal density as

$$E(h(\theta)|y) = \frac{\int \exp(T^*(\theta))d\theta}{\int \exp(T(\theta))d\theta} \approx \frac{\sigma^* exp(T^*(\hat{\theta}^*))}{\sigma exp(T(\hat{\theta}))},$$

where $\sigma^{*2} = -[T^{*\prime\prime}(\hat{\theta}^*)]^{-1}$. Hence, the idea of this method is to (1) locate the modes for the integrands of $E(h(\theta)|y)$, (2) find out the second derivative at these new modes, and (3) approximate these integrals by a second application of Laplace's method. The extension of this method to saddle point approximation is discussed in Kass & Raftery (1995).

- **Markov Chain Sampling.** In Bayesian set up, Markov chain sampling is one of the popular methods for posterior calculation. The concept of Markov chain with the Monte Carlo integration resolves difficult problems in Bayesian context. The idea behind this technique is to construct a Markov Chain which converges to the required probability distribution. Hence, this method allows to draw posterior inference on functions of θ based on the generated samples along the Markov Chain. In this setting, the distribution of the posterior does not need to be known exactly. It is good enough to know the target distribution up to certain proportionality constants. Metropolis-Hastings is one of the popular procedures in the Markov Chain setup. The idea of the Metropolis-Hastings algorithm was first proposed by Metropolis et al. (1953) and was later generalized by Hastings (1970). The Gibbs sampler is a widely used method, especially, with the Metropolis-Hastings algorithm. This was first used in Geman & Geman (1984) and was popularized in the statistics literature by the seminal work of Gelfand & Smith (1990). A generalization of the Gibbs sampler, denoted as, slice sampler, is discussed in Wakefield et al. (1991), Besag & Green (1993), Damien et al. (1999). To compute or approximate complex models, Markov chain sampling methods perform well in Bayesian computation. But it can be difficult in many stochastic models where the likelihood is unknown or there are convergence issues.

1.2 Motivation

In the context of complex models, the likelihood function $L(\theta|y)$ can be analytically unavailable or computationally intractable. In those situations, we can not apply the previous methods. Approximate Bayesian Computation (ABC) provides a computational frame for such models. ABC methods generate samples from a distribution which approximates the posterior distribution of interest. The motivation of this method comes from the levels of intractability in the model. The lev-

els of intractability in Bayesian paradigm (Panek (2015)) https://approximatebayesiancomputational.wordpress.com/) arise in the following situations.

- **Partial intractability:** The likelihood has the form

$$L(\theta|y) = \frac{f(y|\theta)}{C},$$

  where $f(\cdot)$ is the known part of the probability density (or, mass) function and C, the unknown part of the likelihood, is independent of θ. Since, $L(\theta|y)$ is proportional to $f(y|\theta)$, the posterior inference can be handled by MCMC.

- **Full intractability:** If the unknown part of the likelihood depends on the parameter θ, we need to compute or approximate it by using high-dimensional integrals that are hard to compute. So we may not consider MCMC methods in this setup due to the parameter dependent proportionality constant. ABC methods can sample from the posterior distribution to deal with these type of situations. These methods sample from the posterior distribution when the likelihood is unknown or intractable and the calculation of the likelihood function does not require to sample from the posterior.

In Chapter 2, we review the existing ABC methods and illustrate their performance in different examples. Since the models in Bayesian nonparametric setup are complex, ABC method can be proposed in this situation. So, we describe the concept of nonparametric Bayesian models, for example, Dirichlet process and Pitman-Yor process in this chapter. Chapter 3 proposes the Bayesian nonparametric (BNP) models using ABC. Chapter 4 considers the BNP binary generalize linear mixed models and various BNP survival models including recurrent data models are explored in Chapter 5. Chapter 6 deals with the Pitman-Yor process while Chapter 7 focuses on intractable likelihoods, for example, stable distributions using nonparametric Bayesian ABC method. Finally, Chapter 8 provides ideas for future work and conclusion in this general theme.

CHAPTER 2

BACKGROUND

2.1 Introduction

In this chapter, we review approximate Bayesian computation and extend the idea to non-parametric Bayesian inference. The discussion mainly focuses on different methods of ABC including the choices of inputs and the Dirichlet Process (DP) models with a generalization of DP, namely, Pitman-Yor process (PYP). The adaptation of ABC methods in the Dirichlet process mixture (DPM) and Pitman-Yor mixture (PYM) will be discussed in subsequent chapters.

2.2 Approximate Bayesian Computation

Approximate Bayesian Computation method is used to compute the posterior distribution when the likelihood function is intractable or unavailable due to the inaccessibility of the closed form of θ or too expensive to compute. If the likelihood is unavailable, it is assumed that there exists a simulator that returns samples which can be drawn from the sampling distribution. Hence, the underlying idea behind ABC is to provide an approximate posterior distribution without evaluating the likelihood function.

The first ABC-related idea was mentioned by Rubin et al. (1984) in the context of sampling from the posterior distribution. An ABC method was then proposed in population genetics by Tavaré et al. (1997) to discuss the posterior inference. Followed by Pritchard et al. (1999) who produced a generalization of this method, the term Approximate Bayesian Computation was intro-

duced in Beaumont et al. (2002). In the past fifteen years, ABC has been successfully applied in different branches of bioscience and human science, for example, genetics (Beaumont et al. (2002), Tanaka et al. (2006), Lopes & Boessenkool (2010), Lombaert et al. (2011), Estoup et al. (2012), Silk et al. (2013), Foll et al. (2008)), HIV contact tracing (Blum & Tran (2010)), protein networks evolution (Ratmann et al. (2007), Ratmann et al. (2009)), archeology (Wilkinson & Tavaré (2009)), ecology (Jabot & Chave (2009)), molecular biology (Joyce & Marjoram (2008)), and coalescent models (Tavaré et al. (1997)). This method is also applied in other fields, such as operational risk (G. Peters & Sisson (2006)), and engineering (Nevat et al. (2008)).

2.2.1 ABC and Its Extensions

There are different versions of ABC methods that are used to generate observations from the posterior distribution. In this section, we review ABC and some of its extensions and illustrate their performances. We also discuss the choices of the inputs in ABC.

2.2.1.1 Simple ABC

According to Tavaré et al. (1997), the simple method of ABC is based on the rejection method. In this approach, the parameter of interest θ is generated from the prior distribution, $\pi(\theta)$. Then the acceptance of θ is based on the corresponding simulated values being identical with the observed values, y, that is, we have to generate an observation from the sampling distribution given the generated parameter value and if the generated observation is same as the observed one, we accept the generated θ at that step and repeat the process. Hence, for each iteration, the method proceeds as follows

1.i. Generate θ^* from the prior $\pi(\theta)$.

1.ii. Generate z^* from the density $f(z|\theta^*)$.

1.iii. Accept θ^* if $z^* = y$; return to 1.i.

Here, the accepted values of θ are from $\pi(\theta|(z^* - y) = 0)$, i.e., this method gives samples from the exact posterior distribution, $\pi(\theta|y)$ (Marin et al. (2012)) since

$$f(\theta) \propto \sum_{z^* \in \mathcal{D}} \pi(\theta) f(z^*|\theta) \mathbb{I}_y(z^*) = \pi(\theta) f(y|\theta) \propto \pi(\theta|y),$$

where $\mathcal{D}$ is a finite or countable set of y and $\mathbb{I}_y$ denotes the indicator function on y. However, the acceptance rate of observing $z^* = y$ is zero for continuous case and in such situation, the simple ABC method is extremely inefficient. This motivates the following method.

2.2.1.2 ABC-Rejection

The ABC-Rejection method (Pritchard et al. (1999)) handles the data with a set of summary statistics ($\mathcal{S}$), distance metric (ρ), and tolerance level (ε). Instead of matching the exact observations from the observed and simulated data, this method allows summarizing the data using statistics to reduce the dimensionality. Here, the comparison can be done by the lower-dimensional summaries of the data. Additionally, the simple rejection method does not account the continuous distribution, because the probability of exact match between y and z^* is zero. To relax the idea from the simple ABC method, we consider a metric, $\rho > 0$ on the space of summary statistics $\mathcal{S}(\cdot)$ to measure how close the simulated data, z^* is from the observed data, y. Then we might accept the generated parameter value θ with a predefined tolerance level, ε. Thus, each iteration of the previous ABC method can be modified as follows

2.i. Generate θ^* from the prior $\pi(\theta)$.

2.ii. Generate z^* from the density $f(z|\theta^*)$.

2.iii. Calculate the distance, $\rho(\mathcal{S}(y), \mathcal{S}(z^*))$.

2.iv. Accept θ^* if $\rho \leq \varepsilon$, where $\varepsilon > 0$, a tolerance level; return to 2.(i).

This method does not sample from the exact posterior distribution, $\pi(\theta|y)$ but samples data from the marginal of z^*, $\pi^*(\theta|y) = \int \pi^*(\theta, z^*|y)dz^*$ (Marin et al. (2012)), where the integrand is defined as follows

$$\pi^*(\theta, z^*|y) = \frac{\pi(\theta)f(z^*|\theta)\mathbb{I}_{A_{y\varepsilon}(z^*)}}{\int_{A_{y\varepsilon}} \pi(\theta)f(z^*|\theta)dz^*d\theta},$$

with $A_{y\varepsilon} = \{z^* \in \mathcal{D}|\rho(\mathcal{S}(y), \mathcal{S}(z^*)) \leq \varepsilon\}$, $\mathcal{D}$ is a finite or countable set of z^*, and $\mathbb{I}_A$ denotes the indicator function on A.

A particular case of the ABC-Rejection method is equivalent to the simple ABC when the summary statistic is sufficient and the tolerance level is zero.

2.2.1.3 ABC-MCMC

For the high dimensional data, the samples from the prior are rejected with high probability in the ABC rejection method because of the proposals from the prior is in the regions of low posterior probability. A feasible way to overcome this complexity is to combine ABC with the Markov chain Monte Carlo methodology (Marjoram et al. (2003), Turner & Van Zandt (2012), Marin et al. (2012)).

MCMC has been a widely used methods for sampling from the complex models. The Metropolis-Hastings algorithm is the most popular in MCMC setup. It can be used to sample from the posterior distribution without considering $p(y)$ in (1.1). First, we choose an initial value θ, denoted by $\theta^{(0)}$. We then sample a candidate value, θ^* from the proposal distribution, $q(\theta|\theta^{(0)})$ which is the probability of θ^* given the previous value, $\theta^{(0)}$. Now, the acceptance probability of θ depends on the ratio of $\frac{\pi(\theta^*|y)}{q(\theta^*|\theta^{(0)})}$ for the candidate state with the corresponding ratio for the previous state,

$\frac{\pi(\theta^{(0)}|y)}{q(\theta^{(0)}|\theta^*)}$. This form of the acceptance probability ensures that the stationary distribution is the target posterior distribution. Hence the acceptance probability, α is given by

$$\alpha = \min\left\{1, \frac{\pi(\theta^*|y)q(\theta^{(0)}|\theta^*)}{\pi(\theta^{(0)}|y)q(\theta^*|\theta^{(0)})}\right\} = \min\left\{1, \frac{\pi(\theta^*)L(\theta^*|y)q(\theta^{(0)}|\theta^*)}{\pi(\theta^{(0)})L(\theta^{(0)}|y)q(\theta^*|\theta^{(0)})}\right\}.$$

Since we are only interested in the posterior distributions, the marginal distribution cancels out in the calculation of α. Now, we set $\theta^{(1)} = \theta^*$ with probability α, otherwise, $\theta^{(1)} = \theta^{(0)}$. Then, we continue this process until we get a chain of θ, $\{\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(B)}\}$ to estimate the posterior distribution.

MCMC algorithm can be incorporated in ABC method for the target distribution, $\pi^*(\theta, z^*|y)$. The ABC method within the MCMC framework (Marjoram et al. (2003)) proceeds as follows:

3.i. Initialize $\theta^{(0)}, b = 0$.

3.ii. Generate a candidate value $\theta^* \sim q(\theta|\theta^{(b-1)})$, where $q(\cdot)$ is some proposal density.

3.iii. Generate z^* from the density $f(z|\theta^*)$.

3.iv. Set $\theta^{(b)} = \theta^*$ with probability

$$\alpha(\theta^*, \theta^{(b-1)}) = \min\left\{1, \frac{\pi(\theta^*)q(\theta^{(b-1)}|\theta^*)}{\pi(\theta^{(b-1)})q(\theta^*|\theta^{(b-1)})}\mathbb{I}(\rho(\mathcal{S}(z^*), \mathcal{S}(y)) \leq \varepsilon)\right\}$$

otherwise set $\theta^{(b)} = \theta^{(b-1)}$.

3.iv. Repeat until $b \leq B$.

Here, $\pi^*(\theta, z^*|y)$ is a stationary distribution (Marin et al. (2012)), because

$$\begin{aligned} & \frac{\pi^*(\theta^*,z^*|y)}{\pi^*(\theta^{(b-1)},z^*|y)} \times \frac{q(\theta^{(b-1)}|\theta^*)f(z^*|\theta^{(b-1)})}{q(\theta^*|\theta^{(b-1)})f(z^*|\theta^*)} \\ = & \frac{\pi(\theta^*)f(z^*|\theta^*)\mathbb{I}(\rho(\mathcal{S}(z^*),\mathcal{S}(y))\leq\varepsilon)}{\pi(\theta^{(b-1)})f(z^*|\theta^{(b-1)})\mathbb{I}(\rho(\mathcal{S}(z^*),\mathcal{S}(y))\leq\varepsilon)} \times \frac{q(\theta^{(b-1)}|\theta^*)f(z^*|\theta^{(b-1)})}{q(\theta^*|\theta^{(b-1)})f(z^*|\theta^*)} \\ = & \frac{\pi(\theta^*)q(\theta^{(b-1)}|\theta^*)}{\pi(\theta^{(b-1)})q(\theta^*|\theta^{(b-1)})}\mathbb{I}(\rho(\mathcal{S}(z^*),\mathcal{S}(y)) \leq \varepsilon), \end{aligned}$$

where $\theta^{(b)}$ is the generated value of θ at the b^{th} iteration. This method also samples from the approximate posterior distribution and depends on the choices of the proposal density, the initial value of θ, the tolerance level, the summary statistics, and the distance metric.

Other ABC methods include the ABC-partial rejection control (ABC-PRC, Sisson et al. (2007)) has been developed based on sequential Monte Carlo, ABC-PMC (Population Monte Carlo) which is discussed in Beaumont et al. (2009). All the methods in ABC are mainly structured based on the inputs used in the models. In the next section, we discuss the brief idea of the inputs for ABC.

2.2.2 ABC Inputs

The inference from ABC mainly depends on choice of i) the tolerance level (ε), ii) the distance metric (ρ), and iii) the summary statistics ($\mathcal{S}$).

Tolerance Level

The choice of tolerance level ε may affect the acceptance procedure. Smaller values of the tolerance level imply lower acceptance rate of the parameter whereas the higher values of the tolerance are associated with the acceptance of all the parameters from the prior distribution. Also, the tolerance level depends on the choice of the prior as well as different choices of summary

statistics. In particular, if the dimension of the summary statistics is increased, the choice of tolerance level has to be more specific. So the zero tolerance level ensures the required outcome but makes the computations impossible. However, the values larger than zero used in the literature may result in biased results.

Distance Metric

The choice of distance metric measures the closeness of the simulated values with the observed ones. The Euclidean distance is widely used metric in ABC setup. But this choice may not be the best in terms of the error to estimate the posterior. This measure is scale dependent, so changing the scale of measurement implies the change of the results. Since this scale is determined by the summary statistics, the choice of summary statistics plays an important role in choosing the metric. Thus, the choice of summary statistics targets at reducing the dimensions as well as to extract the important information about the parameters of interest. In the most situations, it is quite difficult and impossible to get a suitable set of summary statistics in ABC setup. However, for the categorical variables, we can not use the Euclidean metric. In such situation, we may consider different similarity measures based on categorical variables, for example, Jaccard coefficient.

Summary Statistics

The choice of summary statistics is a crucial step in ABC as it should reflect the reduction of dimensions along while maintaining sufficient information of the parameters of interest. Since the likelihoods are intractable in ABC approach, summary statistics are often not sufficient and hence, the choice of the statistic involves loss of information and reduction of dimensionality. Due to the curse of dimensionality, the rate of convergence of the posterior means with respect

to $\pi(\theta|\mathcal{S}(y))$ decreases as the dimension of summary statistic increases. However, except for the sufficient statistics, the choice of summary statistics depends on the data set as the information of summary statistics may vary within the parameter space. There are different approaches available and Blum et al. (2013) discussed various methods and gave a comparative study based on those methods. Here, we discuss three popular approaches, reviewed in Blum et al. (2013), to choose the best subset, denoted as $s = (s_1, \ldots, s_p)$, of summary statistics.

Regression Adjustment. To avoid the effect of the discrepancy between the observed summary statistic and the summary statistic from the generated observations, Beaumont et al. (2002) proposed two transformations:

- weighting the $\theta^{(b)}$ according to the value of $\rho(\mathcal{S}(y), \mathcal{S}(z^{(b)}))$,
- adjusting the $\theta^{(b)}$ using local linear regression.

Let the regression model be

$$\theta^{(b)} = m(\mathcal{S}(z^{(b)})) + e^{(b)},$$

where $m(\mathcal{S}(z^{(b)})) = E(\theta|\mathcal{S}(y) = \mathcal{S}(z^{(b)})) = \alpha + \beta^T \mathcal{S}(z^{(b)})$ is the mean function and $e^{(b)}$ is the b^{th} random error with mean 0 and common variance. In the neighborhood of $\mathcal{S}(y)$, Beaumont et al. (2002) proposed to approximate the conditional expectation of θ given $\mathcal{S}(y)$ by $\hat{m}(\mathcal{S}(z^{(b)}))$ where

$$\hat{m}(\mathcal{S}(z^{(b)})) = \hat{\alpha} + (\mathcal{S}(z^{(b)}) - \mathcal{S}(y))^T \hat{\beta}.$$

An estimate of the conditional expectation can be obtained by minimizing $\sum_{b=1}^{B} w^{(b)} \| m(\mathcal{S}(z^{(b)})) - \theta^{(b)} \|^2$ with $w^{(b)} = K_e(\|\mathcal{S}(z^{(b)}) - \mathcal{S}(y)\|)$, $K_e(\cdot)$ is generally taken to be Epanechnikov kernel. Hence, the adjustment for the weighted sample is as follows:

$$\theta^{(b)*} = \hat{m}(\mathcal{S}(y)) + (\theta^{(b)} - \hat{m}(\mathcal{S}(z^{(b)})), b = 1, \ldots, B.$$

Blum & François (2010) improved the local linear model by estimating the conditional mean and variance of the nonlinear heteroscedastic regression model. This model is defined as follows:

$$\theta^{(b)} = \alpha + (\mathcal{S}(z^{(b)}) - \mathcal{S}(y))\beta + \sigma(\mathcal{S}(z^{(b)}))e^{(b)}.$$

Here, an estimate of the conditional expectation can be obtained by fitting a nonlinear regression model and the variance term is estimated by using $\log(\theta^{(b)} - \hat{m}(\mathcal{S}(z^{(b)})))^2 = \log\sigma^2(\mathcal{S}(z^{(b)})) + \xi^{(b)}$, $\xi^{(b)}$ is the b^{th} random error.

Neural Network. In the neural network setup, the hidden layers help to decrease the dimension of the summary statistics. Suppose there are $H < p$ hidden units, $x_1, \ldots, x_H$ in the neural network. For $j = 1, \ldots, H, x_j$ can be defined as

$$x_j = h(\sum_{k=1}^{p} w_{(1)jk} s_k + w_{(1)j0}),$$

where $w_{(1)jk}$ are the weights in the initial layer of the neural network and $h(\cdot)$ is a non-linear function. In the second layer, x_j are combined with the regression function $m(\cdot)$ which can be defined as

$$g(\sum_{j=1}^{H} w_{(2)j} z_j + w_{(2)0}),$$

where $w_{(2)j}$ are the weights in the next layer of the neural network and $g(\cdot)$ is a link function. Blum & François (2010) fixed the number of hidden units, H to the dimension of θ and the corresponding weights can be obtained by minimizing the following criterion

$$\sum_{b=1}^{B} w^{(b)} \|m(\mathcal{S}(z^{(b)})) - \theta^{(b)}\|^2 + \lambda\|\omega\|^2,$$

where $w^{(b)} = K_e(\|\mathcal{S}(z^{(b)}) - \mathcal{S}(y)\|)$ is the weight of the sample $(\theta^{(b)*}, \mathcal{S}(z^{(b)}))$, $\lambda > 0$ is the adjusted parameter which is used to reduce the weights towards zero to get instructive summary statistics, and ω is vector of all weights in the neural network. Hence the weighted sample from the posterior distribution is obtained by the adjustment

$$\theta^{(b)*} = \hat{m}(\mathcal{S}(y)) + (\theta^{(b)} - \hat{m}(\mathcal{S}(z^{(b)}))) \frac{\hat{\sigma}(\mathcal{S}(y))}{\hat{\sigma}(\mathcal{S}(z^{(b)}))}, \mathrm{b} = 1, \ldots, B.$$

Semi-automatic Method. Fearnhead & Prangle (2012) proposed to choose summary statistics that are equal to the posterior mean. They used simulation method to estimate appropriate summary statistic. The idea of this method is to run ABC to determine a region of non-negligible posterior mass and then simulate sets of parameter values and data to estimate the summary statistics. Let $f(\cdot)$ be a vector-valued function. The simplest choice of $f(y)$ is y. It can also be vector valued. The explanatory variables in the model are defined as the transformation of the simulated data, $(f(y^1), f(y^2), \ldots, f(y^k))$, where k is the dimension of parameter vector. The responses at the b^{th} step are defined as $\theta^{(b)1}, \ldots, \theta^{(b)k}$. Then the fitted model is

$$\theta^{(b)} = E(\theta^{(b)}|y) + e^{(b)} = \beta_0^{(b)} + \beta^{(b)} f(y) + e^{(b)},$$

where $e^{(b)}$ is a zero-mean noise. Hence, the estimate of $E(\theta^{(b)}|y)$ is $\hat{\beta}_0^{(b)} + \hat{\beta}^{(b)} f(y)$. In ABC, we use the difference in summary statistics, so the constant terms cancel out and hence, the b^{th} summary statistic for ABC is $\hat{\beta}^{(b)} f(y)$. This method is known as Semi-Automatic ABC.

There are also different methods to reduce the dimension of the summary statistics, such as, sufficiency criterion (Joyce & Marjoram (2008)), partial least square regression (Wegmann et al. (2009)), entropy criterion (Nunes & Balding (2010)), AIC-BIC criteria (Blum et al. (2013)), and regularization approach using ridge regression (Blum et al. (2013)).

2.2.3 Examples

Before moving to the next section, let us consider few examples to illustrate different ABC methods discussed above. Here, our main purpose is to compare the ABC methods with the actual posterior distribution or the true mean of the parameters. So we consider one simulation study and one real data examples based on ABC-Rejection and ABC-MCMC, respectively.

Simulation Study

Let Y be a set of independent and identically distributed random variable of size n from the normal distribution with mean θ and variance σ^2. Here, we are interested in the estimation of the mean parameter μ. We assume a normal prior for θ with mean θ_0 and variance σ_0^2. In this example, $n = 10, \sigma^2 = 1, \theta_0 = 3$, and $\sigma_0^2 = 1$. In each iteration of the ABC-rejection method, θ^* is generated from the prior, Normal$(3,1)$. Then using the value of $\theta = \theta^*$, generate z^* from Normal$(\theta^*, 1)$. Here, we consider the summary statistic, $\mathcal{S}$ as the sample mean, $\bar{y} = 3.07$ and the distance metric as Euclidean distance. In the next step of ABC, we calculate the absolute distance between the mean of observed and generated values, that is, y and z^*. Finally, we check if the distance is less than a certain tolerance level. In our case, we consider the tolerance level, ε as 0.01 and we accept θ^* if the distance is less than 0.01. We run the simulation $B = 25000$ times with 5000 burn-in period to get the approximate posterior distribution. In Fig. 2.1, the actual posterior density is compared with the approximate posterior density from the ABC-Rejection. The horizontal axis represents the accepted values of θ from the simulation. Hence, the rejection method performs almost same as the actual posterior for the normal distribution with known variance.

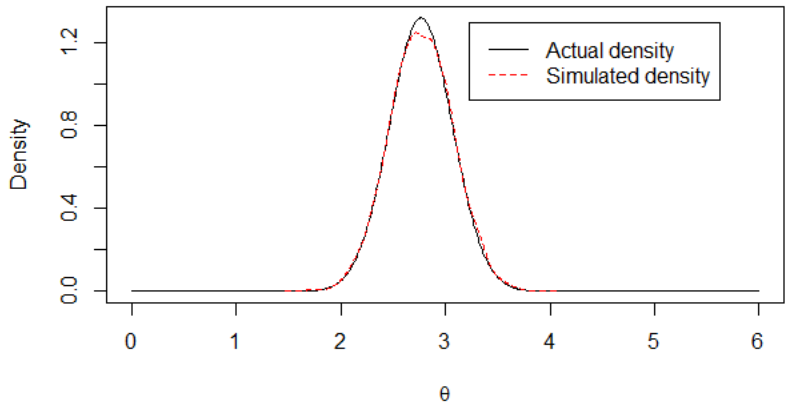


Figure 2.1: Comparison of actual posterior with ABC-Rejection method

Logistic Regression

Now, we illustrate the idea of ABC-MCMC for the logistic regression model using a real life example. We consider the nodal data (taken from `boot` package in R). The data consists of 53 patients who are diagnosed with prostate cancer and 6 predictor variables which are measured before surgery. Here, the goal is determine the relationship between nodal involvement and the predictor variables. The response variable, r, is an indicator of nodal involvement. The predictor variables are m (a column of ones), aged (patients age, less than 60 (0) and 60 or over (1)), stage (size and position of the tumor, 1 indicates a more serious), grade (seriousness of the tumor, 1 indicates a more serious case), xray (X-ray examination, 1 indicates a more serious case), and acid (level of serum acid phosphatase). So, we can use the logistic regression model in this setup. So, the model can be defined as

$$\text{logit}(p_j) = \log(\frac{p_j}{1-p_j}) = X_j^T \beta, p_j = P(y_j = 1 | X_j = x_j), j = 1, 2, \ldots, 53,$$

where X_j is the j^{th} vector of the design matrix X = (m, aged, stage, grade, xray, acid) and $\beta = (\beta_0, \beta_1, \ldots, \beta_5)$ is the corresponding regression parameter vector. Here, we are interested in estimating the regression parameter β. We consider a normal prior for β with mean μ_0 and variance Σ_0. Let $\beta^{(b)}$ denote the b^{th} iteration in the simulation. In ABC-MCMC setup, first we have to initialize $\beta^{(0)} = (0,0,0,0,0,0)$. We then generate β^* from a proposal density $q(\beta|\beta^{(b-1)})$ and generate z^* from the density $f(z|\beta^*)$. Here, we consider a normal proposal distribution with variance Σ_q. Now, we accept $\beta^{(b)} = \beta^*$ with probability,

$$\alpha(\beta^*, \beta^{(b-1)}) = \min\left\{1, \frac{\pi(\beta^*)q(\beta^{(b-1)}|\beta^*)}{\pi(\beta^{(b-1)})q(\beta^*|\beta^{(b-1)})}\mathbb{I}(\rho(\mathcal{S}(z^*), \mathcal{S}(y)) \leq \varepsilon)\right\}$$

and run until we get B vectors of β. In this example, $n = 53$, $B = 100000$, μ_0 = (-3, 0, 1.4, 0.8, 1.8, 1.7), $\Sigma_0 = diag(1,1,\ldots,1)$, $\beta^{(0)} = (0,0,0,0,0,0)$, and $\Sigma_q = 0.2\,\Sigma_0$. Here, we consider the tolerance level as 0.2 and distance metric as Euclidean distance with the summary statistic for the observed data is $y^T X$. Since we run the MCMC sampler, we consider 50000 burn-in to estimate the parameters. In Table 2.1, we compare the estimates of the posterior mean of the parameter using ABC-MCMC with the general MCMC for generalized linear model (using OpenBUGS software). Here, we can see that the estimates from the ABC-MCMC are slightly different from the actual estimate. This may happen due to the choice of the summary statistics as well as the tolerance level.

In this section, we have discussed different methods in ABC to sample from the true or approximate posterior distribution. However, ABC methods require a perfect choice of the summary statistics, the tolerance level, and the distance metric. The tolerance level and the distance metric do not cause any huge problem in the simulation, but the choice of summary statistic creates the difficulty in the ABC methods. It happens because we are not using the full information of the data due to the complex models or intractable likelihood functions and hence the sufficient statistics.

Table 2.1: Comparison of the estimates of β using ABC-MCMC and MCMC

Parameter	ABC-MCMC	MCMC
$\hat{\beta}_0$	-3.195	-3.186
$\hat{\beta}_1$	-0.225	-0.267
$\hat{\beta}_2$	1.682	1.422
$\hat{\beta}_3$	0.586	0.878
$\hat{\beta}_4$	1.531	1.800
$\hat{\beta}_5$	1.611	1.728

In case of the nonparametric Bayesian inference, we have to deal with different complex models. The next section reviews the nonparametric inference under Bayesian setup. The section is divided in to two parts, first part consists of the idea and posterior simulation method depending upon Dirichlet process and the last part is dealing with the Pitman-Yor process.

2.3 Nonparametric Bayesian Models

Bayesian nonparametric approach provides a flexible and at the same time practically feasible approach in which a prior is assigned over the space of all distributions. The Dirichlet process and a generalization, that is, Pitman-Yor process are the popular approaches in the Bayesian nonparametric models. These two stochastic processes are defined as probability measures on the space of probability measures. In next two sections, we review the concepts and properties of both the models.

2.3.1 Dirichlet Process

The Dirichlet process (DP) is a stochastic process. Draws from the DP can be explained as random because it is a distribution over probability measures to allow certain functions to be interpreted as distributions over specific probability space. According to Ferguson (1973), the DP can be defined as follows.

