A Recursive Polya Tree Mixture Model: Computationally Efficient Bayesian Nonparametric Modelling

Li Shujie

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montreal,Quebec

2013-12-09

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Ph.D

DEDICATION

To my parents (Li Jinbao and Lu Qizhen) and my wife (Wang Yun).

ACKNOWLEDGEMENTS

There are many people who deserve thanks for helping me during my study at McGill. First and foremost, I express my sincere gratitude to my supervisors Drs. James Hanley and David Stephens. They not only provided guidance, direction, and funding to me, but also encouraged me during my research at McGill. I am also grateful to my examiners for their thoughtful questions and comments.

I would like to thank my parents (Li Jinbao and Lu Qizhen) and my wife (Wang Yun). Their love and encouragement helped me to overcome many difficulties during my study at McGill.

ABSTRACT

This thesis describes a flexible and computationally efficient Bayesian nonparametric modelling approach based on a recursive Polya tree mixture model. This approach is motivated by the need to capture the heterogeneity observed in many areas of biostatistics such as meta analysis of clinical trials, survival analysis and recurrent event data analysis.

Let $Y_1, \ldots, Y_N$ be mutually independent observations such that Y_i has distribution $h(\cdot|\theta_i)$. It is assumed that the parameters $\theta_1, \ldots, \theta_N$ arise from an unknown distribution F and that the prior on F is a Polya tree distribution. An empirical Bayesian approach is adopted for the choice of the prior's base distribution. As the parameters $\theta_1, \ldots, \theta_N$ are latent, a data augmentation algorithm is used to simulate pseudo values iteratively. The empirical distribution of these pseudo values can then guide the choice of the base distribution of the Polya tree prior. The theoretical properties of this procedure are explored.

Despite its simplicity, the proposed model is practical and computationally efficient. In addition to providing a good approximation for more complicated Bayesian nonparametric models, it can be used to handle difficult problems in classical Bayesian nonparametric modelling. In this thesis, the use of the model is illustrated using the famous data of Brown (2008) and Liu (1996), which are often viewed as test cases for Bayesian nonparametric modelling. It is also shown that the proposed approach

can be applied to density estimation (including in the bivariate case) and meta analysis in biostatistics. Moreover, a Bayesian semi-parametric accelerated failure time (AFT) model based on the proposed approach is considered, and an extension of the AFT model to recurrent event data analysis is introduced.

ABRÉGÉ

Cette thèse décrit une approche flexible et numériquement efficace de modélisation non paramétrique bayésienne au moyen d'un modèle de mélange fondé sur une arborescence de Pólya récursive. Cette approche est motivée par la nécessité de prendre en compte l'hétérogénéité fréquemment observée en biostatistique, notamment lors de la méta-analyse d'essais cliniques ou de l'analyse de durées de vie et d'événements récurrents.

Soient $Y_1, \ldots, Y_N$ des observations mutuellement indépendantes telles que Y_i est de loi $h(\cdot|\theta_i)$. On suppose que les paramètres $\theta_1, \ldots, \theta_N$ proviennent d'une loi F inconnue et que la loi a priori sur F est une arborescence de Pólya. On adopte une approche bayésienne empirique pour le choix de la loi a priori de base. Les paramètres $\theta_1, \ldots, \theta_N$ étant latents, on a recours à un algorithme d'augmentation de données pour en simuler des valeurs de façon itérative. La loi empirique de ces pseudo observations permet alors de guider le choix de la loi a priori de base. Les propriétés théoriques de cette procédure sont explorées.

Malgré sa simplicité, le modèle proposé est pratique et efficace au plan calcul. En plus de fournir une bonne approximation de modèles bayésiens non paramétriques plus complexes, il facilite le traitement de problèmes réputés difficiles en modélisation bayésienne non paramétrique classique. Dans cette thèse, l'emploi du modèle est illustré au moyen des célèbres données de Brown (2008) et de Liu (1996) souvent

considérées comme bancs d'essai pour la modélisation bayésienne non paramétrique. Comme on le fait valoir, l'approche peut aussi servir à estimer une densité (y compris bivariée) et à des fins de méta-analyse en biostatistique. On étudie en outre un modèle bayésien semi-paramétrique à temps de panne accéléré (TPA) fondé sur cette approche et on propose une généralisation du modèle TPA pour l'analyse d'événements récurrents.

STATEMENT OF ORIGINALITY

The original research presented in this thesis was carried out by Li Shujie under the supervision of Drs Hanley and Stephens. The principal contributions will be submitted as papers for publication in statistical or biostatistical journals. These include

`A recursive Polya tree mixture model: computationally efficient Bayesian nonparametric modeling`.

Authors: Li Shujie, David Stephens, and James Hanley.

Chapter: 3

Design of methods: Li Shujie, David Stephens and James Hanley.

Programming: Li Shujie

Data Analysis: Li Shujie, David Stephens and James Hanley.

`A new Bayesian nonparametric regression model for survival analysis.`

Authors: Li Shujie, David Stephens and James Hanley

Chapter: 3 & 5

Design of methods: Li Shujie, David Stephens and James Hanley.

Programming: Li Shujie

Data Analysis: Li Shujie, David Stephens and James Hanley.

`A new Bayesian nonparametric model for meta-analysis.`

Authors: Li Shujie, David Stephens and James Hanley

Chapter: 3 & 4

Design of methods: Li Shujie, David Stephens and James Hanley.

Programming: Li Shujie

Data Analysis: Li Shujie, David Stephens and James Hanley.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

Many statistical applications involve multiple parameters that are related to each other. For example, in a study involving N hospitals, and hospital i ($i = 1, \ldots, N$) has n_i patients. Suppose the measurements of blood pressure $y_{i1}, \ldots, y_{in_i}$ for each hospital i's patients are observed, and the mean of the blood pressure θ_i for a hospital similar to hospital i is of interest. A naive approach is to estimate each θ_i only using the information of hospital i, for example $\hat{\theta}_i = (\sum_{j=1}^{n_i} y_{ij})/(n_i)$. However, it might be reasonable to expect that the estimates of the $\theta_i's$ should be related to each other; thus it is reasonable to view $\theta_i's$ as a sample from a common population distribution F.

The Bayesian hierarchical model (Gelman et al. 2003 [33]) is a natural choice in the statistical applications which involve multiple parameters. Suppose a random sample $Y_1, \ldots, Y_N$ (the random sample can be continuous or count data) are observed, and assumed to have a parametric distribution with parameters θ_i: $Y_i|\theta_i \sim h(Y_i|\theta_i)$ independently (in this thesis, $h(Y_i|\theta_i)$ represents the p.d.f), and the parameters $\theta_1, \ldots, \theta_N$ are assumed to arise from some distribution F, $\theta_1, \ldots, \theta_N \sim F$ `i.i.d`. Parametric models assume that F arises from some parametric distribution. However, this parametric assumption has many restrictions, and may not give useful information from data (Liu 1996 [57]). A more flexible approach is to consider that F arises from some nonparametric distribution. For example, by assuming F to be

a random distribution with a Dirichlet process prior (Ferguson 1973 [28]) or Polya tree prior (Lavine 1992 [54]), we can build a intermediary between nonparametric and parametric models. In recent years, there has been an increase of interest in Bayesian nonparametric models due to their flexibility, and the availability of algorithms (which mostly rely on MCMC) for posterior computation. Dey et al. (1998 [20]) and Hjort et al. (2010 [40]) have summarized the existing computational issues arising in Bayesian nonparametric models.

While Bayesian nonparametric models are extremely powerful and introduce more flexibility, they are still not widely used (Jara et al. 2011 [45]). One possible reason is that they always rely on complicated computing methods. Although an `R` package `DPpackage` (Jara et al. 2011 [45]) has been developed, the daunting computing approaches may still be considered as a restriction. Alternative computing algorithms have been developed, including smoothing by roughening (Shen and Louis 1999 [76]), predictive recursion (Newton and Zhang 1999 [70] and Newton 2002 [69]; Martin and Tokdar 2011 [62]), weighted Chinese restaurant sampling (Ishwaran and James 2003 [44]), sequential imputation (Liu 1996 [57]; MacEachern et al.1999 [59]), and variational Bayes (Blei and Jordan 2006 [8]). Recently, Wang and Dunson (2011 [81]) proposed a fast method of Bayesian inference in Dirichlet process mixture models based on a sequential greedy search algorithm. In our view, all these algorithms represent breakthrough developments.

In this thesis, we propose an alternative Bayesian nonparametric model, called the recursive Polya tree mixture model (`RPTMM`), which not only enjoys the flexibility provided by Bayesian nonparametrics, but is also much easier to implement. We

describe our new model in the framework of a Polya tree. Before introducing the detail of this model in next section, we mention some advantages of the proposed model.

1. In the traditional Bayesian nonparametric hierarchical model, sampling from F is not straightforward (see Liu 1996 and Liu 1999 [57, 58]), since many MCMC algorithms marginalize out F. A direct application of data augmentation to sample F is infeasible. There are alternatives, as described by Doss (1994 [22]), Gelfand and Kottas (2002 [32]), and Liu (1996 [57]). Indeed, even using the stick-breaking process (Sethuraman 1994 [75]) to express F, sampling from F will produce ties, since the support of F in the stick-breaking process is a countable set. In our recursive Polya tree mixture model, sampling from F is very straightforward, and as a result, the calculation of the marginal distribution $\int h(y|\theta)F(\theta)$ is trivial.

2. Bayesian nonparametric models always require users to carry out computation. For example, the derivation of the marginal distribution usually involves complicated or high-dimensional computation. The proposed `RPTMM` does not require any analytical computation, and it can also approximate the complicated Bayesian nonparametric models.

3. The proposed `RPTMM` model is very fast to implement. In many cases, running the recursive algorithms for 100-200 steps is sufficient, and this typically takes a very short time.

The proposed `RPTMM` model makes two significant changes compared with the classical Polya tree model.

1. Simple or mixture of Polya tree models (Hanson and Johnson 2002 [37], Hanson 2006 [36]) partition the support in a hierarchical structure, fix all the partition points, and estimate the "weights" corresponding to each interval. We do the converse; the probability corresponding to each interval is fixed, and we estimate the partition points.

2. Polya trees and Dirichlet processes are nonparametric random distributions centered on a base distribution. The base distribution is always assigned a parametric distribution with parameters ϕ. Users are required to specify the parametric base distribution, and estimate ϕ.

   Empirical ideas have also been developed. Mcauliffe et al. (2006 [63]) proposed a nonparametric empirical bayes for Dirichlet process mixture model. Holmes et al. (2009 [41]) also used the empirical distribution to be the base distribution in a Polya tree-based test statistic. In this thesis, we also bring the empirical Bayes idea to choose the base distribution. In a Bayesian hierarchical model, the parameters are unobservable, and their empirical distribution is unavailable to us. We develop a easy-to-implement data augmentation algorithm to overcome this difficulty.

The overall structure of the thesis is as follows. In Chapter 2, we review some existing Bayesian nonparametric models. In Chapter 3, we introduce the detail of the proposed recursive Polya tree mixture model. Chapter 4 discusses some examples related to Bayesian nonparametric hierarchical model. In Chapter 5, a Bayesian semiparametric AFT model is developed based on the proposed recursive Polya tree

mixture model. In Chapter 6, we give a conclusion of this thesis and discuss some future research topics.

CHAPTER 2
Introduction to Bayesian Nonparametrics

In this chapter, we introduce the basic ideas of Bayesian nonparametrics. In Section 2.1, we introduce the Dirichlet process mixture model. Section 2.2 reviews the Bayesian nonparametrics for survival analysis. In Section 2.3, the Polya Tree model and some examples are discussed. Some other applications are discussed in Section 2.4. In Section 2.5, we introduce other Bayesian nonparametric models, including sequential imputation and predictive recursion.

2.1 The Dirichlet process mixture model

In Bayesian inference, prior distributions are assigned to all unknown quantities (parameters) in a statistical model. In parametric inference, the data generating model is assumed to be a function of a finite dimensional parameter; in nonparametric inference, the parameter dimensionality is not finite (for example, the parameter might be an unknown distribution function). Ferguson (1973 [28]) suggested two principles to guide the construction of prior distributions for nonparametric problems:

1. The support of the prior distribution should be large,
2. The resulting posterior distributions should be tractable analytically.

A third principle that is important for practical implementation is that

3. the hyperparameters defining the priors should be easily interpreted.

The Dirichlet process (Ferguson 1973 [28]) meets all these principles and has become the heart of Bayesian nonparametric solutions. In next section, we outline the definition and properties of the Dirichlet process.

2.1.1 Definition and properties of Dirichlet process

In 1973, Ferguson (1973 [28]) proposed the definition of Dirichlet process.

Definition Dirichlet process (Ferguson 1973 [28]).

Let Ω be the sample space, and $\mathcal{F}$ be the corresponding σ-algebra. A Dirichlet process (DP; Ferguson, 1973 [28]) is a stochastic process G on $(\Omega, \mathcal{F})$ with concentration parameter α and base distribution G_0. For any partition $(B_1, \ldots, B_M)$ on the space of G_0, the random vector $(G(B_1) \ldots G(B_M))$ has a Dirichlet distribution with parameters $(\alpha G_0(B_1) \ldots \alpha G_0(B_M))$

$$(G(B_1) \ldots G(B_M)) \sim Dirichlet(\alpha G_0(B_1) \ldots \alpha G_0(B_M)).$$

We denote a Dirichlet process by

$$G \sim \mathtt{DP}(\alpha, G_0)$$

The existence of a Dirichlet process has been verified in Ferguson (1973 [28]) using the Kolmogorov consistency conditions. Some properties of the Dirichlet process are summarized below.

1. The mean of Dirichlet process G is the base distribution,

$$E(G) = G_0,$$

$$\mathrm{Var}(G) = \frac{G_0(1 - G_0)}{\alpha + 1}$$

Thus G_0 can be considered as a prior guess for G.

2. The concentration parameter α represents the strength of belief in the base distribution. For large values, a random function G drawn from the Dirichlet process will be close to the base distribution G_0. As α diminishes, the variation of the draws around G_0 increases.
3. Suppose N samples have been drawn from G. Then the posterior distribution is also a Dirichlet process

$$G|Y_1 \dots Y_N \sim DP\left(\alpha + N, \frac{\alpha G_0}{\alpha + N} + \frac{\sum_{i=1}^{N} \delta_{Y_i}(.)}{\alpha + N}\right)$$

   This is termed the *conjugacy* property.
4. For small values of α, the posterior distribution based on $Y_1 \dots Y_N$ is close to the empirical distribution, as the contribution from the prior is minimal.
5. Draws from a Dirichlet process (that is, random probability mass functions sampled from the process) are **discrete** with probability 1. This is an unappealing property compared with the Polya tree model (Polya tree can be absolutely continuous with probability 1).

To demonstrate the role of α, we draw $N = 10000$ samples from a Dirichlet process with base distribution to be normal `N(0,1)`, and $\alpha = 1$, 10, 100, and 1000. From Figure 2–1 it is evident that when α becomes larger and larger, the histogram of draws becomes closer and closer to the base distribution. The number, K, of unique values (or atoms) for each draw is

- $\alpha = 1$, the number of atoms = 7,

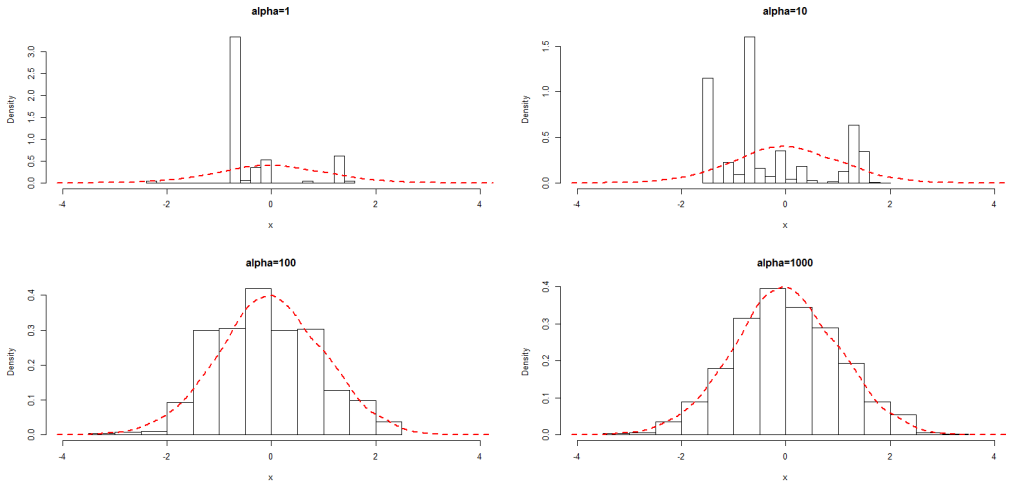

Figure 2–1: Histogram of the draw from a Dirichlet process $\texttt{DP}(\alpha, \texttt{Normal}(0, 1))$. When α becomes larger and larger, the histogram of draws becomes closer and closer to the base distribution.

- $\alpha = 10$, the number of atoms = 73,
- $\alpha = 100$, the number of atoms = 499,
- $\alpha = 1000$, the number of atoms = 2380.

In expectation, larger values of α yield higher numbers of atoms: we have (Antoniak 1974 [2])

$$E[K|\alpha, N] = \sum_{i=1}^{N} \frac{\alpha}{\alpha + i - 1} = \alpha(\psi(\alpha + N) - \psi(\alpha)) \backsimeq \alpha \log(1 + N/\alpha)$$

where $\psi(.)$ is the digamma function .

Ferguson's first construction is now seldom used in real applications, and statisticians have developed several equivalent representations, including the Polya urn model, the Chinese restaurant process and stick-breaking process.

Definition The Polya Urn model.

Blackwell and MacQueen (1973 [7]) developed the Polya urn scheme for representing the Dirichlet process (and other processes) using exchangeability. A sequence of random variables $\theta_1 \dots \theta_N$ is a Polya urn sequence with parameters α and base distribution G_0, if

$$\theta_1 \sim G_0,$$
$$\theta_i \in B | \theta_1 \dots \theta_{i-1} \sim \frac{\alpha G_0(B) + \sum_{j=1}^{i-1} \delta_{\theta_j}(B)}{\alpha + i - 1},$$

for a set B, where $\delta_\theta(y)$ denotes the unit measure concentrating at θ. Imagine an urn with the number of α balls, and there are $\alpha G_0(\theta)$ balls of color θ. The sequence $\theta_1 \dots \theta_N$ can be imagined as: draw a ball from the urn, and after each draw, the ball drawn is replaced and another ball of the same color is added to the urn.

Blackwell and MacQueen (1973 [7]) showed that as N goes to infinity, a Polya urn sequence will converge to a Dirichlet process $G \sim DP(\alpha, G_0)$ almost surely. An important property of the Polya urn model is the explicit clustering property: a new sample has positive probability, $1/(\alpha + i - 1)$ for the ith draw to be equal to one of the previous drawn samples. Here, α determines the probability of choosing a new color from G_0; at step i, this probability is $\alpha/(\alpha + i - 1)$.

Many computing methods using Markov Chain Monte Carlo (MCMC) for posterior inference are based on the Polya urn model: early contributions include Escobar (1994 [26]) and Escobar and West (1995 [27]).

Definition The Chinese restaurant process.

The Chinese restaurant process (CRP) (Aldous 1983 [1]; Pitman 1995 [72]) is a distribution constructed directly on partitions. Imagine a Chinese restaurant

with countably infinitely many circular tables, labeled 1, 2.... Customers enter the restaurant sequentially; the first customer always chooses the first table.

Suppose the first n ($n \geq 1$) customers have occupied K_n tables. Then the $(n+1)$st customer will either

- choose a new table with probability $\alpha/(n+\alpha)$, or
- choose an occupied table k ($1 \leq k \leq K_n$) with probability $n_k/(\alpha+n)$,

where n_k is the number of customers for table k. Suppose the table of customer $i, 1 \leq i \leq n$ is denoted c_i,

$$p(c_{n+1} = k|c_1, \ldots, c_n) = \frac{\alpha}{\alpha+n}\delta_{K_n+1} + \sum_{k=1}^{K_n} \frac{n_k}{\alpha+n}\delta_k, \tag{2.1}$$

where δ_k is the point mass at k. After the Nth label is generated, $K \equiv K_N$ denotes the number of clusters. This process generates the cluster configuration implied by the Dirichlet process, after which we can generate a random discrete distribution on the support of G_0 by sampling variates $\theta_k, k = 1, \ldots, K$ to represent the K cluster centers.

Definition: The Stick-Breaking process.

Sethuraman (1994 [75]) proposed a constructive definition of Dirichlet process, writing for a set B

$$G(B) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(B),$$

where $\theta_k \sim G_0$, and

$$\pi_k = v_k \prod_{j=1}^{k-1}(1 - v_j),$$

with $v_j \sim \texttt{Beta}(1, \alpha)$. Ishwaran and James (2001 [43]) showed that Dirichlet process can be approximated by truncating the number of components to M

$$G(B) \approx \sum_{k=1}^{M} \pi_k \delta_{\theta_k}(B).$$

By setting $v_M = 1$, we can guarantee that $\sum_{j=1}^{M} \pi_j = 1$. The stick-breaking process can be imagined as breaking a unit length stick infinitely many times. Starting with a stick with unit length, we break a piece off the stick and discard it. The breaking point is generated as a random variable $v_j \sim \texttt{Beta}(1, \alpha)$. By continuing this process infinitely many times, we obtain the stick-breaking process.

2.1.2 The Dirichlet process mixture model: the conjugate case

Each draw from a Dirichlet process is almost surely discrete, and this property is undesirable. The *Dirichlet process mixture model* (DPM, Antoniak 1974 [2]) extends the Dirichlet process to deal with the continuous case. For a random sample $Y_1 \dots Y_N$, suppose that

$$Y_i|\theta_i \sim h(Y_i|\theta_i) \text{ independent,}$$

and assume

$$\theta_1, \theta_2, \dots, \theta_N \sim \texttt{DP}(\alpha, G_0(\lambda)) \text{ i.i.d.}$$

Each variate θ_i drawn from G is a draw from a discrete distribution, so that some of the Y_i will share the same values of θ, and the Y_i with the same values of θ belong to the same mixture component. The posterior for θ given Y and λ is proportional to

$$\prod_{i=1}^{N} h(Y_i|\theta_i) \left(\frac{\alpha G_0(\lambda) + \sum_{k<i} I(\theta_i = \theta_k)}{\alpha + i - 1} \right),$$

where $I(.)$ is the indicator function. A simple posterior sampling scheme based on the Polya urn and MCMC (Chapter 1 of Dey et al. [20]) can be summarized as follows: for each i in turn, we have that

$$\theta_i|\theta_{k,k\neq i}, Y, \lambda \sim q_0(Y_i)G_b(\theta_i;\lambda) + \sum_{k\neq i} q_k(Y_i)I(\theta_i = \theta_k), \tag{2.2}$$

where

- $G_b(\theta_i;\lambda) \propto h(Y_i|\theta_i)G_0(\theta_i;\lambda)$, where
  - $G_0(\theta_i;\lambda)$ is the prior base density evaluated at θ_i;
  - $G_b(\theta_i;\lambda)$ is the posterior base density evaluated at θ_i.
- $q_0(Y_i) \propto \alpha \int h(Y_i|\theta_i)G_0(\theta_i;\lambda)\, d\theta_i$;
- $q_k(Y_i) \propto h(Y_i|\theta_k)$ is proportional to the conditional density of Y_i given θ_k;

where the constant of proportionality is determined by the requirement that

$$q_0(Y_i) + \sum_{k\neq i} q_k(Y_i) = 1.$$

Heuristically, the posterior for θ_i is largely based on the θ_j corresponding to Ys near to Y_i. If the hyperparameter of λ in base distribution is known, and a conjugate prior is used, then the integral

$$\int h(Y_i|\theta_i)G_0(\theta_i;\lambda)\, d\theta_i$$

can be computed analytically. As a result, it is easy to use (2.2) to sample from the full conditionals and utilize the Gibbs sampler version of Markov Chain Monte Carlo.

Example 1. The DPMM for mixtures of binomial distributions Consider the binomial random sample $(X_1, Y_1), \ldots (X_N, Y_N)$, where the number of trials is Y_i and the number of successes is X_i. Assume the probability of success is θ_i for trial i; is natural to model the conditional distribution given $Y_i = y_i$ using a binomial distribution:

$$X_i | Y_i = y_i \sim \text{Binomial}(y_i, \theta_i),$$

Now assume that the random effect probabilities θ_i arise from a Dirichlet process $\theta_i \sim \texttt{DP}(\alpha, G_0)$. The base distribution is the conjugate Beta distribution, $G_0 \equiv \texttt{Beta}(a, b)$. It is then very straightforward to implement the algorithm described in (2.2). The posterior base distribution in this case

$$G_b(\theta_i; a, b) \equiv \text{Beta}(x_i + a, y_i - x_i + b).$$

We also have that $q_0(x_i) \propto \alpha \int h(X_i|\theta_i) G_0(\theta_i; \lambda)\, d\theta$, so that

$$q_0(x_i) \propto \alpha \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 t^{x_i+a-1}(1-t)^{y_i-x_i+b-1} dt = \alpha B_{a,b}(y_i, x_i),$$

say. The value of $q_k(x_i)$ is the density value for the binomial distribution with parameters (y_i, θ_k) evaluated at x_i.

Example 2. The DPMM for mixtures of Poisson distributions

Suppose we observe a random sample of count data $Y_1, \ldots Y_N$, and model them using a Poisson distribution, $Y_i|\theta_i \sim \texttt{Poisson}(\theta_i)$, where the θ_i are assumed to arise from a Dirichlet process, $\theta_i \sim \texttt{DP}(\alpha, G_0)$. The base distribution G_0 is a conjugate Gamma distribution, $G_0 \sim \texttt{Gamma}(a, b)$. Escobar and West (1998) (see also Chapter

1 of [20]) derived the posterior base distribution G_b, q_0 and q_k as follows:

$$\begin{aligned} q_0(y_i) &\propto \alpha \frac{(y_i+a-1)!}{y_i!(a-1)!}\left(\frac{b}{1+b}\right)^a\left(\frac{1}{1+b}\right)^{y_i} \\ G_b(\theta_i;a,b) &\equiv \text{Gamma}(a+y_i, b+1) \\ q_k(y_i) &\propto \text{Poisson}(y_i;\theta_k) \end{aligned}$$

Hence for $i = 1, \ldots, N$, we draw θ_i according to:

- $\theta_i \sim \text{Gamma}(a + y_i, b + 1)$ with probability proportional to $q_0(y_i)$.
- $\theta_i = \theta_k$ with probability proportional to

$$\frac{\theta_k^{y_i}\exp\{-\theta_k\}}{y_i!}$$

  for $k \neq i$.

Example 3. The DPMM for mixtures of normal distributions

Suppose that $Y_i|\theta_i \sim N(\theta_i, 1)$, where θ_i is from a Dirichlet process $\theta_i \sim \texttt{DP}(\alpha, G_0)$, and the base distribution is a conjugate standard normal distribution, $G_0 \equiv \texttt{N}(0, 1)$. Then by elementary calculations, if $\phi(.|\mu, \sigma^2)$ denotes the Normal density function for parameters (μ, σ^2), we have

$$\begin{aligned} q_0(y_i) &\propto \alpha\phi(y_i|0,2) \\ G_b(\theta_i|y_i) &\equiv \text{N}(y_i/2, 1/2) \end{aligned}$$

Hence for $i = 1, \ldots, N$, we draw θ_i according to:

- $\theta_i \sim \text{N}(y_i/2, 1/2)$ with probability proportional to $q_0(y_i)$.
- $\theta_i = \theta_k$ with probability proportional to $\phi(y_i|\theta_k, 1)$, for $k \neq i$.

The Bush-MacEachern algorithm

The algorithms described above draw θ_i one at a time, without considering the clustering property. Bush and MacEachern (1996 [13]) made a revision by considering clustering, thereby potentially improving the convergence properties of their algorithm. The Bush-MacEachern algorithm has two steps.

- assign $y_i (1 \leq i \leq N)$ to each component,
- draw parameters for each component.

Let $c = (c_1, \ldots, c_N)$ be the indicator variables which indicate the component to which each data point belongs. Suppose there are K $(1 \leq K \leq N)$ components (excluding c_i) when c_i is to be updated. Then,

$$p(c_i = k | y_i, c_{(-i)}) \propto \frac{n_{-i,k}}{N - 1 + \alpha} h(y_i | \theta_k), \tag{2.3}$$

determines the probability that y_i will be assigned to component k. Here $n_{-i,k}$ denotes the number of data points except y_i in component k. y_i also has the probability to form a new component with

$$p(c_i = K + 1 | y_i, c_{(-i)}) \propto \frac{\alpha}{N - 1 + \alpha} \int h(y_i | \theta) dG_0(\theta) \tag{2.4}$$

where K is the number of components (without c_i) when c_i is ready to be updated (different c_i may have different K). If y_i belongs to a singleton component, this component is completely removed when c_i is ready to be updated. After the indicator variables $c = (c_1, \ldots, c_N)$ are updated, the parameters are sampled according to the

posterior for each component

$$p(\theta_k|y_1,\ldots,y_N,c) \propto G_0(\theta_k) \prod_{i:\, c_i=k} h(y_i|\theta_k),$$

which may not be a standard distribution, but can be sampled using adaptive rejection sampling, the Metropolis-Hastings algorithm or other popular sampling algorithms.

Example 4. Normal example (continued).

Consider the normal example described above. First, the indicator variables are updated using

$$p(c_i = k|y_i, c_{(-i)}) \propto \begin{cases} \dfrac{n_{-i,k}}{N-1+\alpha}\phi(y_i|\theta_k,1) & \text{if } 1 \le k \le K, \\ \dfrac{\alpha}{N-1+\alpha}\phi(y_i|0,2) & \text{if } k = K+1. \end{cases}$$

After the indicator variables are updated, parameters are sampled according to

$$p(\theta_k|y_1,\ldots,y_N,c) \propto \phi(\theta|0,1) \prod_{i:\, c_i=k} \phi(y_i|\theta_k,1).$$

which yields for component k

$$\theta_k|y_1,\ldots,y_N,c \sim \mathrm{N}\left(\frac{n_k\overline{y}_k}{n_k+1}, \frac{1}{n_k+1}\right)$$

where

$$n_k = \sum_{i=1}^{n} I(c_i = k) \qquad \overline{y}_k = \frac{1}{n_k}\sum_{i=1}^{n} y_i I(c_i = k).$$

2.1.3 Posterior computation for the non-conjugate case

If G_0 is a conjugate prior then the integral $\int f(y_i|\theta_i)dG_0(\theta)$ can be calculated analytically. Escobar and West (1998) (Chapter 1 of [20]) contend that it is always reasonable to use conjugate prior as

- they are easy to deal with, especially for hierarchical models;
- although the use of conjugate prior restricts the flexibility, the interest is to find the distributions "around" G_0
- G is sampled from $\texttt{DP}(\alpha, G_0)$, which adds some flexibility

Although Escobar and West(1998) claimed that there is no reason to avoid the use of a conjugate prior, non-conjugate priors are still of interest as they provide additional flexibility. However, in the non-conjugate case, sampling θ_i is much less straightforward. In the next several sections, different sampling algorithms based on Polya urn schemes are introduced.

The No Gap algorithm

The "No Gap" algorithm (MacEachern and Muller 1998 [60]) is designed to deal with non-conjugate cases. As in the Bush-MacEachern algorithm, the No Gap algorithm also contains two steps: the update of indicator variables and the sampling of parameters. Let K denote the number of components (excluding c_i) when c_i is to be updated (different y_i may have different K).

1. **Case I:** If y_i belongs to a singleton component, then

$$c_i \longrightarrow \begin{cases} c_i & \text{with probability } \dfrac{K}{K+1}, \\ K+1 & \text{with probability } \dfrac{1}{K+1}. \end{cases}$$

where

- if $c_i \longrightarrow c_i$, then leave it unchanged and go to update next c_{i+1},
- if $c_i \longrightarrow K+1$, update c_i according to the following formula (2.5).

$$p(c_i = k | y_i, c_{(-i)}) \propto \begin{cases} n_{-i,k} h(y_i|\theta_k) & \text{if } 1 \leq k \leq K, \\ \dfrac{\alpha}{K+1} h(y_i|\theta_{K+1}) & \text{if } k = K+1. \end{cases} \tag{2.5}$$

  where θ_{K+1} is a new draw from G_0.

2. **Case II:** if y_i does not belong to a singleton component, directly use (2.5) to update c_i.

After the indicator variables $c = (c_1, \ldots, c_N)$ are updated, the parameters are sampled to each component with

$$p(\theta_k | y_1, \ldots, y_N, c) \propto G_0(\theta_k) \prod_{i:\, c_i = k} h(y_i|\theta_k) \qquad k = 1, \ldots, K$$

Neal's Algorithm 8

Neal (2000 [68]) updated the No Gap algorithm using auxiliary variables which only exist temporarily during the algorithm. Rather than considering only one new component, Algorithm 8 of Neal (2000 [68]) considered M new components. Suppose there are K components (excluding c_i) when c_i is to be updated, and that the auxiliary components are labeled as $K+1, \ldots, K+M$. If M is large enough, the empirical distribution of the auxiliary variables can be considered as an approximation to G_0. Neal's Algorithm 8 has three steps:

- sampling the auxiliary components;
- updating the indicator variables;

- sampling the parameters.

For $i = 1, \ldots, N$,

1. **Case I:** if y_i belongs to a singleton component,
   - assign the parameter θ_{c_i} to one of the auxiliary components, then
   - draw values from G_0 for the remaining $M - 1$ auxiliary components.
2. **Case II:** If y_i does not belong to a singleton component,
   - draw values from G_0 for all of the M auxiliary components.

The auxiliary components are temporary, which means that when the c_i finish updating, all the auxiliary parameters are deleted. After the auxiliary parameters are updated, the indicator variable c_i is updated according to the rule

$$p(c_i = k | y_i, c_{(-i)}) \propto \begin{cases} \dfrac{n_{-i,k}}{N - 1 + \alpha} h(y_i|\theta_k) & \text{if } 1 \leq k \leq K, \\ \dfrac{\alpha/M}{N - 1 + \alpha} h(y_i|\theta_k) & \text{if } K < k \leq K + M. \end{cases}$$

When updating of the c_i is completed, all the auxiliary components which have not been occupied are deleted. Finally, the parameters are sampled for each component with

$$p(\theta_k | y_1, \ldots, y_N, c) \propto G_0(\theta_k) \prod_{i:c_i=k} h(y_i|\theta_k).$$

2.2 Bayesian nonparametrics for survival analysis

In this section, we introduce the application of Bayesian nonparametrics to estimate survival functions.

2.2.1 Dirichlet process mixture formulation

Doss (1994 [22]) proposed a successive substitution sampling approach based on the Dirichlet process for survival analysis. Suppose observations $Y_1 \dots Y_N$ are drawn independently from F, where Y_i is an exact observation or it can be a censored data (right, left or interval censored) that lie in an interval A_i. The prior on F is a mixture of Dirichlet process models, $F \sim DP(\alpha, G_0(\theta))$, where parameter θ in the base distribution has its own prior distribution $\pi(.)$, that is

$$F|\theta \sim \mathtt{DP}(\alpha, G_0(\theta)) \qquad \theta \sim \pi(d\theta).$$

The objective is to obtain the posterior distribution of F given the incomplete data. The main idea in Doss (1994 [22])'s approach is to generate $F^{(\nu)}$ at each iteration ν, and obtain the quantities of interest through these generated $F^{(\nu)}$.

Doss (1994 [22])'s successive substitution sampling approach is an iterative algorithm in which, at each step, the c.d.f F is estimated. Then the corresponding $Z_1, \dots, Z_N$ are sampled, where $Z_i = Y_i$ if Y_i is an exact observation, and Z_i is a sampled value from interval A_i if $Y_i \in A_i$ is a censored observation.

We start with an initial guess of $F^{(0)}$ and $Z_1^{(0)} \dots Z_N^{(0)}$, and the two steps at each iteration ν are first

$$\text{Generate } F^{(\nu)} \sim DP\left(\alpha + N, \frac{\alpha G_0(\theta) + \sum_{i=1}^{N} \delta_{Z_i^{(\nu-1)}}(.)}{\alpha + N}\right),$$

and second

$$\text{Generate } (Z_1^{(\nu)}, \dots, Z_N^{(\nu)}) \sim H(Z_1 \dots Z_N | F^{(\nu)}, Y_i \in A_i, i = 1 \dots N),$$

where $F^{(\nu)}$ and $(Z_1^{(\nu)}, \ldots, Z_N^{(\nu)})$ are the samples to be drawn at iteration ν , and $H(Z_1, \ldots, Z_N | F^{(\nu)}, Y_i \in A_i, i = 1 \ldots N)$ is the conditional distribution of $Z_1, \ldots, Z_N$ given $F^{(\nu)}$ and Y. After Υ steps are completed, we have obtained Υ estimates $F^{(1)} \ldots F^{(\Upsilon)}$. As result, the distributions for the quantity of interest, like the mean or median of F can be obtained through $F^{(1)} \ldots F^{(\Upsilon)}$.

Next, we introduce the details of the two steps, assuming that θ is known. In the first step, we first draw a large number of samples (say d samples) $\zeta_1, \ldots, \zeta_d$ from the Dirichlet process with precision parameter $\alpha + N$ and base distribution

$$\frac{1}{(\alpha + N)} \left(\alpha G_0(B) + \sum_{i=1}^{N} \delta_{Z_i^{(\nu-1)}}(B) \right).$$

The base distribution can be expressed using a stick-breaking process (Doss 1994 [22]). We first generate

$$\zeta_1, \ldots, \zeta_d \sim \begin{cases} G_0 & \text{with probability } \alpha/(\alpha + N) \\ Z_i & \text{with probability } 1/(\alpha + N) \end{cases}$$

independently, and then generate $v_1, \ldots, v_d \sim \texttt{Beta}(1, \alpha + N)$, and set

$$\pi_i = v_i \prod_{j=1}^{i-1} v_j.$$

So at iteration ν, $\widehat{F}^{(\nu)}$ can be expressed as a mixture distribution of ζ_j with weights π_j,

$$\widehat{F}^{(\nu)} \approx \sum_{j=1}^{d} \pi_j \zeta_j.$$

from which quantities of interest, such as median(F), can be obtained directly.