Definition 1 *Let $\mathcal{X}$ be a space and $\mathcal{A}$ a σ−field of subsets. Let G_0 be a distribution on $(\mathcal{X}, \mathcal{A})$ and α be a non-null finite, non-negative real number. Then G is a Dirichlet process on $(\mathcal{X}, \mathcal{A})$, denoted as $DP(\alpha, G_0)$, with base distribution G_0 and concentration parameter α if for every positive integers n and measurable partition $(A_1, \ldots, A_n)$ of $\mathcal{X}$, i.e., if $A_j \in \mathcal{A}$ for all j, $A_j \cap A_l = \emptyset$ for $j \neq l$, and $\cup_{j=1}^{n} A_j = \mathcal{X}$, then the distribution of $(G(A_1), \ldots, G(A_n))$ has a k-dimensional Dirichlet distribution with parameter $(\alpha G_0(A_1), \ldots, \alpha G_0(A_n))$.*

The DP can be represented in different form. The next section describes the Pólya urn scheme and extend the idea to DP mixture models and predictive distributions.

Pólya Urn Scheme and Dirichlet Process Mixture Models

Let $\theta_1, \ldots, \theta_n$ be a sequence of independent samples from G and θ_j's take values in $\mathcal{X}$ since G is a distribution over $\mathcal{X}$. Now, we are interested in the posterior distribution of G given observed values of $\theta_1, \ldots, \theta_n$. Let $(A_1, \ldots, A_k)$ be a finite measurable partition of $\mathcal{X}$, and let $n_r = \#\{j : \theta_j \in A_r\}$ be the number of observed values in A_r. By the conjugacy property of the Dirichlet and the multinomial distributions, we have:

$$G(A_1), \ldots, G(A_k) | \theta_1, \ldots, \theta_n \sim \mathcal{D}(\alpha G_0(A_1) + n_1, \ldots, \alpha G_0(A_k) + n_k).$$

Since the above is true for all finite measurable partitions, the posterior distribution over G must be a DP as well. The Pólya urn representation of the DP is defined as follows:

$$\begin{aligned} Y_1, \ldots, Y_n &\overset{iid}{\sim} G \\ G &\sim \mathrm{DP}(\alpha, G_0), \end{aligned}$$

where $\mathrm{DP}(\alpha, G_0)$ denotes a DP with concentration parameter α and base distribution G_0.

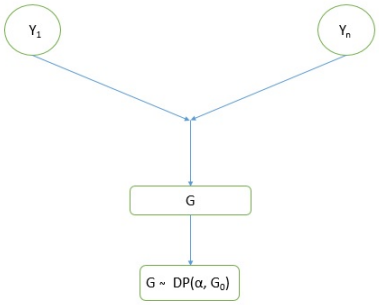


Figure 2.2: Structure of the Dirichlet process

Hence, the marginal distribution of $(Y_1, \ldots, Y_n)$ are

$$\begin{aligned} Y_1 &\sim G_0(Y_1) \\ Y_2|Y_1 &\sim \frac{\alpha}{\alpha+1} G_0(Y_2) + \frac{1}{\alpha+1} \delta_{\{Y_1\}}(Y_2) \\ Y_3|Y_1, Y_2 &\sim \frac{\alpha}{\alpha+2} G_0(Y_3) + \frac{1}{\alpha+2} \delta_{\{Y_1\}}(Y_3) + \frac{1}{\alpha+2} \delta_{\{Y_2\}}(Y_3) \\ &\vdots \\ Y_n|Y_1, \ldots, Y_{n-1} &\sim \frac{\alpha}{\alpha+n-1} G_0(Y_n) + \frac{1}{\alpha+n-1} \sum_{j=1}^{n-1} \delta_{\{Y_j\}}(Y_n). \end{aligned}$$

Under the Dirichlet process mixture (DPM), we have another layer in the model and the model can be defined as follows:

$$\begin{aligned} Y_j|\theta_j &\sim f(y_j|\theta_j) \\ \theta_j|G &\sim G \\ G &\sim \mathrm{DP}(\alpha, G_0). \end{aligned}$$

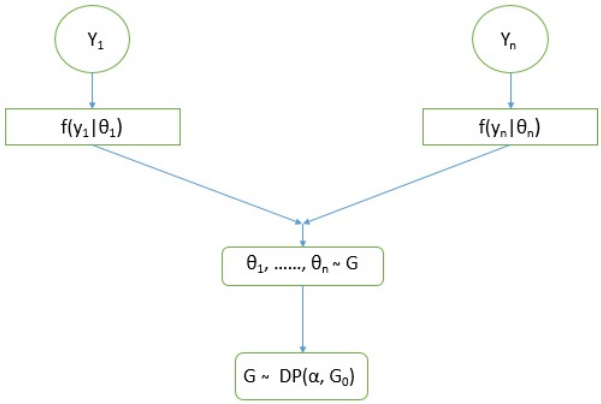


Figure 2.3: Structure of the Dirichlet process mixture

Dirichlet process mixture (DPM) models are discussed in Ferguson (1983), Escobar & West (1995), and MacEachern & Müller (1998). The model is defined on a set of conditionally independent observations, $y = (y_1, \ldots, y_n)$. These observations may be multivariate. Now, the y_j's are drawn from a mixture of distribution, $f(y_j|\theta)$ with the mixing distribution over θ being G. Here, we assume the prior for G is modeled by a DP with concentration parameter α and base distribution G_0. Since the random distribution G is discrete with probability one, these model can be treated as

countably infinite mixture distributions. It is also evident from (2.1) that the representation of θ_j in terms of the successive conditionals is as follows

$$\theta_j|\theta_1,\ldots,\theta_{j-1} \sim \frac{\alpha}{j-1+\alpha}G_0(\theta_j) + \frac{1}{j-1+\alpha}\sum_{l=1}^{j-1}\delta(\theta_l), \tag{2.1}$$

where $\delta(\theta_l)$ is the degenerate distribution at the point θ_l and the full prior conditional distribution of θ_j is as follows:

$$\theta_j|\theta_{-j} \sim \frac{\alpha}{\alpha+n-1}G_0(\theta_j) + \frac{1}{\alpha+n-1}\sum_{k\neq j}\delta_{\{\theta_k\}}(\theta_j),$$

where θ_{-j} is the vector of θ except the j^{th} element. The DP provides a conjugate family of priors over distributions that is closed under posterior updates given observations. Rewriting the posterior DP, we have:

$$G|\theta_1,\ldots,\theta_n \sim \mathrm{DP}\left(\alpha+n, \frac{\alpha}{\alpha+n}G_0 + \frac{n}{\alpha+n}\frac{\sum_j \delta_{\theta_j}}{n}\right).$$

The posterior base distribution is a weighted average between the prior base distribution G_0 and the empirical distribution $\frac{\sum_j \delta_{\theta_j}}{n}$. The weight associated with the prior base distribution is proportional to α, while the empirical distribution has weight proportional to the number of observations n. Thus we can interpret α as the strength or mass associated with the prior. Hence, the full posterior conditional distribution of $\theta_j|\theta_{-j}, y_1,\ldots,y_n$ can be written as

$$\begin{aligned}
\pi(\theta_j|\theta_{-j}, y_1,\ldots,y_n) &\propto \frac{\alpha}{\alpha+n-1}G_0(\theta_j)f(y_j|\theta_j) + \frac{1}{\alpha+n-1}\sum_{k\neq j}\delta_{\{\theta_k\}}(\theta_j)f(y_k|\theta_k) \\
&= \frac{\alpha}{\alpha+n-1}G_0(\theta_j)m(y_j)H(\theta_j|y_j) + \frac{1}{\alpha+n-1}\sum_{k\neq j}\delta_{\{\theta_k\}}(\theta_j)f(y_k|\theta_k),
\end{aligned}$$

where $m(y_j) = \int G_0(\theta_j) f(y_j|\theta_j) d\theta_j$ is the marginal distribution of Y_j and $H(\theta_j|y_j) = \frac{G_0(\theta_j) f(y_j|\theta_j)}{m(y_j)}$ is the posterior distribution of $\theta_j|y_j$.

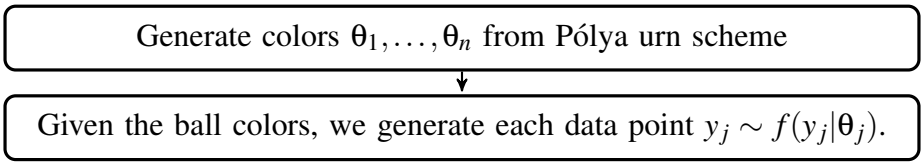
The predictive distribution of θ_{n+1} can be generated using the Pólya urn scheme. In this technique, we assume that an urn contains colored balls and we have to draw balls at random. First we draw a ball and observe its color, we return it back to the urn. Then we add another ball of the same color into the urn. A similar scheme is used by Blackwell & MacQueen (1973) to construct a DP. Suppose each value in $\mathcal{X}$ is a unique color and $\theta_{n+1} \sim G_0$, the color of a ball which is put into the urn. Also, we have an urn containing preselected balls. At first, we have to choose a color drawn from G_0, i.e. draw $\theta_1 \sim G_0$, add a ball with the same color into the urn and repeat the process. In the $(n+1)^{th}$ step, we will choose a new color with probability $\frac{\alpha}{\alpha+n}$ and add a ball with the same color into the urn, or, with probability $\frac{n}{\alpha+n}$, we draw a ball at random, then choose a new ball with the same color and return both balls into the urn. This produces a sequence of $\{\theta_j\}_{j=1}^{\infty}$ with conditional probabilities

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha+n} G_0(\theta_{n+1}) + \frac{\sum_{j=1}^{n} \delta_{(\theta_{n+1}=\theta_j)}}{\alpha+n}. \tag{2.2}$$

Since the values of draws $\{\theta_k\}$ are repeated, let $\eta_1, \ldots, \eta_m$ be the unique values among $\theta_1, \ldots, \theta_n$, and n_k be the number of repeats of η_k. Then the predictive distribution in Eq. (2.2) can be written as:

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha+n} G_0(\theta_{n+1}) + \frac{\sum_{k=1}^{m} n_k \delta_{\eta_k}}{\alpha+n}.$$

Hence, the Pólya urn model proceeds as follows:



Under the DPM models, we can also calculate the predictive distribution of a new observation Y_{n+1}. The corresponding model can be defined as:

$$\begin{aligned} Y_{n+1}|\theta_{n+1} &\sim f(y_{n+1}|\theta_{n+1}) \\ \theta_{n+1}|G &\sim G \\ G &\sim \mathrm{DP}(\alpha, G_0). \end{aligned}$$

Hence, the predictive distribution of Y_{n+1} is

$$\begin{aligned} f(Y_{n+1}) &\propto \frac{\alpha}{\alpha+n}\int G_0(\theta_{n+1})f(y_{n+1}|\theta_{n+1})d\theta_{n+1} + \frac{1}{\alpha+n}\sum_{j=1}^{n} f(y_{n+1}|\theta_j) \\ &= \frac{\alpha}{\alpha+n}m(y_{n+1})H(\theta_{n+1}|y_{n+1}) + \frac{1}{\alpha+n}\sum_{j=1}^{n} f(y_{n+1}|\theta_j). \end{aligned}$$

Integration over θ_{n+1} can be done analytically if $f(\cdot)$ and $G_0(\cdot)$ are conjugate.

An alternative way to represent the DPM model is based on 'latent class'. If the j^{th} and j'^{th} observations belong to the same cluster, $\theta_j = \theta_{j'}$. Here, a set of random variables c_j corresponds to unique component parameters K_c with the data points $\theta_j = K_{c_j}$. The model in this setup (Neal (2000)) can be defined as follows

$$\begin{aligned} Y_j|c_j, K &\sim f(y_j|K_{c_j}) \\ c_j|\mathbf{p} &\sim \mathrm{Discrete}(p_1, \ldots, p_s) \\ K_{c_j} &\sim G_0 \\ \mathbf{p} &\sim \mathcal{D}(\alpha/s, \ldots, \alpha/s), \end{aligned} \tag{2.3}$$

where $\mathbf{p} = (p_1, \ldots, p_s)$, are given a symmetric Dirichlet prior with concentration parameter α/s which tends to zero as $s \to \infty$. This model is used when a cluster with high probability are associated with the parameter θ.

An important generalization of the DPM model, mixtures of DP, arises when the base measure of the DP prior includes unknown hyperparameters η. In this setup, the model has an unknown parameter α and base measure G_η. It can be defined as follows:

$$\begin{aligned} Y_j|\theta_j &\sim f(y_j|\theta_j) \\ \theta_j|G &\sim G \\ G|\alpha,\eta &\sim \mathrm{DP}(\alpha, G_\eta) \\ (\alpha,\eta) &\sim m(\alpha,\eta), \end{aligned}$$

where $m(\alpha,\eta)$ is the joint distribution of (α,η) (Antoniak (1974)). There is another way to represent the DPM by constructing stick-breaking process.

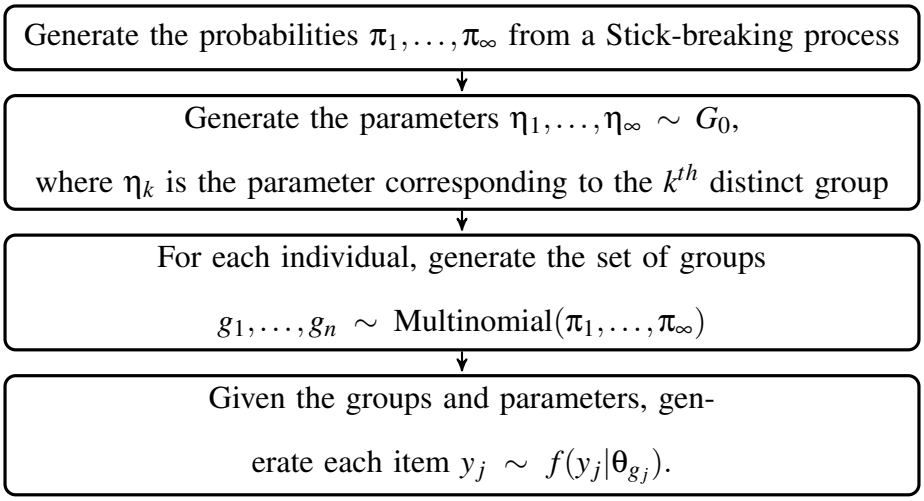
Stick-breaking Construction

The stick-breaking construction is an alternative way to represent a DP which was introduced by Sethuraman (1994). The Pólya urn scheme by Blackwell & MacQueen (1973) generates $\theta \sim G$, not G itself. Stick-breaking is a constructive way to form the measure, $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$. In this representation, we assume that we have a stick of length 1. We generate a random variable β_1 from the Beta distribution with parameters, $(1,\alpha)$ and break it at position β_1 and we assign π_1 equal to the length of the part of the stick that we broke. Now take the stick to the right, and generate $\beta_2 \sim \mathrm{Beta}(1,\alpha)$. Break off the stick β_2 into the stick. Again, π_2 is the length of the stick to the left, that is, $\pi_2 = (1-\beta_1)\beta_2$. We repeat the same process to obtain $\pi_3, \pi_4, \ldots$ and in this way,

we get an explicit construction of G. Then, π_k can be modeled as $\pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l)$, where $\beta_k \sim \text{Beta}(1,\alpha)$. In the previous schemes, θ are sampled directly by the base distribution $\theta_k \overset{iid}{\sim} G_0$. Consequently, the distribution of G can be written as a sum of delta functions weighted with the probabilities, π_k as $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$. Thus the stick-breaking construction gives us a simple and intuitive way to construct a DP.

$$
\begin{aligned}
\beta_k &\sim \text{Beta}(1,\alpha) \\
\pi_1 &= \beta_1 \\
\pi_k &= \beta_k \prod_{l=1}^{k-1}(1-\beta_l), k = 2,3,\ldots \\
\theta_k &\overset{iid}{\sim} G_0 \\
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}.
\end{aligned}
$$

Then $G \sim DP(\alpha, G_0)$. Hence, according to the Stick-breaking construction, the generation process proceeds as follows:



In Bayesian context, we are mainly interested in the posterior inference of the model and if we consider the Bayesian nonparametric models, the computation is challenging. In last 20 years,

there has been notable works on that area and the most popular methods are discussed in Neal (2000).

Posterior Simulations for DPM

In this section, we provide a review of many inference algorithms (Neal (2000)) for DPM models. There are different simulation-based methods for conjugate and non-conjugate priors to get inference on posterior distribution. These methods are mainly based on the Markov chain with few modifications. For conjugate prior models, we can apply Gibbs sampling. But for non-conjugate priors, it is difficult to perform the numerical integration. MacEachern & Müller (1998) and West & Escobar (1993) developed Monte Carlo based approach to handle the non-conjugate priors. Depending upon the representation (2.1) or (2.3) used, algorithms will either sample θ_j directly, or sample the indicators c_j. These algorithms are briefly discussed in Neal (2000).

Conjugate Prior. The first algorithm is introduced for DPM models in Escobar (1994) and has been used in Escobar & West (1995). It directly samples θ_j from the distribution in (2.1). Using (2.1), parameters are sampled one at a time from the following distribution:

$$\theta_j|\theta_{-j}, y_j, \alpha, G_0 \sim n_c \frac{1}{\alpha+n-1}\sum_{l\neq j} f(y_j|\theta_l)\delta(\theta_l) + n_c \frac{\alpha}{\alpha+n-1} r_j H_j, \tag{2.4}$$

where n_c is a normalizing constant, $f(y_j|\theta_l)$ is the density of y_j, $r_j = \int f(y_j|\theta) dG_0(\theta)$, and H_j is the posterior distribution for θ based on G_0 and $f(y_j|\theta)$. Here, the probability of a new component is proportional to the posterior of the parameter provided the present observation weighted by the marginal density of the observation. Hence, the algorithm can be summarized as

Algorithm 1 *(Neal (2000)) Let* $\theta = (\theta_1, \ldots, \theta_n)$.

- *Initialize* θ.

- *Generate* θ_j *from (2.4), for* $j = 1, \ldots, n$.

This algorithm may be inefficient due to the slow convergence rate. It happens because sometimes a group of observations is associated with the same parameter value. Since this algorithm can change only one observation at a time, the chance of occurrence is very low for the change of θ in such a group. For a parameter vector to change when more than one observation is attached to it, the low probability transition is required.

This difficulty can be handled by using Gibbs sampling to the model (2.3). The next algorithm was first used by Bush & MacEachern (1996). For the finite K, the Gibbs sampling consists of choosing a new value for each unique component c_j from its conditional distribution. When K tends to infinity, we choose the K_c that is associated with certain observations and perform the Gibbs sampling. Hence, the posterior distribution over c_j is as follows:

$$P(c_j = c | c_{-j}, y_j, K) = \begin{cases} n_c \frac{n_{-j,c}}{n-1+\alpha} f(y_j|K_c) & \text{if } c = c_l \text{ for some } l \neq j; \\ n_c \frac{\alpha}{n-1+\alpha} \int f(y_j|K_c) dG_0(K) & \text{if } c \neq c_l \forall l \neq j \end{cases} \tag{2.5}$$

where c_{-j} is the vector of c without the j^{th} term, n_c is a normalizing constant, $n_{-j,c}$ is the number of c_l for $l \neq j$, and K is a set of K_c. Hence, the algorithm can be written as:

Algorithm 2 *(Neal (2000)) Let* $c = (c_1, \ldots, c_n)$ *and* $K = (K_c : c \in \{c_1, \ldots, c_n\})$.

- *Initialize* $(c_1, \ldots, c_n)$ *and* K.

- *Remove* j *from* K_{c_j} *and generate a new sample from (2.5). If* c_j *is not same as any other values, draw a value from the posterior* H_j.

- *Generate a new value of* $K_c | y_j, \forall j$ *for which* $c_j = c$.

In conjugate setup, instead of using K_c, we may use the Markov chain that consists of c_j. This algorithm is discussed in MacEachern (1994) and Neal (1992) in the context of normal and categorical

models, respectively. Here, all parameter vectors are integrated out of the state of the Markov chain, without c_j and other hyper-parameters. Hence, the posterior probability corresponding to c_j is

$$P(c_j = c|c_{-j}, y_j) = \begin{cases} n_c \frac{n_{-j,c}}{n-1+\alpha} \int f(y_j|K) dH_{-j,c}(K) & \text{if } c = c_l \text{ for some } l \neq j; \\ n_c \frac{\alpha}{n-1+\alpha} \int f(y_j|K) dG_0(K) & \text{if } c \neq c_l \forall l \neq j \end{cases} \tag{2.6}$$

where n_c is a normalizing constant, $n_{-j,c}$ is the number of c_l for $l \neq j$, and $H_{-j,c}$ is the posterior distribution of K based on G_0 and for all y_j for which $c_j = c$. Hence, the algorithm can be summarized as follows:

Algorithm 3 *(Neal (2000)) Let* $c = (c_1, \ldots, c_n)$.

- *Initialize* $c_1, \ldots, c_n$.
- *Generate* c_j *from (2.6), for* $j = 1, \ldots, n$.

This algorithm is a very efficient sampling algorithm for the conjugate case because it removes the noise from the random parameters and gives a precise estimate of the likelihood.

Non-conjugate Prior. If G_0 is not conjugate, simulations from the previous algorithms can not be preformed. Because, the integrals associated with each algorithm are usually analytically intractable. West & Escobar (1993) proposed a Monte Carlo based approximation to deal with the integrals. But, this method is not accurate. MacEachern & Müller (1998) proposed another algorithm, *no gap* algorithm, based on a valid Markov chain sampler.

In the *no gap* algorithm, $c_1 \ldots, c_n$ consists of k distinct elements and the idea of this algorithm is to extend the **Algorithm 2**. MacEachern & Müller (1998) modified the method as follows:

Algorithm 4 *(Neal (2000)) Let* $c = (c_1, \ldots, c_n)$, $K = (K_c : c \in \{c_1, \ldots, c_n\}) = (K_1, \ldots, K_k)$.

- *Initialize* $(c_1, \ldots, c_n)$, $(K_1, \ldots, K_k)$.

- *If $c_j \neq c_l$ for all $l \neq j$, then with probability $q/(q+1)$, c_j would be unchanged. Otherwise, label c_j as $q+1$, or draw a value for K_{q+l} from G_0 if $c_j = c_l$ for some $l \neq j$. Then draw a new value for c_j from $\{1,\ldots,q+1\}$ using the following probabilities*

$$P(c_j = c|c_{-j}, y_j, K_1, \ldots, K_{q+1}) = \begin{cases} n_c n_{-j,c} f(y_j|K_c) & \text{if } 1 \leq c \leq q; \\ n_c \frac{\alpha}{q+1} f(y_j|K_c) & \text{if } c = q+1 \end{cases}$$

where q is the number of distinct c_l for $l \neq j$ with the values of c_l consist of $\{1,\ldots,q\}$.

- *Generate a new value of $K_c|y_j, \forall j$ for which $c_j = c$.*

This algorithm does not require an evaluation of the integral that featured in the MCMC algorithm for conjugate DPM models. The algorithm can be implemented to any model irrespective of non-conjugate G_0. Neal (2000) described another approach based on the Metropolis-Hastings algorithms to update c_j using the conditional prior as the proposal distribution. This algorithm cancels the factors and it does not contain c_j when computing the following acceptance probability,

$$a(c_j^*, c_j) = \min\left[1, \frac{f(K_{c_j^*}|y_j)}{f(K_{c_j}|y_j)}\right]. \tag{2.7}$$

Hence, the algorithm can be written as follows:

Algorithm 5 *(Neal (2000)) Let $c = (c_1,\ldots,c_n)$ and $K = (K_c : c \in \{c_1,\ldots,c_n\})$.*

- *Initialize $(c_1,\ldots,c_n)$ and K.*

- *Repeat the following R times*

    1. *Generate a candidate $K_{c_j^*}$ from the following conditional prior distribution*

$$P(c_j = c|c_{-j}) = \begin{cases} \frac{n_{-j,c}}{n-1+\alpha} & \text{if } c = c_l \text{ for some } l; \\ \frac{\alpha}{n-1+\alpha} & \text{if } c \neq c_l \text{ for all } l \end{cases}$$

2. *If c_j^* is not in $\{c_1,\ldots,c_n\}$, generate value from G_0 and set the new value of c_j as c_j^* with probability (2.7). Otherwise, set the new value of c_j as the previous value.*

- *Generate a new value of $K_c|y_j, \forall j$ for which $c_j = c$.*

If the updates of the K_c in the **Algorithm 5** is excluded, the algorithm can be written in terms of θ_j instead of K_{c_j}.

Algorithm 6 *(Neal (2000)) Let $\theta = (\theta_1,\ldots,\theta_n)$.*

- *Initialize θ.*

- *For $j = 1,\ldots,n$, repeat the following R times*

1. *Generate a candidate θ_j^* from the following distribution*

$$\frac{1}{n-1+\alpha}\sum_{l\neq j}\delta(\theta_l) + \frac{\alpha}{n-1+\alpha}G_0.$$

2. *Compute the acceptance probability,*

$$a(\theta_j^*, \theta_j) = \min\left[1, \frac{f(\theta_j^*|y_j)}{f(\theta_j|y_j)}\right]. \tag{2.8}$$

3. *Set the new value of θ_j as θ_j^* with probability in (2.8). Otherwise, set the new value of θ_j as the previous value.*

A modification to **Algorithm 6** improves mixing by proposing new clusters for non-singletons, that is, more than one components are associated with the data vector and proposing non-singletons for singletons. This modification can be done using the combination of the Metropolis-Hasting with the partial Gibbs sampling. The algorithm proceeds as follows:

Algorithm 7 *(Neal (2000)) Let $c = (c_1,\ldots,c_n)$ and $K = (K_c : c \in \{c_1,\ldots,c_n\})$.*

- *Initialize* $(c_1, \ldots, c_n)$ *and* K.

- *For* $j = 1, \ldots, n$, *if* c_j *is not singleton,*

    1. *Generate a candidate* $K_{c_j^*}$ *from* G_0.

    2. *Set the new value of* c_j *as* c_j^* *with probability*

$$a(c_j^*, c_j) = \min\left[1, \frac{\alpha}{n-1}\frac{f(K_{c_j^*}|y_j)}{f(K_{c_j}|y_j)}\right].$$

    Otherwise, draw c_j^* *from* c_{-j} *choosing* $c_j^* = c$ *with probability* $n_{-j,c}/(n-1)$. *Set the new value of* c_j *as* c_j^* *with probability*

$$a(c_j^*, c_j) = \min\left[1, \frac{n-1}{\alpha}\frac{f(K_{c_j^*}|y_j)}{f(K_{c_j}|y_j)}\right].$$

    Otherwise, set the new value of c_j *as the previous value.*

- *For* $j = 1, \ldots, n$, *if* c_j *is singleton,*

    1. *Keep the same* c_j.

    2. *Otherwise, choose the new value of* c_j *from* $(c_1, \ldots, c_n)$ *with probability*

$$n_c \frac{n_{-j,c}}{n-1} f(y_j|K_c).$$

- *Generate a new value of* $K_c|y_j, \forall j$ *for which* $c_j = c$.

Neal (2000) proposed an algorithm based on the Gibbs sampler with auxiliary variables for generating new clusters in the algorithm. This sampler is a simple and efficient way to deal with the non-conjugate priors. In this algorithm, there are m auxiliary components. As $m \to \infty$, this algo-

rithm tends to the **Algorithm 2** because it uses m samples to estimate the marginal density. Here, the probability distribution corresponding to c_j is as follows:

$$P(c_j = c|c_{-j}, y_j, K_1, \ldots, K_{q+m}) = \begin{cases} n_c \frac{n_{-j,c}}{n-1+\alpha} f(y_j|K_c) & \text{if } 1 \leq c \leq q; \\ n_c \frac{\alpha/m}{n-1+\alpha} f(y_j|K_c) & \text{if } q < c \leq q+m \end{cases} \tag{2.9}$$

Hence, the algorithm proceeds as follows:

Algorithm 8 *(Neal (2000)) Let* $c = (c_1, \ldots, c_n)$ *and* $K = (K_c : c \in \{c_1, \ldots, c_n\})$.

- *Initialize* $(c_1, \ldots, c_n)$ *and* $(K_1, \ldots, K_k)$.
- *Let* q *be the number of distinct* c_l *for* $l \neq j$ *with the values of* c_l *consist of* $\{1, \ldots, q\}$. *Then we have to choose one of the followings:*
  1. *Draw the values for* K_c *from* G_0 *for which* $q < c \leq q+m$ *if* $c_j = c_l$ *for some* $l \neq j$.
  2. *If* $c_j \neq c_l$ *for all* $l \neq j$ *with label* c_j *as* $q+1$, *draw the values for* K_c *from* G_0 *for which* $q+1 < c \leq q+m$.
- *Draw a new value for* c_j *from* $\{1, \ldots, q+m\}$ *using (2.9).*
- *Generate a new value of* $K_c|y_j, \forall j$ *for which* $c_j = c$.

There are some other algorithms available for both conjugate and non-conjugate model distributions. Jain et al. (2007) and Jain & Neal (2012) have proposed split-merge algorithms to focus on splitting a cluster into two or merging two clusters into one with the others remain same. The main idea of these algorithms is to draw a conclusion about the better formation of one or more clusters based on the proposed clusters. There are also different methods based on slice sampling (Walker et al. (1999), Walker (2007)), particle filtering (Fearnhead (2004)), variational methods (Blei & Jordan (2004), Kurihara et al. (2007)), and expectation propagation (Minka & Ghahramani (2003)).