In the second step, after $\widehat{F}^{(\nu)}$ is obtained, new simulated data $Z_1 \dots Z_N$ are drawn from $\widehat{F}^{(\nu)}$. If Y_i is an exact observation, then $Z_i = Y_i$; if $Y_i \in A_i$ is censored, then generate

$$U_i^{(1)} \sim U(0,1).$$

Choose $J_i^{(1)}$ such that

$$\sum_{j=1}^{J_i^{(1)}-1} \pi_j \leq U_i^{(1)} \leq \sum_{j=1}^{J_i^{(1)}} \pi_j.$$

If $\zeta_{J_i^{(1)}} \in A_i$, set

$$Z_i^{(k)} = \zeta_{J_i^{(1)}}.$$

Otherwise, repeat the process using independent uniforms $U_i^{(2)}, \dots, U_i^{e_i}$ until observe one $\zeta_{J_i^{(e_i)}} \in A_i$. After all the $Z_1^{(\nu)}, \dots Z_N^{(\nu)}$ are sampled, they are used in the next step as part of the base distribution in the Dirichlet process. The two steps are repeated many times to produce Υ estimators of the quantities of interest.

We have reviewed the Dirichlet process mixture model, and discussed some applications. In next section, we give a brief introduction to another popular Bayesian nonparametric model: Polya tree.

2.3 The Polya tree

This section reviews the formulation of the *Polya tree* model, for which full details can be found from Lavine (1992 [54], 1994 [55]), Hanson and Johnson (2002 [37]) and Hanson (2006 [36]). Suppose a random sample $Y = (Y_1, \dots, Y_N)$ is observed. The traditional parametric approach models the random sample through a parametric distribution with cumulative distribution function $G_0(\theta)$ (the corresponding density $g_0(\theta)$). Let the support of the distribution be Ω.

The Polya tree model is specified through a hierarchical, recursive partition of the support of Ω. At the first level, Ω is partitioned into two intervals $B(0)$ and $B(1)$, with $\Omega = B(0) \cup B(1)$ and $B(0) \cap B(1) = \phi$. At the second level $B(0)$ is split into $B(00)$ and $B(01)$, with $B(0) = B(00) \cup B(01)$ and $B(00) \cap B(01) = \phi$. $B(1)$ is partitioned in the similar way. Following Hanson (2006 [36]), at level j, denote $e_j(k)$ to be the j-fold binary sequence representation of partition component $k-1$. For example, suppose

$$e_j(k) = \epsilon_1 \dots \epsilon_j;$$

we have $e_2(1) = 00$, and $e_4(6) = 0101$ and so on. Note that at level j, there will be 2^j intervals. Following the recursive binary partition, we know $B(\epsilon_1 \dots \epsilon_j) = B(\epsilon_1 \dots \epsilon_j 0) \cup B(\epsilon_1 \dots \epsilon_j 1)$.

The partition points at each level are constructed through the *canonical partition* of Lavine (1992 [54]), in which the partition points are specified to be the quantiles of the parametric distribution $G_0(\theta)$. At the first level,

$$B(0) = (G_0^{-1}(\theta)(0), G_0^{-1}(\theta)(1/2)) \qquad B(1) = (G_0^{-1}(\theta)(1/2), G_0^{-1}(\theta)(1)).$$

At each level j, define the intervals

$$B(e_j(k)) = (G_0^{-1}(\theta)((k-1)/2^j), G_0^{-1}(\theta)(k/2^j)) \qquad k = 1, \dots, 2^j - 1$$

and

$$B(e_j(2^j)) = (G_0^{-1}(\theta)((2^j-1)/2^j), G_0^{-1}(\theta)(1)),$$

where $G_0^{-1}(\theta)(q)$ denotes the qth quantile of G_0. Let

$$\Pi_j \equiv \{B(e_j(k)) : k = 1 \dots 2^j\}$$

be the set of partition intervals of Ω at level j. Note that, under the canonical partition, every set in Π_j contains probability 2^{-j} by construction.

Definition Polya tree (Lavine 1992 [54]; Hanson 2006 [36]).

A random probability measure G on a separable measurable space (Ω, F) has a *Polya tree distribution*, with parameter $(\alpha, \rho, G_0(\theta))$, written as $G \sim \mathtt{PT}(\alpha, \rho, G_0(\theta))$, if there exist random vectors

$$\Upsilon = \{(\Xi_{e_j(k)0}, \Xi_{e_j(k)1}) : k = 1 \dots 2^{j-1}; j = 1 \dots M\}$$

such that

1. All the random vectors in Υ are independent;
2. $\{(\Xi_{e_j(k)0}, \Xi_{e_j(k)1})\} \sim \text{Beta}(\alpha\rho(j), \alpha\rho(j))$, more precisely

$$\Xi_{e_j(k)0} \sim \text{Beta}(\alpha\rho(j), \alpha\rho(j)) \qquad \Xi_{e_j(k)1} = 1 - \Xi_{e_j(k)0};$$

3. For every $B(\epsilon_1 \dots \epsilon_j) \in \Pi_j$,

$$G(B(\epsilon_1 \dots \epsilon_j)) = \prod_{k=1}^{j} \Xi_{\epsilon_1 \dots \epsilon_k}.$$

Polya trees enjoy a conjugacy result. Suppose a random sample $Y_1 \dots Y_N \sim G$, with $G \sim PT(\alpha, \rho, G_0(\theta))$. In light of the data, the posterior on G, $G|Y$, is also a

Polya tree with parameters updated as

$$(\Xi_{e_j(k)0}, \Xi_{e_j(k)1})|Y \sim \text{Beta}(\alpha_{e_j(k)0}, \alpha_{e_j(k)1}) = \text{Beta}(\alpha\rho(j)+n(e_j(k)0), \alpha\rho(j)+n(e_j(k)1)), \tag{2.6}$$

where $n(\epsilon_1 \dots \epsilon_j)$ is the number of elements of the observations in the interval $B(\epsilon_1 \dots \epsilon_j)$. In this formulation, the partition points do not change, but the parameters $(\Xi_{e_j(k)0}, \Xi_{e_j(k)1})$ are updated through (2.6). We term the $\alpha\rho(j) + n(e_j(k))$ for interval $B(e_j(k))$ as the "beta weight" value.

Quantity $\rho(j)$ is a function of level j of the hierarchical partition. Walker and Mallick (1997 [66]) suggest the choice $\rho(j) = j^2$. Ferguson (1974 [29]) pointed out that the parameters set to j^2 place prior probability one on absolutely continuous distributions (see also Kraft 1964 [15]). The Dirichlet process is recovered if $\rho(j)$ is set proportional to 2^{-j} (Lavine (1992,1993) [54, 55]).

The partition points are decided through the parametric distribution $G_0(\theta)$, termed the base distribution. The Polya tree approach generalizes the "guess" distribution, and it is also centered at the base distribution.

$$E[G(B(\epsilon_1 \dots \epsilon_j))] = \frac{1}{2^j} = G_0(B(\epsilon_1 \dots \epsilon_j))$$

Fixing α at a large value reflects the strong belief that the true distribution should be close to the parametric base distribution $G_0(\theta)$. As $\alpha \to \infty$, the Polya tree G equals $G_0(\theta)$ with probability 1. Conversely, if small values of α are used, the Polya tree posterior will be close to the empirical distribution of data.

Hanson (2006 [36]) pointed out that "The Polya tree model provides an "intermediary" between the empirical c.d.f and a given parametric distribution in a manner

similar to the Dirichlet Process", and found that adding levels to a Polya tree can not guarantee the improvement of model fit or predictive utility. He suggested a rule of thumb

$$M = -\log_2 \left(\frac{E}{N} \right),$$

where N is the sample size, E is a "...'typical' number of observations falling into each set at level M" (Hanson 2006 [36]), that is, the number of observations expected to fall within each category at the final level of the hierarchical partition; E/N is the corresponding proportion of observations.

The probability density function (p.d.f) value of Y_i can be easily obtained. Suppose $Y_i \in B(\epsilon_1 \dots \epsilon_M)$ at level M, the p.d.f value can be calculated through

$$g(Y_i) = 2^M g_0(Y_i) G(B(\epsilon_1 \dots \epsilon_j)),$$

where $g_0(Y_i)$ is the p.d.f value of Y_i from the base distribution, and $G(B(\epsilon_1 \dots \epsilon_j))$ is the posterior probability measure corresponding to the interval $B(\epsilon_1 \dots \epsilon_M)$ in which Y_i lies.

Example 5. Two component mixture of normals

Consider a random sample of size 100,

$$Y_i \sim 0.5\,\mathrm{N}(-3, 1) + 0.5\,\mathrm{N}(3, 1).$$

The mean of the sample is 0.053, and the standard deviation is 3.18. A naive parametric approach may model the data through a single Normal distribution

$$G_0(\theta) = \mathrm{N}(0.053, 3.18).$$

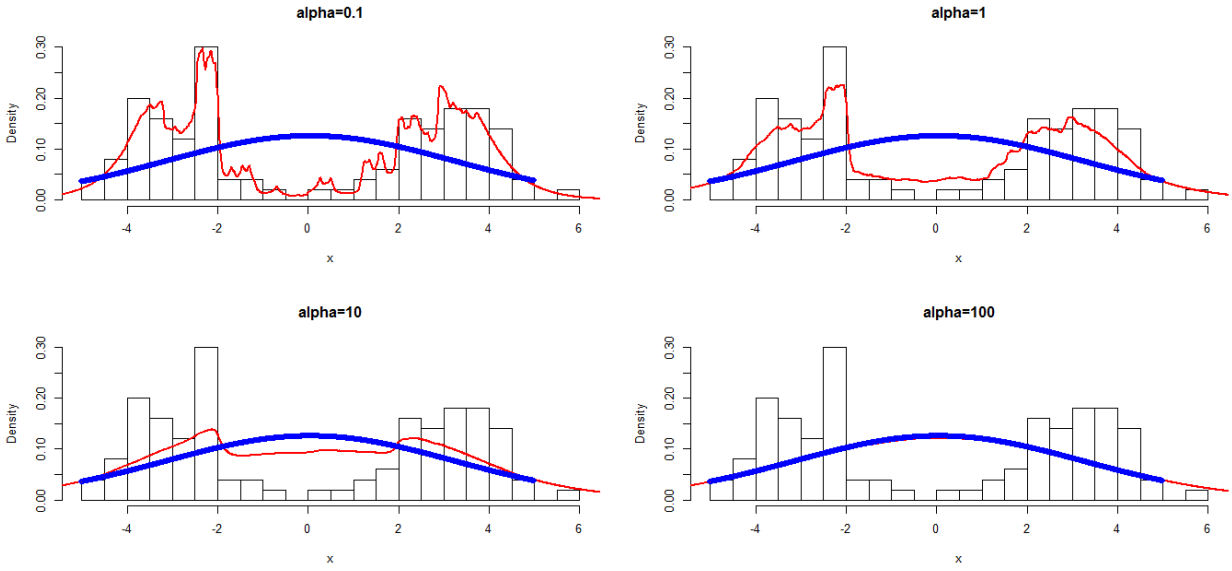

Figure 2–2: The histogram, Polya tree fit (red line), and parametric model fit, N(0.053,3.18) (blue line). The values of α are fixed as 0.1, 1, 10 and 100

The histogram and the density curve of N(0.053,3.18) (blue line) are plotted in Figure (2–2). The parametric model fails to capture the bimodal structure of the original data. Consider instead a Polya tree model with maximum level six ($\log_2(100) \approx 6$). The partition points and intervals at the first two levels are

$$B(e_1(1)) = (\infty, 0.053); B(e_1(2)) = (0.053, \infty),$$

$$B(e_2(1)) = (\infty, -2.09); B(e_2(2)) = (-2.09, 0.053); B(e_2(3)) = (0.053, 2.2); B(e_2(4)) = (2.2, \infty).$$

Note that $G_0^{-1}(\theta; 0.25) = -2.09$, $G_0^{-1}(\theta; 0.5) = 0.053$ and $G_0^{-1}(\theta; 0.75) = 2.2$.

We fit the Polya tree using the `R` package "`DPpackage`", and the fit is plotted in Figure 2–2. When the α value is small (upper left plot, $\alpha = 0.1$), the Polya tree fit is close to the empirical distribution, and the bimodal structure is captured. When

the α value is large (lower right plot, $\alpha = 100$), the Polya tree fit is almost the same as the parametric model (the blue and red lines coincide). Another finding from Figure 2–2 is that as α becomes larger and larger, the curve becomes smoother and smoother.

2.4 Applications

In next several subsections, we introduce some biostatistical and other applications using the Polya Tree model.

2.4.1 Meta-analysis (Modeling the log of odds ratio)

We consider N independent case-control studies examining the relationship between an exposure and disease. We consider an analysis based on the standard frequentist asymptotic approximation of the log odds ratio; denote the observed log odds ratios as $Y_1 \dots Y_N$, and assume they arise from normal distributions

$$Y_i|\theta_i, \sigma_i \sim \mathrm{N}(\theta_i, \sigma_i^2).$$

where σ_i is known to us, and it is study dependent, and a function of the study sample size for study i. Branscum and Hanson (2008 [9]) considered modeling θ as a random distribution F with a mixture of Polya trees prior.

$$\theta_i \sim \mathtt{PT}(\alpha, \rho, G_0),$$

with base distribution $G_0 \equiv \mathtt{N}(\mu, \tau^2)$. The parameters in the base distribution μ and τ^2 have their own prior distribution,

$$\mu \sim \mathrm{N}(\mu_b, S_b^2),$$

$$\tau^{-2} \sim \text{Gamma}\left(\frac{\tau_1}{2}, \frac{\tau_2}{2}\right),$$

where $\mu_b, S_b^2, \tau_1, \tau_2$ are specified in advance. Since a conjugate prior is used, the implementation of the Metropolis-Hastings method is straightforward, and it is introduced in Branscum and Hanson (2008 [9]).

1. Draw
$$\theta_i^* \sim \text{N}\left(\frac{\tau^2 Y_i + \sigma_i^2 \mu}{\tau^2 + \sigma_i^2}, \frac{\sigma_i^2 \tau^2}{\sigma_i^2 + \tau^2}\right),$$
and accept it with probability
$$\min\left(1, \frac{\texttt{PT}(\alpha, \rho, \text{N}(\mu, \tau^2))(\theta_i^*)}{\texttt{PT}(\alpha, \rho, \text{N}(\mu, \tau^2))(\theta_i)}\right),$$
where $\texttt{PT}(\alpha, \rho, \text{N}(\mu, \tau^2))(\theta_i^*)$ denotes the probability measure of the set containing θ_i^* of Polya tree. The precision is α and base distribution is $\text{N}(\mu, \tau^2)$:
2. Draw
$$\mu^* \sim \text{N}\left(\frac{S_b^2 \overline{\theta} + \tau^2 \mu_b}{\tau^2 + S_b^2}, \frac{\tau^2 S_b^2}{\tau^2 + S_b^2}\right),$$
and accept it with probability
$$\min\left(1, \frac{\prod_{i=1}^{N} \texttt{PT}(\alpha, \rho, \text{N}(\mu^*, \tau^2))(\theta_i)}{\prod_{i=1}^{N} \texttt{PT}(\alpha, \rho, \text{N}(\mu, \tau^2))(\theta_i)}\right).$$
3. Draw
$$(\tau^*)^{-2} \sim \text{Gamma}\left(\tau_1 + \frac{N+1}{2}, \tau_2 + \frac{1}{2}(\mu - \mu_b)^2 + \frac{1}{2}\sum_{i=1}^{N}(\theta_i - \mu)^2\right),$$

and accept it with probability

$$\min\left(1, \frac{\prod_{i=1}^{N} \mathtt{PT}(\alpha, \rho, \mathrm{N}(\mu, (\tau^*)^2))(\theta_i)}{\prod_{i=1}^{N} \mathtt{PT}(\alpha, \rho, \mathrm{N}(\mu, \tau^2))(\theta_i)}\right).$$

After Υ iterations, we can obtain Υ samples for $\theta_i, i = 1 \ldots N$, μ, and τ. As a result the standard error, mean and other quantity of interest of these estimates are available.

Note that we can fix the mean of the base distribution μ to be the median of F in a straightforward way, and it will have a clear interpretation. However, the τ^2 is just the variance for the base distribution, and not the variance for F, so it does not have a clear interpretation. In Chapter 4, we will show that in the proposed recursive Polya tree mixture model, both μ and τ have clear interpretations.

2.4.2 Survival analysis

Muliere and Walker (1997 [66]) presented a Bayesian nonparametric approach to survival analysis using a Polya tree. Their analysis assigns a Polya tree prior to the space of survival curves, and based survival statements on the posterior predictive probabilities for future observations. They review some previous work based on the extended gamma process and the beta process. Their approach is easy to implement, easy to obtain posterior samples, and can be thought of as a generalization of the Kaplan-Meier estimator. Muliere and Walker's work is the first one to apply Polya tree model to survival analysis. We will review their approach through a real data set. There are two important steps to apply Polya trees to estimate survival functions: the partitioning approach and the construction of predictive distributions.

Partitioning approach: The partitioning approach of Muliere and Walker (1997 [66]) still relies on the recursive binary partitions of Lavine (1992 [54]). All the censored data are set to the partition points at each level. We take an example to explain their partition approach. Consider the data from Kaplan and Meier (1958 [47])

Deaths occurred: 0.8, 3.1, 5.4, 9.2;

Censoring occurred: 1.0, 2.7, 7.0, 12.1.

At the first level, consider the first censored time 1.0;

$$B(0) = [0, 1.0); B(1) = [1.0, \infty).$$

At the second level, consider the second censored time 2.7 and first death time 0.8;

$$B(00) = [0, 0.8); B(01) = [0.8, 1.0); B(10) = [1.0, 2.7); B(11) = [2.7, \infty).$$

At the third level, since no observations occur in $B(00)$ and $B(01)$, they do not need further partition. The third censored time 7.0 occurs in $B(11)$, so

$$B(110) = [2.7, 7.0); B(111) = [7.0, \infty).$$

At the fourth level;

$$B(1100) = [2.7, 3.1); B(1101) = [3.1, 7.0); B(1110) = [7.0, 12.1); B(1111) = [12.1, \infty).$$

Finally at the fifth level;

$$B(11010) = [3.1, 5.4); B(11011) = [5.4, 7.0); B(11100) = [7.0, 9.2); B(11101) = [9.2, 12.1)$$

Predictive distributions: If there are no censored observations, the posterior predictive distribution is

$$P(Y_{N+1} \in B(\epsilon_1 \dots \epsilon_M)) = \frac{\alpha_{\epsilon_1} + N_{\epsilon_1}}{\alpha_{\epsilon_1} + \alpha_{\epsilon_0} + N} \frac{\alpha_{\epsilon_1\epsilon_2} + N_{\epsilon_1\epsilon_2}}{\alpha_{\epsilon_1 0} + \alpha_{\epsilon_1 1} + N_{\epsilon_1}} \dots \frac{\alpha_{\epsilon_1 \dots \epsilon_M} + N_{\epsilon_1 \dots \epsilon_M}}{\alpha_{\epsilon_1 \dots \epsilon_{M-1} 0} + \alpha_{\epsilon_1 \dots \epsilon_{M-1} 1} + N_{\epsilon_1 \dots \epsilon_{M-1}}}, \tag{2.7}$$

where $\alpha_\epsilon + N_\epsilon$ is the "beta weight" (see section 2.3) for interval $B(\epsilon)$. If right censored observations exist, Muliere and Walker (1997 [66]) update (2.7) by deleting the censored data from the risk set (which is the denominator) in the denominator. The probability $P(Y_{N+1} \in B(\epsilon_1 \dots \epsilon_M))$

$$P(Y_{N+1} \in B(\epsilon_1 \dots \epsilon_M)) = \frac{\alpha_{\epsilon_1} + N_{\epsilon_1}}{\alpha_{\epsilon_1} + \alpha_{\epsilon_0} + N} \dots \frac{\alpha_{\epsilon_1 \dots \epsilon_M} + N_{\epsilon_1 \dots \epsilon_M}}{\alpha_{\epsilon_1 \dots \epsilon_{M-1} 0} + \alpha_{\epsilon_1 \dots \epsilon_{M-1} 1} + N_{\epsilon_1 \dots \epsilon_{M-1}} - Nc_{\epsilon_1 \dots \epsilon_{M-1}}}, \tag{2.8}$$

where N_ϵ and Nc_ϵ are the number of deaths and censored observations in interval $B(\epsilon)$. Muliere and Walker (1997 [66]) observed the fact that (2.8) reduces to the Kaplan-Meier estimator if all the α values are set to 0 and the convention $\frac{0}{0} = 0$ is adopted. A parametric distribution is assigned to the base distribution G_0; for example, an exponential distribution may be appropriate for the data presented above (see Muliere and Walker 1997 [66]). Muliere and Walker (1997 [66]) suggest

$$\alpha_{\epsilon_1 \dots \epsilon_j} = \alpha \times j^2 \times G_0(B(\epsilon_1 \dots \epsilon_j)),$$

where, recall, $G_0(B(\epsilon_1 \dots \epsilon_j))$ is the probability measure of base distribution G_0 on the set $B(\epsilon_1 \dots \epsilon_j)$. In the above example, Muliere and Walker (1997 [66]) assigned

a exponential distribution with parameter 0.12 to the base distribution, so that

$$G_0(B(\epsilon_1 \ldots \epsilon_j)) = \int_{B(\epsilon_1 \ldots \epsilon_j)} 0.12 \exp(-0.12z)dz$$

The hyperparameter can be adjusted to capture genuine prior knowledge.