2.3.2 Pitman-Yor Process

An extension to the DP is the Pitman-Yor process. Two parameter Poisson-Dirichlet process defined by Pitman & Yor (1997), also referred as the Pitman-Yor process, $\mathcal{PY}(d,\alpha,G_0)$ is a related probability measure on the space of probability measures. The parameters of this process are: (1) discount parameter d, $0 \leq d < 1$; (2) concentration parameter α, $\alpha > -d$; (3) a base distribution, G_0. Let $V_1, V_2, \ldots$ be the set of independent random variables which are drawn from Beta$(1-d, \alpha+jd)$ and p_j be the corresponding probabilities. Under the stick-breaking construction, p_j are defined as follows:

$$\begin{aligned} p_1 &= V_1 \\ p_2 &= V_2(1-V_1) \\ &\vdots \\ p_j &= V_j \prod_{l=1}^{j-1}(1-V_j) \\ &\vdots \end{aligned} \tag{2.10}$$

Hence, the PYP can be written as follows:

$$\begin{aligned} V_j &\sim \text{Beta}(1-d, \alpha+jd) \\ (\theta_1,\ldots,\theta_n) &\overset{iid}{\sim} G_0 \\ p_1 &= V_1 \\ p_j &= V_j \prod_{l=1}^{j-1}(1-V_j), \quad j = 2,3,\ldots, \\ \mathcal{PY}(d,\alpha,G_0) &\overset{\mathcal{D}}{=} \sum_{j=1}^{\infty} p_j \delta_{\theta_j}. \end{aligned}$$

If $d = 0$, the PYP reduces to the DP. This process can also be defined in terms of the unique values of $\theta_1, \ldots, \theta_n$, denoted as, $\eta_1, \ldots, \eta_m$. Other than stick braking construction, PYP can be redefined in terms of Pólya urn scheme.

2.3.2.1 Pólya urn scheme

The Pólya urn scheme is another representation of the PYP. Let d, $(0 \leq d < 1)$ be the discount parameter, α, $(\alpha > -d)$ denote the concentration parameter, and G_0 be the base distribution. Suppose $(\theta_1, \ldots, \theta_n)$ be the sequence of parameters. In this setup, the predictive distribution of the model can be defined as

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha + dm}{\alpha + n} G_0(\theta_{n+1}) + \frac{1}{\alpha + n} \sum_{k=1}^{m} (n_k - d)\delta_{\eta_k}(\theta_{n+1}),$$

where $\{\eta_1, \ldots, \eta_m\}$ are the unique values of $\{\theta_1, \ldots, \theta_n\}$ with corresponding frequency n_k. For the DP, the previous process is reduced to the following form

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} G_0(\theta_{n+1}) + \frac{1}{\alpha + n} \sum_{k=1}^{n-1} \delta_{\theta_k}(\theta_{n+1}).$$

Let $\theta_1, \ldots, \theta_n|G \sim G$ with $E(G) = G_0$. According to Fall & Barat (2014), the posterior distribution of PYP can be expressed as follows:

$$G|\theta_1, \ldots, \theta_n \stackrel{\mathcal{D}}{=} \sum_{l=1}^{m} p_j \delta_{\eta_j} + r_m G_m, \tag{2.11}$$

where $\eta_1, \ldots, \eta_m$ are the unique values of $\theta_1, \ldots, \theta_n$ with corresponding frequencies are $n_1, \ldots, n_m$,

$$\begin{aligned} p_1, \ldots, p_m, r_m &\sim \text{Dirichlet}(n_1 - d, \ldots, n_m - d, \alpha + dm) \\ G_m &\sim \mathcal{PY}(d, \alpha + dm, G_0) \end{aligned}$$

with $E(G_m) = G_0$ and G_m is independent of $p_1, \ldots, p_m, r_m$.

2.3.2.2 Pitman-Yor mixture models

Now, the extension of the DPM is the Pitman-Yor mixture (PYM) models and it can be defined as follows:

$$\begin{aligned} Y_j|\theta_j &\sim f(y_j|\theta_j) \\ \theta_j|G &\sim G \\ G &\sim \mathcal{PY}(d, \alpha, G_0). \end{aligned}$$

For the PYM, it is difficult to calculate the posterior distributions. There are different approximate techniques are available. Fall & Barat (2014) reviewed some of the methods. Also, the techniques from Neal (2000) can be extended in this context. In the next chapter, we propose an approximate Bayesian computation method based on DP and PYP nonparametric models.

CHAPTER 3

APPROXIMATE BAYESIAN COMPUTATION FOR BAYESIAN NONPARAMETRIC MODELS (ABC-BNP)

3.1 Introduction

In recent years, a large literature has developed on Bayesian nonparametric models due to the versatility and availability of simple and efficient way to compute the posterior distribution. Most of the literature discuss about the DPM models and a substantial amount of study has been done on MCMC for posterior calculation. ABC provides a fast and flexible method for posterior computation and intractable and the intractable likelihoods are easily handled by this computation method. In addition, it is also applicable to PYP. In this chapter, we propose a general method for Bayesian nonparametric models using ABC, denoted as ABC-BNP.

3.2 Method

We consider a Bayesian nonparametric model of the form

$$\begin{aligned} Y_j|\theta_j &\sim f(y_j|\theta_j) \\ \theta_j|G &\sim G \\ G &\sim \text{BNP}(\zeta, G_0), \end{aligned} \tag{3.1}$$

where BNP stands for Bayesian nonparametric models, G is the random mixing measure with base distribution G_0, $f(y_j|\theta_j)$ denotes a parametric distribution, and ζ is the parameter vector corresponding to G. For DPM models, $\zeta = \alpha$ and for PYM models, ζ becomes (d, α), where d is the discount parameter and α represents the concentration parameter. Here, for each observation, $j = 1, \ldots, n$, θ_j can be obtained by using the Pólya urn scheme and the corresponding prior distribution is defined by the full conditionals:

$$\pi(\theta_j|\theta_{-j}) \quad = \quad \frac{\alpha_k}{\alpha+n-1} G_0(\theta_j) + \frac{1}{\alpha+n-1} \sum_{k \neq j} l_k, \tag{3.2}$$

where $\theta_{-j} = \{\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_n\}$, α_k is the constant based on the different nonparametric Bayesian models, l_k is a coefficient depending upon the condition on $\delta(\cdot)$ corresponding to Bayesian nonparametric models and j^{th} observation, and $\delta(x)$ is the degenerate distribution at the point x. So, the full conditional posterior distribution can be expressed as

$$\theta_j|\theta_{-j}, y_j \quad \sim \quad n_c \alpha_k P_j \int f(y_j|\theta) dG_0(\theta) + n_c \sum_{k \neq j} f(y_j|\theta_j) l_k,$$

where P_j is the posterior density based on the base measure. For the non-conjugate case, the posterior may not be in the closed form. So we combine the idea of Bayesian nonparametric models with ABC-MCMC method which can be applied to both the conjugate and non-conjugate cases. For j^{th} step of Bayesian nonparametric models for ABC (ABC-BNP), we sample a candidate value, θ_j^* from the prior in (3.2) and using the value of θ_j, generate a new sample z_j^* from the sampling distribution $f(z_j|\theta_j^*)$. Now, the acceptance probability of the Metropolis-Hastings step in the ABC-MCMC (see Sec. 2.2.1.3) procedure is defined as

$$\alpha(\theta_j, \theta_j^*) = \min \left\{ 1, \frac{q(\theta_j|\theta_j^*)\pi(\theta_j^*)}{q(\theta_j^*|\theta_j)\pi(\theta_j)} \mathbb{I}(\rho(\mathcal{S}(y), \mathcal{S}(z^*)) \leq \varepsilon) \right\}, \tag{3.3}$$

where ρ represents the distance metric, $\mathcal{S}$ indicates the summary statistic, $\mathbb{I}$ denotes the indicator function, $q(\cdot)$ is the proposal density and $\pi(\cdot)$ is the prior distribution. We assign Euclidean distance throughout the study. The fully intractable models imply unavailability of the sufficient statistics and hence, it is difficult to use any ABC type methods that are discussed in Section 2.2 to estimate the posterior. Thus we update the parameter one at a time and do not need to use the summary statistic in the model. So, the last part of the acceptance probability in (3.3) reduces to $\mathbb{I}(A_j)$ which is a model dependent condition based on y_j and z_j^*. For example, if the response is binary, the choice of A_j would be $(y_j - z_j^*) = 0$, that is, in the binary case, $I(A_j) = 1$ if and only if $y_j* = z_j*$ and $I(A_j) = 0$ if otherwise. For the DPM models under the ABC-BNP setup, if we use the DP or PYP prior as the proposal density, that is, $q(\theta_j|\theta_j^*) = \pi(\theta_j|\theta_{-j})$, $q(\theta_j^*|\theta_j) = \pi(\theta_j^*|\theta_{-j})$, $\pi(\theta_j) = \pi(\theta_j|\theta_{-j})$, and $\pi(\theta_j^*) = \pi(\theta_j^*|\theta_{-j})$, the factors cancel out in the probability and (3.3) becomes

$$\alpha(\theta_j, \theta_j^*) = \min\left\{1, \frac{\pi(\theta_j|\theta_{-j})\pi(\theta_j^*|\theta_{-j})}{\pi(\theta_j^*|\theta_{-j})\pi(\theta_j|\theta_{-j})}\mathbb{I}(A_j)\right\} = \mathbb{I}(A_j),$$

Therefore, we set $\theta_j = \theta_j^*$ with probability

$$\alpha(\theta_j, \theta_j^*) = \mathbb{I}(A_j). \tag{3.4}$$

Accordingly, the transition kernel T of θ can be written as

$$T(\theta^{(b)}|\theta^{(b-1)}) = \prod_{j=1}^{n} s(\theta_j^{(b)}|\theta_{-j}^{(b-1)}),$$

where $\theta_{-j}^{(b-1)} = (\theta_1^{(b)}, \ldots, \theta_{j-1}^{(b)}, \theta_{j+1}^{(b-1)}, \ldots, \theta_n^{(b-1)})$ and $s(\cdot)$ can be defined in terms of (3.2) and (3.4). Hence, for the computational purpose, the method can be implemented as follows:

I. Initialize $\theta^{(0)}$, $b = 0$.

II. Updating the parameters vector based on the transition kernel, T, which can be defined as

$$T(\theta^{(b)}|\theta^{(b-1)}) = \prod_{j=1}^{n} s(\theta_j^{(b)}|\theta_{-j}^{(b-1)}),$$

where for $j = 1,\ldots,n$, repeat the following steps.

1. Generate a candidate value,

$$\theta_j^*|\theta_1^{(b)},\ldots,\theta_{j-1}^{(b)},\theta_{j+1}^{(b-1)},\ldots,\theta_n^{(b-1)} \sim \pi(\theta_j^*|\theta_{-j}^{(b-1)})$$

as described in (3.2).

2. Generate a dataset

$$z_j^* \sim \mathrm{f}(z_j|\theta_j^*).$$

3. Set

$$\theta_j^{(b)} = \begin{cases} \theta_j^* & \text{if } \mathbb{I}(A_j); \\ \theta_j^{(b-1)} & \text{otherwise} \end{cases}$$

where $\mathbb{I}(C) = 1$ if C holds, and 0 otherwise.

III. Repeat the procedure for $b = 1,\ldots,B$.

Based on the ABC-MCMC (Marjoram et al. (2003)), the conditional posterior distribution converges to the parameter vector $(\theta_1,\ldots,\theta_n)$. Since the ABC-BNP is based on MCMC, the parameters of interest also converge to the target posterior distribution.

3.3 An Example

This section deals with an real data example based on ABC-BNP for a DP conjugate model. Suppose $y = (y_1, \ldots, y_n)$ represents continuous response variable on n observations. We assume that each $y_j, j = 1, \ldots, n$, are normal distribution with unknown mean μ and unknown variance σ^2. While working with the nonparametric Bayesian models, we introduce $\theta_j = (\mu_j, \sigma_j^2)$ and assume that θ_j are sampled independently and identically from a distribution G. Here, G is DP prior with concentration parameter α and base distribution G_0. Under the conjugacy of the DPM models, we can specify a conjugate base measure for μ_j and σ_j^2. Now, the simplest choice for G_0 is the normal-inverse-gamma distribution where the normal distribution has the mean μ_0, variance σ_0^2 and the inverse-gamma distribution has the shape a_0 and scale b_0. Hence the conjugate DPM model for ABC-BNP is defined as follows:

$$\begin{aligned} Y_j|\theta_j &\sim \text{Normal}(y_j|\mu_j, \sigma_j^2) \\ \theta_j|G &\sim G \\ G &\sim \text{DP}(\alpha, G_0) \\ G_0 &\equiv \text{Normal}(\mu_j|\mu_0, \sigma_0^2) \cdot IG(\sigma_j^2|a_0, b_0). \end{aligned}$$

Here, we have used the galaxy data which consists of velocities of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. This data is used in Escobar & West (1995) and we scaled the data by dividing the speed by 10000. In this example, $n = 82, \mu_0 = 2, \sigma_0^2 = 1, \alpha = 10, a_0 = 2$, and $b_0 = 5$. For ABC-BNP, we choose the j^{th} condition as $A_j = |y_j - z_j^*| \leq \varepsilon$, where the tolerance level, ε is 0.001. We run the simulation $B = 100000$ times with 20000 burn-in period to get the approximate posterior distribution.

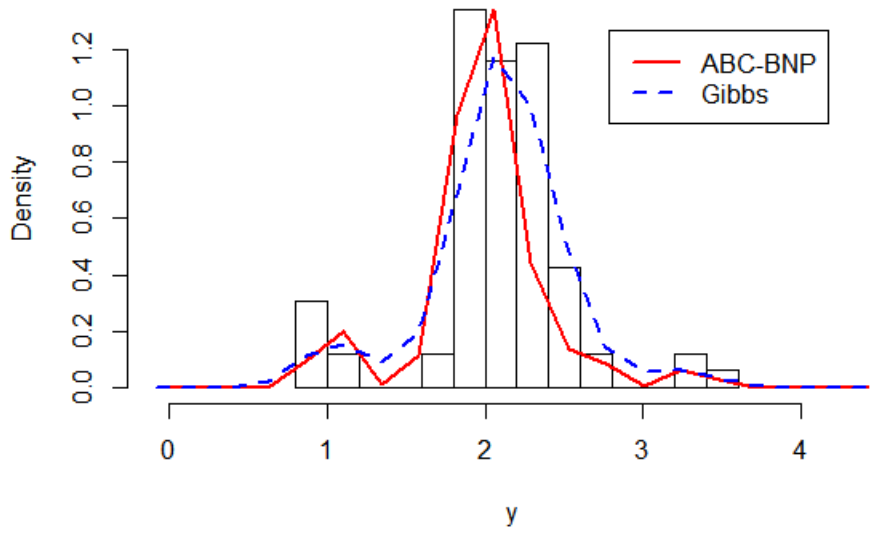


Figure 3.1: Comparison of predictive distribution based on Gibbs and ABC-BNP sampler

For ABC-BNP, the average acceptance rate is 2.3%. Now, we want to compare the result of ABC-BNP with the Gibbs sampler using the predictive distribution. In Fig. 3.1, the ABC-BNP and Gibbs sampling provide almost same predictive distribution of a new observation Y_{83}. Since ABC-BNP method is more simpler, it took less time for computation compared to Gibbs method (Table 3.1).

Table 3.1: Time comparison of ABC-BNP and Gibbs method for normal model

Method	Time (in minutes)
ABC-BNP	3.71
Gibbs	30.02

CHAPTER 4

ABC-BNP AND GENERALIZED LINEAR MIXED MODELS FOR BINARY RESPONSES

4.1 Introduction

Generalized Linear Models (GLM) are most commonly used in statistical inference for the fixed effects. GLM consists of three components: random component, linear predictor, and a link function. Let $Y = (Y_1, \ldots, Y_n)$ be a set of independent data. The random component is the probability distribution of Y_j, the linear predictor of GLM, denoted as γ_j, is a linear function of predictor variables, $\gamma_j = X_j^T \beta$, where X_j is the vector of predictors for j^{th} subject with fixed effects β, and the link function $g(\cdot)$ transforms the expectation of the response variable to the linear predictor γ_j, that is, $g(\cdot) = \gamma_j$. These models are derived and explained in McCullagh & Nelder (1989). An extension to the GLM is the Generalized Linear Mixed Models (GLMM, see, for example, McCulloch & Searle (2001)) which represent an important class of regression models. Here the random effects are combined with the fixed effects in GLM to account for the correlation effect. These models include different types of responses, such as continuous, categorical, and counts. For the binary categorical case, logistic and probit regression models are most commonly used in GLMM.

Let $Y = (Y_1, \ldots, Y_n)$ denote binary random variable of n subjects, $X = (X_1, \ldots, X_p)$, X_j be the vector of p covariates corresponding to the j^{th} observation, $\beta = (\beta_1, \ldots, \beta_p)$ be the regression coefficient vector of dimension p, $\Psi = (\psi_1, \ldots, \psi_n)$ be the vector of random effects, where ψ_j is the random effect corresponding to j^{th} subject, and $\gamma = (\gamma_1, \ldots, \gamma_n)$ be the vector of linear predictors.

In GLMM, for $j = 1, \ldots, n$, we assume that $Y_j|\psi$ are conditionally independent and drawn from the exponential family of distribution. For each j, the GLMM model under binary setup can be defined as

$$\begin{aligned} Y_j &\sim \text{Bernoulli}(p_j) \\ g(p_j) &= \gamma_j \\ \gamma_j &= f(X_j, \beta, \psi_j). \end{aligned}$$

where $f(\cdot)$ can be additive or multiplicative models. For the Bayesian DPM model, Kyung et al. (2011) assumed that

$$\begin{aligned} \psi_j &\sim G \\ G &\sim DP(\alpha, G_0). \end{aligned} \tag{4.1}$$

where α is the concentration parameter and G_0 is the base distribution. Hence, the full conditional of ψ_j given $\psi_1, \ldots, \psi_{j-1}, \psi_{j+1}, \ldots, \psi_n$ can be defined as

$$\psi_j|\psi_1, \ldots, \psi_{j-1}, \psi_{j+1}, \ldots, \psi_n \sim \frac{\alpha}{\alpha+n-1} G_0(\psi_j) + \frac{1}{\alpha+n-1} \sum_{k \neq j} \delta_{\{\psi_k\}}(\psi_j), \tag{4.2}$$

where $\delta_{\{\psi_k\}}(\psi_j)$ is the degenerate distribution at the point ψ_j and the sampling distribution for $y = (y_1, \ldots, y_n)$ is as follows:

$$f(y) = \int \prod_{j=1}^{n} [g^{-1}(f(X_j, \beta, \psi_j))]^{y_j} [1 - g^{-1}(f(X_j, \beta, \psi_j))]^{1-y_j} dG_0(\Psi),$$

where $g^{-1}(\cdot)$ is the inverse of $g(\cdot)$. Now, suppose there are k distinct groups, $\{S_1, \ldots, S_k\}$, for n samples of the random effect Ψ. So, if $j \in S_l$, $\psi_j = \eta_l$ and hence $\Psi = A\eta$, where A is a matrix of order $n \times k$ with each row vector a_l represents the vector of all zeros except for a 1 in the position

indicating which group the observation belongs to. Hence, $\eta = (\eta_1, \ldots, \eta_k)$ with $\eta_l \overset{iid}{\sim} G_0$ for $l = 1, \ldots, k$. Hence, the sampling distribution for y given A for distinct values of Ψ becomes:

$$f(y|A) = \int \prod_{j=1}^{n} [g^{-1}(f(X_j, \beta, (A\eta)_j))]^{y_j} [1 - g^{-1}(f(X_j, \beta, (A\eta)_j))]^{1-y_j} dG_0(\eta).$$

Now, the choice of $f(\cdot)$ depends on the link function of the models. For example, if the model is related to the logit link, the corresponding model could be the random intercept and for the probit link, we can think of scale mixture model.

In Bayesian framework, when the random effects appeared in the model due to the DP, MCMC methods play an important role to simulate observations from the posterior. In the case of the conjugate prior, Gibbs sampling in the Pólya urn set up are commonly used. On the other hand, for the non-conjugate priors, the algorithms that are described in Neal (2000) are widely applied for the posterior calculations. The complexity of Markov chain sampling and Bayesian inference in GLMM have been extensively studied in the recent literature. In this chapter, we apply our proposed method to handle the binary GLMM and compare with the existing MCMC procedures.

4.2 Random Intercept Model

In the binary setting, the logit link function is represented as the random intercept model and the model can be defined as

$$\begin{aligned} g(p_j) &= \log\left(\frac{p_j}{1-p_j}\right) = \gamma_j \\ \Longrightarrow g^{-1}(\gamma_j) &= \frac{\exp(\gamma_j)}{1+\exp(\gamma_j)} = p_j \\ \gamma_j &= f(X_j, \beta, \psi_j) = X_j^T \beta + \psi_j. \end{aligned}$$

Now, under the DPM models in (4.1), suppose G_0 is a normal distribution with mean 0 and variance σ^2 and the concentration parameter is fixed, for example, $\log(n)$. Also, we can assume a prior on the concentration parameter, for example, gamma distribution. The prior for σ^2 is an inverse-gamma distribution with shape parameter a_0, the scale parameter b_0 with fixed (a_0, b_0) and the prior for β is assigned to be Normal$(\mu_0, \sigma_0^2 I)$. Hence, the model can be represented as follows:

$$\begin{aligned} P(Y_j = 1|X_j, \beta, \psi_j) &= \frac{\exp(X_j^T\beta + \psi_j)}{1+\exp(X_j^T\beta + \psi_j)}, j = 1, \ldots, n \\ \psi_j|G &\sim G \\ G &\sim \mathrm{DP}(\alpha, G_0) \\ G_0 &\equiv \mathrm{Normal}(\psi_j|0, \sigma^2) \\ \sigma^2 &\sim IG(a_0, b_0) \\ \beta &\sim \mathrm{Normal}(\mu_0, \sigma_0^2 I). \end{aligned} \tag{4.3}$$

In most cases, it is not possible to sample from the complex models. MCMC methods play an important role to sample from such models. There are different methods are available based on this structure.

4.2.1 Sampling Methods

4.2.1.1 Slice Sampler

For the random intercept models, Kyung et al. (2011) proposed the slice sampler to update the model parameters. In this case, the MCMC procedure overcomes the problem of small steps for random walk and high steps for high rejection rates in the Metropolis-Hastings algorithm are

adjusted by the number of steps. This sampling method has been discussed in Damien et al. (1999), Neal (2003), Dittmar (2013) for improving the efficiency of the MCMC methods.

The idea of the slice sampling is based on Damien et al. (1999) which implies that if it is hard to generate sample from $f(\theta) \propto L(\theta)\pi(\theta)$, where $L(\theta)$ is the likelihood function and $\pi(\theta)$ is the prior distribution, we can introduce a latent variable U, defined on $(0, L(\theta))$ and define the joint distribution with θ as $f(\theta, u) \propto I\{u < L(\theta)\}\pi(\theta)$. Hence a Markov chain is performed to sample from $f(\theta, u)$ that will converge to the uniform distribution. An easier way to handle this situation is to use the Gibbs sampling which samples from θ and u, given the value of other variables. In the slice sampling, u can de drawn from Uniform distribution with range $(0, L(\theta))$ and θ is sampled uniformly from the points given by $\{\theta : L(\theta) > u\}$.

Kyung et al. (2011) used the idea of the slice sampling and implemented to DP generalized linear mixed model. They proposed normal base distribution in (4.1) with mean 0 and variance σ^2. They have considered two latent variables, $U = (U_1, \ldots, U_n)$ and $V = (V_1, \ldots, V_n)$. The prior for σ^2 is taken to be an inverse gamma (IG) distribution with shape parameter a_0 and the scale parameter is b_0 and the prior for β is a normal distribution with mean 0 and variance $d^*\sigma_0^2$, that is, $\beta|\sigma_0^2 \sim \text{Normal}(0, d^*\sigma_0^2 I), \sigma^2 \sim IG(a_0, b_0), d^* > 1$ and (a_0, b_0) are fixed. Then for fixed α and A, the Gibbs sampler for $(\beta, \sigma^2, \eta, U, V)$ is as follows:

- for $d = 1, \ldots, p$,

$$\beta_d|\beta_{-d}, \sigma^2, \eta, U, V, A, y \sim \begin{cases} \mathcal{N}\left(0, d^*\sigma_0^2\right) \text{ if } \beta_d \in \\ \left[\max\left\{\left(\max_{X_{jd}>0}\left(\frac{\alpha_{jd}}{X_{jd}}\right)\right), \left(\max_{X_{jd}\leq 0}\left(\frac{\gamma_{jd}}{X_{jd}}\right)\right)\right\}, \right. \\ \left. \min\left\{\left(\min_{X_{jd}\leq 0}\left(\frac{\alpha_{jd}}{X_{jd}}\right)\right), \left(\min_{X_{jd}>0}\left(\frac{\gamma_{jd}}{X_{jd}}\right)\right)\right\}\right] \\ 0, \text{otherwise.} \end{cases}$$

where

$$\alpha_{jd} = -\log(u_j^{-\frac{1}{y_j}} - 1) - \sum_{l \neq d} X_{jl}\beta_l - (A\eta)_j \quad \text{for} \quad j \in S$$

$$\gamma_{jd} = -\log(v_j^{-\frac{1}{y_j - 1}} - 1) - \sum_{l \neq d} X_{jl}\beta_l - (A\eta)_j \quad \text{for} \quad j \in F$$

where $S = \{j : y_j = 1\}$ and $F = \{j : y_j = 0\}$.

- $$\sigma^2|\beta,\eta,U,V,A,y \sim IG\left(\frac{k}{2} + a_0, \frac{1}{2}\|\eta\|^2 + b_0\right)$$

- for $l = 1,\ldots,k$,

$$\eta_l|\beta,\sigma^2,U,V,A,y \sim \begin{cases} \text{Normal}\left(0,\sigma^2\right) & \text{if } \eta_l \in \left(\max_{j\in S_l}\{\alpha_j^*\}, \min_{j\in S_l}\{\gamma_j^*\}\right) \\ 0 & \text{otherwise.} \end{cases}$$

where

$$\alpha_j^* = -\log(u_j^{-1} - 1) - X_j\beta \quad \text{for} \quad j \in S$$

$$\gamma_j^* = \log(v_j^{-1} - 1) - X_j\beta \quad \text{for} \quad j \in F$$

- for $j = 1,\ldots,n$,

$$\pi_k(U_j|\beta,\sigma^2,\eta,V,A,y) \propto I\left[u_j < \left\{\frac{1}{1+\exp(-X_j^T\beta - (A\eta)_j)}\right\}^{y_j}\right] \quad \text{for} \quad j \in S$$

$$\pi_k(V_j|\beta,\sigma^2,\eta,U,A,y) \propto I\left[v_j < \left\{\frac{1}{1+\exp(X_j^T\beta + (A\eta)_j)}\right\}^{1-y_j}\right] \quad \text{for} \quad j \in F$$

The formulation of the slice sampler in the DPM generalized linear mixed model as described by Kyung et al. (2011) is very complicated. So, we propose an alternative approach for GLMM using ABC to generate the parameters.

4.2.1.2 Proposed Method for Random Intercept Model

In this setup, we use the ABC-BNP which is more easier to compute than other nonparametric methods. Here, the ABC-BNP for GLMM is used to get the probability distribution of the random effect Ψ. The transition kernel for this model is

$$\begin{aligned} T(\sigma^{2(b)}, \Psi^{(b)}, \beta^{(b)} | \sigma^{2(b-1)}, \Psi^{(b-1)}, \beta^{(b-1)}) &= s_1(\sigma^{2(b)} | \sigma^{2(b-1)}, \Psi^{(b-1)}, \beta^{(b-1)}) \\ &\times \prod_{j=1}^{n} s_2(\psi_j^{(b)} | \sigma^{2(b)}, \psi_{-j}^{(b-1)}, \beta^{(b-1)}) \\ &\times s_3(\beta^{(b)} | \sigma^{2(b)}, \Psi^{(b)}, \beta^{(b-1)}). \end{aligned} \tag{4.4}$$

Here $s_1(\cdot)$ indicates the distribution of $\sigma^{2(b)} | \sigma^{2(b-1)}, \Psi^{(b-1)}, \beta^{(b-1)}$, that is, given the values of $(\sigma^{2(b-1)}, \Psi^{(b-1)}, \beta^{(b-1)})$, $\sigma^{2(b)}$ is sampled from the following inverse-gamma distribution,

$$IG\left(\frac{K}{2} + a_0, \frac{1}{2}\|\eta^{(b-1)}\|^2 + b_0\right),$$

where $\|\cdot\|$ represents the Euclidean norm and $\eta^{(b-1)}$ defines the vector of unique values of $\Psi^{(b-1)}$ with length K.

$s_2(\cdot)$ is defined based on (4.2) and (3.4) for $\Psi^{(b)}$ conditional on $\psi_{-j}^{(b-1)} = (\psi_1^{(b)}, \ldots, \psi_{j-1}^{(b)}, \psi_{j+1}^{(b-1)}, \ldots, \psi_n^{(b-1)})$, $\beta^{(b-1)}$ and $\sigma^{2(b)}$. It is derived from the ABC-BNP method as discussed in Chapter 3.3. Here, for each $j = 1, \ldots, n$, we set

$$\psi_j^{(b)} = \begin{cases} \psi_j^* & \text{if } \mathbb{I}(A_j); \\ \psi_j^{(b-1)} & \text{otherwise,} \end{cases} \tag{4.5}$$

where $\mathbb{I}(C) = 1$ if C holds, ψ_j^* is sampled from (3.2), that is,

$$\psi_j^* | \psi_{-j}^{(b-1)}, \sigma^{2(b)}, \beta^{(b-1)} \sim \frac{\alpha_k}{\alpha + n - 1} G_0(\psi_j^*) + \frac{1}{\alpha + n - 1} \sum_{k \neq j} l_k,$$

where $\alpha_k = \alpha$ and $l_k = \delta_{\psi_k}(\psi_j^*)$. Then using the value of ψ_j^*, the probability,

$$p_j^{(b)} = \frac{\exp(X_j^T \beta^{(b-1)} + \psi_j^*)}{1 + \exp(X_j^T \beta^{(b-1)} + \psi_j^*)}$$

is calculated based on the logit link and generate a new sample z_j^* from the Bernoulli distribution with probability $p_j^{(b)}$. For the binary responses, the condition A_j in (4.5) can be defined as $y_j = z_j^*$.

Finally, the distribution of $\beta^{(b)}$ conditional on $\sigma^{2(b)}, \Psi^{(b)}, \beta^{(b-1)}$ is denoted as $s_3(\cdot)$. This distribution is updated using the Metropolis-Hastings algorithm. Let $q(\cdot)$ be the proposal distribution. So, we can generate β^* from the proposal density and set

$$\beta^{(b)} = \begin{cases} \beta^* & \text{with probability } \min\left\{1, \frac{\pi(\beta^* | \Psi^{(b)}, \sigma^{2(b)}) q(\beta^{(b-1)})}{\pi(\beta^{(b-1)} | \Psi^{(b)}, \sigma^{2(b)}) q(\beta^*)}\right\}; \\ \beta^{(b-1)} & \text{otherwise,} \end{cases}$$

where $\pi(\beta | \Psi, \sigma^2)$ is the full conditional distribution of β and the corresponding prior is Normal$(0, d^* \sigma_0^2 I)$, $d^* > 1$ as defined by Kyung et al. (2011).

Another way to define the binary model is to use the scale mixture model and we can develop ABC-BNP method for this. In the next section, we provide a detailed structure of the model and method in terms of probit regression.

4.3 Scale Mixture Model

The scale mixture model under DPM structure is discussed in Basu & Chib (2003). Here, the link function is defined as follows:

$$\begin{aligned} g(p_j) &= \Phi^{-1}(p_j) = \gamma_j \\ \Longrightarrow g^{-1}(\gamma_j) &= \Phi(\gamma_j) = p_j \\ \gamma_j &= f(X_j, \beta, \psi_j) = X_j^T \beta \sqrt{\psi_j}. \end{aligned}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution. Now, under the DPM models, we assume the random effects Ψ sample from the DP with base measure G_0 and the concentration parameter α. Here, we consider a gamma prior for G_0 with shape parameter a_0, scale parameter b_0 and a fixed value of the concentration parameter α. For the regression coefficient β, we assume multivariate normal with mean vector μ and variance-covariance matrix Σ. Hence the model can be written as

$$\begin{aligned} P(Y_j = 1 | X_j, \beta, \psi_j) &= \Phi(X_j^T \beta \sqrt{\psi_j}), j = 1, \ldots, n \\ \psi_j | G &\sim G \\ G &\sim \mathrm{DP}(\alpha, G_0) \\ G_0 &\equiv \mathrm{Gamma}(\psi_j | a_0, b_0) \\ \beta &\sim \mathrm{Normal}(\mu_0, \Sigma_0). \end{aligned} \tag{4.6}$$

In both the cases, we can define the models in terms of the vector of unique values (η) of the random effects(Ψ) to consider the effect of clustering in DPM models. The next section proposes ABC-BNP method for binary scale mixture models.

4.3.1 Proposed Method for Scale Mixture Model

In this setup, we use the ABC-BNP within Gibbs sampler. Here, the ABC-BNP for GLMM is used to obtain the distribution of random effect Ψ. The transition kernel for the scale mixture model is

$$T(\Psi^{(b)},\beta^{(b)}|\Psi^{(b-1)},\beta^{(b-1)}) = \left[\prod_{j=1}^{n} s_1(\psi_j^{(b)}|\psi_{-j}^{(b-1)},\beta^{(b-1)})\right] \times s_2(\beta^{(b)}|\Psi^{(b)},\beta^{(b-1)}), \qquad (4.7)$$

where $\psi_{-j}^{(b-1)} = (\psi_1^{(b)},\ldots,\psi_{j-1}^{(b)},\psi_{j+1}^{(b-1)},\ldots,\psi_n^{(b-1)})$ and $s_1(\cdot)$ is derived from (4.2) and (3.4). According to the ABC-BNP method, for each $j = 1,\ldots,n$, the parameter $\psi_j^{(b)}$ can be defined as

$$\psi_j^{(b)} = \begin{cases} \psi_j^* & \text{if } \mathbb{I}(y_j = z_j^*); \\ \psi_j^{(b-1)} & \text{otherwise,} \end{cases} \qquad (4.8)$$

where $\mathbb{I}(C) = 1$ if C holds and since the response is binary, the condition can be constructed as $y_j = z_j^*$. We sample the candidate value, ψ_j^* given $\psi_{-j}^{(b-1)}$, $\beta^{(b-1)}$ from (3.2) with base distribution G_0 as Gamma(shape = a_0, scale = b_0). So, in the expression (3.2), $\alpha_k = \alpha$, $\theta_j = \psi_j^*$. and $l_k = \delta(\psi_k)$, that is,

$$\psi_j^*|\psi_{-j}^{(b-1)},\beta^{(b-1)} \sim \frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{k\neq j}\delta_{\psi_k}(\psi_j^*).$$

Then using the value of ψ_j^*, we calculate the probability,

$$p_j^{(b)} = \Phi(X_j\beta^{(b-1)}\sqrt{\psi_j^*})$$

based on the probit link and generate a new sample z_j^* from the Bernoulli distribution with probability $p_j^{(b)}$ to check the condition in (4.8).

$s_2(\cdot)$ represents the distribution of the regression coefficient $\beta^{(b)}$ conditional on $\Psi^{(b)}$. This parameter is updated using the random walk within Metropolis-Hastings framework. Given the values of $\Psi^{(b)}$ and $\beta^{(b-1)}$, we generate δ^* from the proposal density $q(\cdot)$ and calculate β^*, that is,

$$\begin{aligned} \delta^* &\sim \text{Normal}(\mu, c\Sigma^{-1}) \\ \beta^* &= \beta^{(b-1)} + \delta^*. \end{aligned}$$

where c is a randomly chosen constant and $\Sigma^{-1} = X^T X$. Now, we set

$$\beta^{(b)} = \begin{cases} \beta^* & \text{with probability } \min\{1, \frac{\pi(\beta^*|\Psi^{(b)})q(\beta^{(b-1)})}{\pi(\beta^{(b-1)}|\Psi^{(b)})q(\beta^*)}\}; \\ \beta^{(b-1)} & \text{otherwise.} \end{cases}$$

where $\pi(\beta|\Psi)$ is the full conditional distribution of β with Normal(μ_0, Σ_0) prior density.

In the next section, we illustrate two examples for binary DP generalized linear mixed models and compare the performance of ABC-BNP with the existing methods.

4.4 Examples

For the random intercept model, we consider the data based on the voting responses of social attitudes in Scotland and the prostate cancer data, known as 'nodal' data is used for scale mixture regression model.

4.4.1 Scottish Social Attitude Study

In the light of the recent political turmoil in Europe centering "BREXIT", it is an interesting study to analyze the response of Scotland to compare the Dirichlet process generalized linear mixed models using slice sampler (as described in Kyung et al. (2011)) and the ABC-BNP. This example is taken from a social science research on voting behavior study of social attitudes in Scotland. The data is available in the Scottish Social Attitudes Survey, 2006 (UK Data Archive Study Number 5840). We used the dataset, `ssas`, from the `glmdm` package in R. In the actual study, 1594 Scottish females, between the age group 18 and 25 had face to face interview based on 669 computer and paper-based questionnaire. However, in our example, we use 13 covariates from the `glmdm` package and randomly choose 200 observations for less computing time.

Here, the response variable is a binary variable which explains if the voters are supported full freedom for Scotland with or without enrollment in the European Union (EU) versus remaining in the UK under changing circumstances. In 2006, approximately 30% respondents voted for full freedom without a part of EU and the rest of voters wanted to ba a part of UK with or without their own elected parliament. The independent variables consist of nominal, binary, ordinal, and count variables. There are one nominal variable,

- `unionsa` union membership at work;

five binary variables:

- `relgsums` identification with the Church of Scotland versus another or no religion,
- `idlosem` the voter agreed with the statement that increased numbers of Muslims in Scotland would erode the national identity,
- `marrmus` the voter would be unhappy or very unhappy if a family member married a Muslim,

- `nhssat` satisfaction or dissatisfaction with the National Health Service,
- `whrbrn` born in Scotland or not;

six ordinal variables:

- `ptyallgs` party allegiance with the ordering of parties given from more conservative to more liberal,
- `ukintnat` agreement that the UK government works in Scotland's long-term interests,
- `natinnat` agreement that the Scottish Executive works in Scotland's long-term interests,
- `voiceuk3` the voter believes that the Scottish Parliament gives Scotland a greater voice in the UK,
- `hincdif2` the degree to which the voter is living comfortably on current income or not (better in the positive direction),
- `hedqual2` the voter's education level;

and one count variable,

- `household` the number of people living in the voter's household.

In this example, we ran two methods, MCMC for DP-GLMM for 50000 iterations with 25000 burn-in periods and ABC-BNP for GLMM for 100000 iterations after 25000 burn-in periods. For both methods, we set the parameters: $\alpha = 10, \mu = 0, \sigma_0^2 = 1, d^* = 5, a_0 = 3$ and $b_0 = 2$. Acceptance rate for ABC-BNP method is 13.3%. Table 4.1 shows the parameter estimates for both methods.

Also, note that ABC-BNP for GLMM took less computing time than MCMC for DP-GLMM (Table 4.2). For a subsample of 200 observations, ABC-BNP method takes around 10 minutes whereas, without ABC method, the simulation takes around 40 minutes. We have also reported

Table 4.1: Comparison of the parameter estimates of β using ABC-BNP and slice sampler based on 200 observations for random intercept model

Coefficient	ABC-BNP		Slice	
	Mean (SD)	95% C.I	Mean (SD)	95% C.I
Intercept	-2.233 (0.865)	(-3.885, -0.850)	-1.582 (0.473)	(-2.519, -0.656)
househld	0.266 (0.199)	(-0.150, 0.634)	0.181 (0.106)	(-0.023, 0.394)
relgsums	-0.273 (0.536)	(-1.420, 0.759)	-0.069 (0.256)	(-0.579, 0.433)
ptyallgs	0.074 (0.062)	(-0.057, 0.197)	0.053 (0.027)	(0.001, 0.107)
idlosem	1.112 (0.693)	(-0.239, 2.505)	0.574 (0.330)	(-0.073, 1.221)
marrmus	0.078 (0.582)	(-1.094, 1.171)	0.043 (0.278)	(-0.511, 0.577)
ukintnat	-1.220 (0.491)	(-2.497, -0.445)	-0.584 (0.190)	(-0.959, -0.218)
natinnat	0.745 (0.332)	(0.090, 1.402)	0.463 (0.171)	(0.135, 0.807)
voiceuk3	0.310 (0.281)	(-0.214, 0.894)	0.133 (0.129)	(-0.118, 0.389)
nhssat	1.114 (0.510)	(0.173, 2.189)	0.571 (0.236)	(0.112, 1.029)
hincdif2	-0.269 (0.295)	(-0.899, 0.283)	-0.125 (0.128)	(-0.376, 0.127)
unionsa	0.028 (0.531)	(-1.050, 1.038)	0.018 (0.256)	(-0.495, 0.513)
whrbrn	-0.868 (0.785)	(-2.554, 0.546)	-0.415 (0.364)	(-1.151, 0.291)
hedqual2	-0.294 (0.122)	(-0.521, -0.017)	-0.132 (0.059)	(-0.247, -0.018)

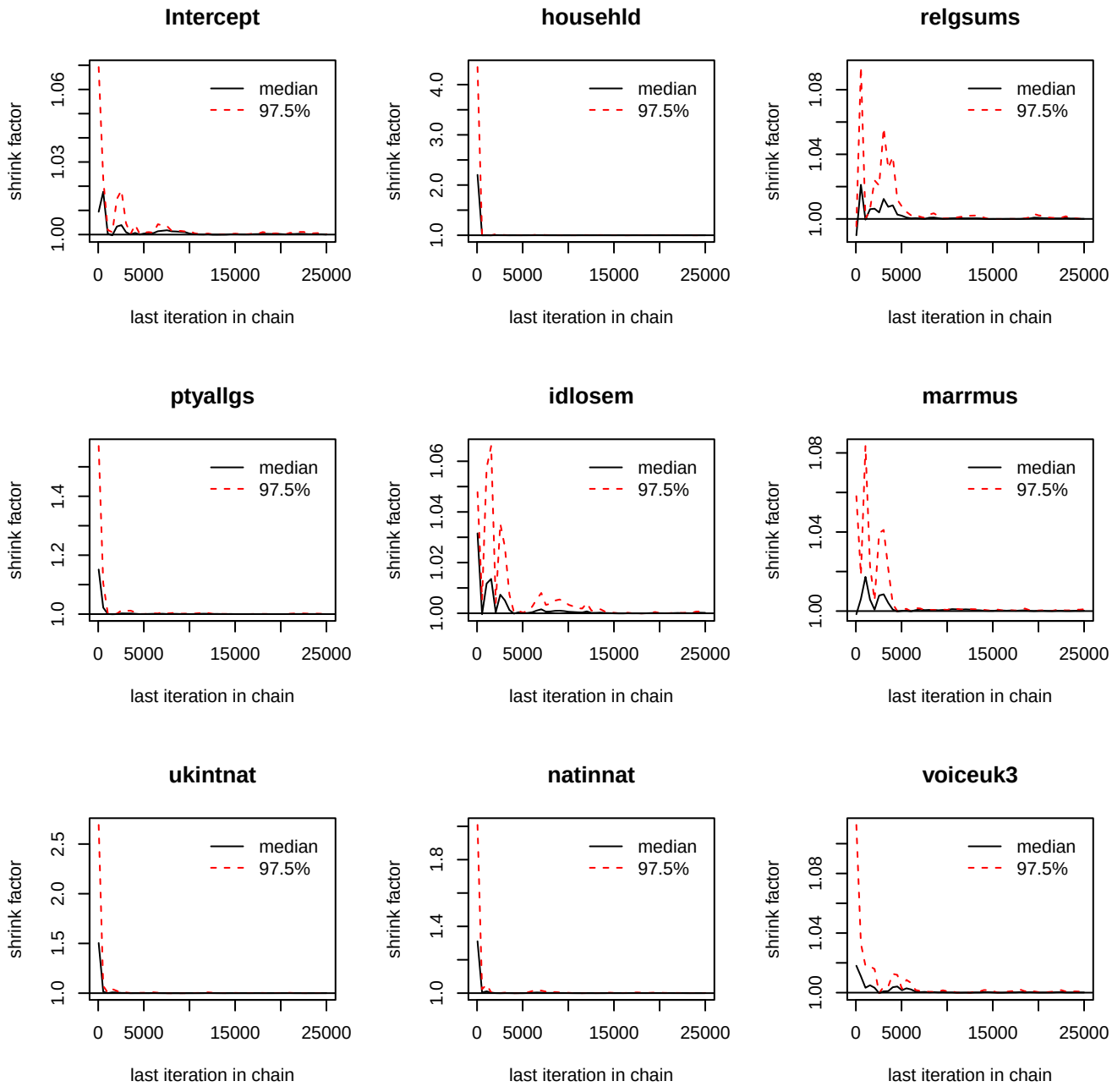
the maximum of log likelihood over MCMC for both sampling procedures in Table 4.3. To know whether this sample is sufficiently close to the posterior, we use Gelman-Rubin plot to see if there is a significant difference between the variance within several chains and we can see from Fig. 4.1 and Fig. 4.2 with 25000 simulation after 25000 burn-in period, the chains are converged after a certain period of time. So, in GLMM setup, ABC-BNP method performs better than Markov chain simulation.

Table 4.2: Computing time for ABC-BNP and slice sampler for random intercept model

Number of samples	ABC-BNP	Slice
Subsample of n= 200	9.3 minutes	40.2 minutes
Full sample of n=1594	3.5 hours	$>$ 10 hours

Table 4.3: Comparison of the maximum log likelihood over the MCMC for random intercept model

Method	Max log likelihood
ABC-BNP	-378.157
Slice	-379.064



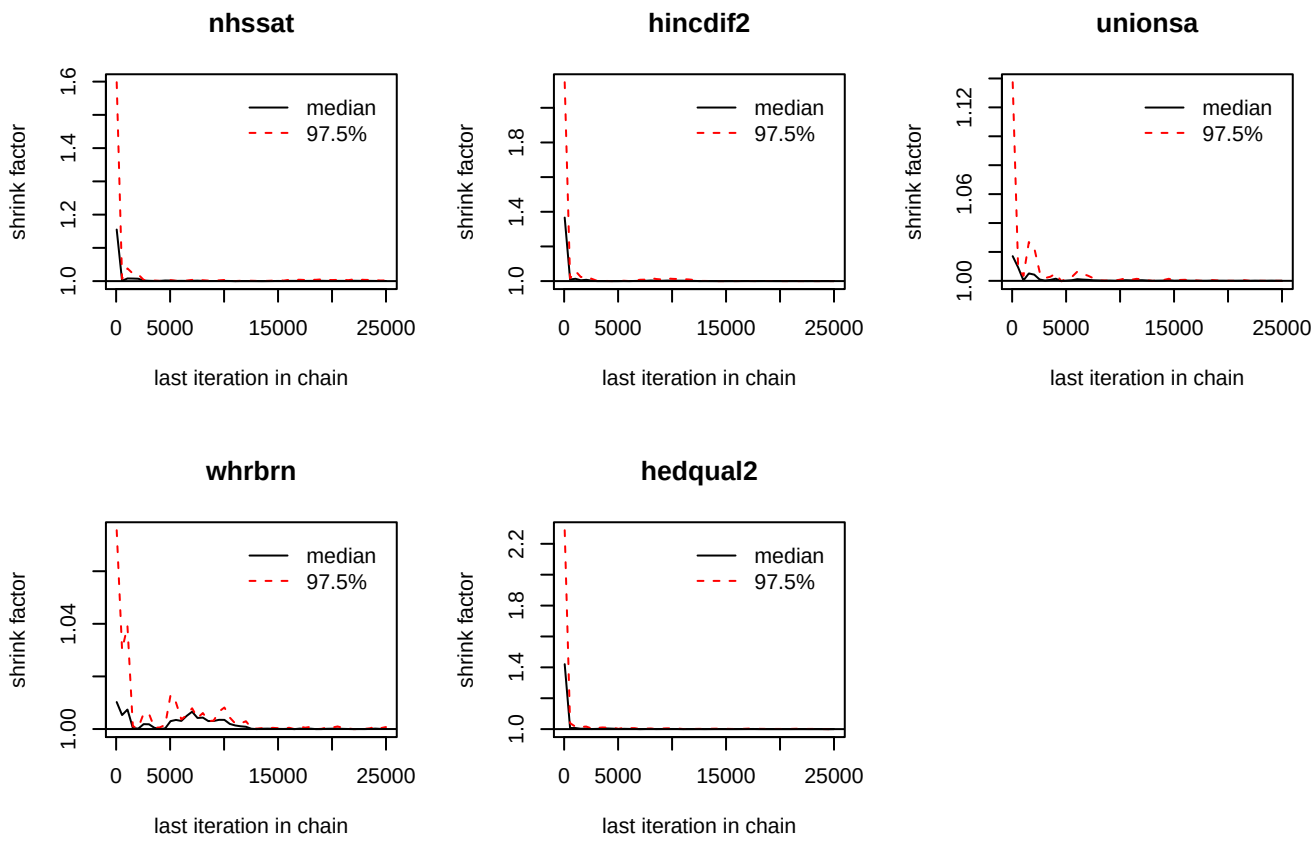
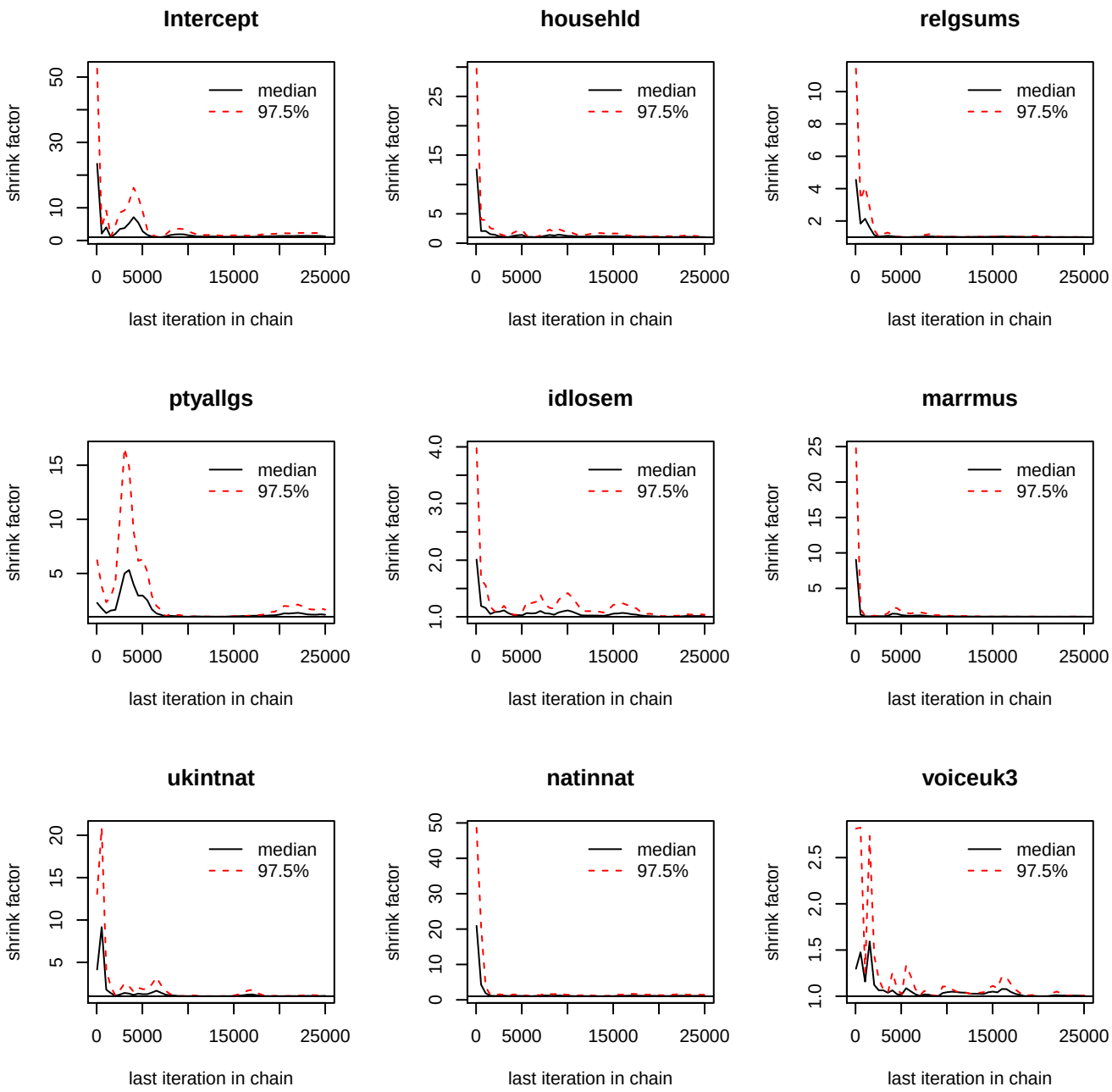


Figure 4.1: Gelman-Rubin plot using slice sampler for 200 observations with 25000 simulations after 25000 burn-in period for random intercept model


Intercept
household
relgsums
ptyallgs
idlosem
marrmus
ukintnat
natinnat
voiceuk3
shrink factor
median
97.5%
last iteration in chain
0
5000
15000
25000

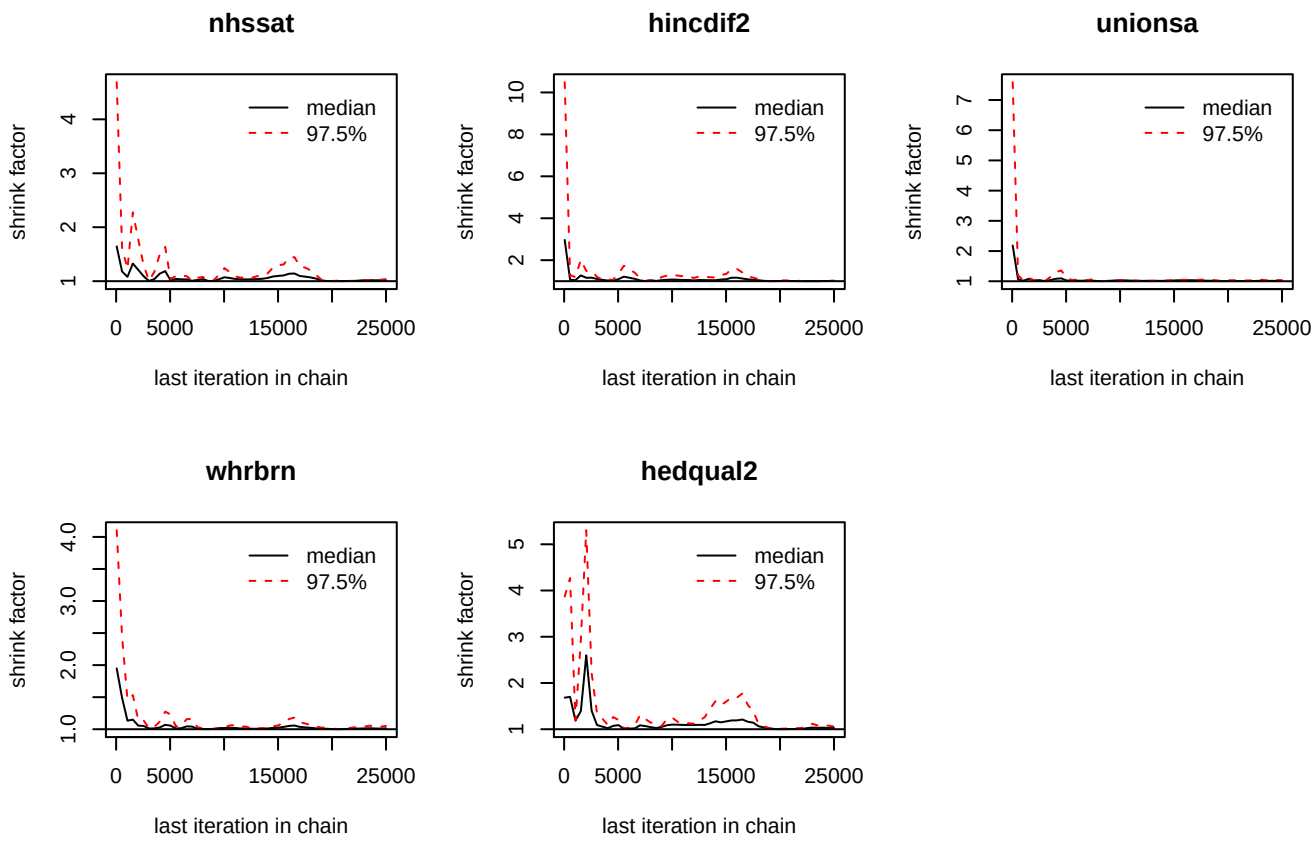


Figure 4.2: Gelman-Rubin plot using ABC-BNP sampler for 200 observations with 25000 simulations after 25000 burn-in period for random intercept model

4.4.2 Nodal Data

Now, we illustrate the idea of generalized linear mixed effects model for the scale mixture regression using a real life example. We consider the nodal data (taken from `DPpackage` in R). The data consists of 53 patients who are diagnosed with prostate cancer and 5 predictor variables which are measured before surgery. Here, the goal is to determine the relationship between nodal involvement and the predictor variables. The response variable, ssln, is an indicator of nodal involvement (1 if cancer had spread to the surrounding lymph nodes and 0 otherwise). The predictor variables are

- `m` intercept term for the design matrix X,
- `acid` level of serum acid phosphatase,
- `xray` X-ray reading, 0 if negative and 1 if positive,
- `size` size and position of the tumor, 0 if small and 1 if large,
- `grade` seriousness of tumor, 1 indicates a more serious case, and
- `age` log of patients age in years at diagnosis.

Here, X_j is the j^{th} vector of the design matrix, X = (m, acid, x-ray, size, grade, age) and $\beta = (\beta_0, \beta_1, \ldots, \beta_5)$ is the corresponding regression parameter vector. Here, we are interested in estimating the regression parameter β. In this example, $n = 53$, $a_0 = 5, b_0 = 1/2, \alpha = 10, \mu = (0, \ldots, 0)$, and c is randomly chosen from $\{0.2, 0.02, 0.002\}$. In our case, we consider the tolerance level, ε as 0.1 and run the simulation $B = 200000$ times with 20000 burn-in periods to get the approximate posterior distribution. In Table 4.4, we compare the estimate of the posterior mean using ABC-BNP with the estimates of β using stick breaking Gibbs sampler (RJAGS software). Here, we can see that the estimates from the ABC-BNP are slightly different from the estimates from the

MCMC. To know whether this sample is sufficiently close to the posterior, we use Gelman-Rubin plot to see if there is a significant difference between the variance within several chains and we can see from Fig. 4.3 and Fig. 4.4 with 45000 simulation after 10000 burn-in period, the chains are converged after a certain period of time. The maximum of log likelihood over MCMC for both sampling procedures are stated in Table 4.5 and for ABC-BNP, posterior reached the maximum point.