2.4.3 Simulating the Polya tree prior or posterior

Simulating data from posterior distribution is easy in Polya tree. At first level, select $B(0)$ or $B(1)$ with probability Ξ_0 and Ξ_1. At level j, if a set $B(\epsilon_1 \ldots \epsilon_{j-1})$ is chosen, the probability to select $B(\epsilon_1 \ldots \epsilon_{j-1}0)$ or $B(\epsilon_1 \ldots \epsilon_{j-1}1)$ is $(\Xi_{\epsilon_1 \ldots \epsilon_{j-1}0}, \Xi_{\epsilon_1 \ldots \epsilon_{j-1}1})$.

At the maximum level M, if $B(\epsilon_1 \ldots \epsilon_M)$ is selected, a sample is taken (uniformly) from $B(\epsilon_1 \ldots \epsilon_M)$.

Example 6. Kaplan-Meier's data survival analysis

The following results are obtained to compute posterior quantities of interest for the Kaplan-Meier data: Doss (1994 [22])'s approach (see section 2.2) and Muliere and Walker (1997 [66])'s Polya Tree model are used. Table 2–1 show the results. The first

t	0.8	1.0+	2.7+	3.1	5.4	7.0+	9.2	12.1+
`KME`	0.125	0.125	0.125	0.300	0.475	0.475	0.738	0.738
`Doss`	0.108	0.120	0.207	0.300	0.472	0.526	0.682	0.750
`Prior`	0.090	0.110	0.270	0.300	0.470	0.560	0.660	0.760
`M & W` $\alpha = 1$	0.050	0.110	0.190	0.220	0.390	0.520	0.600	0.760
`M & W` $\alpha = 10$	0.080	0.110	0.260	0.290	0.460	0.550	0.650	0.760
`M & W` $\alpha = 0.01$	0.110	0.110	0.110	0.300	0.480	0.480	0.740	0.740

Table 2–1: Estimator of c.d.f values of Kaplan and Meier's data using Doss (1994) and Muliere & Walker's (1997) approaches

row shows the result from the Kaplan-Meier estimator. The second row summarizes

the estimates from Doss (1994 [22])'s approach. In the third row, a parametric model with a exponential distribution is fit to the data. It is interesting to note that Doss's approach is a intermediary between the nonparametric Kaplan-Meier estimator and the parametric estimator. In rows 4 to 6, the Polya tree with different α values is implemented. As expected, large α values give the results close to parametric estimates; and small α values give the results close to the nonparametric estimates.

Comments: Muliere and Walker's work is the first one to apply Polya tree to survival analysis based on arbitrarily right censored data. Their approach is easy to implement, and provides a intermediary between the nonparametric Kaplan Meier estimator and parametric approaches. However, their partition approaches relies on the recursive binary split, and all the censoring observations are set to the partition points at each level. As a result, the maximum level M must be greater than the number of distinct censored observations, and this is not an appealing feature. The recursive binary partition approach is also a difficult way to handle the case of interval censoring. In Chapter 5 of this thesis, we will present a new partition approaches that can overcome the difficulties caused by recursive binary partition approach.

2.5 Other Bayesian nonparametric models

2.5.1 Sequential imputation

Liu (1996 [57]) developed a sequential imputation algorithm for binomial problems. Consider the binomial random sample $(X_1, Y_1), \ldots (X_N, Y_N)$. The number of trials is Y_i and the number of successes is X_i. Assume the probability of success is

θ_i for (X_i, Y_i); it is natural to model it using a binomial distribution:

$$X_i \sim \text{Binomial}(Y_i, \theta_i),$$

$$\theta_1, \ldots, \theta_N \sim F \text{ i.i.d}$$

In Example 1, Section 2.1.2, we introduced an MCMC algorithm which can be used to estimate each individual θ_i. If we are further interested in the random distribution F, it is not that easy. As Liu (1999 [58]) comments that "Since F is an infinite-dimensional parameter, there is no easy way of displaying its full posterior distribution." In this section we introduce the sequential imputation algorithm from Liu (1996 [57]).

As discussed in Section 2.1.2, assume the random effect θ_i arises from a Dirichlet process,

$$\theta_i \sim \texttt{DP}(\alpha, G_0).$$

The base distribution is the conjugate Beta distribution,

$$G_0 \equiv \text{Beta}(a, b).$$

The quantities G_b, q_0 and q_k are

$$G_b(\theta_i; a, b) = \text{Beta}(X_i + a - 1, Y_i - X_i + b - 1).$$

$$q_0(x_i) \propto \alpha \frac{\Gamma_{a+b}}{\Gamma(a)\Gamma(b)} \int_0^1 t^{x_i+a-1}(1-t)^{y_i-x_i+b-1} dt = \alpha B_{ab}(y_i, x_i),$$

The quantity $q_k(x_i)$ is the density value for the binomial distribution with parameters (y_i, θ_k) evaluated at x_i.

The sequential imputation strategy still samples each θ_i in the same way, but Liu (1996 [57]) also considered the weight corresponding to each draw. For the lth draw,

$$w_l = p(x_1) \prod_{t=2}^{N} p(x_t | \theta_{l1}, \ldots, \theta_{l(t-1)}). \tag{2.9}$$

Each draw consists of N values of θ: $\theta_{l1}, \ldots, \theta_{lN}$. For the binomial mixture model, each term in (2.9) is:

$$p(x_i | \theta_{l1}, \ldots, \theta_{l(i-1)}) = \frac{\alpha}{\alpha + i - 1} B_{ab}(y_i, x_i) + \frac{1}{\alpha + i - 1} \sum_{t=1}^{i-1} \theta_{l(t)}^{x_i} (1 - \theta_{l(t)})^{(y_i - x_i)}$$

It is possible to estimate the posterior distribution by a weighted mixture. From Theorem 5 of Liu (1996 [57]), the posterior distribution of G can be approximated by a weighted mixture of Dirichlet processes:

$$p(G|X) \approx \frac{1}{W} \sum_{l=1}^{L} w_l \mathtt{DP} \left(\alpha G_0 + \sum_{i=1}^{N} \delta(\theta_{li}) \right),$$

where $W = w_1 + \cdots + w_L$ for L draws, and θ_{li} denotes the ith value in the lth draw.

The posterior expectation of G, which is also the predictive distribution for a future θ is expressed as a weighted mixture of αG_0 and point masses (Theorem 5 of Liu 1996 [57])

$$E(G|X) \approx \frac{1}{\int_0^1 \alpha G_0(du) + N} \left\{ \alpha G_0 + \frac{1}{W} \sum_{l=1}^{L} \sum_{i=1}^{N} w_l \delta(\theta_{li}) \right\}.$$

Posterior means and variances of the θ_i are approximated by

$$E(\theta_i | X) \approx \frac{1}{W} \sum_{l=1}^{L} w_l \theta_{li},$$

$$\mathrm{Var}(\theta_i|X) \approx \frac{1}{W}\sum_{l=1}^{M} w_l\theta_{li}^2 - \{E(\theta_i|X)\}^2$$

Liu (1996 [57]) also derives the approach to obtain the optimal value of α.

As Liu (1996 [57]) mentioned: "direct application of Gibbs sampling is difficult since drawing the infinite-dimensional parameter can not be done cheaply". Sequential imputation provides an approach to approximate G, which is a significant improvement of Polya urn Gibbs sampling. Although Liu (1996 [57]) only discussed the Binomial mixture model, it is very easy to extend the sequential imputation approach to other kinds of models, like Gaussian mixture or Poisson mixture. In below, we briefly introduce an analysis of the rolling thumbtacks in Liu (1996 [57]).

Example 7. Tossing Thumbtacks Beckett and Diaconis (1994 [5]) generated binary strings from rolls of common thumbtacks. There were 320 tacks which were flicked, 9 times each, and the numbers of times that the tacks landed point up were recorded. This process produces 320 data points which can be modeled as a binomial distribution.

$$X_i \sim \text{Binomial}(9, \theta_i), i = 1 \ldots 320 \text{ independently}$$

The number of trials Y_i are 9 for $i = 1 \ldots 320$, and the number of successes X_i were recorded and are shown in Table 2–2.

Number of "success"	0	1	2	3	4	5	6	7	8	9
Frequencies	0	3	13	18	48	47	67	54	51	19

Table 2–2: Thumbtacks data. The first row shows the number of tacks landed point up; the second row shows the corresponding number of data points.

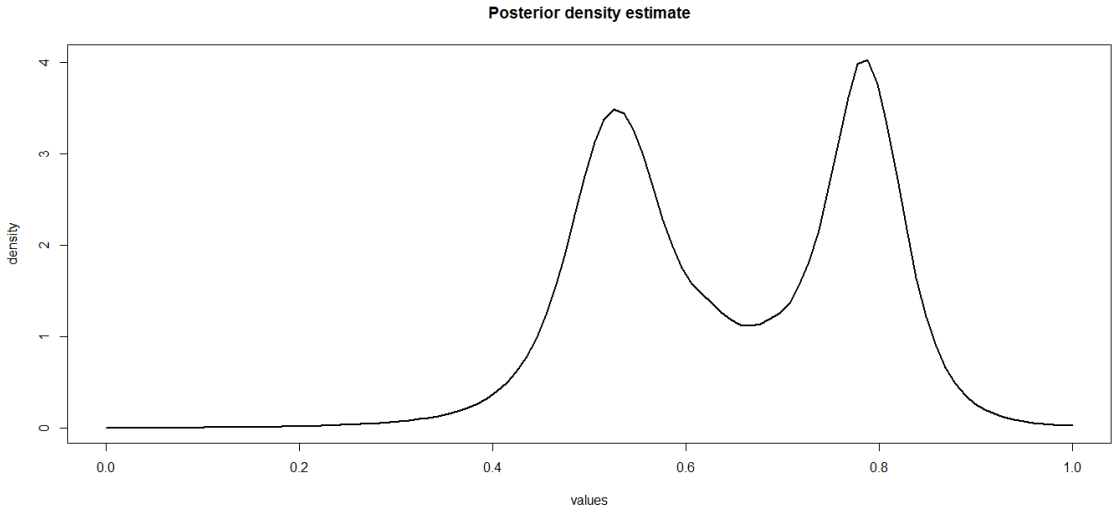

Figure 2–3: Example 7 (Thumbtack data): Density estimation using the prior from Liu (1996)

Liu (1996 [57]) analyzed these data by assuming θ_i arise from a nonparametric random distribution,

$$\theta_i \sim F \text{ i.i.d.}$$

The objective of Liu (1996 [57])'s work is to estimate the posterior mean of F, which is $E(F|\text{data})$. Liu (1996 [57]) used sequential imputation and found an unusual feature, that is, that the estimate of F demonstrated a surprising bi-modal pattern. He pointed out that "this feature is unexpected and cannot be revealed by a regular parametric hierarchical analysis using the Beta-binomial priors". Liu (1996 [57])'s result is plotted in Figure 2–3 computed through the `R` package "`DPpackage`". The finding of this unusual feature reflects the advantage of Bayesian nonparametrics. Through assuming F arise from a nonparametric random distribution, we can utilize information that can not be found through a parametric Bayesian hierarchical model.

2.5.2 Predictive recursion

Newton (1999 [70], 2002 [69]) proposed a fast recursive algorithm to estimate the unknown F. Predictive recursion requires users to define a finite support of F, and starts from a large number, K, of fixed mass points whose corresponding probability density function, f, is estimated in a sequential fashion.

1. Choose an initial guess f_0 for f, and a sequence of weights $w_1, \ldots, w_N \in (0, 1)$.
2. Set
$$f_i(\theta) = (1 - w_i) f_{i-1}(\theta) + w_i \frac{h(Y_i|\theta) f_{i-1}(\theta)}{\int_\Theta h(Y_i|\theta') f_{i-1}(\theta') d\theta'}, i = 1, \ldots, N$$
3. Repeat the processes many times and average the results.

Newton (2002 [69]) starts with a finite support for θ, defines K mass points uniformly in this support, and provides an initial estimate of the density values $f_0(\theta_1), \ldots, f_0(\theta_K)$. In the first step, these values are updated as

$$f_1(\theta_j) = (1 - w_1) f_0(\theta_j) + w_1 \frac{h(Y_1|\theta_j) f_1(\theta_j)}{\sum_{h=1}^{K} h(Y_1|\theta_h) f_1(\theta_h)}, j = 1 \ldots K$$

Only Y_1 is used to update the p.d.f. values for all of the mass points. In the next step, Y_2 is used, and this process continues from Y_1 to Y_N.

Predictive recursion is very fast, efficient and robust. However, the order dependence (that is, the order in which the data is introduced) is a limitation. Newton (2002 [69]) showed that the order dependence is weak, and can be minimized by repeating the algorithm many times and averaging the result. However, this order dependence is still an unappealing property.

In this chapter, we briefly introduce the basic ideas and some applications of Bayesian nonparametric models. In next chapter, we introduce the proposed recursive Polya tree mixture model and discuss its advantages.

CHAPTER 3
The Recursive Polya tree mixture model

In this Chapter, I introduce the recursive Polya tree mixture model that is computationally efficient and does not suffer from the drawbacks identified in the previous Chapter.

3.1 Introduction

In Bayesian hierarchical model,

$$Y_i|\theta_i \sim h(Y_i|\theta_i), i = 1, \ldots, N, \text{independently}$$

$$\theta_i \sim F, i = 1, \ldots, N \text{ i.i.d.}$$

The marginal distribution is a mixture distribution with

$$Y_i \sim \int h(Y_i|\theta_i) dF(\theta_i).$$

and the Y_i are conditionally independent, but not independent.

In general, there are two goals for this hierarchical model:

- estimating the distribution function F,
- estimating each individual random effect θ_i corresponding to Y_i.

To construct a Bayesian solution to these two problems, I will assume that F has a Polya tree distribution and apply the "empirical Bayes" idea to choose the base

distribution. Specifically, the empirical distribution is regarded to be the base distribution. However, note that the θ are unobservable or latent variables, and their empirical distribution is unavailable to us.

I develop what I term the *recursive Polya tree mixture model* (`RPTMM`) based on an iterative algorithm. At each iteration step, the empirical distribution of the θs is re-estimated. In the next step, the estimated empirical distribution from the last step is taken as the Polya tree base distribution. Three steps are involved in the implementation of the `RPTMM`: the `updating` step, the `augmentation` step, and the `construction` step.

3.2 Constructing the recursive Polya tree mixture model

3.2.1 Updating step

The `RPTMM` starts with initial distribution $F^{(0)}$ with a broad support $[S_1, S_2]$ (in most cases, a uniform distribution on $[S_1, S_2]$ is taken as $F^{(0)}$). A Polya tree is built to be centered on $F^{(0)}$, and truncated at level M. At level M, there are 2^M intervals, and I label them as $I_{M1}, \ldots, I_{M2^M}$. Denote $F(I_{Mk})$ by $P_{Mk}, k = 1, \ldots, 2^M$. The conditional probability density function of Y can be written as

$$g(Y|\theta, P_{M1}, \ldots, P_{M2^M}) = \sum_{j=1}^{2^M} P_{Mj} h(Y|\theta),$$

where $\theta \in I_{Mj}$ with probability P_{Mj} under F. Using the "canonical partition" of Lavine (1992 [54]), $P_{Mk} = 1/2^M, k = 1, \ldots, 2^M$. Letting L_k, R_k be partition points for interval I_{Mk}, the probability density function of θ, $f^{(0)}$ can be written as a mixture

of uniform distributions:

$$f^{(0)}(\theta) = \sum_{k=1}^{2^M} P_{Mk} U_\theta(L_k, R_k),$$

where $U_\theta(a, b)$ represents uniform distribution on (a, b). After a data point $Y_i = y_i$ is observed, the posterior probability associated with interval I_{Mk} is

$$\begin{aligned} P(\theta_i \in I_{Mk}|y_i) = \omega_{ik} &\propto P(\zeta_{Mk}|y_i)(R_{Mk} - L_{Mk}) \\ &= \frac{P_{Mk} h(y_i|\zeta_{Mk})}{R_{Mk} - L_{Mk}}(R_{Mk} - L_{Mk}) = P_{Mk} h(y_i|\zeta_{Mk}), \quad (3.1) \end{aligned}$$

where ζ_{Mk} is the midpoint of interval I_{Mk}. The first line of (3.1) is from the trapezoidal rule. Thus, through normalizing (3.1),

$$P(\theta_i \in I_{Mk}|y_i) = \omega_{ik} = \frac{P_{Mk} h(y_i|\zeta_{Mk})}{\sum_{l=1}^{2^M} P_{Ml} h(y_i|\zeta_{Ml})}.$$

The quantity ω_{ij} represents the posterior probability that $\theta_i \in I_{Mj}$ given data Y_i.

3.2.2 Augmentation step

For θ_i, I generate an interval I_{Mj} from the discrete distribution over intervals according to the probabilities $\omega_{i1}, \ldots, \omega_{i2^M}$, and define an indicator variable z_{ij},

$$z_{ij} = \begin{cases} 1 \texttt{ if } \theta_i \in I_{Mj}, \\ 0 \texttt{ otherwise}. \end{cases}$$

Then probabilities $P_{Mj}^{(0)} = 1/2^M, j = 1, \ldots, 2^M$ can be updated to

$$P_{Mj}^{(1)} = \frac{\sum_{i=1}^{N} z_{ij}}{N} \quad \text{or} \quad P_{Mj}^{(1)} = \frac{\sum_{i=1}^{N} \omega_{ij}}{N} \qquad (3.2)$$

Note that equation (3.2) allows some prior information to be added. For example, in the density estimation problem, I could consider prior information as

$$\text{Prior}_{Mj} = \frac{\sum_{i=1}^{N} I(y_i \in I_{Mj})}{N},$$

where $I(.)$ is indicator function. Then (3.2) can be replaced as

$$P_{Mj}^{(1)} = \frac{\sum_{i=1}^{N}(z_{ij} + I(y_i \in I_{Mj}))}{N+N} \quad \text{or} \quad P_{Mj}^{(1)} = \frac{\sum_{i=1}^{N}(\omega_{ij} + I(y_i \in I_{Mj}))}{N+N} \tag{3.3}$$

After $P_{Mj}^{(0)}$ has been updated to $P_{Mj}^{(1)}$, $f^{(0)}(\theta)$ has been updated to

$$f^{(1)}(\theta) = \sum_{k=1}^{2^M} P_{Mk}^{(1)} U_\theta(L_k, R_k).$$

In the Augmentation step, I sample pseudo parameters from the mixture of uniform distributions $f^{(1)}$; the sampling process is very straightforward:

- draw an interval I_{Mi} from $I_{M1}, \ldots I_{M2^M}$ based on the probabilities $P_{M1}^{(1)}, \ldots, P_{M2^M}^{(1)}$,
- sample a pseudo data uniformly from I_{Mi}.

The process is repeated d times to produce pseudo data $\theta_1^*, \ldots, \theta_d^*$. The empirical distribution function of these pseudo data $\theta_1^*, \ldots, \theta_d^*$ can be obtained.

In next step, the base distribution is estimated by the empirical distribution of these pseudo data. Note that the support of sampled pseudo θ at iteration υ will always be included in the support at iteration $\upsilon - 1$, which means

$$[\min(\theta_1^{*(\upsilon)}, \ldots, \theta_d^{*(\upsilon)}), \max(\theta_1^{*(\upsilon)}, \ldots, \theta_d^{*(\upsilon)}))] \in [\min(\theta_1^{*(\upsilon-1)}, \ldots, \theta_d^{*(\upsilon-1)}), \max(\theta_1^{*(\upsilon-1)}, \ldots, \theta_d^{*(\upsilon-1)}))]$$

As a result, the support of pseudo data $(\theta_1^*, \ldots, \theta_d^*)$ will continue to shrink. To overcome this problem, I add a buffer to the support of these pseudo data. So the support will be

$$[\min(\theta_1^*, \ldots, \theta_d^*) - \text{buffer}, \max(\theta_1^*, \ldots, \theta_d^*) + \text{buffer}]$$

The rule of thumb to choose the buffer is

$$\text{buffer} = \frac{\max(Y_1, \ldots, Y_N) - \min(Y_1, \ldots, Y_N)}{2^M}.$$

In practice, the `RPTMM` starts with a very broad support $[S_1^{(0)}, S_2^{(0)}]$, which ultimately will shrink to the area $[S_1, S_2]$ such that

$$F((-\infty, S_1] \cup [S_2, \infty)) \approx 0.$$

The support will become stable when the algorithm converges.

3.2.3 Construction step

In the construction step, a new Polya tree is constructed, and its base distribution is the empirical distribution of pseudo data from last iteration step. To construct the new Polya tree, I have to decide the partition points at each level, and from the "canonical partition" of Lavine (1992 [54]), they should be the quantiles of the base distribution.

3.2.4 Algorithm summary

The three steps `updating`, `augmentation`, and `construction` are repeated many times until convergence. I summarize the recursive Polya tree mixture model

below. Firstly under the Bayesian hierarchical model

$$Y_i|\theta_i \sim h(Y_i|\theta_i) \text{ independent,}$$

$$\theta_i \sim F \text{ i.i.d.}$$

1. Start with a guess $F^{(0)}$ on a broad support (S_1, S_2). A Polya tree truncated at level M is built and centered on $F^{(0)}$.
2. For $l = 1, \ldots, \Upsilon$,
   - for each data point Y_i, calculate $P(\theta_i \in I_{Mk}|Y_i) = \omega_{ik}$ $(k = 1, \ldots, 2^M)$, and draw a interval for θ_i from I_{M1}, ..., I_{M2^M} with probability ω_{i1} , ..., ω_{i2^M};
   - compute
   $$P_{Mj} = \frac{\sum_{i=1}^{N} z_{ij}}{N} \quad \text{or} \quad P_{Mj} = \frac{\sum_{i=1}^{N} \omega_{ij}}{N};$$
   - draw a large number of pseudo data $\theta_1^*, \ldots, \theta_d^*$;
   - build a new Polya tree using the empirical distribution of $\theta_1^*, \ldots, \theta_d^*$ as the base distribution.

3.2.5 Parameter estimation

The algorithm described above is a data augmentation method. Suppose there are in total Υ steps, and produce Υ estimates of generic parameters ψ: $\psi^{(1)}, \ldots, \psi^{(\Upsilon)}$. These $\psi^{(\upsilon)}$ values can be easily sampled from pseudo data at each iteration. Then I estimate each ψ_i as

$$\widehat{\psi}_i = \frac{1}{\Upsilon} \sum_{\upsilon=1}^{\Upsilon} \psi_i^{(\upsilon)}.$$

The standard deviation of $\widehat{\psi}_i$ is

$$\mathrm{sd}(\widehat{\psi}_i) = \sqrt{\frac{1}{\Upsilon - 1}\sum_{\upsilon=1}^{\Upsilon}(\psi_i^{(\upsilon)} - \widehat{\psi}_i)^2}.$$

In the next section, I discuss the theoretical justification of this approach.

3.3 Theoretical properties

3.3.1 Convergence

To study the convergence of the iterative algorithm, I apply the duality principle introduced in Diebolt and Robert (1994 [21]). The recursive Polya tree mixture model can be written in a typical form of data augmentation algorithm,

$$\begin{aligned} &\text{(a)} \quad \text{generate } z^{(\nu)} \sim q(z|y, \lambda^{(\nu)}), \\ &\text{(b)} \quad \text{generate } \lambda^{(\nu+1)} \sim \pi(\lambda|y, z^{(\nu)}), \end{aligned}$$

where the step (a) corresponds to the generation of the indicator variables z_i according to a multinomial distribution with probabilities ω_{ij}, where

$$\omega_{ik} = \frac{P_{Mk}h(Y_i|\zeta_{Mk})}{\sum_{l=1}^{2^M} P_{Ml}h(Y_i|\zeta_{Ml})}.$$

In step (b), the $\lambda^{(\nu)}$ represent the simulated pseudo data generated at iteration ν. Diebolt and Robert (1994 [21]) describe this iterative algorithm using two dual Markov kernels

$$\begin{aligned} H(z'|z) &= \int f(z'|y, \lambda)\pi(\lambda|y, z)d\lambda, \\ K(\lambda'|\lambda) &= \int \pi(\lambda'|y, z)f(z|y, \lambda)dz. \end{aligned}$$

At iteration ν, the posterior distributions are derived from these kernels as

$$\pi^{(\nu)}(\lambda'|y) = \int K(\lambda'|\lambda)\pi^{(\nu-1)}(\lambda|y)d\lambda,$$

$$f^{(\nu)}(z'|y) = \int H(z'|z)f^{(\nu-1)}(z|y)dz.$$

Theorem 3.1 (Diebolt and Robert 1994 [21])**.** If the chain for z is geometrically ergodic, with distribution $f(z|y)$, the chain for λ derived by $\lambda \sim \pi(\lambda|z)$ is also geometrically ergodic, with invariant distribution $\pi(\lambda|y)$. Moreover, there exists $\rho \in (0,1)$ and a constant $C > 0$ such that

$$\int_\Lambda |\pi^{(\nu)}(\lambda|y) - \pi(\lambda|y)|d\lambda \leq C\rho^\nu.$$

The importance of this theorem is that if the chain for λ are derived from a second chain for z by simulation from $\pi(\lambda|z)$, the properties of the chain for λ can be gathered from those of the chain for z.

In the recursive Polya tree mixture model, since the indicator variables z have a finite support and the Markov chain is aperiodic and irreducible, the chain is geometrically ergodic. Note that, at iteration ν, there are d pseudo data $\theta_1^*, \ldots, \theta_d^* \sim F^{(\nu)}$ sampled according to

$$\begin{aligned}
\theta_1^* &\sim \pi^{(\nu)}(\theta^*|z) = \sum_{j=1}^{2^M} \left\{ \frac{1}{N} \sum_{i=1}^{N} z_{ij} U(L_j, R_j) \right\}, \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
\theta_d^* &\sim \pi^{(\nu)}(\theta^*|z) = \sum_{j=1}^{2^M} \left\{ \frac{1}{N} \sum_{i=1}^{N} z_{ij} U(L_j, R_j) \right\}.
\end{aligned}$$

Since Theorem 3.1 holds for any conditional distribution $\pi^{(\nu)}(\theta^*|z)$, the chain $\theta^{*(\nu)}$ derived by $\theta^{*(\nu)} \sim \pi(\theta^*|z)$ is geometrically ergodic. In step ν, I also know that these pseudo data $\theta^{*(\nu)}$ have empirical cumulative distribution function $F^{(\nu)}(\theta|y)$. From Theorem 3.1, for each specific θ, $F^{(\nu)}(\theta|y)$ converge in distribution to the stationary posterior distribution $F(\theta|y)$.

3.3.2 Large-sample analysis

To study the large-sample behavior, I follow Shen and Louis (1999 [76]). The updating process of the recursive Polya tree mixture model is related to the *smoothing by roughening* (SBR) of Shen and Louis (1999 [76]). There are two differences. Firstly, the finite support in SBR is fixed, but is changing in `RPTMM`. Secondly, the mass points in SBR are fixed, and the "weights" corresponding to each mass point are updated, but the `RPTMM` treats the mass points as the parameters to be estimated.

I begin with the first iteration, $g^{(0)}(y) = \int_{-\infty}^{\infty} h(y|\theta)dF^{(0)}(\theta)$. In the first iteration, for any value $t \in I_{MJ}(J = 1, \ldots, 2^M)$

$$F^{(1)}(t|Y) = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{j=1}^{J-1} P_{Mj}^{(0)} h(Y_i|\zeta_{Mj}) + \left(\frac{t - L_J}{R_J - L_J}\right) P_{MJ}^{(0)} h(Y_i|\zeta_{MJ})}{\sum_{l=1}^{2^M} P_{Ml}^{(0)} h(Y_i|\zeta_{Ml})}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{\int_{-\infty}^{t} h(Y_i|\theta)dF^{0}(\theta)}{g^{(0)}(Y_i)}.$$

When the sample size $N \to \infty$,

$$\begin{aligned} F^{(1)}(t|Y) &\longrightarrow \int_{-\infty}^{\infty} \frac{\int_{-\infty}^{t} h(Y|\theta) dF^0(\theta)}{g^{(0)}(Y)} dG(Y) \\ &= \int_{-\infty}^{t} \int_{-\infty}^{\infty} \frac{h(Y|\theta)}{g^{(0)}(Y)} dG(Y) dF^{(0)}(\theta) \\ &= F^{(1)}(t), \end{aligned}$$

where $G(Y)$ is the marginal cumulative distribution function of Y. Shen and Louis (1999 [76]) take the derivative of $F^{(1)}(t)$ with respect to t, and obtain

$$f^{(1)}(\theta) = f^{(0)}(\theta) \int_{-\infty}^{\infty} \frac{h(Y|\theta) g(Y)}{g^{(0)}(Y)} dY.$$

Shen and Louis (1999 [76]) show that, in general, at iteration ν,

$$\begin{aligned} f^{(\nu+1)}(\theta) &= f^{(\nu)}(\theta) \int_{-\infty}^{\infty} \frac{h(Y|\theta) g(Y)}{g^{(\nu)}(Y)} dY \\ &= \int \frac{h(Y|\theta) f^{(\nu)}(\theta)}{g^{(\nu)}(Y)} g(Y) dY \\ &= \int \frac{h(Y|\theta) f^{(\nu)}(\theta)}{\int h(Y|\theta) f^{(\nu)}(\theta)} g(Y) dY \\ &= \int f^{(\nu)}(\theta|Y) g(Y) dY \end{aligned} \tag{3.4}$$

So f is a fixed point of Equation (3.4) (Shen and Louis 1999 [76]).

3.4 Discussion

3.4.1 Comparison with the classical Polya tree model

In the classical Polya tree model, the partition points are fixed, and the "weights" corresponding to each interval are updated in each step. In the recursive Polya tree mixture model, I do the converse, and update the partition points.

I describe our model in the framework of a Polya tree, since the class of Polya trees is overwhelmingly large. In practice, one could even ignore the "hierarchical" partition structure, and directly focus on the $K = 2^M$ partitions in the support, and treat it as a mixture of uniform distribution. So the parameters to be estimated are the partition points.

3.4.2 Convergence

The `RPTMM` relies on the data augmentation algorithm, so it is necessary to check whether the Markov chain has become stable. All the existing methods to diagnose convergence in the literature can be used. I introduce a simple way to check the convergence, based on a criterion mentioned in Laud and Ibrahim (1995 [53]) and Krnjajic et al. (2008 [51]): the *full-sample log score* ($LSFS$)

$$LSFS = \frac{1}{N}\sum_{i=1}^{N} \log p(Y_i|\theta) \tag{3.5}$$

The p.d.f value of $p(Y_i|\theta)$ can be easily computed through the Monte Carlo method. If the Markov chain has converged to a stationary distribution, I should observe stable $LSFS$ values, and the $LSFS$ values should not oscillate too much. I plot the $LSFS$ values in the following example. The left panel in Figure 3–1 shows an example of good convergence.

3.4.3 The choice of the buffer and the support

In the **Augmentation** step, I introduce the use of a buffer. If the buffer is too small, it can not prevent the continuous shrinkage of the support; if the buffer is too large, the support will oscillate too much, and has poor mixing property. One may use trial-and-error method to choose the value of the buffer. Start with a

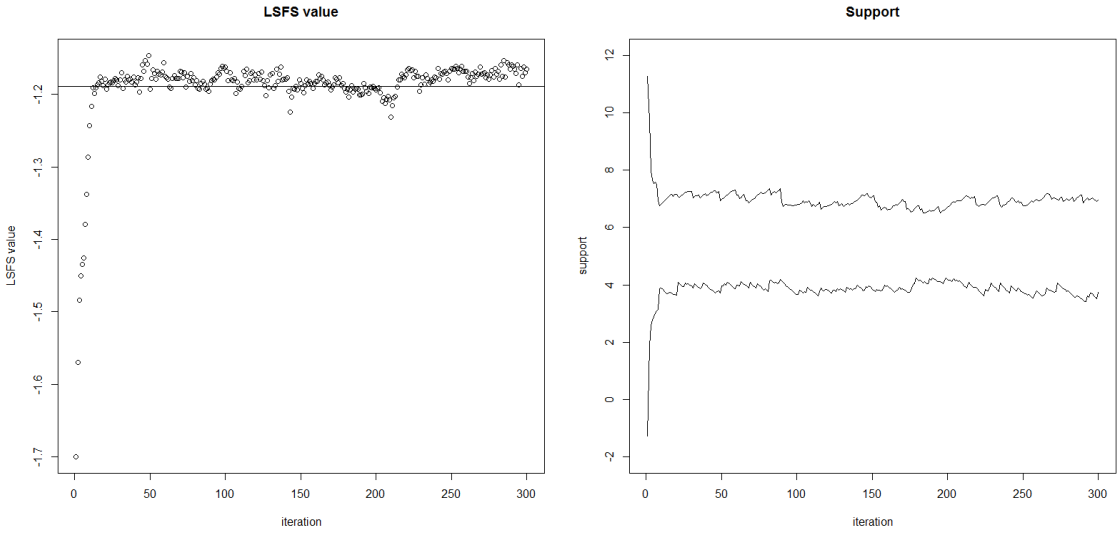

Figure 3–1: Convergence: The left panel shows the $LSFS$ values, and the right panel shows the support of pseudo data at each iteration. After 10-15 iterations, the $LSFS$ values become stable.

guess buffer value (say $0.1 \times (\max(Y) - \min(Y))$), run the `RPTMM` algorithm and plot $\widehat{S}_1 = \min(\theta_1^*, \ldots, \theta_d^*)$ and $\widehat{S}_2 = \max(\theta_1^*, \ldots, \theta_d^*)$ at each step to judge the choice of buffer. If the difference of $\widehat{S}_2 - \widehat{S}_1$ continues to become smaller, it means that the buffer is too small, and a larger one should be used. If the difference oscillates too much, which means the buffer is too large, and a smaller one should be used.

In this thesis, I consider a rule of thumb buffer $= \text{range}(Y_i, i = 1, \ldots, N)/2^M$, and find that this rule of thumb works very well in real applications. The right panel in Figure 3–1 shows a good example of such a plot. I find that the curve becomes stable quickly (at iteration between 10 and 15), which indicates that the choice of buffer is a good one.

One may also run the algorithm many times, estimate the support $[S_1, S_2]$ and fix it, and regard the process as a burn-in process. Then re-run the algorithm with the fixed estimated support. I find the two algorithms perform similarly in practice, and in this thesis I focus on the first one (not fix the support).

Although the theoretical support for θ is $(-\infty, \infty)$, using finite support $[S_1, S_2]$ to approximate $F([S_1, S_2]) \approx 1$ is also considered. For example, the finite mixture model (chapter 9 of Bishop 2006 [6]) assumes the support for θ to be the finite number of points. Shrinkage methods, like the James-Stein estimator (Efron and Morris, 1975 [25]) shrink the extreme values towards the mean. As a result, the hidden assumption of the James-Stein estimator is that $[S_1, S_2] \in (\min(Y_1, \ldots, Y_n), \max(Y_1, \ldots, Y_n))$.

Shen and Louis (1999 [76]) proposed a "smoothing by roughening (SBR)" algorithm, and Newton and Zhang (1999 [70]) and Newton (2002 [69]) developed a predictive recursion algorithm. Both of them fix a finite support and many mass points, and then update the p.d.f values corresponding to each mass point. The SBR relies on an iterative algorithm, and the predictive recursion is a sequential updating method. Although their approaches are fast and powerful, it is difficult for the SBR algorithm and predictive recursion to choose the finite support.

3.4.4 The choice of the number of levels

As mentioned in Chapter 2, Hanson (2006 [36]) found that adding levels to a Polya tree did not guarantee the improvement of model fit or predictive utility. Hanson (2006 [36]) suggested a rule of thumb

$$M = -\log_2(5/N) \text{ or } \log_2(N),$$

and discussed its rationale. Note that Hanson (2006 [36]; 2008 [9])'s rule is consistent with the theoretical result of Barron et al. (1999 [3]) who studied the consistency of posterior distributions in nonparametric problems. For the Polya tree, to satisfy the conditions of their Corollary 2, they required that $M \leq \log_2(\varepsilon^2 N/10)$ for some $\varepsilon > 0$. Hanson (2006 [36]; 2008 [9])'s rule corresponds to $\varepsilon^2 = 2$ or 10. In this thesis, I follow their rule of thumb.

3.5 Nuisance parameters

In a typical Normal setting, $Y_i|\theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$ independently, and $\theta_1, \ldots, \theta_N \sim F$, i.i.d. If σ_i is unknown to us, and heterogeneity is believed to exist, each σ_i should be estimated also. I first consider the classical finite mixture of normal distributions with K components,

$$g(y) = \sum_{j=1}^{K} \pi_j \mathrm{N}(\mu_j, \sigma_j^2).$$

If μ_j is known, I can estimate each individual σ_j as (see Chapter 9 of Bishop 2006 [6])

$$\widehat{\sigma}_j^2 = \frac{1}{\gamma_j} \sum_{i=1}^{N} \omega_{ij} (y_i - \mu_j)^2,$$

where

$$\gamma_j = \sum_{i=1}^{N} \omega_{ij} = \sum_{i=1}^{N} \frac{\pi_j \mathrm{N}(y_i|\mu_j, \sigma_j)}{\sum_{k=1}^{K} \pi_k \mathrm{N}(y_i|\mu_k, \sigma_k)},$$

Bishop (2006 [6]) denoted γ_j as the "effective degree of freedom".