Table 4.4: Comparison of the estimates of β for scale mixture model

Coefficient	ABC-BNP		Stick breaking Gibbs	
	Mean (SD)	95% C.I	Mean (SD)	95% C.I
Intercept	7.916 (0.706)	(6.544, 9.178)	7.880 (0.686)	(6.562, 9.232)
Acid	1.631 (0.904)	(-0.121, 3.976)	1.618 (0.474)	(0.713, 2.580)
X-ray	1.222 (1.001)	(-0.761, 3.194)	1.374 (0.338)	(0.744, 2.043)
Size	0.987 (0.991)	(-0.812, 2.924)	1.078 (0.325)	(0.459, 1.741)
Grade	0.472 (0.920)	(-1.244, 2.401)	0.494 (0.323)	(-0.122, 1.140)
Age	-2.579 (0.978)	(-4.451, -0.593)	-2.599 (0.207)	(-3.012, -2.196)

Table 4.5: Comparison of maximum log likelihood over MCMC for scale mixture model

Method	Max log likelihood
ABC-BNP	-20.28
Stick breaking Gibbs	-21.03

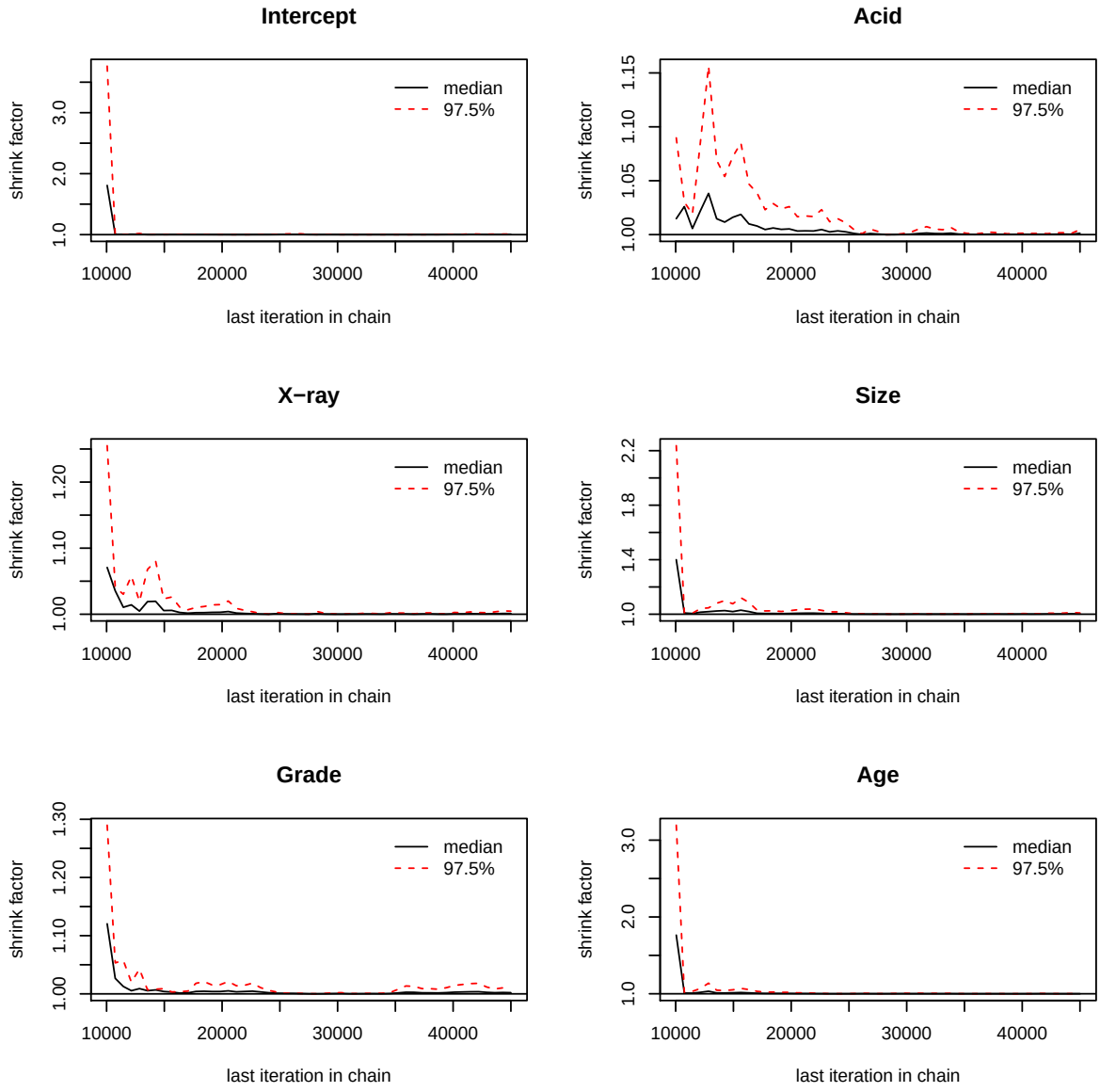


Figure 4.3: Gelman-Rubin plot for scale mixture model using stick breaking Gibbs with 35000 simulation after 10000 burn-in period

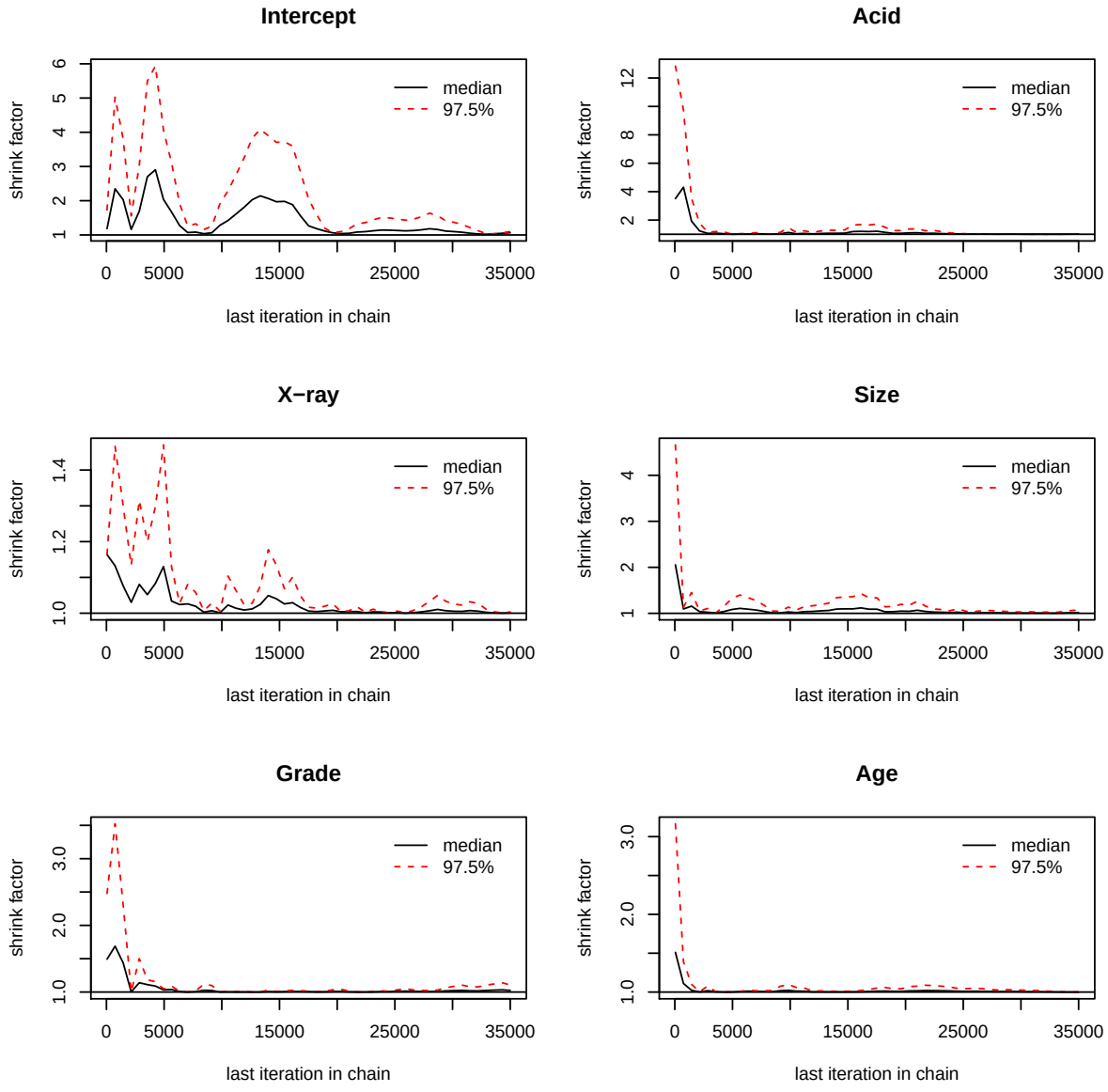


Figure 4.4: Gelman-Rubin plot for scale mixture model using ABC-BNP with 35000 simulation after 10000 burn-in period

CHAPTER 5

ABC-BNP AND SURVIVAL MODELS

5.1 Introduction

Survival data analysis is one of the most extensively used statistical method in Biostatistics and epidemiology. This analysis includes the time to an event, for example, time to death, spread or recurrence of a disease. First, we introduce the basic structure of the survival function and the form used in modeling Bayesian survival data and then, extend the idea to ABC-BNP for different nonparametric survival models.

According to Klein & Moeschberger (2005), suppose T denotes a nonnegative random variable of survival time from a homogeneous population. In the survival models, the probability density (or mass) function $f(t)$ represents the probability of occurrence of an event at time t. Now, the survival function is the probability of an individual surviving after time t and this function is defined as

$$S(t) = 1 - F(t),$$

where $F(t)$ the distribution function.

Censoring is a common characteristics in survival data. It occurs when the events are known to have happened within certain period of time. Right censored and interval censored are two most common feature of the survival data. For the right censored data, an individual followed until the event has occurred, but then leaves the study. In case of interval censored data, the occurrence of the exact time of the event is not known but an interval time is noticed. However, left censored data occurs when an event has already occurred for an individual before that subject is observed in

the study at a certain time. The Likelihood corresponding to various types of censored data can be written as

$$H \propto \prod_{i \in Obs} f(t_i) \prod_{i \in R} S(R_i) \prod_{i \in L} (1 - S(L_i)) \prod_{i \in I} (S(I_{L_i}) - S(I_{R_i})),$$

where R corresponds to the right censored observation, L refers to the left censored data, and (I_L, I_R) indicates the interval censored data.

The study of censored survival data analysis is not straightforward in the frequentist approach. Bayesian procedures are flexible and can easily handle these type of data by using the MCMC methods. Also, in Bayesian context, the prior information can be easily implemented that improves the estimates of the model parameters. In the nonparametric setup, Susarla & Van Ryzin (1976) Ferguson & Phadia (1979). and Kuo et al. (1992) provide the posterior analysis for censored data, especially based on DP prior using the Gibbs sampler. Different survival models under Bayesian setup are discussed in Ibrahim et al. (2005). Since the model structure is complicated, an alternative to the MCMC approach is ABC. In this chapter, we introduce two survival models in Bayesian point of view and implement the ABC-BNP for these models.

5.2 Bayesian Nonparametric Survival Models

Suppose T is a random variable denoting the event times of the subjects defined on $\mathbb{R}^+$. Let $S(t) = P(T > t)$ denote the survival function denoting the probability of a subject surviving until time t. Let $t_1, \ldots, t_n$ be n independent and identically distributed survival times, where t_j is the interval censored data for j^{th} subject in the interval $(a_j, b_j]$. The model is discussed in De Iorio et al. (2009). They have used log transformation on T, that is, $y_j = \log(t_j)$ with y_j are sampled from normal distribution. Here, $X = (X_1, \ldots, X_p)$ and X_j is the vector of covariate corresponding to j^{th} observation. They have considered DPM prior on G for θ_j with base mea-

sure G_0 and concentration parameter α. Here, $\theta_j = (\beta_j, \sigma_j^2)$. The base measure corresponding to β_j is normal with mean vector μ_g, variance-covariance matrix s_g and for σ_j^2, it is inverse-gamma distribution with shape parameter $\tau_1/2$ and rate parameter $\tau_2/2$. A gamma prior on α is taken with shape parameter a_0 and scale parameter b_0. Also, the hyper-priors are assumed for $\mu_g(\text{Normal}(\mu_0, s_0)), s_g(\text{Inverse-Wishart}(\nu, \psi))$, and $\tau_2(\text{Gamma}(\text{shape} = \tau_{s_1}/2, \text{rate} = \tau_{s_2}/2))$. Hence the model can be formalized as follows:

$$\begin{aligned}
T_j &\in (a_j, b_j] \\
Y_j &= \log(T_j) \\
Y_j|X_j, \beta_j, \sigma_j^2 &\sim \text{Normal}(y_j|X_j\beta_j, \sigma_j^2) \\
(\beta_j, \sigma_j^2)|G &\sim G \\
G &\sim DP(\alpha, G_0) \\
G_0 &\equiv \text{Normal}(\beta_j|\mu_g, s_g) \cdot IG(\sigma_j^2|\tau_1/2, \tau_2/2) \\
\alpha &\sim \text{Gamma}(a_0, b_0) \\
\mu_g &\sim \text{Normal}(\mu_0, s_0) \\
s_g &\sim \text{Inverse-Wishart}(\nu, \psi) \\
\tau_2 &\sim \text{Gamma}(\tau_{s_1}/2, \tau_{s_2}/2).
\end{aligned} \tag{5.1}$$

This section develops the ABC-BNP method for nonparametric survival models.

5.2.1 Proposed Method for Bayesian Nonparametric Survival Models

In the Bayesian nonparametric survival models, the random effects are $\beta = (\beta_1, \ldots, \beta_n)$ and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_n^2)$ where β is the functions of mean and σ^2 the variance term of the model. Under nonparametric Bayesian models, $\theta = (\theta_1, \ldots, \theta_n)$ with $\theta_j = (\beta_j, \sigma_j^2), j = 1, \ldots, n$, is assumed to

follow the probability measure G with concentration parameter α and base distribution G_0. For the model in (5.1), the transition kernel can be expressed as

$$\begin{aligned}
&T(\theta^{(b)},\alpha^{(b)},\mu_g^{(b)},s_g^{(b)},\tau_2^{(b)}|\theta^{(b-1)},\alpha^{(b-1)},\mu_g^{(b-1)},s_g^{(b-1)},\tau_2^{(b-1)}) \\
&= \left[\prod_{j=1}^{n} s_1(\theta_j^{(b)}|\theta_{-j}^{(b-1)},\alpha^{(b-1)},\mu_g^{(b-1)},s_g^{(b-1)},\tau_2^{(b-1)})\right] \times s_2(\alpha^{(b)}|\theta^{(b)},\alpha^{(b-1)},\mu_g^{(b-1)},s_g^{(b-1)},\tau_2^{(b-1)}) \\
&\times s_3(\mu_g^{(b)}|\theta^{(b)},\alpha^{(b)},\mu_g^{(b-1)},s_g^{(b-1)},\tau_2^{(b-1)}) \times s_4(s_g^{(b)}|\theta^{(b)},\alpha^{(b)},\mu_g^{(b)},s_g^{(b-1)},\tau_2^{(b-1)}) \\
&\times s_5(\tau_2^{(b)}|\theta^{(b)},\alpha^{(b)},\mu_g^{(b)},s_g^{(b)},\tau_2^{(b-1)}),
\end{aligned} \tag{5.2}$$

where $\theta_{-j}^{(b-1)} = (\theta_1^{(b)},\ldots,\theta_{j-1}^{(b)},\theta_{j+1}^{(b-1)},\ldots,\theta_n^{(b-1)})$ and $s_1(\cdot)$ is derived from (3.2) and (3.4) based on $(\alpha^{(b-1)},\mu_g^{(b-1)},s_g^{(b-1)},\tau_2^{(b-1)})$. Here, for each $j = 1,\ldots,n$, $\theta_j^{(b)}$ can be expressed as follows

$$\theta_j^{(b)} = \begin{cases} \theta_j^* & \text{if } \mathbb{I}(A_j); \\ \theta_j^{(b-1)} & \text{otherwise,} \end{cases} \tag{5.3}$$

where $\mathbb{I}(C) = 1$ if C holds and we sample a candidate value, θ_j^* from the DP prior with base measure as normal-inverse-gamma distribution. So, in (3.2), $\alpha_k = \alpha$ and $l_k = \delta_{\theta_k}(\theta_j^*)$, that is,

$$\theta_j^*|\theta_{-j}^{(b-1)},\alpha^{(b-1)},\mu_g^{(b-1)},s_g^{(b-1)},\tau_2^{(b-1)} \sim \frac{\alpha}{\alpha+n-1}G_0(\theta_j^*) + \frac{1}{\alpha+n-1}\sum_{k\neq j}\delta_{\theta_k}(\theta_j^*).$$

The choice of A_j in (5.3) is based on j^{th} observation of the data. If the response is observed and continuous, the condition is defined as

$$|y_j - z_j^*| < \varepsilon,$$

where ε is a predefined threshold value and

$$z_j^* \sim f(z_j|\theta_j^*).$$

For the censored term, the condition $\mathbb{I}(A_j)$ in (3.4) is constructed depending upon the choices of censored survival observation. Let L_{y_j}, R_{y_j}, and $(I_{L_{y_j}}, I_{R_{y_j}})$ be the left, right, and interval censored data corresponding to observed observations. $L_{z_j^*}, R_{z_j^*}$, and $(I_{L_{z_j^*}}, I_{R_{z_j^*}})$ denote the respective generated observations. Hence the choices of A_j for the censored survival data are as follows:

$$\begin{aligned} \text{Right censored:} \quad & R_{y_j} < R_{z_j^*} \\ \text{Left censored:} \quad & L_{y_j} > L_{z_j^*} \\ \text{Interval censored:} \quad & I_{L_{y_j}} > I_{L_{z_j^*}} \text{ and } I_{R_{y_j}} < I_{R_{z_j^*}}. \end{aligned} \tag{5.4}$$

Hence for the censored data,

$$z_j^* \sim S(z_j|\theta_j^*),$$

where $S(\cdot)$ is the survival sampling distribution of the censored terms and the condition is taken from (5.4). In (5.2), $s_2(\cdot)$, $s_3(\cdot)$, $s_4(\cdot)$, and $s_5(\cdot)$ denote the full conditional distributions of $(\alpha^{(b)}|\theta^{(b)}, \alpha^{(b-1)}, \mu_g^{(b-1)}, s_g^{(b-1)}, \tau_2^{(b-1)})$, $(\mu_g^{(b)}|\theta^{(b)}, \alpha^{(b)}, \mu_g^{(b-1)}, s_g^{(b-1)}, \tau_2^{(b-1)})$, $(s_g^{(b)}|\theta^{(b)}, \alpha^{(b)}, \mu_g^{(b)}, s_g^{(b-1)}, \tau_2^{(b-1)})$, and $(\tau_2^{(b)}|\theta^{(b)}, \alpha^{(b)}, \mu_g^{(b)}, s_g^{(b)}, \tau_2^{(b-1)})$, respectively. In this step, $\alpha, \mu_g, s_g, \tau_2$ are updated using the Metropolis-Hastings algorithm. Let $\phi = (\alpha, \mu_g, s_g, \tau_2)$ and $q(\cdot)$ be the proposal density. So, $\phi^{(b)}$ can be defined as

$$\phi^{(b)} = \begin{cases} \phi^* & \text{with probability } \min\left\{1, \frac{\pi(\phi^*|\theta^{(b)}) q(\phi^{(b-1)})}{\pi(\phi^{(b-1)}|\theta^{(b)}) q(\phi^*)}\right\}; \\ \phi^{(b-1)} & \text{otherwise,} \end{cases}$$

where $\pi(\phi|\theta)$ is the full conditional posterior distribution of ϕ.

Another popular model under survival setting is the recurrent data models which introduce the random effects through frailty term to represent association and unobserved heterogeneity in the model.

5.3 Bayesian Nonparametric Survival Models for Recurrent Data

One of the popular models is the recurrent survival model with gap time between two consecutive events. This type of models include recurrent occurrence of events. Cook & Lawless (2007) explore the idea of gap time models of recurrent data and analyze these models under different circumstances. Let $W_{jk}, k = 1,\ldots,m_j, j = 1,\ldots,n$ be the k^{th} gap time of j^{th} subject and we are considering log-normal models as described in Cook & Lawless (2007). So, we can define $Y_{jk} = \log(W_{jk})$ with $Y_j = (Y_{j1},\ldots,Y_{jn_j})$ is the vector of log of gap time events of j^{th} subject. The length of $Y_j, j = 1,\ldots,n$ may be different for each subject. So the likelihood of the k^{th} gap time of j^{th} subject is normal distribution with variance σ_e^2. In this step, we introduce a random effect u_j to the model. Hence, $Y_{jk} = \mu_y + u_j + \varepsilon_{jk}$ and the mean becomes $\mu_y + u_j$, where μ_y is the intercept of the model and ε_{jk} is the independent and identically distributed random error with Normal(0, σ_e^2).

In this model, u_j is defined as a frailty term. The frailty, an unobserved random effect, illustrates the hazard function of a cluster or group of clusters having multiplicative frailty factor. The concept of random effect in the survival model was introduced by Beard (1959) to get better model of mortality. In their model, they have attached a frailty term for each subject. Vaupel et al. (1979) mentioned the term frailty for the first time to explain the mortality data more accurately and proposed a model based on univariate frailty for each individuals. In practice, two classifications of frailty models are used in survival analysis. One is the univariate frailty models which deal with univariate survival times and another one is the multivariate frailty models which consider

multivariate survival times. In both cases, the models account for the unobserved heterogeneity in the population. Duchateau & Janssen (2007) discusses different choices of frailty models and few techniques to calculate estimators.

Now, in this study, we suggest DPM models for the frailty terms. Alternatively, PYP structure can be used for this model. So, for each subject, the frailty term can be modeled as

$$\begin{aligned} u_j &\sim G \\ G &\sim DP(M, G_0), \end{aligned}$$

where M represents the concentration parameter and G_0 is the base measure. In this case, the conditional distribution of u_j given $u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_n$ can be defined as

$$u_j | u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_n \sim \frac{\alpha}{\alpha + n - 1} G_0(u_j) + \frac{1}{\alpha + n - 1} \sum_{k \neq j} \delta_{\{u_k\}}(u_j), \tag{5.5}$$

where $\delta_{\{u_k\}}(u_j)$ refers to the degenerate distribution at the point u_j. In this step, we have also assigned prior for the models parameters. The prior for μ is Normal(μ_0, σ_0^2) and inverse-gamma distributions are attached to the scale parameters. Hence, the DPM frailty model for recurrent data can be expressed as follows:

$$
\begin{aligned}
Y_{jk} &= \log(W_{jk}), j = 1,\ldots,n; k = 1,\ldots,m_j \\
Y_{jk} &= \mu_y + u_j + \varepsilon_{jk} \\
\varepsilon_{jk} &\overset{iid}{\sim} \text{Normal}(0,\sigma_e^2) \\
u_j|G &\sim G \\
G &\sim DP(M,G_0) \\
G_0 &\equiv \text{Normal}(u_j|\mu,\sigma^2) \\
\mu_y &\sim \text{Normal}(\mu_0,\sigma_0^2) \\
\sigma^2 &\sim IG(a_u,b_u) \\
\sigma_e^2 &\sim IG(a_e,b_e).
\end{aligned} \tag{5.6}
$$

5.3.1 Proposed Method for Recurrent Data Model

For this model, the ABC-BNP is used to estimate the posterior analysis of the frailty term $u_j, j = 1,\ldots,n$. For this model, the transition kernel can be stated as

$$
\begin{aligned}
&T(u^{(b)},\mu_y^{(b)},\sigma_e^{2(b)},\sigma^{2(b)}|\theta^{(b-1)},\mu_y^{(b-1)},\sigma_e^{2(b-1)},\sigma^{2(b-1)}) = \left[\prod_{j=1}^{n} s_1(u_j^{(b)}|u_{-j}^{(b-1)},\mu_y^{(b-1)},\sigma_e^{2(b-1)},\sigma^{2(b-1)})\right] \\
&\times s_2(\mu_y^{(b)}|u^{(b)},\mu_y^{(b-1)},\sigma_e^{2(b-1)},\sigma^{2(b-1)}) \times s_3(\sigma_e^{2(b)}|u^{(b)},\mu_y^{(b)},\sigma_e^{2(b-1)},\sigma^{2(b-1)}) \\
&\times s_4(\sigma^{2(b)}|u^{(b)},\mu_y^{(b)},\sigma_e^{2(b)},\sigma^{2(b-1)}).
\end{aligned} \tag{5.7}
$$

where $u_{-j}^{(b-1)} = (u_1^{(b)},\ldots,u_{j-1}^{(b)},u_{j+1}^{(b-1)},\ldots,u_n^{(b-1)})$ and $s_1(\cdot)$ is derived from (5.5) and (3.4). For each j, the recurrent model consists of m_j observed values and corresponding censored observation,

denoted by c_{y_j}. For $j = 1, \ldots, n$, given the values of $u^{(b-1)}, \mu_y^{(b-1)}, \sigma_e^{2(b-1)}$, and $\sigma^{2(b-1)}$, the frailty term can be defined as

$$u_j^{(b)} = \begin{cases} u_j^* & \text{if } \mathbb{I}(A_j); \\ u_j^{(b-1)} & \text{otherwise,} \end{cases} \tag{5.8}$$

where $\mathbb{I}(C) = 1$ if C holds and we sample the candidate value, u_j^* from the DP prior with concentration parameter M and base distribution G_0 as the normal density. So, in the expression (3.2), $\alpha_k = M$, $\theta_j = u_j^*$, and $l_k = \delta(u_k)$, that is,

$$u_j^* | u_{-j}^{(b-1)}, \mu_y^{(b-1)}, \sigma_e^{2(b-1)}, \sigma^{2(b-1)} \sim \frac{M}{M+n-1} G_0(u_j^*) + \frac{1}{M+n-1} \sum_{k \neq j} \delta_{u_k}(u_j^*).$$

Let z_j^* and $c_{z_j^*}$ be the vector of the generated values and corresponding censored observation, respectively. These observations are drawn from the following function:

$$(f(z_j^*))^{(\delta_j = 1)} (S(c_{z_j^*}))^{(\delta_j = 0)}, \tag{5.9}$$

where $\delta_j = 1$ if the j^{th} observation is observed, otherwise $\delta_j = 0$ for the censored data with sampling distribution $S(\cdot)$. The condition in (5.8) is based on $y_j, z_j^*, c_{y_j}, c_{z_j^*}$ and

$$\begin{aligned} z_j^* &\sim f(z_j | u_j^*, \phi_j^*), \text{ if observed} \\ c_{z_j^*} &\sim S(c_{z_j} | u_j^*, \phi_j^*), \text{ if censored.} \end{aligned}$$

We update the frailty term u_j in ABC step conditioning on the fact that the absolute value of the observed and the generated observations must be less than a predefined threshold value ε, that is, if the response is observed and continuous, the condition is defined as

$$|y_j - z_j^*| < \varepsilon,$$

where ε is a predefined threshold value for ABC-BNP and for the censored term, the condition in (5.4) is applied based on the data.

In (5.7), $s_2(\cdot)$ represents the full conditional of $(\mu_y^{(b)}|u^{(b)},\mu_y^{(b-1)},\sigma_e^{2(b-1)},\sigma^{2(b-1)})$, $s_3(\cdot)$ defines the full conditional of $(\sigma_e^{2(b)}|u^{(b)},\mu_y^{(b)},\sigma_e^{2(b-1)},\sigma^{2(b-1)})$, and $s_4(\cdot)$ is the full conditional distribution of $(\sigma^{2(b)}|u^{(b)},\mu_y^{(b)},\sigma_e^{2(b)},\sigma^{2(b-1)})$. Here, $\phi=(\mu_y,\sigma^2,\sigma_e^2)$ is updated using the Metropolis-Hastings algorithm with acceptance probability

$$\min\left\{1,\frac{\pi(\phi^*)L(\phi^*|y)q(\phi^*|\phi)}{\pi(\phi)L(\phi|y)q(\phi|\phi^*)}\right\},$$

where

$$\begin{aligned}
\pi(\cdot) &\equiv \text{Normal}(\mu_0,\sigma_0^2)\text{Inverse-Gamma}(a_u,b_u)\text{Inverse-Gamma}(a_0,b_0),\\
L(\cdot|y) &= \prod_{j=1}^{n}[\{\prod_{k=1}^{m_j} Normal(\mu_y+u_k,\sigma^2)\}\cdot Normal(\mu_y+u_j,\sigma^2)],\\
q(\cdot) &\equiv \text{Normal}(\mu_{0q},\sigma_{0q}^2)\text{Inverse-Gamma}(a_{uq},b_{uq})\text{Inverse-Gamma}(a_{0q},b_{0q}),
\end{aligned}$$

and ϕ^* is generated from the proposal density $q(\cdot)$. Given the values of $u^{(b)}$, $\mu_y^{(b-1)}$, $\sigma_e^{2(b-1)}$, and $\sigma^{2(b-1)}$, the distribution of ϕ can be stated as

$$\phi^{(b)}=\begin{cases}\phi^* & \text{with probability } \min\{1,\frac{\pi(\phi^*|u^{(b)})q(\phi^{(b-1)})}{\pi(\phi^{(b-1)}|u^{(b)})q(\phi^*)}\};\\ \phi^{(b-1)} & \text{otherwise.}\end{cases}$$

where $\pi(\phi|u)$ is the full conditional posterior distribution of $\phi=(\mu_y,\sigma_e^2,\sigma^2)$. This method also converges to the target posterior distribution.

5.4 Examples

5.4.1 Deterioration Data

The 'deterioration' data in `DPpackage` in R examines the time to cosmetic deterioration of the breast for women with early stage of breast cancer. The patients had experienced a lumpectomy, for two treatments, radiation and radiation with chemotherapy as both of these treatments are considered powerful in preventing recurrence of the early stage of cancer. The data is taken from a retrospective study with 46 patients who experienced radiation only and 48 who received radiation with chemotherapy. Then the patients went to a clinic to determine the occurrence of retraction. If the result is positive, the time of interval of the present and last visits are taken to be the time of retraction. The data set consists of three variables,

- `left` left limit of the interval
- `right` right limit of the interval
- `trt` treatment (0 = radiation only, 1 = radiation with chemotherapy)

Here, the unknown limits are coded as -999. In this example, we assume $a_0 = 10, b_0 = 1, \tau_1 = 6.01, \tau_{s_1} = 6.01, \tau_{s_2} = 2.01, \mu_0 = (3, -.5), s_0 = diag(1,1), \nu = 4$, and $\psi^{-1} = diag(1,1)$. For ABC-BNP, we run the simulation $B = 100000$ times with 20000 burn-in period to get the approximate posterior density. The estimates and the distribution of the parameters are given in Table 5.1 and Fig. 5.1. Here, the estimates of ABC-BNP are compared with the stick braking Gibbs sampler (RJAGS software). The estimates are very close to each other and 95% credible intervals are slightly smaller for ABC-BNP. Gelman Rubin diagnostic is used to check for lack of convergence. It calculates both the between and within chain variance and assesses whether they are different

from each other. We can see from Fig. 5.2 and Fig. 5.3 for 20000 simulation with 10000 burn-in period, the chains are converged after a certain period of time.