In the `RPTMM`, the probability density function of Y can also be written as a finite mixture of normal distributions with 2^M components,

$$g(Y|\theta, P_{M1}, \ldots, P_{M2^M}) = \sum_{j=1}^{2^M} P_{Mj} h(Y|\theta),$$

$$\theta \in I_{Mj} \text{ with probability } P_{Mj}.$$

The difference is that in the RPTMM the support is 2^M intervals; for a finite mixture of normal distributions the support is K points. So I could also estimate each individual σ_{Mj} corresponding to each component $j(j = 1, \ldots, 2^M)$ as

$$\widehat{\sigma}^2_{Mj} = \frac{1}{\gamma_j} \sum_{i=1}^{N} \omega_{ij}(y_i - \zeta_j)^2 \qquad j = 1, \ldots, 2^M$$

where ζ_j is the midpoint of interval I_{Mj} and

$$\gamma_j = \sum_{i=1}^{N} \omega_{ij}.$$

The σ_i corresponding to each data value Y_i is estimated as

$$\widehat{\sigma}^2_i = \sum_{j=1}^{2^M} \omega_{ij} \widehat{\sigma}^2_{Mj},$$

where $\omega_{ij} = P(\theta_i \in I_{Mj} | y_i)$.

3.6 Summary

In this chapter, I introduce the detail of the proposed recursive Polya tree mixture model (RPTMM). Instead of fixing partition points and updating "weights" corresponding to each interval, the RPTMM update the partition points. Sampling pseudo data from F is very straightforward. As a result, the empirical distribution of these pseudo data is available to us, and is treated as the base distribution of Polya tree.

I discuss the related problems, such as convergence diagnostics, and the choice of the buffer and number of levels. The estimation of nuisance parameters in the normal case is provided. I also discuss some theoretical aspects related to the proposed

`RPTMM`. In the next chapter, I discuss some examples related to Bayesian nonparametric hierarchical models.

CHAPTER 4
Bayesian Nonparametric Hierarchical Models: Examples

4.1 Example 1: Density estimation

I consider the simulated data from Kottas and Gelfand (2001 [50]), which consist of 250 observations generated from a mixture of normal distributions

$$y = (y_1, \ldots, y_{250}) \sim 0.435\mathrm{N}(-4, 1) + 0.43\mathrm{N}(0, 1.5^2) + 0.135\mathrm{N}(5, 2^2).$$

Hanson and Johnson (2002 [37]) examined several Bayesian nonparametric models to estimate the density of the 250 simulated observations. These models include the simple Polya tree, a mixture of Polya tree, and Dirichlet process mixture model. Hanson and Johnson (2002 [37]) considered two criteria to compare the performance of each model: integrated squared error (`ISE`) and weighted integrated squared error (`WISE`):

$$\mathrm{ISE}(\widehat{f}) = \int (\widehat{f}(y) - f(y))^2 dy \quad \texttt{and} \quad \texttt{WISE}(\widehat{f}) = \int (\widehat{f}(y) - f(y))^2 f(y) dy,$$

where $\widehat{f}(y)$ is the estimated p.d.f value and $f(y)$ is the true p.d.f value. I use a Monte Carlo method to calculate the two measurements. For `ISE`, I can rewrite the formula as

$$\mathrm{ISE}(y) = \int \frac{(\widehat{f}(y) - f(y))^2}{f(y)} f(y) dy.$$

Since $y_1, \ldots, y_{250} \sim f(y)$, I can approximate `ISE` as

$$\text{ISE}(y) = \frac{1}{250} \sum_{i=1}^{250} \frac{(\widehat{f}(y_i) - f(y_i))^2}{f(y_i)}.$$

`WISE` is straightforward to evaluate using Monte Carlo:

$$\text{WISE}(y) = \frac{1}{250} \sum_{i=1}^{250} (\widehat{f}(y_i) - f(y_i))^2.$$

I assume each observation arises from a normal distribution $Y_i|\theta_i \sim \texttt{Normal}(\theta_i, \sigma_i^2)$, and $\theta_1, \ldots, \theta_N \sim F$. I simulate 200 data sets and calculate the `ISE` and `WISE` values for each data set.

The average values of `ISE` and `WISE` from `RPTMM` are listed in the first line of Table 4–1. Hanson and Johnson (2002 [37]) implemented three Bayesian nonparametric models (simple Polya tree, mixture of Polya tree, and Dirichlet process mixture models) and a parametric model. All the `ISE` and `WISE` values are listed from Line 2 to Line 13 in Table 4–1. The S_0 in Table 4–1 are the fixed standard deviations for base distributions.

Our `RPTMM` is shown to perform better than simple Polya tree and mixture of Polya trees. Note also that, among all the 13 models in Table 4–1, `RPTMM` ranks 3rd (slightly worse than the Dirichlet process mixture model with normal base density). In Figure 4–1, I draw the density estimation curve from `RPTMM` and the true density for one of the simulated data sets. The `RPTMM`'s density curve is shown to be close to the true one. Density curves from classical kernel density estimation and Dirichlet process mixture model also perform very well (close to the true curve). The density

Models	WISE	ISE
RPTMM	**0.000564**	**0.003130**
MPT		
$\alpha = 1$, normal base	0.000730	0.004230
$\alpha = 1$, logistic base	0.000740	0.004090
Simple PT		
$S_0 = 6.81$, $\alpha = 0.1$	0.001530	0.007940
$S_0 = 22.7$, $\alpha = 0.1$	0.004300	0.019340
DPM		
split densities, less informative	0.002770	0.014990
split densities, more informative	0.002500	0.013480
uniform densities, less informative	0.002890	0.014370
uniform densities, more informative	0.000900	0.004060
Gaussian, $S_0 = 0.375$	0.001880	0.010790
Gaussian, $S_0 = 0.75$	0.000310	0.002400
Gaussian, $S_0 = 1$	0.000140	0.001330
Parametric normal model	0.005420	0.030860

Table 4–1: ISE and WISE values for several models, from Hanson & Johnson (2002). RPTMM represents recursive Polya tree mixture model; MPT represents the mixture of Polya tree; Simple PT represents simple Polya tree. DPM represents the Dirichlet process mixture model. The S_0 are the fixed standard deviations for the base distributions.

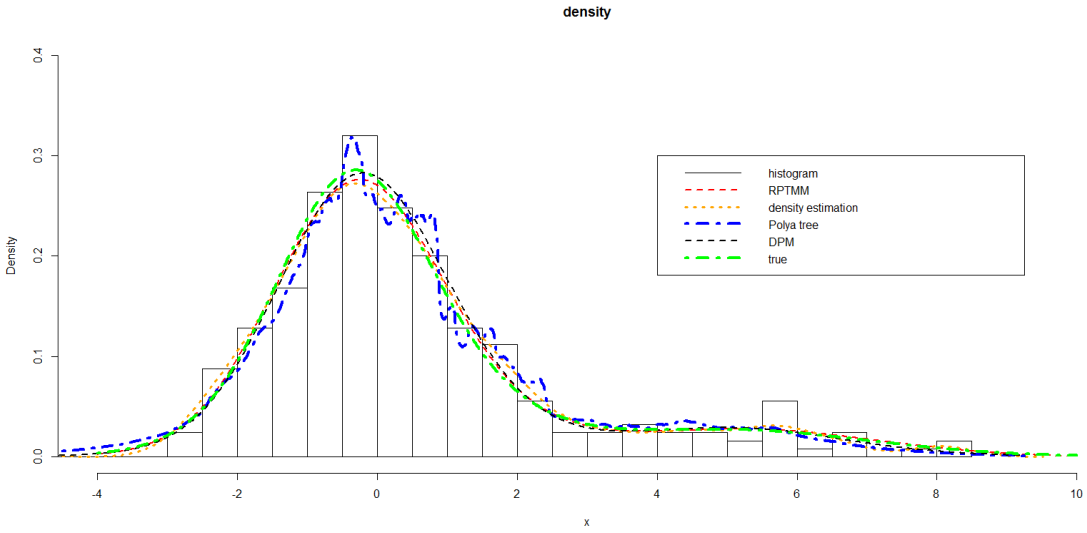

Figure 4–1: The density estimates from `RPTMM`, kernel density estimation, Polya tree, Dirichlet process mixture model and the true density.

curve from the mixture of Polya tree model (with precision value 1) seems to be close to histogram.

4.2 Density estimation for real data sets

In this section, I consider the density estimation problem for two real data sets. They were analyzed in Richardson and Green (1997 [35]) using a complicated reversible jump MCMC.

Example 8. Galaxy data The first data set is the galaxy data, and it consists of the velocities of 82 distant galaxies. Escobar & West (1995 [27]) used a Dirichlet process mixture model and Richardson & Green (1997 [35]) considered an infinite mixture of normals using reversible jump MCMC.

The density estimates using classical kernel density estimation, Dirichlet process mixture model (DPM), mixture of Polya tree (PT), and `RPTMM` are shown in Figure 4–2.

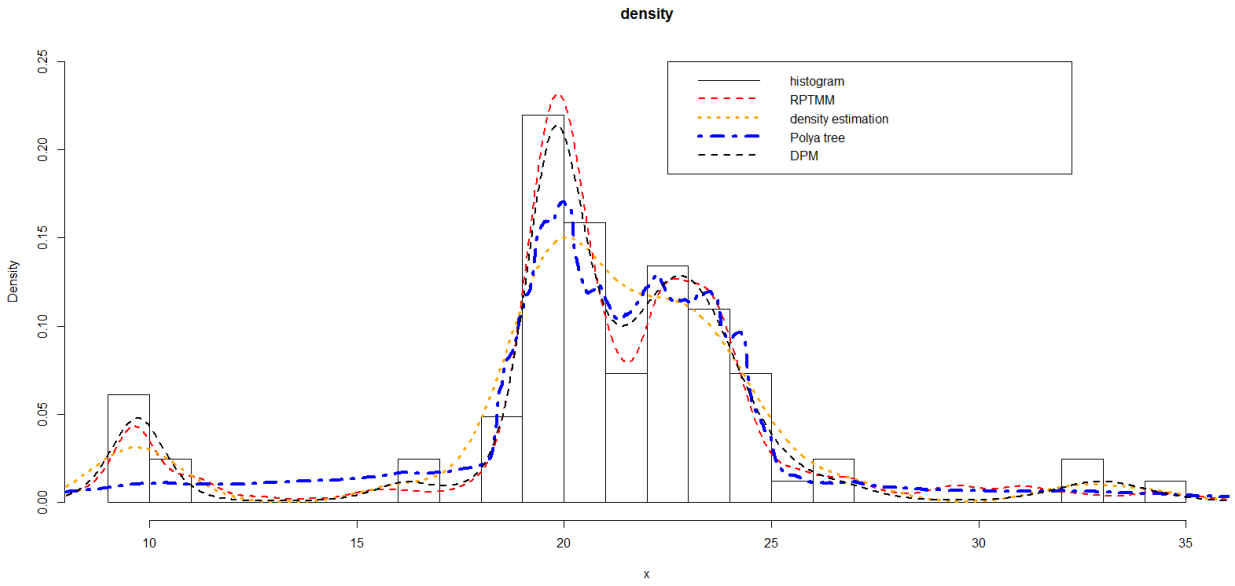

Figure 4–2: Galaxy data: Density estimates from `RPTMM`, Polya tree, Dirichlet process mixture model, and classical density estimate

The `RPTMM` curve is close to the DPM curve. It is interesting to note that simple Polya tree (PT) fails to capture the cluster on the leftmost end of the range, when the observations are around 10. The kernel density also did well, but fails to clearly show the cluster effects.

Example 9. Acidity data

The second data set consists of 155 acidity index measurements for lakes in north-central Wisconsin. All the methods indicate that there are two clusters.

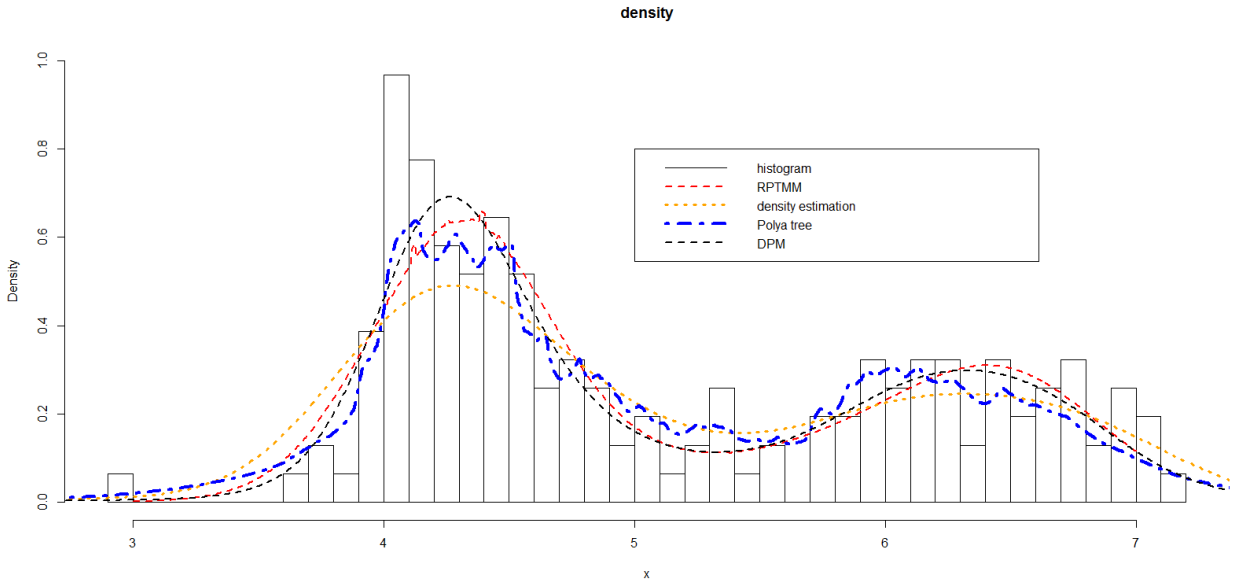

Figure 4–3: Acidity data: Density estimates from `RPTMM`, Polya tree, Dirichlet process mixture model, and classical density estimation

I discuss how well `RPTMM` performs on density estimation, not intending to provide new insights, but to provide a numeric comparison with other Bayesian nonparametric models. In the simulated data sets, the `RPTMM` performed better than most of the other Bayesian nonparametric models. In the two real data sets, `RPTMM` also did very well. In real applications, kernel density estimation or histograms are the easiest approach to density estimation. With the help of the bootstrap (Efron 1979 [24]), the uncertainty of density estimation is also very easy to obtain. However, in some cases (e.g., the galaxy data), Bayesian nonparametric models can help to identify some additional information (such as clustering of elements).

4.3 Bivariate density estimation

It is easy to generalize the recursive Polya tree mixture model to solve the bivariate density estimation problem.

Example 10. Ozone data

Suppose we observe a bivariate random sample $(X_1, Y_1), \ldots, (X_N, Y_N)$, and assume

$$(X_i, Y_i)|\vec{\mu}_i, \Sigma_i \sim \mathrm{N}_2(\vec{\mu}_i, \Sigma_i), \text{ independently}$$

where $\mathtt{N}_2$ denotes a bivariate normal distribution, $\vec{\mu}_i$ is a 2×1 mean vector and Σ_i is a 2×2 covariance matrix. I further assume that $\vec{\mu}_i \sim F$ i.i.d. Instead of 2^M "intervals", I consider 2^{2M} "bins", labeled as $B_1, \ldots, B_{2^{2M}}$. The probability $P(\vec{\mu}_i \in B_j) = \omega_{ij}$ $(j = 1, \ldots, 2^{2M})$ is

$$\omega_{ij} = \frac{P_{Mj}\phi_2((X_i, Y_i)|\zeta_j, \Sigma_i)}{\sum_{k=1}^{2^{2M}} P_{Mk}\phi_2((X_i, Y_i)|\zeta_k, \Sigma_i)},$$

where ϕ_2 denotes a bivariate normal density function and $\zeta_j (j = 1, \ldots, 2^{2M})$ is the midpoint of the jth bin. The algorithm in Section 3.5 can be used to estimate Σ_i, and I only need to change the normal density function to the bivariate normal density function. Thus, I estimate Σ_i as

$$\hat{\Sigma}_i = \sum_{j=1}^{2^{2M}} \omega_{ij}\hat{\Sigma}_{Mj},$$

where

$$\hat{\Sigma}_{Mj} = \frac{1}{\gamma_j}\sum_{i=1}^{N} \omega_{ij}([X_1, Y_1]^T - \vec{\mu})([X_1, Y_1]^T - \vec{\mu})^T,$$

and

$$\gamma_j = \sum_{i=1}^{N} \omega_{ij}.$$

Hanson (2006 [36]) discuss an ozone data set, where $N = 111$ bivariate observations on the cube root of ozone concentration (x_i) and radiation level (y_i) are modeled. Hanson (2006 [36]) considered two Bayesian nonparametric models: a mixture of Polya tree model (MPT) and a Dirichlet multinomial allocation (DMA) model. Hanson (2006 [36]) assessed these model using the log pseudo marginal likelihood (LPML), with larger value indicating better performance. The parametric model gives $LPML \approx -796$, while the mixture of Polya tree (MPT) improves the $LPML \approx -782$. Dirichlet multinomial allocation (DMA) has $LPML \approx -774$, and gives a significant improvement over the MPT.

From the recursive Polya tree mixture model (`RPTMM`), the $LMPL \approx -772$. The performance of `RPTMM` is similar to the Dirichlet multinomial allocation (DMA), and significantly better than MPT. The contour plot and perspective plot are shown in Figure 4–4

Example 11. Kidney data

In this section, I discuss a data set from McGilchrist and Aisbett (1991 [64]) that consists of bivariate measures $(T_{i1}, T_{i2})(i = 1, \ldots, N)$ of times from insertion of catheter to infection for $N = 38$ dialysis patients. There are 6 censored observations in the first time, and 12 censored observations in the second, and 3 censored observations on both times. If censored data occur, it is straightforward to impute values in each step based on the truncated bivariate normal distribution. The bivariate density estimation problem using this data set was also considered in Barajas and

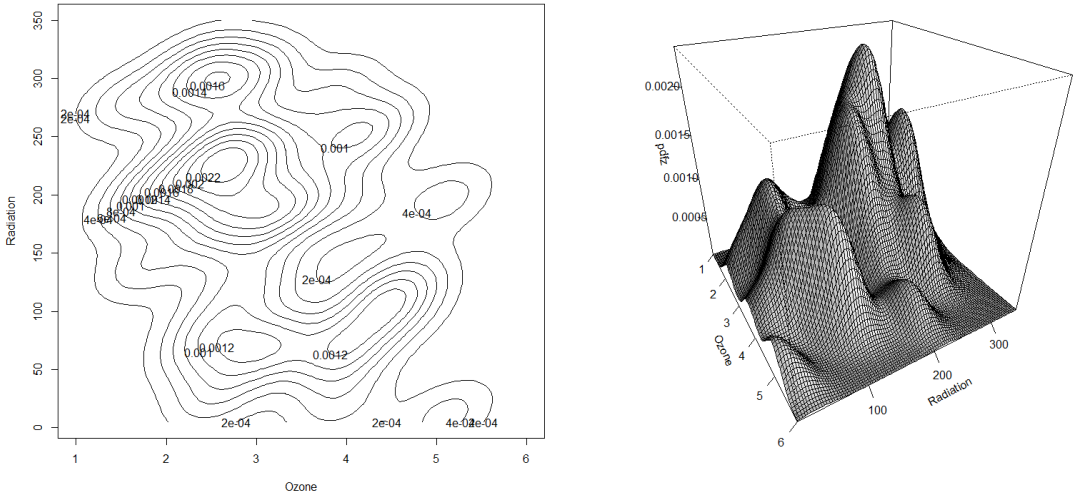

Figure 4–4: Ozone Concentration and Radiation: contour plot and perspective plot

Walker (2007 [71]) and Walker and Stephens (1999 [80]). Figure 4–5 shows the perspective plot for female kidney patients. The plots from the Dirichlet process mixture model and recursive Polya tree mixture model are very similar. When the "true" bivariate density is parametric (bivariate normal), although the `RPTMM` is based on a nonparametric assumption, it can still make valid density estimation.