Table 5.1: Comparison of the estimates of the parameters for nonparametric Bayesian survival model

Coefficient	ABC-BNP		Stick braking Gibbs	
	Mean (SD)	95% C.I	Mean (SD)	95% C.I
α	10.04 (3.23)	(4.86, 17.58)	9.99 (3.18)	(4.83, 17.10)
$\mu_g[1]$	3.11 (1.00)	(1.22, 5.06)	2.99 (1.01)	(1.01, 4.98)
$\mu_g[2]$	-0.39 (0.99)	(-2.34, 1.58)	-0.50 (0.99)	(-2.44, 1.43)
$s_g[1,1]$	2.01 (1.56)	(0.63, 5.89)	1.97 (1.98)	(0.04, 7.41)
$s_g[1,2]$	0.01 (1.09)	(-2.06, 2.04)	0.01 (1.39)	(-2.96, 2.93)
$s_g[2,2]$	2.06 (1.86)	(0.61, 6.34)	1.98 (1.97)	(0.05, 7.36)
τ_2	3.01 (1.73)	(0.65, 7.24)	3.01 (1.73)	(0.62, 7.25)

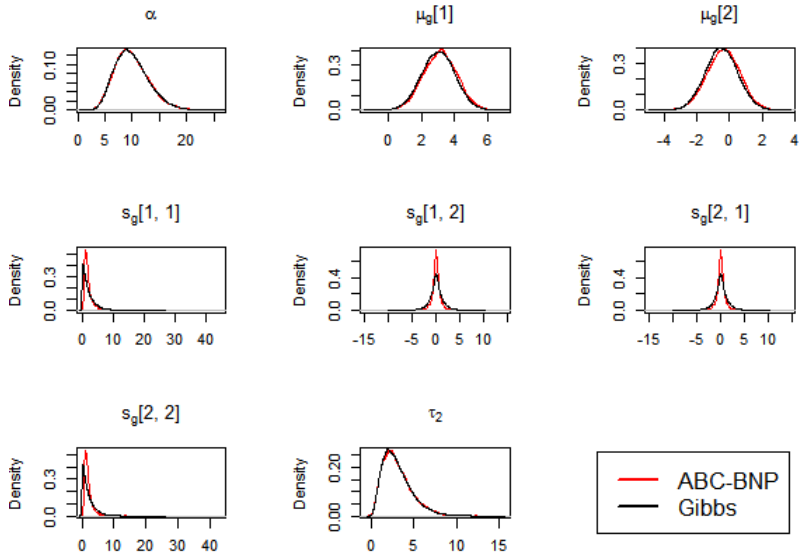


Figure 5.1: Comparison of the distributions of the parameters for the nonparametric Bayesian survival model

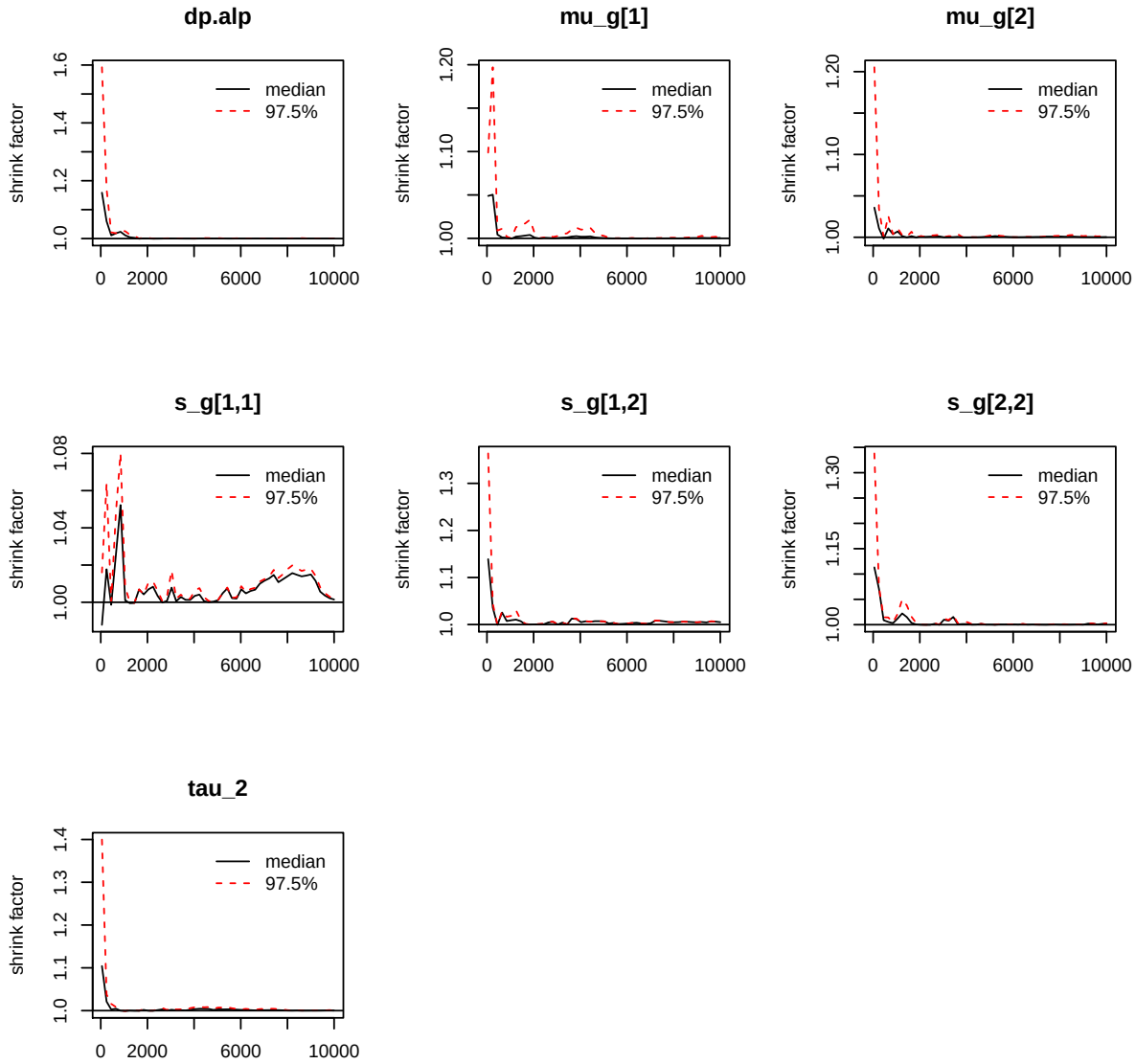


Figure 5.2: Gelman-Rubin plot for nonparametric Bayesian survival model using ABC-BNP with 10000 simulation after 10000 burn-in period

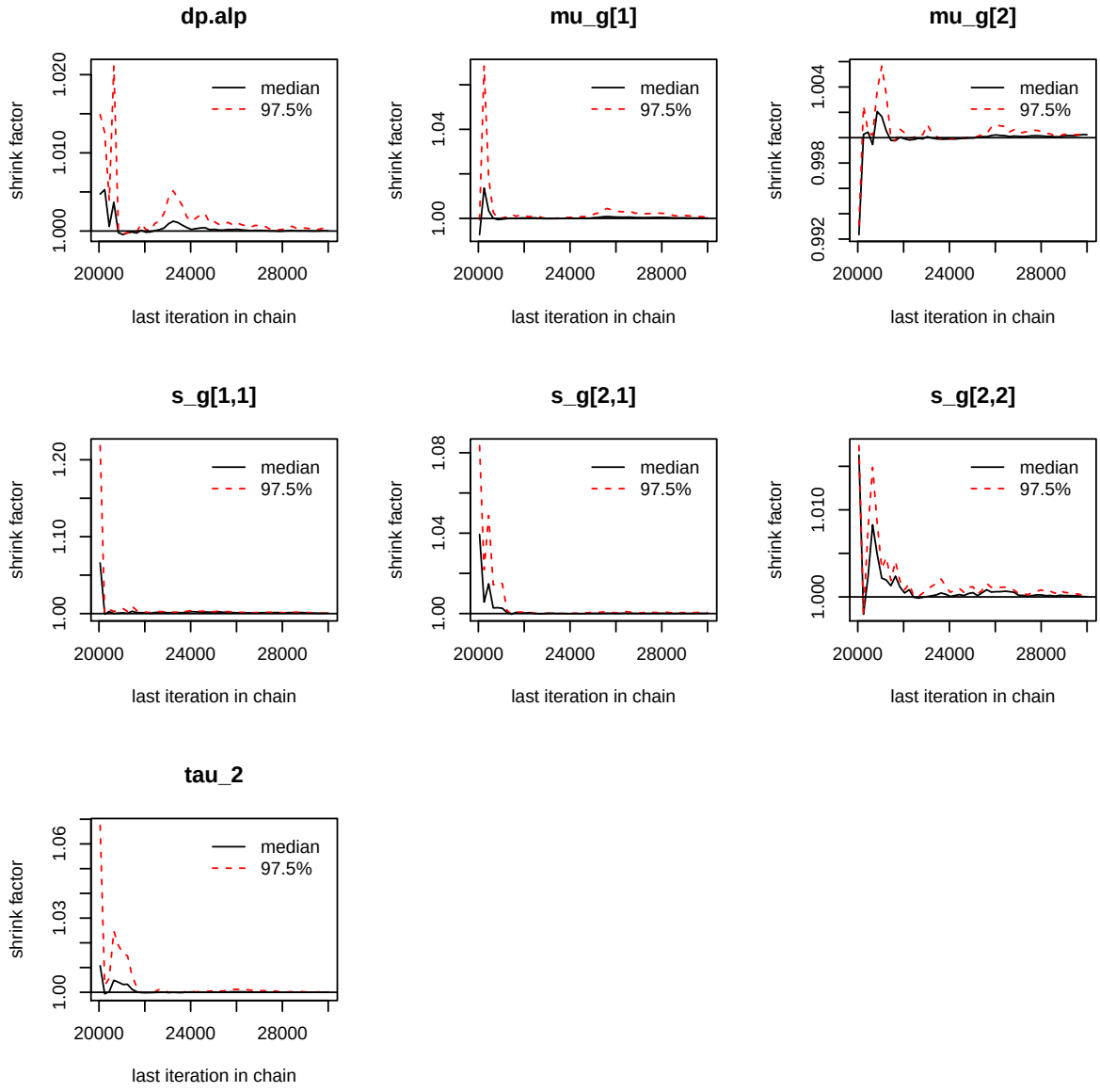


Figure 5.3: Gelman-Rubin plot for nonparametric Bayesian survival model using stick breaking Gibbs with 10000 simulation after 10000 burn-in period

5.4.2 Bowel Motility Cycles

Here, we have used the bowel motility data which is discussed in Cook & Lawless (2007) for the recurrent data model. 19 healthy individuals were given a standard meal at 6:00 pm to induce a "fed state" and then intraluminal pressure was observed for each individual overnight for 13 hours and 40 minutes. The migrating motor complex (MMC) which is the time between two successive fasting cycles were observed for each individual and then calculated the time gaps between two consecutive cycles. The gaps are defined as $W_{jk}, j = 1,\ldots,19; k = 1,\ldots,n_j$. In the data, the last MMC differences are censored. The data set consists of the following variables

- `time` duration of a cycle,
- `status` Censored or not,
- `enum` number of cycles for an individual.

In this example, we choose $\mu = 0, \mu_0 = 0, \sigma_0^2 = 1, a_u = 1, b_u = 1/10, a_0 = 1, b_0 = 1/10, \mu_{0q} = \text{mean}(Y), \sigma_{0q}^2 = 3, a_{uq} = 4, b_{uq} = 1/10, a_{0q} = 4, b_{0q} = 1/10$, and $\varepsilon = 0.005$. The simulation for ABC-BNP is iterating $B = 100000$ times with 20000 burn-in period to get the estimates of the parameters. It is observed that the acceptance rate 11.7% and the computing time is 1.4 minutes for ABC-BNP. Table 5.2 shows the comparison of estimated parameter values. Table 5.3 and Fig. 5.4 provide the comparison and distribution of log likelihoods over MCMC. It is noted that maximum of log likelihood reached for ABC-BNP.

Table 5.2: Comparison of the parameters for recurrent data model

	ABC-BNP	Stick braking Gibbs
$\hat{\mu_y}$	4.11	4.26
$\hat{\sigma}_u$	0.19	0.16
$\hat{\sigma}$	0.80	0.79

Table 5.3: Comparison of maximum log likelihood over MCMC for recurrent data model

Method	Max log likelihood
ABC-BNP	-113.9
Stick braking Gibbs	-114.9

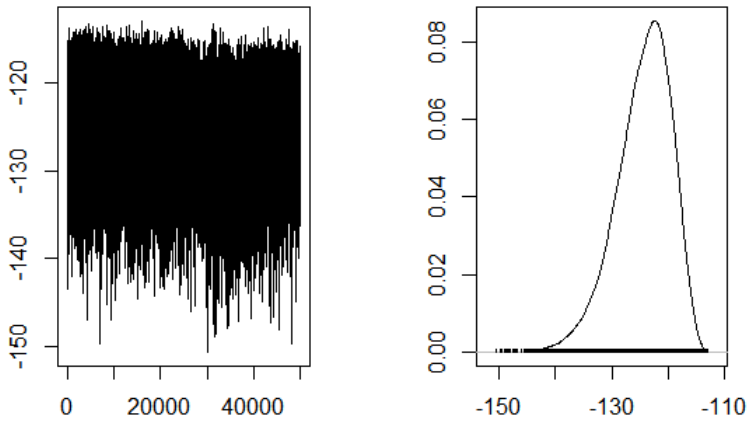


Figure 5.4: Trace plot and the distribution of log likelihood for ABC-BNP recurrent data model

CHAPTER 6

ABC-BNP FOR PITMAN-YOR PROCESS

6.1 Introduction

Pitman-Yor process (PYP) is a generalization of the Dirichlet process prior. As we discussed in Section 2.3.2, $\mathcal{PY}(d, \alpha, G_0)$ is a probability measure in the space of probability measures as DP. Here, the discount parameter d, $0 \leq d < 1$ and the concentration parameter α tune the variability of generated measures around the base measure G_0.

DP provides a flexible model for modeling data with clusters. But, DP models are not appropriate if the number of clusters follow a power-law. PYP is useful tool to cluster data that captures power law behavior. The PYP is an extension to the DP that allows heavier-tailed distributions over partitions. Let m be the number of unique values in the process. The probabilities associated with each cluster decrease almost exponentially for DP and the expected value can be expressed as $E(m) = O(\alpha \log n)$. Instead, in PYP, the probabilities do not drop down exponentially and $E(m) = O(\alpha n^d)$ which follows the power law that has a heavier tail than DP.

Suppose $Y = (Y_1, \ldots, Y_n)$ represents continuous response variable on n observations. We assume that each $Y_j, j = 1, \ldots, n$, are generated from the distribution $f(\cdot)$ with parameter θ_j. We assume that θ_j's are sampled independently and identically from a random distribution G. Here,

G is PYP prior with discount parameter d, concentration parameter α, and base distribution G_0. Hence the PYM model is defined as follows:

$$\begin{aligned} Y_j|\theta_j &\sim \mathrm{f}(y_j|\theta_j) \\ \theta_j|G &\sim G \\ G &\sim \mathcal{PY}(d,\alpha,G_0). \end{aligned}$$

6.2 ABC-BNP for Pitman-Yor Process

For the posterior inference, we also combine the idea of ABC-BNP method with PYM which can be applied in both conjugate and non-conjugate cases. So, for each observation, $j = 1,\ldots,n$, θ_j can be obtained by using the Pólya urn scheme and the corresponding prior distributions can be defined by the full conditionals:

$$\pi(\theta_j|\theta_{-j}) = \frac{\alpha+dm}{\alpha+n-1}G_0(\theta_j) + \frac{1}{\alpha+n-1}\sum_{k\neq j}(n_k-d)\delta_{\eta_k}(\theta_j), \tag{6.1}$$

where $\theta_{-j} = \{\theta_1,\ldots,\theta_{j-1},\theta_{j+1},\ldots,\theta_n\}$, $\delta(\theta_l)$ is the degenerate distribution at the point θ_l and $\{\eta_1,\ldots,\eta_m\}$ are the unique values of $\{\theta_1,\ldots,\theta_n\}$ with corresponding frequency n_k. So, the conditional posterior distribution can be expressed as

$$\theta_j|\theta_{-j},y_j \sim n_c(\alpha+dm)P_j\int f(y_j|\theta)dG_0(\theta) + n_c\sum_{k\neq j}(n_k-d)f(y_k|\eta_k)\delta_{\eta_k}(\theta_j),$$

where P_j is the posterior density based on the base measure. For j^{th} step of ABC-BNP, we sample a candidate value, θ_j^* from the prior in (6.1) and using the value of θ_j, generate a new sample from the sampling distribution $f(z_j|\theta_j)$. Finally, set $\theta_j^{(b)} = \theta_j^*$ with probability

$$\alpha(\theta_j^{(b)}, \theta_j) = \mathbb{I}(A_j), \tag{6.2}$$

where $\mathbb{I}(\cdot)$ is an indicator function and A_j is a model dependent condition based on y_j and z_j^*, for example, if the data is binary, A_j can be defined as $y_j = z_j^*$. So, for this model, the transition kernel of θ can be written as

$$T(\theta^{(b)}|\theta^{(b-1)}) = \prod_{j=1}^{n} s(\theta_j^{(b)}|\theta_{-j}^{(b-1)}), \tag{6.3}$$

where $\theta_{-j}^{(b-1)} = (\theta_1^{(b)}, \ldots, \theta_{j-1}^{(b)}, \theta_{j+1}^{(b-1)}, \ldots, \theta_n^{(b-1)})$ and $s(\cdot)$ can be defined in terms of (6.1) and (6.2). Since the method is based on ABC-MCMC, the posterior converges to $\theta_1, \ldots, \theta_n$ (as described in Marjoram et al. (2003)).

6.3 Simulation Study

In this section, we consider three simulation studies for PYM models.

6.3.1 Data Generation: Normal

For this data, we simulate the data from the normal distribution with mean 0 and variance 2. Suppose $Y = (Y_1, \ldots, Y_n)$ represents continuous response variable on n observations. We assume that each Y_j are normal distribution with unknown mean θ_j and known variance σ^2. We assume that θ_j are sampled independently and identically from a distribution G. Here, G is PYP prior with

discount parameter d, concentration parameter α, and base distribution G_0. For the normal model, we can specify a conjugate base measure for θ_j. Now, the simplest choice is the normal distribution with mean μ_0 and variance σ_0^2. The prior for μ_0 is assigned to be normal(μ_p, σ_p^2). Hence the models is defined as follows:

$$\begin{aligned} Y_j|\theta_j &\sim \text{Normal}(y_j|\theta_j, \sigma^2) \\ \theta_j|G &\sim G \\ G &\sim \mathcal{PY}(d, \alpha, G_0) \\ G_0 &\equiv \text{Normal}(\theta_j|\mu_0, \sigma_0^2) \\ \mu_0 &\sim \text{Normal}(\mu_0|\mu_p, \sigma_p^2). \end{aligned}$$

In this example, $n = 100, \sigma^2 = 1, \sigma_0^2 = 1, \mu_p = 0, \sigma_p^2 = 1, d = 0.5$, and $\alpha = 1$. For ABC-BNP, we choose the tolerance level, ε as 0.05 and run the simulation $B = 100000$ times with 10000 burn-in period to get the approximate posterior distribution. The predictive density is displayed in Fig. 6.1 and Fig. 6.2 compares the distribution of μ based on PYM and DPM. This shows that PYM has the longer tail than DPM. Fig. 6.3, 6.4, and Table 6.1 provide the summaries of the cluster distribution.

Table 6.1: Summary of the number of clusters for normal data

	Minimum	Median	Mean	Maximum
PYM	1.00	20.00	20.65	59.00
DPM	1.00	2.00	2.11	8.00

6.3.2 Data Generation: Student's t

In this example, we sample the data from the Student's t distribution with 2 degrees of freedom. Suppose $Y = (Y_1, \ldots, Y_n)$ represents continuous response variable on n observations and the

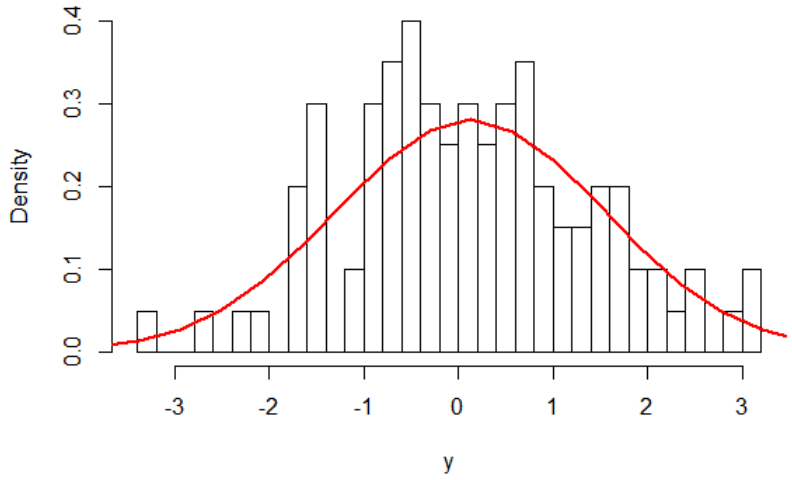


Figure 6.1: Comparison of the predictive distribution using PYM for the data generated from normal distribution

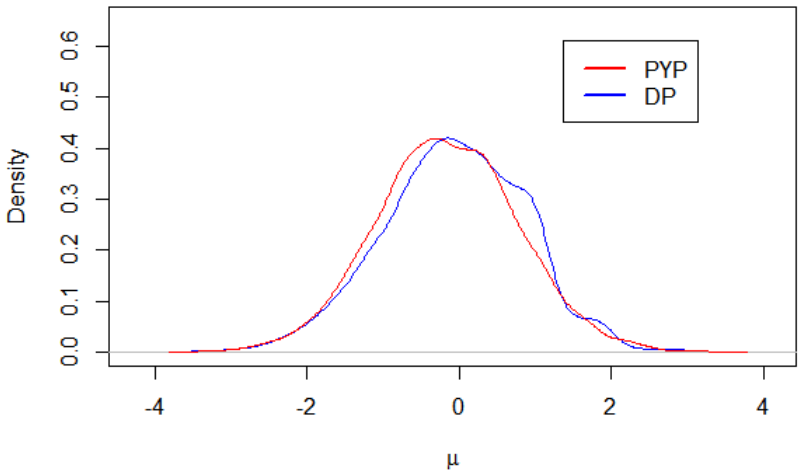


Figure 6.2: Comparison of the distributions of μ using PYM and DPM for the data generated from normal distribution

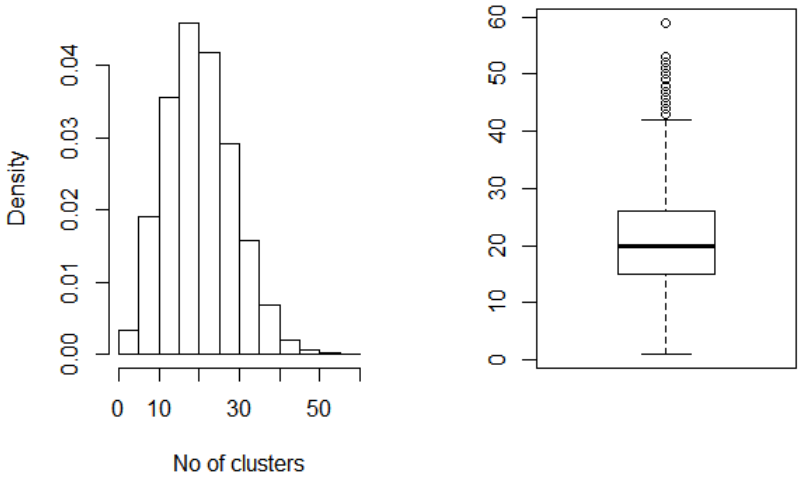


Figure 6.3: Distribution of cluster size using PYM for normal data

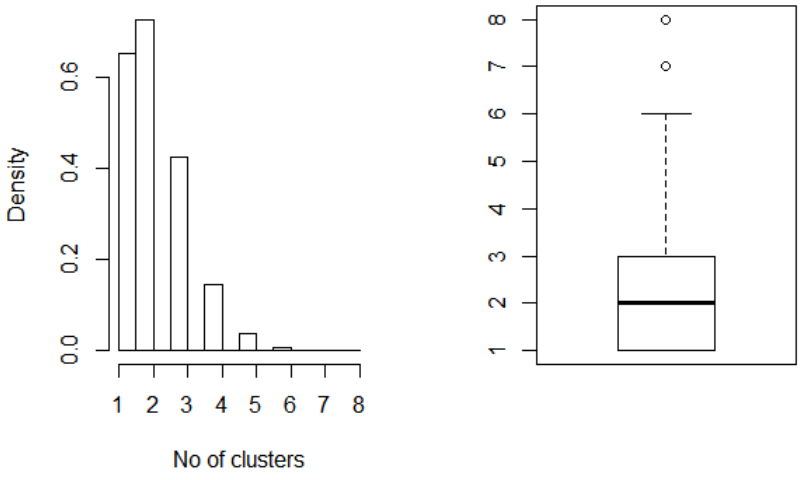


Figure 6.4: Distribution of cluster size using DP for normal data

sampling distribution of Y_j are normal with known mean 0 and unknown precession θ. We assume that θ_j are sampled independently and identically from a random distribution G. Here, G is PYP process prior with discount parameter d, concentration parameter α and base distribution G_0. For this model, we can specify a conjugate base measure for θ_j. One choice of G_0 is the gamma distribution with shape parameter a_0 and scale parameter b_0. Hence the model can be written as

$$\begin{aligned} Y_j|\theta_j &\sim \text{Normal}(y_j|0,\theta_j) \\ \theta_j|G &\sim G \\ G &\sim \mathcal{PY}(d,\alpha,G_0) \\ G_0 &\equiv \text{Gamma}(\theta_j|a_0,b_0). \end{aligned}$$

In this example, $n = 100, a_0 = 1, b_0 = 2, d = 0.5$, and $\alpha = 1$. For ABC-BNP, we choose the tolerance level ε as 0.05, the condition of A_j is taken to be $|y_j - z_j^*| < \varepsilon$, and run the simulation $B = 100000$ times with 20000 burn-in period to get the approximate posterior distribution. The predictive distribution of θ is presented in Fig. 6.5 which shows almost same pattern as the data. Fig. 6.6 shows the comparison of PYM and DPM based on the distribution of θ and PYM has the longer tail than DPM. Also, we compared the mixing distribution plot of PYM with the Gamma mixing distribution for the Student's t. Fig. 6.7 shows that PYM can recover the mixing distribution for θ much better than DPM. The summaries of the clusters are reported in Fig. 6.8, Fig. 6.9, and Table 6.2. Gelman-Rubin plot in Fig. 6.10 for 50000 simulation with 10000 burn-in shows that the sample is sufficiently close to the posterior and we can see that the chains are converged very quickly.