Figure 4–6 shows the posterior estimates of the survival functions for females. The plots here are very similar to Figure 1 and 2 in Barajas and Walker (2007 [71]).

4.4 Thumbtacks data analysis revisited

This data set is from Beckett et al. (1994 [5]), and studied previously in Example 7. Some 320 tacks were flicked, 9 times each, and the number of times that the tacks landed point up were recorded. This process produces 320 data points which can be modeled as Binomial distributions: $X_i \sim \texttt{Binomial}(9, \theta_i), i =$

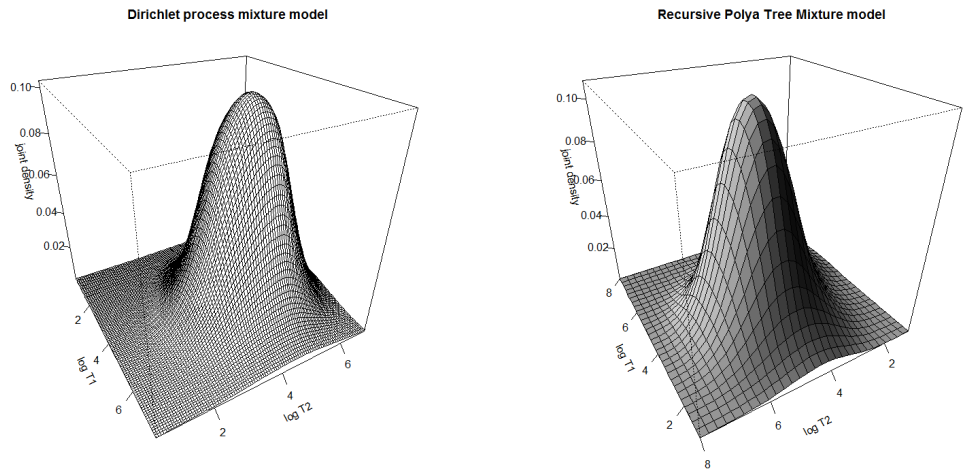

Figure 4–5: Perspective plot for data from female kidney patients. Left panel shows the plot from the Dirichlet process mixture model; right panel shows the plot from recursive Polya tree mixture model.

$1, \ldots, 320$ `independently`. The number of trials Y_i are 9 for $i = 1, \ldots, 320$, and the number of "successes" X_i were recorded and are shown in Table 4–2.

Number of "success"	0	1	2	3	4	5	6	7	8	9
Frequencies	0	3	13	18	48	47	67	54	51	19

Table 4–2: Thumbtacks data. The first row shows the number of tacks landed point up; the second row shows the corresponding number of data points. For example, there are 3 data points in which only 1 tack landed point up; and there are 19 data points in which 9 tacks landed point up.

I first consider the data $x_1/9, \ldots, x_{320}/9$, and fit the data using kernel density estimation, shown as the blue curve in Figure 4–7. It is interesting to note that there is no unusual feature that can be found directly from kernel density estimation, or

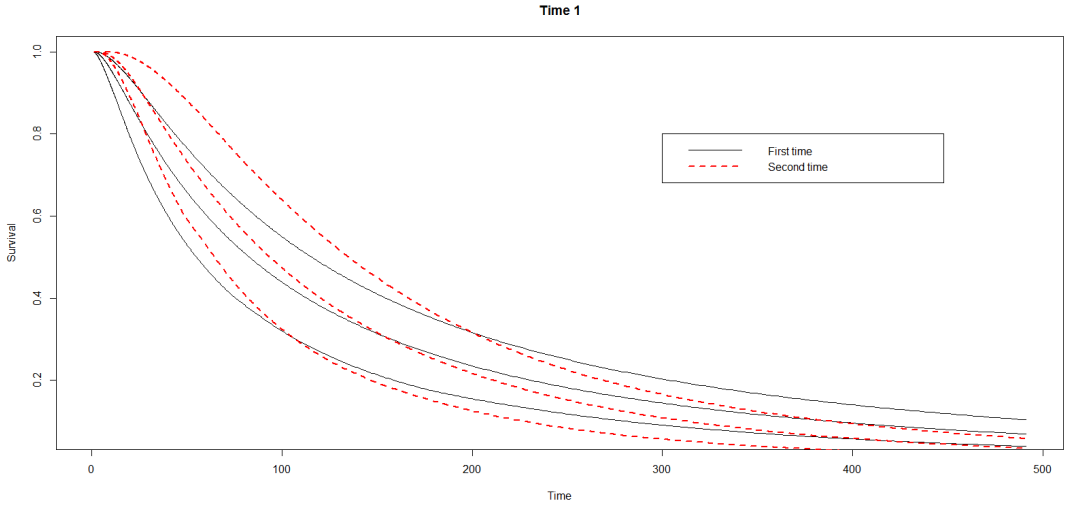

Figure 4–6: Posterior estimates of the survival functions for females. Time to first infection T_1 and time to second infection T_2. The 95% credible intervals are also shown.

the histogram. In a superficial analysis, one may think the kernel density estimation is enough to explain the data.

Liu (1996 [57]) analyzed these data by assuming θ_i arise from a nonparametric random distribution, $\theta_i \sim F$ `i.i.d`. The objective of Liu's (1996 [57]) work is to estimate the posterior distribution $F|$data; to do this Liu (1996 [57]) used sequential imputation, and found an unusual feature.

The unusual feature is that the estimate of F demonstrated bimodality. He pointed out that "this feature is unexpected and cannot be revealed by a regular parametric hierarchical analysis using the Beta-binomial priors". The finding of the unusual feature reflects an advantage of Bayesian nonparametrics. By assuming F arises from a nonparametric random distribution, we could "mine" some information that can not be found through a regular Bayesian hierarchical model. Newton (2002

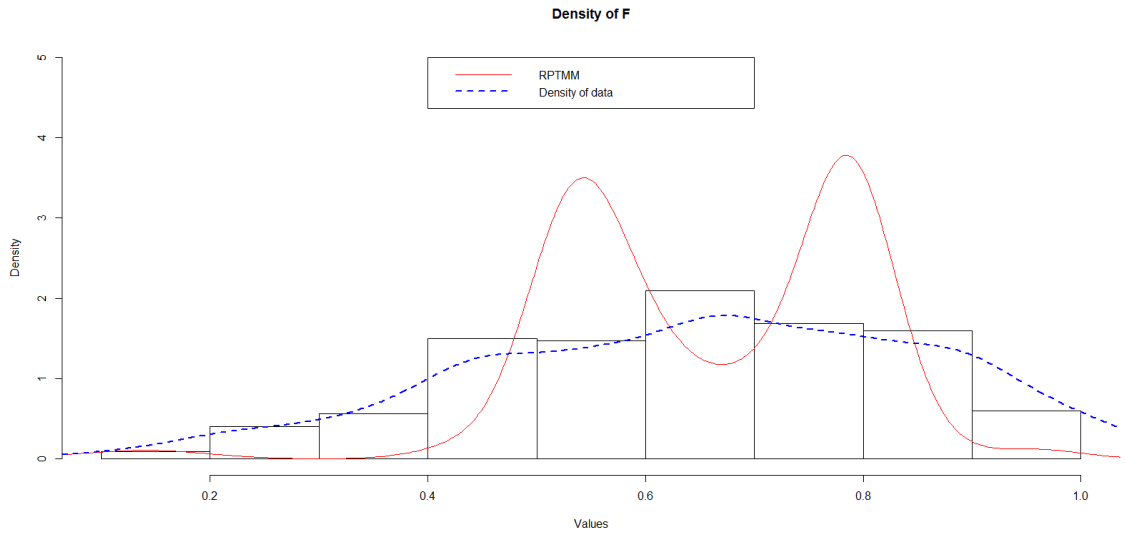

Figure 4–7: The histogram, density estimates for the thumbtacks data. The red line shows the density estimate for F from the `RPTMM`.

[69]) reanalyzed these data using predictive recursion, and found a similar conclusion to Liu (1996 [57]).

I also assume $\theta_1, \ldots, \theta_N \sim F$ `i.i.d`, and use a recursive Polya tree mixture model to estimate F. The density estimation curve of F is shown in Figure 4–7. When I compare with the plot reported in Liu (1996 [57]), I find both of them are very similar and both indicate the surprising bimodality. The recursive Polya tree mixture model is easier to implement than the sequential imputation proposed in Liu (1996 [57]) and does not require any analytical computation.

4.5 Baseball data

The Major League Baseball data set is from Brown (2008 [10]), and consists of batting records for 567 Major League Baseball players in the 2005 season. For an individual player, given a observed number of attempts (called "at-bats"), an

observed number of successes ("hits") is available. Denote the y_i as the "at-bats" for player i, and x_i as the "hits".

4.5.1 Model formulation

It is natural to model these data as a binomial distribution: $x_i \sim \texttt{Binomial}(y_i, \theta_i)$, where θ_i, unknown to us, represents the player's latent ability. Batting average is a performance measure for each player, and is typically measured as $\texttt{BA}_i = x_i/y_i$, the the maximum likelihood estimator of θ_i. Efron and Morris (1975 [25]) has shown that this naive estimator where each θ_i is estimated individually in isolation from the others can have poor performance.

Given I have the batting records from an early part of the season, the objective is to estimate each individual θ_i, and to predict the batting averages for the remainder of the season (denoted as BAr_i). The batting records from an early part of the season can be considered as training data, and I estimate each θ_i using the training data. The batting records for the remainder of the season are also available to us, and they can be considered as test data. So I can test the performance of our estimates by comparing the predicted values $\widehat{\theta_i}$ to the "true" values.

Brown (2008 [10]) considered a variance stabilizing transformation,

$$T = \arcsin \sqrt{\frac{X + 0.25}{Y + 0.5}}.$$

and discussed the advantages of this transformation, instead of Efron and Morris (1975 [25])'s transformation approach. Brown (2008 [10]) showed that using the

reparameterization $\mu_i = \arcsin\sqrt{\theta_i}$, I have

$$T_i \dot{\sim} \mathrm{N}\left(\mu_i, \frac{1}{4Y_i}\right),$$

4.5.2 Performance measurements

Brown (2008 [10]) proposed three performance measurements. Denote the training data in the first half of the season to be $(X_{11}, Y_{11}), \ldots, (X_{1N_1}, Y_{1N_1})$, and the test data in the second half of the season to be $(X_{21}, Y_{21}), \ldots, (X_{2N_2}, Y_{2N_2})$. The three performance measurements discussed in Brown (2008 [10]) are

$$\begin{aligned}
\mathrm{TSE} &= \sum_{i=1}^{N_2}(\widehat{\mu}_{2i} - T_{2i})^2 - \sum_{i=1}^{N_2}\frac{0.25}{Y_{2i}}, \\
\mathrm{TSER} &= \sum_{i=1}^{N_2}\left(\frac{X_{2i}}{Y_{2i}} - \widehat{\mu}_{2i}\right)^2 - \sum_{i=1}^{N_2}\frac{X_{2i}}{Y_{2i}^2}\left(1 - \frac{X_{2i}}{Y_{2i}}\right), \\
\mathrm{TWSE} &= \sum_{i=1}^{N_2} Y_{1i}(\widehat{\mu}_{2i} - T_{2i})^2 - \sum_{i=1}^{N_2}\frac{0.25Y_{1i}}{Y_{2i}}.
\end{aligned}$$

Smaller values of `TSE, TSER, TWSE` indicate better performance. Brown (2008 [10]) regard the naive estimator (directly use the batting average in the training data to be the estimate) as the baseline method, and report the ratio of `TSE`, `TSER` and `TWSE` relative to those of the naive estimator. The performance measures reported in Brown (2008 [10]) are

$$\frac{\mathrm{TSE}}{\mathrm{TSE}_0} \qquad \frac{\mathrm{TSER}}{\mathrm{TSER}_0} \qquad \frac{\mathrm{TWSE}}{\mathrm{TWSE}_0},$$

where $\mathtt{TSE}_0$, $\mathtt{TSER}_0$ and $\mathtt{TWSE}_0$ are the quantities corresponding to the naive estimators.

Models	TSE	TSER	TWSE
N_1	567	567	567
N_2	499	499	499
Naive	1.000	1.000	1.000
Group mean	0.852	0.877	0.741
EB(MM)	0.593	0.606	0.626
EB(ML)	0.902	0.925	0.607
NP EB	0.508	0.509	0.560
Harmonic prior	0.884	0.905	0.600
James-Stein	0.525	0.540	0.502
Bayesian estimator	0.884		
Muralidharan	0.588		
Zhang & Liu: Bernstein	0.663	0.683	0.532
Zhang & Liu: Dirichlet	0.697	0.725	0.597
RPTMM	**0.510**	**0.510**	**0.536**

Table 4–3: Baseball data: Performance measures for different estimates using all the players' data.

4.5.3 Results

Brown (2008 [10]) analyzed the baseball players data using several empirical Bayes approaches. 1: naive estimator: $\widehat{\mu}_i = T_i$; 2: grand mean: $\widehat{\mu}_i = \frac{\sum_{i=1}^{N} T_i}{N}$; 3: parametric empirical Bayes: assume the prior for $\mu_i \sim \mathtt{N}(\mu, \tau^2)$, then estimate (μ, τ^2) using data through method of moments or maximum likelihood; 4: nonparametric empirical Bayes; 5: James-Stein estimator; 6: Bayesian estimator: assume that $\mu_i \sim \mathtt{N}(\mu, \tau^2)$ and $\mu \sim \mathtt{Uniform}(-\infty, \infty)$ and $\tau^2 \sim \mathtt{Uniform}(0, \infty)$. Muralidharan (2010 [67]) considered a new mixture model, and Zhang and Liu (2012 [84]) used a nonparametric Bayesian model with Bernstein polynomial priors. All the results are listed in Tables 4–3 to 4–5.

Models	TSE	TWSE
N_1	486	486
N_2	435	435
Naive	1.000	1.000
Group mean	0.378	0.561
EB(MM)	0.387	0.494
EB(ML)	0.398	0.477
NP EB	0.372	0.527
Harmonic prior	0.391	0.473
James-Stein	0.359	0.469
Bayesian estimator	0.391	
Muralidharan	0.314	
RPTMM	**0.330**	**0.489**

Table 4–4: Baseball data: Performance measures for different estimates using only non-pitchers' data.

I implement the RPTMM model for the baseball players data. If following Brown (2008 [10]), I assume

$$T_i \sim \mathrm{N}(\mu_i, \sigma_i^2), \text{ independent}$$

$$\mu_i \sim \texttt{RPTMM}.$$

When using all players, the RPTMM is always among the top three models (see Table 4–3). Only considering non-pitchers (see Table 4–4), Muralidharan (2010 [67])'s approach did best in terms of TSE, RPTMM did just a little bit worse, but still better than the other approaches. In terms of TWSE, three approaches (parametric empirical Bayes, harmonic prior, and James-Stein's estimator) perform better then RPTMM. Our RPTMM model seems to perform moderately for non-pitchers. In Table 4–5 for pitchers, RPTMM is also among the top three models in terms of TSE and TWSE.

Models	TSE	TWSE
N_1	81	81
N_2	64	64
Naive	1.000	1.000
Group mean	0.127	0.262
EB(MM)	0.129	0.191
EB(ML)	0.117	0.180
NP EB	0.212	0.266
Harmonic prior	0.128	0.190
James-Stein	0.164	0.226
Bayesian estimator	0.128	
Muralidharan	0.156	
RPTMM	**0.119**	**0.200**

Table 4–5: Baseball data: Performance measures for different estimates using only pitchers' data.

In general, no approach can perform best in terms of every measurement, but RPTMM is among one of the top models to estimate the individual θ_i in the baseball players data.

4.6 Biostatistical meta-analysis

In many medical settings, several studies are carried out to address the same medical issue. For example, a typical problem is to compare results with a new treatment and an old treatment. Each study gathers a summary statistic D_i (for example, a log odds ratio), with estimated standard error $\widehat{\sigma}_i$. These studies may have inconsistent conclusions: some studies may be favorable to the new treatment and the others are not. Meta-analysis is a systematic review approach to combine all the information across these studies, and provide an overall conclusion.

A challenge in meta-analysis is the existence of heterogeneity across all the studies. Different studies may be conducted in different times and places. The participants may also have very different characteristics, like age, sex, health situation etc. The existence of heterogeneity prompts the use of random effects meta-analysis: see Smith et al. (1995 [77]). Random-effects models assume that each study has a "true effect", which should be obtained if the sample sizes of each study were infinite. Let the observed log odds ratios be $D_1, \ldots, D_N$, from classical asymptotic theory,

$$D_i \dot{\sim} \mathrm{N}(\theta_i, \widehat{\sigma}_i) \tag{4.1}$$

independently, and σ_i is assumed to be known. A popular parametric random effects model assumes that the random effects $\theta_1, \ldots, \theta_N$ arise from a normal distribution (DerSimonian and Laird 1986 [19])

$$\theta_1, \ldots, \theta_N \sim \mathrm{N}(\mu, \tau^2). \tag{4.2}$$

The (4.2) is a modelling assumption, and is made just for the sake of convenience. Higgins et al. (2009 [39]) summarize five tasks in meta analysis: estimating heterogeneity, estimating mean effect, estimating study effects, prediction, and testing. Theses tasks are easily performed for the parametric model.

However, it is very possible that the distribution of random effects are non-normal. One possibility is the presence of potential outliers, which cause heavy tails. Another possibility is that the participants can be clustered to several groups (e.g.,"young" vs "senior"), which make a mixture of normals a better choice. The possibility of the existence of non-normality encourages statisticians to develop more

flexible models, and Bayesian nonparametric models is one of them. Instead of making the normality assumption, the random effects in Bayesian nonparametric models are assumed to arise from a nonparametric distribution $\theta_1, \ldots, \theta_N \sim F$. The random distribution F is commonly assumed to have a Dirichlet process prior or Polya tree prior.

Burr and Doss (2005 [12]) proposed a "conditional Dirichlet process", which extend the ordinary mixture of Dirichlet process by fixing μ to be the median. Branscum and Hanson (2008 [9]) considered the mixture of Polya trees prior. Both of them rely on complicated Markov chain Monte Carlo method to estimate parameters. I apply our `RPTMM` to a meta-analysis introduced in Burr and Doss (2005 [12]). The `RPTMM` does not require any analytical calculation and computation is straightforward, yet it can approximate the ordinary Dirichlet process and conditional Dirichlet process results well.

4.6.1 Example: Decontamination of the digestive tract

The data consist of the results of 14 clinical trials carried out by the Digestive Tract Trialists' Collaborative Group (1993). The patients were randomized to either a dual treatment group (who received both topical and systemic antibiotics), or a control group.