Table 6.2: Summary of the number of clusters for Student's t data

	Minimum	Median	Mean	Maximum
PYM	1.00	20.00	20.64	52.00
DPM	1.00	2.00	2.12	8.00

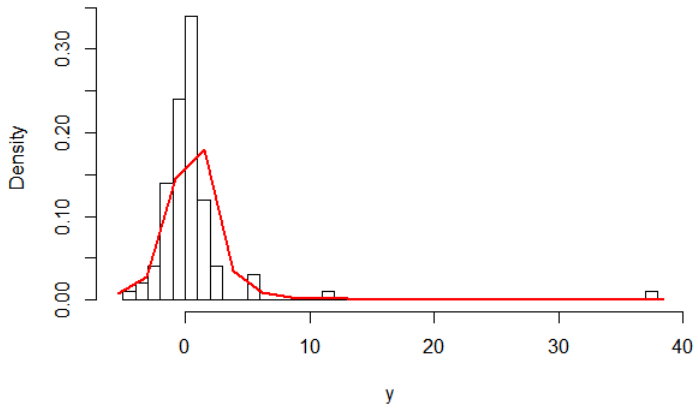


Figure 6.5: Comparison of the predictive distribution for the data generated from t-distribution

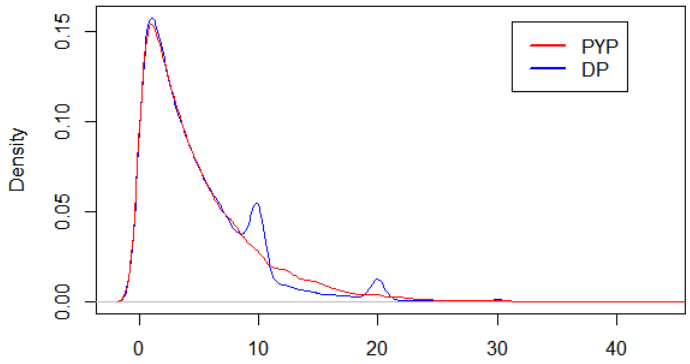


Figure 6.6: Comparison of the distributions for θ using PYM and DPM for the data generated from Student's *t* distribution

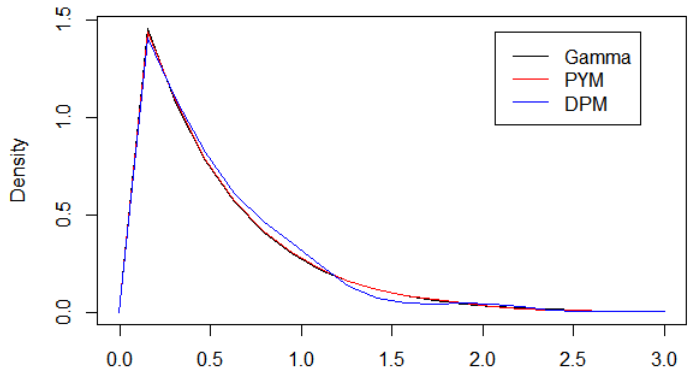


Figure 6.7: Comparison of the mixing distributions of θ for the data generated from Student's *t* distribution

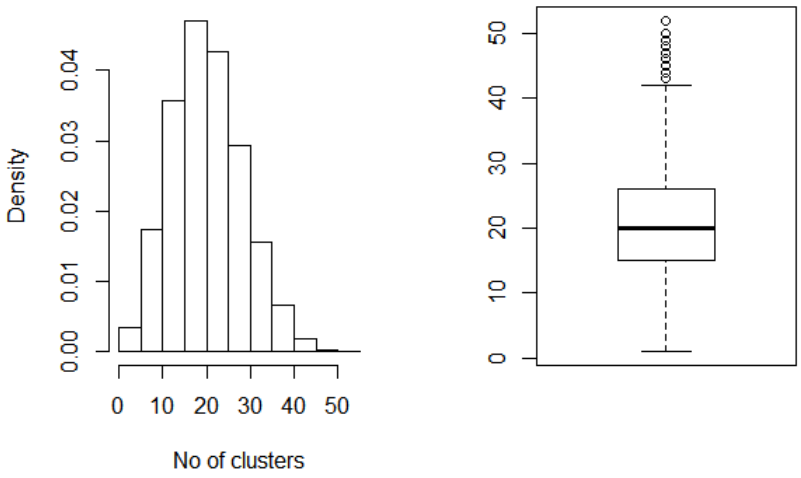


Figure 6.8: Distribution of cluster size using PYM for Student's *t* data

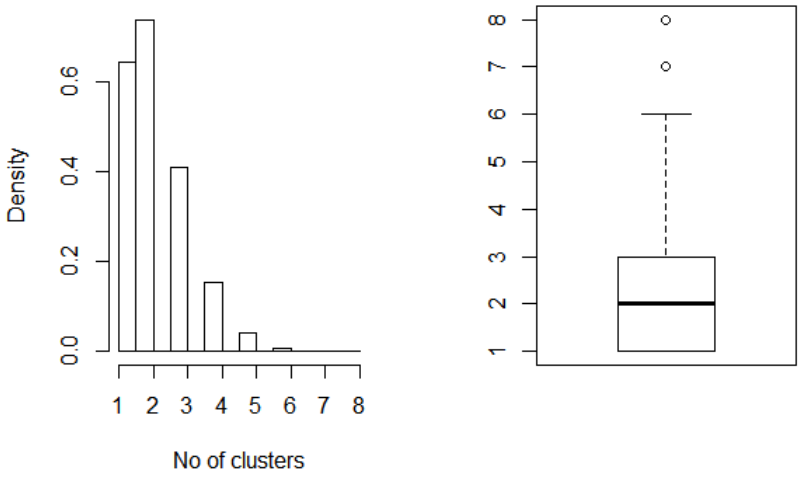


Figure 6.9: Distribution of cluster size using DP for Student's *t* data

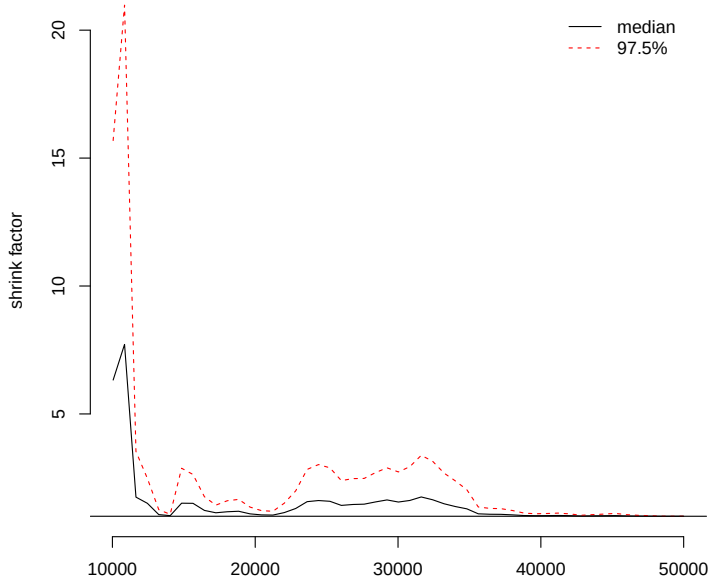


Figure 6.10: Gelman-Rubin plot using ABC-BNP with 50000 simulations after 10000 burn-in period for PYM model of the Student's t data

6.3.3 Data Generation: PYP

First, we generate the mean parameters μ from the PYP with discount parameter 0.5, concentration parameter 1, and base distribution Normal$(0,1)$. Then, the data is sampled from the normal distribution with mean μ and variance 0.1. Now, the sampling distribution of $Y_j, j = 1,\ldots,n$, are normal distribution with unknown mean θ_j and known variance 1. We assume that θ_j are sampled independently and identically from a random measure G. Here, G is PYP prior with discount parameter $d = 0.5$, concentration parameter $\alpha = 1$, and base distribution G_0. For the normal model, we can specify a conjugate base measure for θ_j. Now, in this case, the simplest choice of G_0 is normal distribution with mean μ_0, variance σ_0^2. Hence the model can be defined as follows:

$$\begin{aligned} Y_j|\theta_j &\sim \text{Normal}(y_j|\theta_j,1) \\ \theta_j|G &\sim G \\ G &\sim \mathcal{PY}(d,\alpha,G_0) \\ G_0 &\equiv \text{Normal}(\theta_j|\mu_0,\sigma_0^2). \end{aligned}$$

In this example, $n = 100, \mu_0 = 0, \sigma_0^2 = 1$, and the parameters of the PYM are same as the values in the data generation step. For ABC-BNP, we choose the tolerance level, ε as 0.05 and run the simulation $B = 50000$ times with 10000 burn-in period to get the approximate posterior distribution. We can compare the predictive density with the data. In Fig. 6.11, the data and the predictive distribution using PYM provide almost same densities. The distribution and the summaries of the clusters are reported in Fig. 6.12 and Table 6.3, respectively.

Table 6.3: Summary of the number of clusters for PYM data

Minimum	Median	Mean	Maximum
6.00	20.00	20.24	48.00

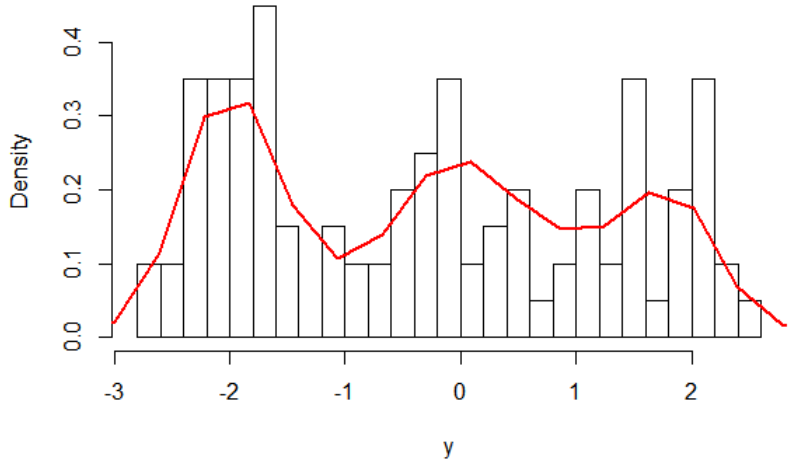


Figure 6.11: Comparison of the predictive distribution for the data generated from PYM

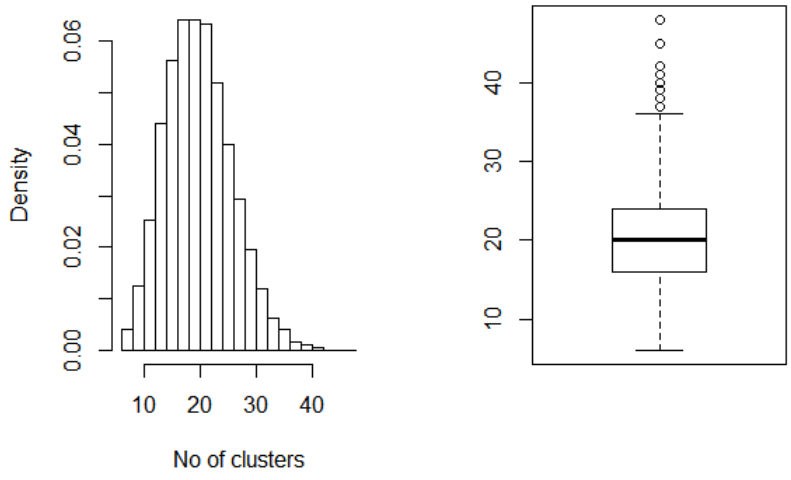


Figure 6.12: Distribution of cluster size for PYM data

6.4 Analysis of Galaxy Data

This data is used in Chapter 3 for DP prior. Now, for the PYM model, the discount parameter assigned to be 0.5. For ABC-BNP, we consider the j^{th} condition as $A_j = |y_j - z_j^*| \leq \varepsilon$, where the tolerance level, ε is 0.05. We run the simulation $B = 100000$ times with 20000 burn-in period to get the approximate posterior distribution. The predictive distribution of a new observation Y_{83} is displayed in Fig. 6.13 and the data shows almost same pattern as the predictive distribution. The distribution and the summary of the clusters are reported in Fig. 6.14 and Table 6.4, respectively. To know whether this sample is sufficiently close to the posterior, we use Gelman-Rubin plot to see if there is a significant difference between the variance within several chains and we can see from Fig. 6.15 for 50000 simulations with 10000 burn-in period, the chains are converged after a certain period of time.

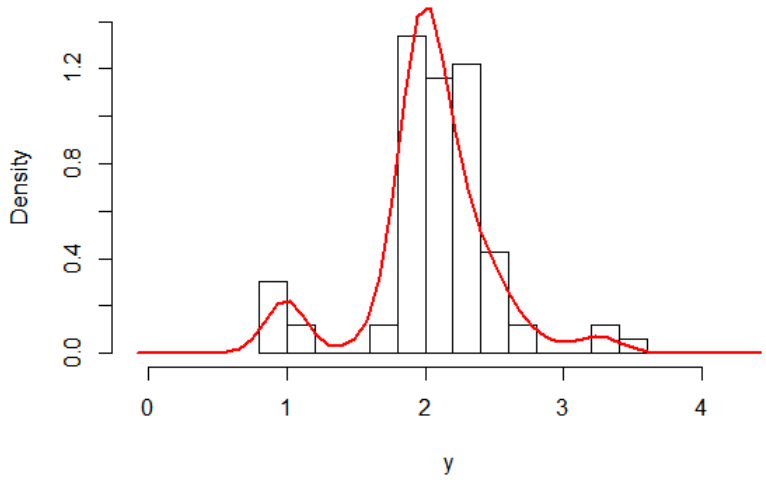


Figure 6.13: The predictive distribution for the galaxy data

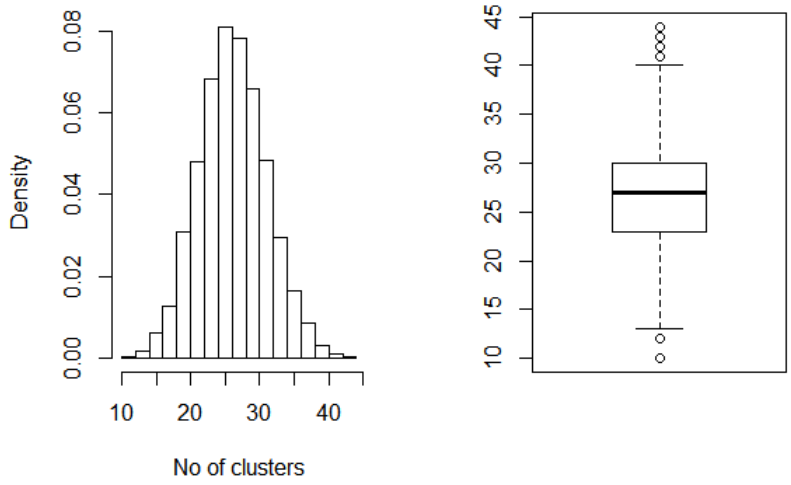


Figure 6.14: Distribution of cluster size for galaxy data

Table 6.4: Summary of the number of clusters for galaxy data

Minimum	Median	Mean	Maximum
10.00	27.00	26.67	44.00

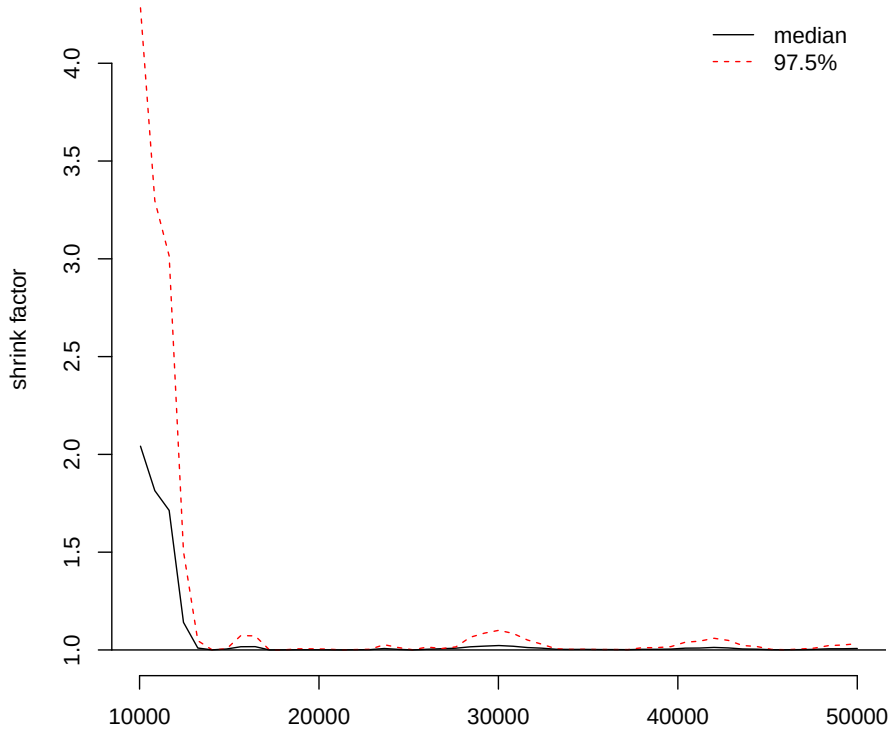


Figure 6.15: Gelman-Rubin plot using ABC-BNP with 50000 simulations after 10000 burn-in period for PYM model of the Galaxy data

CHAPTER 7

ABC-BNP AND STABLE DISTRIBUTION

7.1 Introduction

Stable distributions are the class of probability distributions that provide a rich class of tail behaviors. Lévy (1924) studied the stable family of distribution in the context of sums of independent identically distributed variables. The rich properties of this class of distributions are discussed in books by Samoradnitsky & Taqqu (1994), Nolan (1997), and Feldman & Taqqu (1998). These distributions do not provide any general analytic expression for the probability density but are defined in terms of domain of attraction and characteristic function. The following definition of the stable distribution is based on the domain of attraction as described in Samoradnitsky & Taqqu (1994).

Definition 2 *A random variable X is said to follow stable distribution if it has a domain of attraction, that is, if there is a sequence of independent and identically distributed random variables $Y_1, Y_2, \ldots$ and a sequence of positive numbers $\{d_n\}$ and real numbers $\{a_n\}$, such that*

$$\frac{Y_1 + \ldots + Y_n}{d_n} + a_n \stackrel{d}{\Longrightarrow} X, \tag{7.1}$$

where $\stackrel{d}{\Longrightarrow}$ denotes the convergence of distribution.

Precisely, the stable distribution can be obtained as limits of normalized sum of independent and identically distributed random variables. Another equivalent definition of the stable distribution is in terms of characteristic function. In this setup, a stable distribution has four parameters: i) an index of stability $\alpha \in (0, 2]$, denotes as characteristic exponent, which determines the rate at

which the tails of the distribution gradually decrease, ii) a location parameter $\mu \in \mathbb{R}$ which shifts the distribution to the left or right, iii) a scale parameter $\sigma \geq 0$ which disperses the distribution about μ, and iv) an asymmetry parameter $\beta \in [-1,1]$ which determines the skewness of the distribution, that is, if β is positive, the distribution is skewed to the right and if β is negative, the distribution is skewed to the left, and it is symmetric if $\beta = 0$. Now, the stable distribution can be defined as follows:

Definition 3 *(Samoradnitsky & Taqqu (1994), Nolan (1997)) A random variable X is said to follow stable distribution if there are parameters* $\alpha \in (0,2]$, $\mu \in \mathbb{R}$, $\sigma \geq 0$, *and* $\beta \in [-1,1]$ *such that the characteristic function is defined as*

$$E \exp i\theta X = \begin{cases} \exp\{-\sigma^{\alpha}|\theta|^{\alpha}(1 - i\beta(sign\ \theta)\tan\frac{\pi\alpha}{2}) + i\mu\theta\} & if\ \alpha \neq 1 \\ \exp\{-\sigma|\theta|(1 - i\beta\frac{2}{\pi}(sign\ \theta)\ln|\theta|) + i\mu\theta\} & if\ \alpha = 1 \end{cases} \tag{7.2}$$

where

$$sign\ \theta = \begin{cases} 1 & if\ \theta > 0 \\ 0 & if\ \theta = 0 \\ -1 & if\ \theta < 0 \end{cases}. \tag{7.3}$$

The parameters σ, β, μ *are unique, except when* $\alpha = 2, \beta$ *is arbitrary. Here, we denote the stable distribution as* $S(\alpha, \sigma, \beta, \mu)$.

Non-degenerate stable distributions are unimodal and the probability density function have infinitely differentiable property. Normal $(\alpha = 2, \beta = 0)$ and Cauchy $(\alpha = 1, \beta = 0)$ distributions provide the only analytically tractable distribution of this family. The lack of closed form for densities and distribution functions has been a major technical difficulty for handling of stable distributions. In Bayesian context, it is hard to estimate the parameters based on MCMC since we need to know the likelihood up to a proportionality constant. In this situation, ABC is an easier

way to handle the situation and it can be managed if we are able to generate from the distribution although the form of the distribution is intractable. The next section will focus on the stable distribution as base measure for BNP priors and extend the idea to recurrent data models.

7.2 Stable Model

Suppose we have a set of n heavy tailed random variables $Y = (Y_1, \ldots, Y_n)$ with corresponding parameters $\theta_1, \ldots, \theta_n$ and we can assume that $\theta_1, \ldots, \theta_n$ are drawn from independent and identically distributed random mixing measure G which follows DP or PYP prior with base distribution G_0. Let ζ be the part of the parameter vector corresponding to G. For DPM models, $\zeta = M$ and in case of PYP models, ζ becomes (d, M), where d is the discount parameter and M refers to the concentration parameter. To incorporate heavy tail in the model, we can assume stable distribution as the base measure and the corresponding parameters are i) characteristic component α, ii) location parameter μ, iii) scale parameter σ, and iv) asymmetry parameter β. Here θ_j is the model parameter of the j^{th} step. Different priors can be assigned for the parameter vector of the stable distribution. In this study, we have fixed the values of $\beta = 0$ and $\sigma = 1$. In comparison with the normal model, the prior for the characteristic component α can be assigned as uniform distributions with different choices of intervals, (a_0, b_0). The normal prior is attached to the location parameter μ with the mean μ_p and variance σ_p^2. Under the nonparametric Bayesian models structure, the model can be formalized as follows:

$$
\begin{aligned}
Y_j|\theta_j &\sim f(y_j|\theta_j),\ j=1,\ldots,n \\
\theta_j|G &\sim G \\
G &\sim \text{BNP}(\zeta, G_0) \\
G_0 &\equiv \text{Stable}(\theta_j|\alpha,\sigma,\beta,\mu) \\
\alpha &\sim \text{Uniform}(a_0, b_0) \\
\mu &\sim \text{Normal}(\mu_j|\mu_p,\sigma_p^2).
\end{aligned} \tag{7.4}
$$

In general, the absence of a closed-form density of stable distribution is prevented from evaluating the likelihood function and thus constructing posterior inference. Conditionally on an auxiliary variable, it is possible to express the density function in a closed form (Buckle (1995)) and the Gibbs sampling can be used for estimating the stable distribution. Lombardi (2007) proposed another MCMC approach to draw from the posterior distribution. Also, there are different likelihood free methods are available to analyze the posterior distribution and to estimate the parameters of the stable distribution. G. W. Peters et al. (2012) proposed an approach based on ABC-PRC method. We developed Bayesian inferential methods to fit the models depending upon ABC-BNP method. Here, the ABC method allows approximate posterior simulation for Bayesian models without knowing the form of the likelihood function. The next segment will emphasize on the proposed methodology for two models which have been discussed in previous section.

7.2.1 Proposed Method for Stable Models

In this case, we can use the basic method mentioned in Chapter 3. Here, the ABC-BNP is applied to get the probability distribution of the parameter θ_j. The transition kernel for stable models is defined as

$$T(\theta^{(b)},\alpha^{(b)},\mu^{(b)}|\theta^{(b-1)},\alpha^{(b-1)},\mu^{(b-1)}) = \prod_{j=1}^{n} s_1(\theta_j^{(b)}|\theta_{-j}^{(b-1)},\alpha^{(b-1)},\mu^{(b-1)}) \times s_2(\alpha^{(b)}|\theta^{(b)},\alpha^{(b-1)},\mu^{(b-1)}) \times s_3(\mu^{(b)}|\theta^{(b)},\alpha^{(b)},\mu^{(b-1)}), \tag{7.5}$$

where $\theta_{-j}^{(b-1)} = (\theta_1^{(b)},\ldots,\theta_{j-1}^{(b)},\theta_{j+1}^{(b-1)},\ldots,\theta_n^{(b-1)})$ and $s_1(\cdot)$ is derived from (3.2) and (3.4). As stated in the ABC-BNP method, for each $j = 1,\ldots,n$, $\theta_j^{(b)}$ can be constructed as

$$\theta_j^{(b)} = \begin{cases} \theta_j^* & \text{if } \mathbb{I}(A_j); \\ \theta_j^{(b-1)} & \text{otherwise,} \end{cases} \tag{7.6}$$

where $\mathbb{I}(C) = 1$ if C holds and if we assume continuous response, the condition in (7.6) can be constructed as

$$|y_j - z_j^*| < \varepsilon,$$

where ε is a predefined threshold value and z_j^* is sampled from the sampling distribution $f(z_j|\theta_j^*)$. Now, we sample a candidate value θ_j^* from the DP or PYP prior with base measure G_0 as stable distribution, that is,

$$\theta_j^*|\theta_{-j}^{(b-1)},\alpha^{(b-1)},\mu^{(b-1)} \sim \pi(\theta_j^*|\theta_{-j}),$$

where $\pi(\theta_j^*|\theta_{-j})$ is defined in (3.2).

Finally, $s_2(\cdot)$, and $s_3(\cdot)$ indicate the full conditionals of $(\alpha^{(b)}|\theta^{(b)},\alpha^{(b-1)},\mu^{(b-1)})$, and $(\mu^{(b)}|\theta^{(b)},\alpha^{(b)},\mu^{(b-1)})$, respectively. we update the parameters $\alpha^{(b)},\mu^{(b)}$ conditional on $\theta^{(b)}$. In this case, $\alpha^{(b)}$ is updated using slice sampler. The Metropolis-Hastings method is performed to obtain the conditional posterior of $\mu^{(b)}$. First, we generate the candidate values μ^* from the proposal density $q(\cdot)$ and

$$\mu^{(b)} = \begin{cases} \mu^* & \text{with probability } \min\{1, \frac{\pi(\mu^*|\theta^{(b)},\alpha^{(b)})q(\mu^{(b-1)})}{\pi(\mu^{(b-1)}|\theta^{(b)},\alpha^{(b)})q(\mu^*)}\}; \\ \mu^{(b-1)} & \text{otherwise.} \end{cases}$$

where $\pi(\mu|\theta,\alpha)$ is the full conditional posterior distribution of μ.

7.2.2 Simulation Study

Suppose $Y = (Y_1,\ldots,Y_n)$ represents continuous response variable on n observations. We assume that each Y_j follows normal distribution with unknown mean θ_j and known variance σ_y^2. We assume that θ_j are sampled independently and identically from a distribution G. Here, G is PYP prior with discount parameter d, concentration parameter M and base distribution G_0. For this model, we can specify a base measure for θ_j as stable distribution with index of stability α, asymmetric parameter $\beta = 0$, location parameter μ, and scale parameter $\sigma = 1$.

Here, we have simulated the data from Normal$(0,\sigma_y^2)$. We consider two choices of the index of stability, for example, $\alpha = 2$ and $\alpha = (1.5,2)$. In this example, $n = 100, \sigma_y^2 = 1, \mu_p = 0, \sigma_p^2 = 1, d = 0.5$, and $M = 1$. For ABC-BNP, we choose the tolerance level, ε as 0.005 and run the simulation $B = 100000$ times with 25000 burn-in period to get the approximate posterior density. Here the α is updated using the slice sampler. Fig. 7.1 shows the distribution of μ for $\alpha = 2$ and $\alpha = (1.5,2)$ for PYP process and for $\alpha = (1.5,2)$, the distribution of μ captures the tail probabilities. The corresponding acceptance rates are 3.93% and 6.7%. The summaries of the clusters are reported in Fig. 7.2 and Table 7.1.

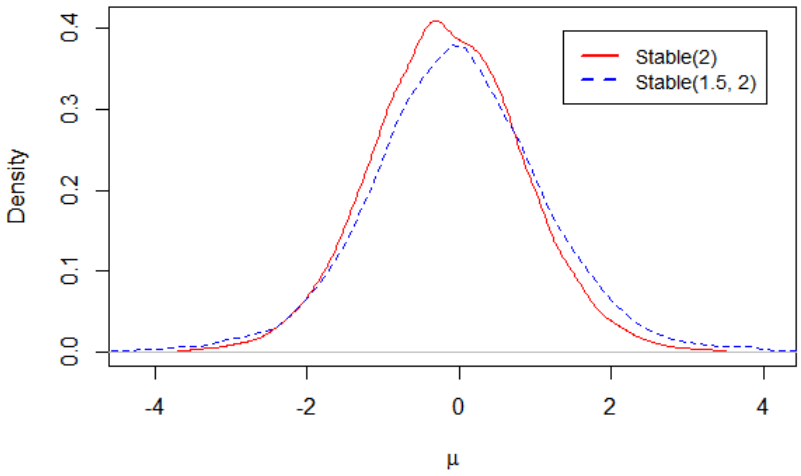


Figure 7.1: Distributions of μ for stable model

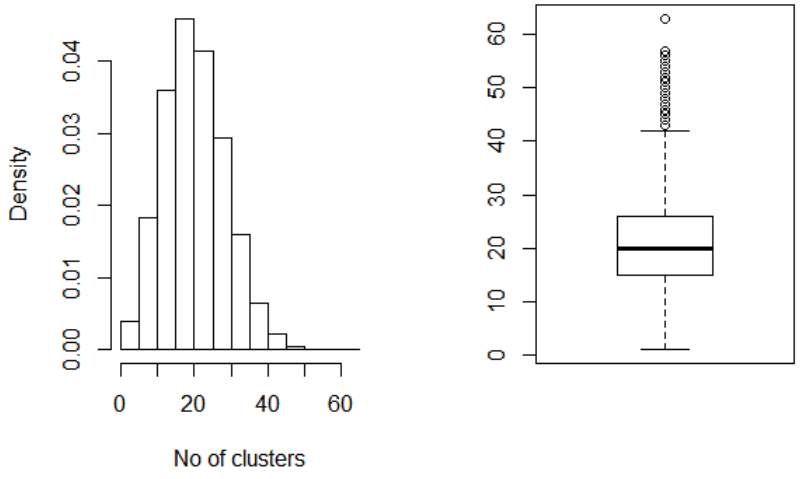


Figure 7.2: Distribution of cluster size for stable(2) model

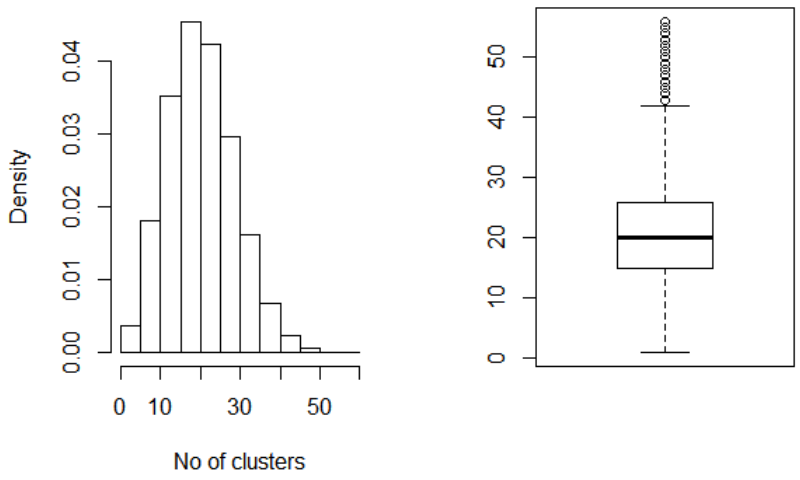


Figure 7.3: Distribution of cluster size for stable(1.5, 2) model

Table 7.1: Summaries of the number of clusters for stable model

Model	Minimum	Median	Mean	Maximum
Stable(2)	1.0	20.0	20.6	63.00
Stable(1.5, 2)	1.0	20.0	20.7	56.00

7.3 Stable Recurrent Data Model

Now, the idea of the previous model has been extended to the recurrent data models. One extension of such models is the recurrent survival model with gap time between two consecutive events. We have discussed this model in Chapter 5. For the heavy tailed data, stable distribution is one of the choices for base measure. In this step, we have also assigned prior for the models parameters. The prior for μ is the normal distribution with mean μ_0 and variance σ_0^2, uniform (a_0, b_0) distribution is assigned for the index of stability, and inverse-gamma distributions are attached to the scale parameters. Hence, the DPM frailty model for recurrent data can be expressed as follows:

$$\begin{aligned}
Y_{jk} &= \log(W_{jk}), j = 1,\ldots,n; k = 1,\ldots,m_j \\
Y_{jk} &= \mu_y + u_j + \varepsilon_{jk} \\
\varepsilon_{jk} &\overset{iid}{\sim} \text{Normal}(0, \sigma_e^2) \\
u_j|G &\sim G \\
G &\sim DP(M, G_0) \\
G_0 &\equiv \text{Stable}(u_j|\alpha, \sigma, \beta, \mu) \\
\mu_y &\sim \text{Normal}(\mu_0, \sigma_0^2) \\
\alpha &\sim \text{Uniform}(a_0, b_0) \\
\sigma^2 &\sim IG(a_u, b_u) \\
\sigma_e^2 &\sim IG(a_e, b_e).
\end{aligned} \tag{7.7}$$

ABC-BNP method has been applied to recurrent data model as described in Chapter 5 and we implement the idea of stable distribution to this model.

7.3.1 Proposed method for Stable Recurrent Data Models

For this model, the ABC-BNP is used to update the frailty term u_j and the other parameters, $\alpha, \mu_y, \sigma_e^2, \sigma^2$, are updated using Metropolis-Hastings algorithm. The transition kernel for this model can be expressed as

$$\begin{aligned} T(u^{(b)}, \alpha^{(b)}, \mu_y^{(b)}, \sigma_e^{2(b)}, \sigma^{2(b)} | \theta^{(b-1)}, \alpha^{(b-1)}, \mu_y^{(b-1)}, \sigma_e^{2(b-1)}, \sigma^{2(b-1)}) = \prod_{j=1}^{n} s_1(u_j^{(b)} | u_{-j}^{(b-1)}, \alpha^{(b-1)}, \\ \mu_y^{(b-1)}, \sigma_e^{2(b-1)}, \sigma^{2(b-1)}) \times s_2(\alpha^{(b)} | u^{(b)}, \alpha^{(b-1)}, \mu_y^{(b-1)}, \sigma_e^{2(b-1)}, \sigma^{2(b-1)}) \\ \times s_3(\mu_y^{(b)} | u^{(b)}, \alpha^{(b)}, \mu_y^{(b-1)}, \sigma_e^{2(b-1)}, \sigma^{2(b-1)}) \times s_4(\sigma_e^{2(b)} | u^{(b)}, \alpha^{(b)}, \mu_y^{(b)}, \sigma_e^{2(b-1)}, \sigma^{2(b-1)}) \\ \times s_5(\sigma^{2(b)} | u^{(b)}, \alpha^{(b)}, \mu_y^{(b)}, \sigma_e^{2(b)}, \sigma^{2(b-1)}). \end{aligned} \tag{7.8}$$

Here, the method is almost same as we described in Chapter 5. Instead of normal base distribution, stable distribution is implemented in this model and we assign a prior on α. The distribution of α is updated using the slice sampler and the Metropolis-Hastings algorithm is applied to $\phi = (\mu_y, \sigma^2, \sigma_e^2)$ with acceptance probability

$$\min\left\{1, \frac{\pi(\phi^*)L(\phi^*|y)q(\phi^*|\phi)}{\pi(\phi)L(\phi|y)q(\phi|\phi^*)}\right\},$$

where $\pi(\cdot)$, $L(\cdot|y)$, and $q(\cdot)$ are same as indicated in (5.10). ϕ^* is generated from the proposal density $q(\cdot)$. Given the values of $u^{(b)}$, $\alpha^{(b)}$, $\mu_y^{(b-1)}$, $\sigma_e^{2(b-1)}$, and $\sigma^{2(b-1)}$, the distribution of $\phi^{(b)}$ is as follows

$$\phi^{(b)} = \begin{cases} \phi^* & \text{with probability } \min\{1, \frac{\pi(\phi^*|u^{(b)}, \alpha^{(b)})q(\phi^{(b-1)})}{\pi(\phi^{(b-1)}|u^{(b)}, \alpha^{(b)})q(\phi^*)}\}; \\ \phi^{(b-1)} & \text{otherwise.} \end{cases}$$

where $\pi(\phi|u, \alpha)$ is the full conditional of $\phi = (\mu_y, \sigma_e^2, \sigma^2)$. Since the method is based on the ABC-MCMC, the conditional posterior converges to the parameter of interests.

7.3.2 Analysis of Bowel Motility Cycles

We have discussed this data in Sec. 5.4.2. In this example, the values of the hyper-parameters are same as before. But we consider a prior for α for the stable model. The simulation has been iterated $B = 100000$ times with 50000 burn-in period to get the estimates of the parameters. Here, the acceptance rate 11.2% and the computing time is 4.3 minutes. We use the slice sampler to update the stable parameter α and the Metropolis-Hastings algorithm is used to approximate other parameters. Table 7.2 shows the comparison of estimated parameters using ABC-BNP and stick breaking Gibbs (RJAGS software). Table 7.3 and Fig. 7.4 provide the comparison and distribution of log likelihoods over MCMC.

Table 7.2: Comparison of the parameters for stable recurrent data model

	ABC-BNP(Stable $\alpha = (1,2)$)	ABC-BNP (Normal)	Stick breaking Gibbs
$\hat{\mu_y}$	4.22	4.11	4.26
$\hat{\sigma}$	0.82	0.80	0.79
$\hat{\sigma}_u$	0.20	0.19	0.16

Table 7.3: Comparison of maximum log likelihood over MCMC for recurrent data model

Method	Max log likelihood
ABC-BNP(stable(1, 2))	-113.5
ABC-BNP(stable(2))	-113.9
Stick breaking Gibbs	-114.9

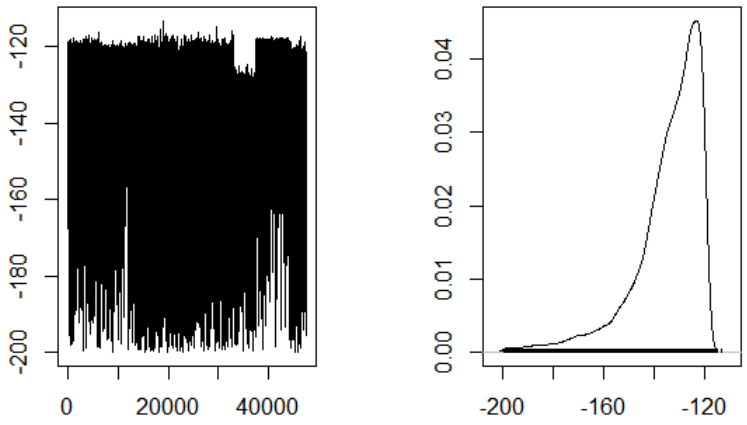


Figure 7.4: Trace plot and the distribution of log likelihood for stable(1, 2) recurrent data model

CHAPTER 8

FUTURE EXTENSIONS AND CONCLUSION

We conclude by discussing some possible extensions of the different models suggested by our proposed method.

8.1 Future Extensions

8.1.1 GLMM for Count Data

Let $Y = (Y_1, \ldots, Y_n)$ denote count random variable of n subjects, $X = (X_1, \ldots, X_p)$, X_j be the vector of p covariates corresponding to the j^{th} observation, $\beta = (\beta_1, \ldots, \beta_p)$ be the regression coefficient vector of dimension p, $\Psi = (\psi_1, \ldots, \psi_n)$ be the vector of random effects, where ψ_j is the random effect corresponding to j^{th} subject, and $\gamma = (\gamma_1, \ldots, \gamma_n)$ be the vector of linear predictors. For each j, the GLMM model under count data can be defined as

$$\begin{aligned} Y_j|\lambda_j &\sim \text{Poisson}(y_j|\lambda_j) \\ g(\lambda_j) &= \log\left(\lambda_j\right) = \gamma_j \\ \implies g^{-1}(\gamma_j) &= \exp(\gamma_j) = \lambda_j \\ \gamma_j &= X_j^T\beta + \psi_j. \end{aligned}$$

Kyung et al. (2011) used this model under DP. We can apply this model under PYP. Hence, the model can be defined as follows:

$$\begin{aligned} P(Y_j = 1|X_j, \beta, \psi_j) &= \exp(X_j^T \beta + \psi_j), j = 1, \ldots, n \\ \psi_j|G &\sim G \\ G &\sim \text{BNP}(\zeta, G_0). \end{aligned}$$

8.1.2 Joint Modeling

One extension to the survival model is the joint modeling. In clinical research, it is very common to notice both longitudinal and time to event data. If there is no correlation between these two, one can easily analyze the data separately. But if the longitudinal variable is correlated with the survival point, it is not appropriate to use two independent studies. As a remedy, we can use the joint modeling for longitudinal and time to event data. The detailed idea of this type of models is given by Guo & Carlin (2004) and Rizopoulos (2012).

Let y_{jk} be the longitudinal response at time t_{jk} and t_j^* be the time to event data for j^{th} subject with $k = 1, \ldots, m_j$ and $j = 1, \ldots, n$. Now, the basic structure of joint model (Henderson et al. (2000)) is defined as

$$\begin{aligned} y_{jk} &= x_{1j}\beta_1 + W_{1j}(t_{jk}) + \varepsilon_{jk} \\ \lambda_j(t_j^*) &= \lambda_0(t_j^*) \exp(x_{2j}\beta_2 + W_{2j}(t_j^*)), \end{aligned}$$

where $\varepsilon_{jk} \sim \text{Normal}(0, \sigma_\varepsilon^2)$ is the random error for the first model and $\lambda_0(t_j^*)$ represents the baseline hazard function. x_1, x_2 denote the design matrices and β_1, β_2 represent the corresponding regression coefficients for longitudinal and survival models. W_1 and W_2 indicate the functions of shared terms

between two models. One simple choice of W_1 and W_2 at the time point t is defined as (Guo & Carlin (2004))

$$\begin{aligned} W_{1j}(t) &= U_{1j} + U_{2j}(t) \\ W_{2j}(t) &= U_{1j} + U_{2j}(t) + \gamma(U_{1j} + U_{2j}(t)), \end{aligned}$$

where $U_j = (U_{1j}, U_{2j})$ is the vector of frailty terms or random effects and γ denotes the measure of association between the two models. In this model, we can proceed by assigning the nonparametric Bayesian models for the frailty term and use the ABC-BNP method to update it.

We would also like to investigate other nonparametric models based on ABC. In last twenty years, DP have become extremely popular and useful for nonparametric models in the Bayesian studies due to the efficient computation methods. Here, the model is used as a prior for the unknown distribution. An area of research is extending this to a wider class of models where the unknown distribution depends on different objectives. Extension of the DP provides a class of models that are not only computationally feasible, it also allows many basic modeling designs. The popular extensions of the DP include dependent DP (MacEachern (1999), MacEachern (2000), De Iorio et al. (2004), De Iorio et al. (2009)), hierarchical DP (Teh et al. (2006)) and nested DP (Rodriguez et al. (2008)).

8.1.3 Dependent Dirichlet Process

MacEachern (1999) generalizes the idea of DP to dependent DP (DDP). In DDP, the stick-breaking process of DP can be generalized by taking θ_k instead of distinct θ_k^*. Now, the DDP prior

for the collection of random distributions, $\mathcal{G} = \{G_s : s \in S\}$, where S is a covariate space, can be defined as follows:

$$\begin{aligned}
\beta_k(s) &\sim \text{Beta}(1, \alpha(s)) \\
\pi_1(s) &= \beta_1(s) \\
\pi_k(s) &= \beta_k(s) \prod_{l=1}^{k-1} (1 - \beta_l(s)), \;\; k = 2, 3, \ldots, \\
\theta_{kx}(s) &\overset{iid}{\sim} G_{0,S} \\
G_s &= \sum_{k=1}^{\infty} \pi_k(s) \delta_{\theta_{(kx)}(s)}.
\end{aligned}$$

where $G_{0,S}$ is a stochastic process defined on S. Hence, for any fixed s, the DDP construction yields a DP prior distribution for G_s. An application of the DDP is the ANOVA type dependent model, denoted as ANOVA-DDP (De Iorio et al. (2004), De Iorio et al. (2009)). Let y_i, $i = 1, \ldots, n$ be i^{th} data point and x_i be the corresponding covariate vector. In this setup, $\theta_{kx}(s)$ is modeled as

$$\theta_{kx}(s) = m_s + A_{vs} + B_{ws},$$

where $m_s \overset{iid}{\sim} p^0_m(m_s)$, $A_{vs} \overset{iid}{\sim} p^0_{Av}(A_{vs})$, and $B_{ws} \overset{iid}{\sim} p^0_{Bw}(A_{ws})$ and independent across s, v, and w. Hence, the ANOVA-DDP model with concentration parameter M and base measure $p^0 = (p^0_m, p^0_{Av}, p^0_{Bw})$, can be written as

$$\begin{aligned}
y_i | x_i &= x \sim H_x(y_i) \\
H_x(y_i) &= \int \mathcal{N}(y_i | \mu, \sigma^2) dG_s(\mu) \\
G_s, s \in S &\sim \text{ANOVA DDP}(M, p^0).
\end{aligned} \tag{8.1}$$

Let $\alpha_s = [m_s, A_{2s}, \ldots, A_{Vs}, B_{2s}, \ldots, B_{Ws}]$ and d_i be the design vector corresponding to x_i, that is, $\theta_{xk}(s) = \alpha_s d_i$ for $x = x_i$. Then the model (8.1) can be written as

$$\begin{aligned} y_i|x_i &= x \sim H_x(y_i) \\ H_x(y_i) &= \int \mathcal{N}(y_i|\alpha d_i, \sigma^2) dG(\alpha) \\ G &\sim \text{DP}(M, p^0). \end{aligned}$$

8.1.4 Hierarchical Dirichlet Process

If we wish to model a grouped data, in which each group is associated with a mixture model, an extension to DPM model, denoted as hierarchical DP (HDP), is appropriate in this situation. Teh et al. (2006) proposed this model to use as the prior over the factors for grouped data. Let $y_j = (y_{j1}, y_{j2}, \ldots)$ be the observations in j^{th} group, θ_{ji} be a factor corresponding to the observation y_{ji}, $f(y_{ji}|\theta_{ji})$ be the distribution of $y_{ji}|\theta_{ji}$, G_j be the random probability measure for j^{th} group, and G_0 be the global random probability measure. G_0 and G_j are distributed as DP with concentration parameters γ and α_0, respectively and the base measures as H and G_0, respectively. Hence, the HDP can be defined as

$$\begin{aligned} G_0 &\sim \text{DP}(\gamma, H) \\ G_j &\sim \text{DP}(\alpha_0, G_0) \\ \theta_{ji}|G_j &\sim G_j \\ y_{ji}|\theta_{ji} &\sim f(y_{ji}|\theta_{ji}). \end{aligned}$$

This model can be extended to more than two levels. Hence, in general, we can extend the HDP mixture model as a tree with each node associated with DPM models.

8.1.5 Nested Dirichlet Process

The nested DP (NDP) (Rodriguez et al. (2008)) is mainly used for clustering probability distributions and simultaneous multilevel clustering in nested settings. Let y_{ij} be the i^{th} observation of the j^{th} group, for $i = 1,\ldots,n_j$, $j = 1,\ldots,J$ and $y_{ij} \overset{i.i.d}{\sim} f_j$ for $j = 1,\ldots,J$. Now, consider a collection of distribution $\{G_1,\ldots,G_J\}$ with $G_j \sim Q$ and $Q = DP(\alpha DP(\beta H))$ with $f_j(\cdot|\phi) = \int p(\cdot|\theta,\phi)G_j d\theta$, where $p(\cdot|\theta,\phi)$ is a distribution for given θ,ϕ. Hence, $\{f_1,\ldots,f_J\}$ is said to follow NDP with

$$
\begin{aligned}
\beta_k^* &\sim \text{Beta}(1,\alpha) \\
u_{lk}^* &\sim \text{Beta}(1,\beta) \\
\pi_k^* &= \beta_k^* \prod_{s=1}^{l-1}(1-\beta_s^*) \\
w_{lk}^* &= u_{lk}^* \prod_{s=1}^{l-1}(1-u_{sl}^*) \\
\theta_{lk}^* &\sim H \\
G_j &\sim Q = \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*} \\
G_k^* &= \sum_{k=1}^{\infty} w_{lk}^* \delta_{\theta_{lk}^*} \\
y_{ji} &\sim f_j.
\end{aligned}
$$

NDP can be characterized as DDP models. There is a difference between the HDP and the NDP models (see Figure. 8.1, taken from Rodriguez et al. (2008)). In HDP, the collection of distribution $\{G_1,\ldots,G_J\}$ has the same structure with different weights, whereas, for NDP models, the different distributions have either the same structure with the same weight or completely different.

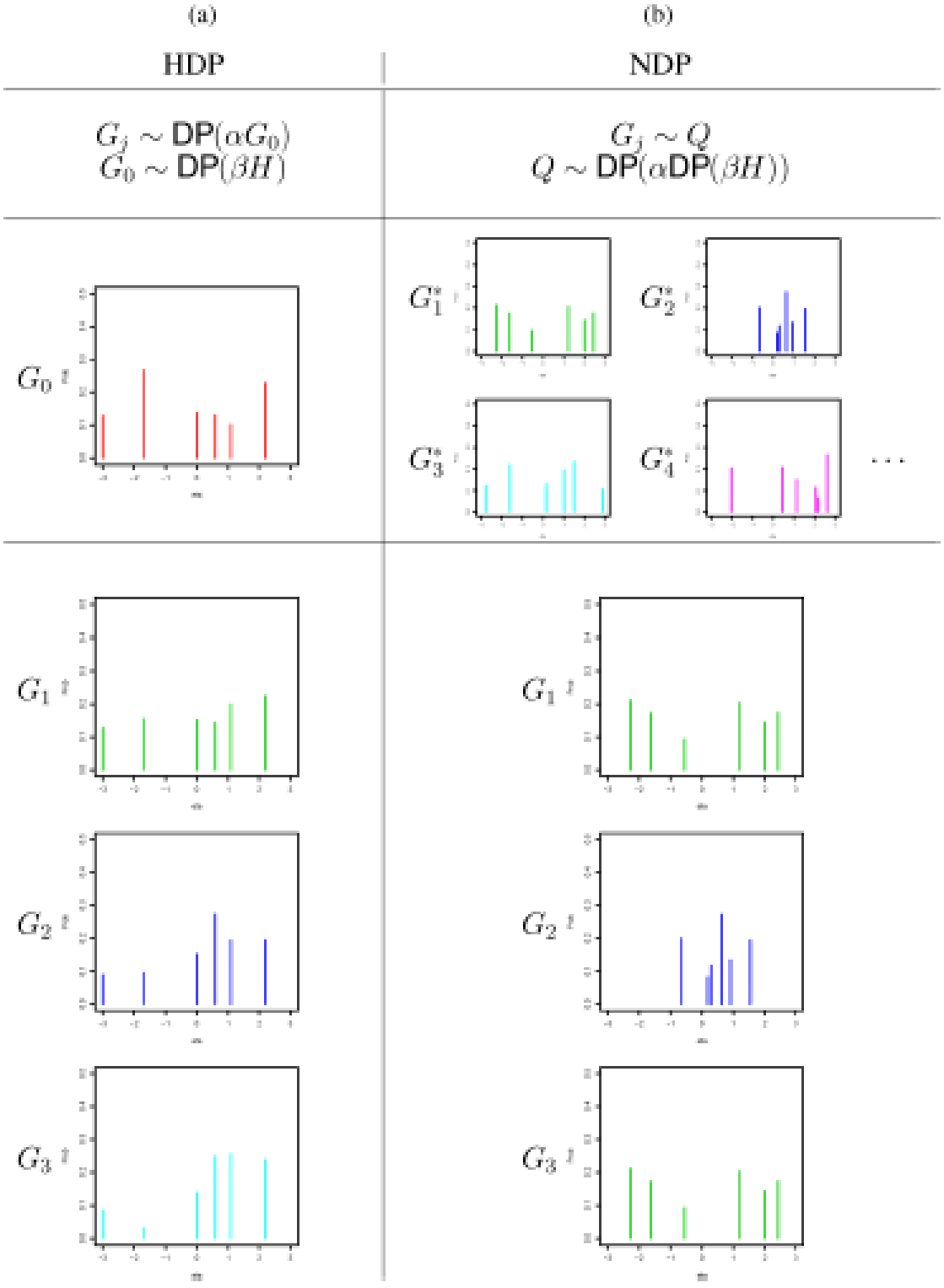


Figure 8.1: Comparing NDP and HDP models

8.2 Conclusion

This dissertation proposed a nonparametric Bayesian computation method that takes into account an easier way to estimate the posterior distribution when the likelihood is intractable or computationally intensive. First, we reviewed the existing ABC methods and the nonparametric Bayesian models. After that, the basic structure of our proposed ABC-BNP method has been discussed and developed some new methods based on that.

Different Bayesian models have been explored in this study. In practice, we implement the MCMC method to estimate the posterior summaries. However, ABC-BNP provides a straightforward way to handle the situation. Chapter 4 presented ABC-BNP for binary GLMM models. For the random intercept and scale mixture models, the form of the posterior is non-conjugate and it is easier to implement our proposed method. The results show that the posterior means of ABC-BNP are almost same as for both the binary models. As well as, the random intercept model provided the less computing time than the preexisting technique. We have constructed on the already developed method for Bayesian nonparametric survival and recurrent data models in Chapter 5. The modification was incorporated by defining the random effect in the model and the ABC-BNP method was used to update the parameters. This method showed a significant improvement based on the log likelihood over MCMC of the model. We further proposed ABC-BNP method under the PYP in Chapter 6 and used one real data and three simulated data to establish the method. Chapter 7 suggested dealing with an intractable likelihood, the stable distribution. We considered a stable model and the survival recurrent data model to compare the ABC-BNP with the existing MCMC method.

It is observed that our proposed method performed better based on the maximum of log likelihood estimates over the MCMC because the maximum values reached in ABC-BNP. Also, this method can be easily implemented in the model and it is fast. In most of the cases, the computing

time is very less than the other existing method. Most of the situations, it is notable that the resulted posterior estimates are almost similar and the 95% credible intervals became wider for the regression parameters of the ABC-BNP method relative to MCMC methods. Also, the predictive distributions are nearly same as the data.

REFERENCES

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

Basu, S., & Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association*, *98*(461), 224–235.

Beard, R. E. (1959). Note on some mathematical mortality models. *The lifespan of animals, Ciba colloquium on Aging*, *G.E.W Wolstenholme and M, O'Connor (eds). Little, Brown, Boston*, 302–311.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., & Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, asp052.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035.

Besag, J., & Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25–37.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, 353–355.

Blei, D. M., & Jordan, M. I. (2004). Variational methods for the Dirichlet process. In *Proceedings of the twenty-first international conference on machine learning* (p. 12).

Blum, M. G., & François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, *20*(1), 63–73.

Blum, M. G., Nunes, M. A., Prangle, D., Sisson, S. A., et al. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, *28*(2), 189–208.

Blum, M. G., & Tran, V. C. (2010). HIV with contact tracing: a case study in approximate Bayesian computation. *Biostatistics*, *11*(4), 644–660.

Buckle, D. (1995). Bayesian inference for stable distributions. *Journal of the American Statistical Association*, *90*(430), 605–613.

Bush, C. A., & MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, *83*(2), 275–285.

Cook, R. J., & Lawless, J. (2007). *The statistical analysis of recurrent events*. Springer Science & Business Media.

Damien, P., Wakefield, J., & Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 331–344.

De Iorio, M., Johnson, W. O., Müller, P., & Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, *65*(3), 762–771.

De Iorio, M., Müller, P., Rosner, G. L., & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, *99*(465), 205–215.

Dittmar, D. (2013). Slice sampling. *TU Darmstadt. URL http://www. ausy. informatik. tu-darmstadt. de/uploads/Teaching/RobotLearningSeminar/Dittmar_RLS_2013. pdf*.

Duchateau, L., & Janssen, P. (2007). *The frailty model*. Springer Science & Business Media.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, *89*(425), 268–277.

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

Estoup, A., Lombaert, E., MARIN, J.-M., Guillemaud, T., Pudlo, P., Robert, C. P., & CORNUET, J. (2012). Estimation of demo-genetic model probabilities with approximate Bayesian computation using linear discriminant analysis on summary statistics. *Molecular Ecology Resources*, *12*(5), 846–855.

Fall, M. D., & Barat, É. (2014). Gibbs sampling methods for Pitman-yor mixture models.

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, *14*(1), 11–21.

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 419–474.

Feldman, R., & Taqqu, M. (1998). *A practical guide to heavy tails: statistical techniques and applications*. Springer Science & Business Media.

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

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent Advances in Statistics*, *24*(1983), 287–302.

Ferguson, T. S., & Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, 163–186.

Foll, M., Beaumont, M. A., & Gaggiotti, O. (2008). An approximate Bayesian computation approach to overcome biases that arise when using amplified fragment length polymorphism markers to study population structure. *Genetics*, *179*(2), 927–939.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398–409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*(6), 721–741.

Guo, X., & Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, *58*(1), 16–24.

Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, *1*(4), 465–480.

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2005). *Bayesian survival analysis*. Wiley Online Library.

Jabot, F., & Chave, J. (2009). Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecology letters*, *12*(3), 239–248.

Jain, S., & Neal, R. M. (2012). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*.

Jain, S., Neal, R. M., et al. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, *2*(3), 445–472.

Joyce, P., & Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, *7*(1).

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Klein, J. P., & Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kuo, L., Smith, A. F., MacEachern, S., & West, M. (1992). Bayesian computations in survival models via the Gibbs sampler. In *Survival analysis: State of the art* (pp. 11–24). Springer.

Kurihara, K., Welling, M., & Teh, Y. W. (2007). Collapsed variational Dirichlet process mixture models. In *Ijcai* (Vol. 7, pp. 2796–2801).

Kyung, M., Gill, J., & Casella, G. (2011). Sampling schemes for generalized linear Dirichlet process random effects models. *Statistical Methods & Applications*, *20*(3), 259–290.

Lévy, P. (1924). Théorie des erreurs. la loi de gauss et les lois exceptionnelles. *Bulletin de la Société mathématique de France*, *52*, 49–85.

Lombaert, E., Guillemaud, T., Thomas, C., Lawson Handley, L., Li, J., Wang, S., ... others (2011). Inferring the origin of populations introduced from a genetically structured native range by approximate Bayesian computation: case study of the invasive ladybird Harmonia axyridis. *Molecular Ecology*, *20*(22), 4654–4670.

Lombardi, M. J. (2007). Bayesian inference for α-stable distributions: A random walk MCMC approach. *Computational Statistics & Data Analysis*, *51*(5), 2688–2700.

Lopes, J. S., & Boessenkool, S. (2010). The use of approximate Bayesian computation in conservation genetics and its application in a case study on yellow-eyed penguins. *Conservation Genetics*, *11*(2), 421–433.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, *23*(3), 727–741.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *Asa proceedings of the section on Bayesian statistical science* (pp. 50–55).

MacEachern, S. N. (2000). Dependent Dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, 1–40.

MacEachern, S. N., & Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, *7*(2), 223–238.

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, *22*(6), 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *100*(26), 15324–15328.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). CRC press.

McCulloch, C., & Searle, S. (2001). Linear mixed models. *McCulloch CE, Searle SR, New York, John Wiley & Sons*, 156–186.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American Atatistical Association*, *44*(247), 335–341.

Minka, T., & Ghahramani, Z. (2003). Expectation propagation for infinite mixtures. In *Nips workshop on nonparametric bayesian methods and infinite models* (Vol. 19).

Naylor, J. C., & Smith, A. F. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, 214–225.

Neal, R. M. (1992). Bayesian mixture modeling. In *Maximum entropy and bayesian methods* (pp. 197–211). Springer.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 705–741.

Nevat, I., Peters, G. W., & Yuan, J. (2008). *Bayesian inference in linear models with a random gaussian matrix: algorithms and complexity*.

Nolan, J. P. (1997). Numerical calculation of stable densities and distribution functions. *Communications in Statistics. Stochastic models*, *13*(4), 759–774.

Nunes, M. A., & Balding, D. J. (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, *9*(1).

Peters, G., & Sisson, S. (2006). Bayesian inference, Monte Carlo sampling and operational risk. *Journal of Operational Risk*, *1*(3), 27–50.

Peters, G. W., Sisson, S. A., & Fan, Y. (2012). Likelihood-free Bayesian inference for α-stable models. *Computational Statistics & Data Analysis*, *56*(11), 3743–3756.

Pitman, J., & Yor, M. (1997). The two-parameter Poisson- Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 855–900.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, *16*(12), 1791–1798.

Ratmann, O., Andrieu, C., Wiuf, C., & Richardson, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, *106*(26), 10576–10581.

Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M., Richardson, S., & Wiuf, C. (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of h. pylori and p. falciparum. *PLoS Comput Biol*, *3*(11), e230.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in r*. CRC Press.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods (second edition)*. Springer-Verlag, New York.

Rodriguez, A., Dunson, D. B., & Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, *103*(483), 1131–1154.

Rubin, D. B., et al. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

Samoradnitsky, G., & Taqqu, M. S. (1994). *Stable non-gaussian random processes: stochastic models with infinite variance* (Vol. 1). CRC press.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica sinica*, 639–650.

Silk, D., Filippi, S., & Stumpf, M. P. (2013). Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Statistical Applications in Genetics and Molecular Biology*, *12*(5), 603–618.

Sisson, S. A., Fan, Y., & Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *104*(6), 1760–1765.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Susarla, V., & Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, *71*(356), 897–902.

Tanaka, M. M., Francis, A. R., Luciani, F., & Sisson, S. (2006). Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, *173*(3), 1511–1520.

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, *145*(2), 505–518.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82–86.

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69–85.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*(3), 439–454.

Von Neumann, J. (1951). Various techniques used in connection with random digits. *J. Resources of the National Bureau of Standards - Applied Mathematics Series*, *12*, 36–38.

Wakefield, J., Gelfand, A., & Smith, A. (1991). Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing*, *1*(2), 129–133.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, *36*(1), 45–54.

Walker, S. G., Damien, P., Laud, P. W., & Smith, A. F. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 485–527.

Wegmann, D., Leuenberger, C., & Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, *182*(4), 1207–1218.

West, M., & Escobar, M. D. (1993). *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University.

Wilkinson, R. D., & Tavaré, S. (2009). Estimating primate divergence times by using conditioned birth-and-death processes. *Theoretical Population Biology*, *75*(4), 278–285.