The objective of the meta-analysis is to check whether the dual treatment can prevent an infection that is an important cause of death for decontamination of the digestive tract. Table 4–6 give the number infected and the total number in the treatment group, and the same information in control group. Note that the observed odds ratio in Table 4–6 seem to give inconsistent conclusions.

	Treated		Control		Est.
Trial	Infected	Total	Infected	Total	OR
1	14	45	23	46	0.46
2	22	55	33	57	0.49
3	27	74	40	77	0.54
4	11	75	16	75	0.64
5	4	28	12	60	0.71
6	51	131	65	140	0.74
7	33	91	40	92	0.74
8	24	161	32	170	0.76
9	14	49	15	47	0.86
10	14	48	14	49	1.03
11	15	51	14	50	1.07
12	34	162	31	160	1.10
13	45	220	40	220	1.16
14	47	220	40	220	1.22

Table 4–6: Decontamination of the digestive tract. First two columns list the number infected and total number in the treatment group, and column 3 and 4 provide the same information for the control group. The fifth column gives the estimated odds ratio.

Higgins et al. (2009 [39]) give a re-evaluation of random effects meta-analysis. They point out that "a sole focus on estimating the mean of a random effects distribution may be misleading", and there are "five aspects are likely to be relevant and useful results from a random effects meta-analysis". I will focus on three aspects: estimating heterogeneity, estimating the mean effect, and estimating study effects.

Burr and Doss (2005 [12]) proposed a Bayesian semiparametric model for random effects meta-analysis based on so called "conditional Dirichlet process". The random effects $\theta_1, \ldots, \theta_N$ are assumed to arise from a nonparametric distribution with Dirichlet process prior. The base distribution in Dirichlet process is taken to be $\mathtt{N}(\mu, \tau^2)$. Burr and Doss (2005 [12]) noticed that the μ in the base distribution does not have a clear interpretation, since the mean of F does not equal to μ. So the overall effects can not be clearly expressed.

Burr and Doss (2005 [12]) did add a key modification, and fixed the median F to be μ: conditional on this restriction, F is a Dirichlet process. Then μ can be used to represent the "overall effect". Burr and Doss also developed complicated MCMC approaches to estimate parameters.

Task 1: Estimating heterogeneity

In the random effects meta-analysis, from the asymptotic theory, the observed log odds-ratios D_i have normal distribution,

$$D_i \sim \mathrm{N}(\theta_i, \widehat{\sigma}_i),$$

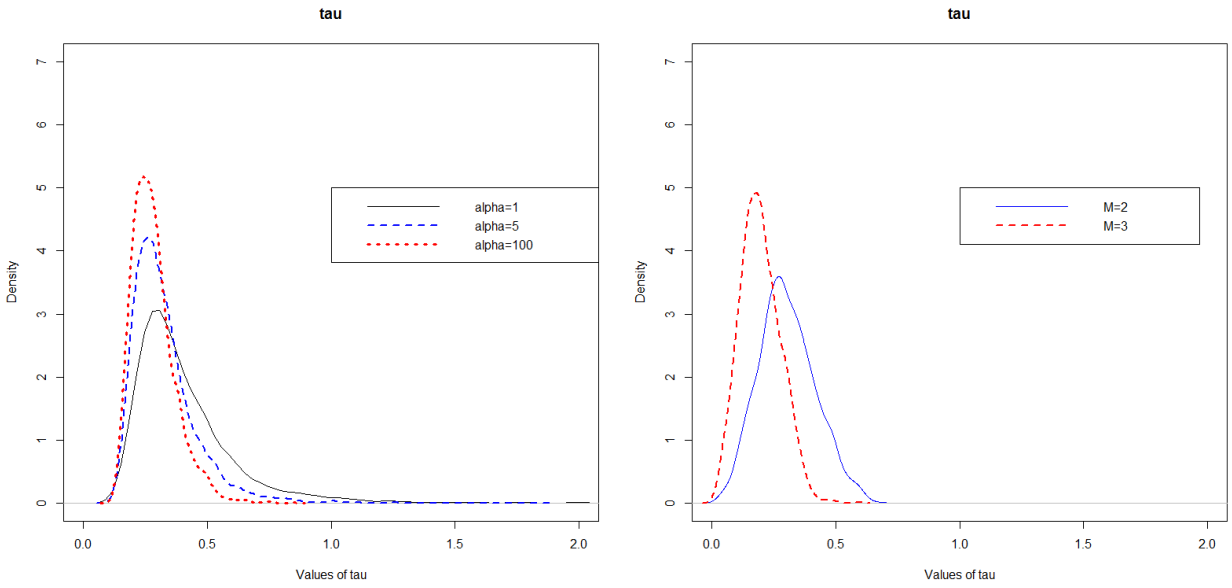

Figure 4–8: Decontamination of the Digestive tract data: Density of τ for the base distribution of conditional Dirichlet process (`CDP`) and `RPTMM`. Left panel: `CDP` with precision $\alpha = 1$, 5, or 100. Right panel: `RPTMM` with level $M = 2$ or 3.

independently with σ_i is assumed to be known, and

$$E[\theta_i] = \mu \qquad \text{var}(\theta_i) = \tau^2.$$

In practice, one of the major objectives in random effects meta-analysis is to make inference to the mean or "overall" effect, μ. Higgins et al. (2009 [39]) pointed out that "a single parameter cannot adequately summarize heterogeneous effects." They conclude that "estimation of τ^2 is just as important", and "this variance explicitly describes the extent of the heterogeneity and has a crucial role in assessing the degree of consistency of effects across studies." In their analysis, they considered the classical parametric model, and assume $\theta_1, \ldots, \theta_N \sim \texttt{N}(\mu, \tau^2)$. In this parametric case, estimating τ^2 is easy by using some standard methods, like empirical Bayes or fully Bayesian inference via MCMC.

However, if the random effects are assumed to arise from a nonparametric random distribution F, the estimation of τ is not straightforward, and a Dirichlet process or mixture of Polya trees prior may be used. Both priors have a base distribution, such as $\texttt{N}(\mu, \tau^2)$. The τ^2 in the base distribution does not equal $\texttt{var}(\theta_i)$. The first two columns in Table 4–7 summarize the value of τ in the base distribution in conditional Dirichlet process and Dirichlet process.

It is very easy to estimate the heterogeneity in our `RPTMM` model (without any analytical computation), and the $\texttt{var}(\theta_i) = \tau^2$ is directly estimated. In each run, in the `Augmentation` step, pseudo data $\theta_1^*, \ldots, \theta_d^*$ are sampled. These pseudo data are considered as samples from F. So I just need to calculate the standard deviation of

Model	τ	Model	τ	Model	τ
CDP($\alpha = 1$)	0.43 (0.39)	DP($\alpha = 1$)	0.52 (0.18)	RPTMM($M = 2$)	0.30 (0.11)
CDP($\alpha = 5$)	0.33 (0.13)	DP($\alpha = 5$)	0.48 (0.15)	RPTMM($M = 3$)	0.20 (0.08)
CDP($\alpha = 100$)	0.28 (0.09)	DP($\alpha = 20$)	0.45 (0.13)		

Table 4–7: Decontamination of the digestive tract data. First column summarizes the τ from base distribution in conditional Dirichlet process (CDP). Second column summarizes the τ from base distribution in Dirichlet process (DP). Third column summarizes the standard deviation τ in RPTMM, and M is the truncated level. The "()" indicates the corresponding standard deviation.

these pseudo data, say $\tau^{(\upsilon)}$ at iteration υ. Then I can estimate τ as

$$\widehat{\tau} = \sum_{\upsilon=1}^{\Upsilon} \tau^{(\upsilon)} \qquad \text{var}(\widehat{\tau}) = \frac{1}{\Upsilon - 1} \sum_{\upsilon=1}^{\Upsilon} (\tau^{(\upsilon)} - \widehat{\tau})^2$$

There are totally 14 trials, and according to Hanson (2006 [36])'s rule, $\log_2(14/5)$ suggests that the level should be truncated at $M = 2$; $\log_2(14)$ suggests that the level should be truncated at $M = 3$. The estimates of τ with their standard errors are listed in the third column of Table 4–7.

In the conditional and Dirichlet process models, the estimated τ is just for the base distribution. However, in RPTMM, the estimated τ is directly for F, and has a more clear interpretation. The estimated τ in RPTMM is close to that in conditional Dirichlet process (CDP) when the precision value is large (note that when the precision value is very large, the conditional Dirichlet process would be close to the parametric model).

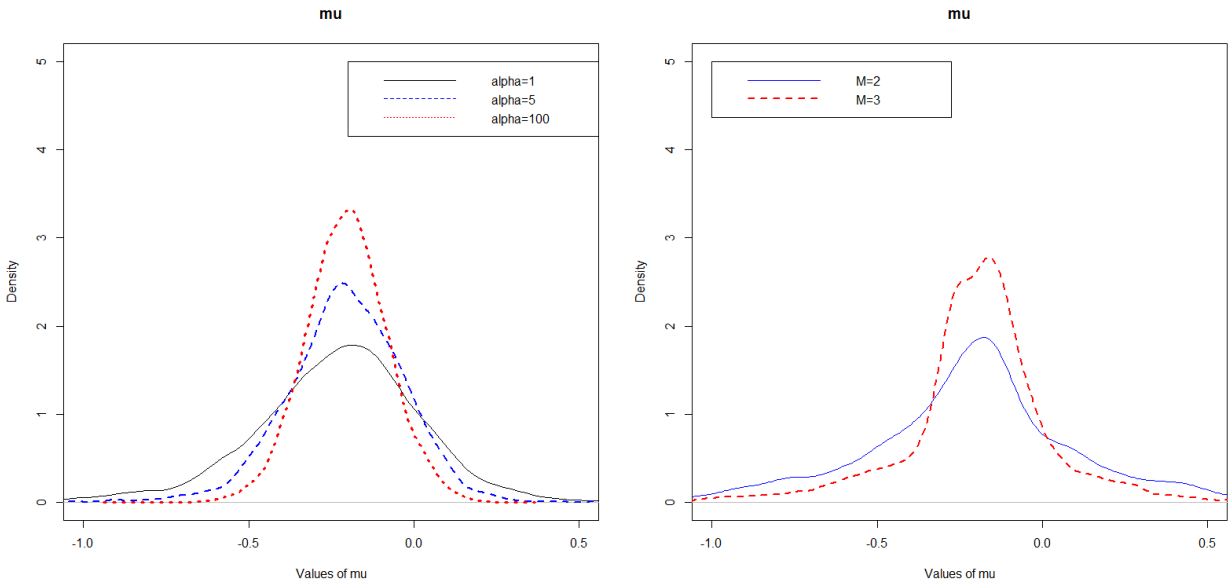

Figure 4–9: Decontamination of the Digestive tract data: Density of μ for the conditional Dirichlet process (`CDP`) and `RPTMM`. Left panel: `CDP` with precision $\alpha = 1$, 5, or 100. Right panel: `RPTMM` with level $M = 2$ or 3.

Model	μ	Model	μ	Model	μ
`CDP`($\alpha = 1$)	-0.23 (0.31)	DP ($\alpha = 1$)	-0.23 (0.21)	`RPTMM`($M = 2$)	-0.22 (0.33)
`CDP`($\alpha = 5$)	-0.21 (0.18)	DP ($\alpha = 5$)	-0.23 (0.24)	`RPTMM`($M = 3$)	-0.20 (0.22)
`CDP`($\alpha = 100$)	-0.20 (0.12)	DP ($\alpha = 20$)	-0.24 (0.26)		

Table 4–8: The first column lists the estimates of μ from conditional Dirichlet process, and α is the precision value. The second column contains the estimates from ordinary Dirichlet process mixture model, and α is the precision value. The third column lists the estimates from recursive Polya tree mixture model, and M is the truncated level. The "()" indicates the corresponding standard deviation.

Model	Prob.	Model	Prob.
CDP($\alpha = 1$)	0.840	RPTMM($M = 2$)	0.800
CDP($\alpha = 5$)	0.900	RPTMM($M = 3$)	0.874
CDP($\alpha = 100$)	0.950	.	

Table 4–9: The probability that the odds ratio is smaller than 1 from conditional Dirichlet process and recursive Polya Tree mixture model.

Task 2: Estimating the mean effect

This is one of the main objects in meta-analysis, which reflects the overall effects. Recall that in the RPTMM, in each run, in the Augmentation step, pseudo data $\theta_1^*, \ldots, \theta_d^*$ are sampled, and considered as samples from F. So I just need to calculate the mean of these pseudo data, say $\mu^{(\upsilon)}$ at iteration υ. Then I can estimate μ as

$$\widehat{\mu} = \sum_{\upsilon=1}^{\Upsilon} \mu^{(\upsilon)} \qquad \text{var}(\widehat{\mu}) = \frac{1}{\Upsilon - 1} \sum_{\upsilon=1}^{\Upsilon} (\mu^{(\upsilon)} - \widehat{\mu})^2.$$

Table 4–8 lists the estimates from conditional Dirichlet process, ordinary Dirichlet process, and recursive Polya tree mixture model. I find almost all the approaches give similar estimates.

Researchers may be interested in the estimates of the posterior probability that the odds ratio is smaller than 1, $P(\exp(\mu) < 1|D)$. From the Υ values $\mu^{(1)}, \ldots, \mu^{(\Upsilon)}$ in the Υ MCMC runs, I can easily compute the probability $P(\exp(\mu) < 1|D)$ as

$$P(\exp(\mu) < 1|D) = \frac{\sum_{l=1}^{\Upsilon} I(\exp(\mu^{(l)}) < 1)}{\Upsilon}.$$

In the meta-analysis of decontamination of the digestive tract, $P(\exp(\mu) < 1|D) = 0.8$ when truncated at level $M = 2$; and $P(\exp(\mu) < 1|D) = 0.874$ when truncated

Trial	CDP($\alpha = 1$)	CDP($\alpha = 5$)	CDP($\alpha = 100$)	RPTMM ($M = 2$)
1	0.66	0.68	0.69	0.65
2	0.66	0.68	0.69	0.64
3	0.66	0.69	0.69	0.65
4	0.74	0.76	0.76	0.73
5	0.78	0.80	0.79	0.78
6	0.77	0.77	0.77	0.76
7	0.78	0.79	0.78	0.77
8	0.79	0.79	0.79	0.78
9	0.84	0.84	0.84	0.84
10	0.88	0.88	0.87	0.89
11	0.90	0.89	0.89	0.92
12	0.96	0.95	0.95	0.98
13	0.99	0.99	0.98	1.04
14	1.00	1.01	1.02	1.07

Table 4–10: The first three columns list the study effects from a conditional Dirichlet process when precision values are set to 1, 5 and 100. The last column contains the estimates from recursive Polya tree mixture model at truncated level $M = 2$

at level $M = 3$. This suggests that the dual treatment protocol led to a significant reduction in mortality. Branscum and Hanson (2008 [9]) analyzed the same data, and find $P(\exp(\mu) < 1|D) = 0.95$ from mixture of Polya trees, and $P(\exp(\mu) < 1|D) = 0.87$ from the Dirichlet process. I used the `R` package `bspmma` to implement the conditional Dirichlet process, and find $P(\exp(\mu) < 1|D) = 0.81, 0.84, 0.95$ for precision $\alpha = 1, 5, 100$.

Task 3: Estimating study effects

Estimating study effects $\theta_1, \ldots, \theta_N$ may be of interest, especially "if studies are distinguishable from each other in ways that cannot be quantified" (Higgins et al. 2009 [39]). The estimated study effects from the conditional Dirichlet process,

Trial	DP ($\alpha = 1$)	DP ($\alpha = 5$)	DP ($\alpha = 20$)	RPTMM ($M = 3$)
1	0.79	0.74	0.72	0.73
2	0.78	0.73	0.72	0.74
3	0.78	0.73	0.72	0.74
4	0.82	0.80	0.79	0.79
5	0.84	0.83	0.83	0.81
6	0.82	0.80	0.79	0.81
7	0.83	0.81	0.80	0.81
8	0.83	0.82	0.81	0.81
9	0.85	0.85	0.85	0.83
10	0.86	0.89	0.89	0.87
11	0.87	0.89	0.90	0.87
12	0.89	0.94	0.96	0.90
13	0.91	0.98	1.00	0.94
14	0.93	1.00	1.03	0.97

Table 4–11: The first three columns list the study effects from ordinary Dirichlet process mixture model when precision values are set to 1, 5 and 20. The last column contains the estimates from recursive Polya tree mixture model truncated at level $M = 3$

Trial	CDP($\alpha = 1$)	CDP($\alpha = 5$)	CDP($\alpha = 100$)	RPTMM($M = 2$)	RPTMM($M = 3$)
1	0.93	0.92	0.93	0.94	0.97
2	0.94	0.94	0.94	0.96	0.97
3	0.94	0.94	0.96	0.97	0.98
4	0.85	0.85	0.88	0.88	0.92
5	0.77	0.77	0.81	0.81	0.89
6	0.88	0.89	0.91	0.92	0.95
7	0.84	0.85	0.87	0.88	0.93
8	0.83	0.85	0.87	0.89	0.93
9	0.74	0.74	0.77	0.78	0.88
10	0.68	0.69	0.72	0.71	0.83
11	0.67	0.68	0.70	0.66	0.82
12	0.59	0.59	0.62	0.60	0.79
13	0.53	0.52	0.54	0.51	0.71
14	0.51	0.48	0.47	0.48	0.63

Table 4–12: The estimated probability that odds ratio is smaller than 1 for each individual study.

ordinary Dirichlet process and recursive Polya tree mixture model are listed in Table 4–10 and 4–11. When truncated at level $M = 3$, I find that recursive Polya tree mixture model gives similar estimates to those from ordinary Dirichlet process. When truncated at level $M = 2$, the recursive Polya tree mixture model gives estimates similar to those from conditional Dirichlet process.

Table 4–12 lists the probability that the odds ratio is smaller than 1 for each study. The RPTMM when truncated at $M = 2$ again gives results similar to the conditional Dirichlet process when $\alpha = 1$ or 5. I find that though the recursive Polya tree mixture model does not require any analytical computation, it can approximate conditional Dirichlet process if truncated at level $M = \log_2(N/5)$, and approximate ordinary Dirichlet process mixture model if truncated at level $M = \log_2(N)$.

Trial	L_j	e.s.e.(L_j)
1	-0.04	0.161
2	-0.13	0.179
3	-0.36	0.172
4	0.01	0.119
5	-0.45	0.180
6	-0.34	0.103
7	-0.24	0.092
8	-0.51	0.247
9	-0.22	0.147
10	-0.37	0.207
11	-0.22	0.422
12	-0.36	0.172
13	-0.27	0.096
14	-0.31	0.092
15	0.10	0.091
16	0.00	0.114
17	-0.92	0.147

Table 4–13: Data on possible risk-lowering effect of NSAIDS on breast cancer: Summary data from 17 studies on aspirin and breast cancer: the observed log risk ratio L_j and the corresponding estimated standard error e.s.e.(L_j).

4.6.2 Example: Effect of NSAIDS on risk of breast cancer

The use of non-steroidal anti-inflammatory drugs (NSAIDs) is believed to reduce the risk of breast cancer. There have been some studies on the effect of NSAIDs on risk of breast cancer in the literature. Some studies have strongly suggested that long-term use of NSAIDs can decrease the risk of breast cancer significantly; while others suggested just a slight risk reduction or no risk reduction. These kinds of inconsistent results may be due to the heterogeneity in the subjects (age, ethnicity, and health status).

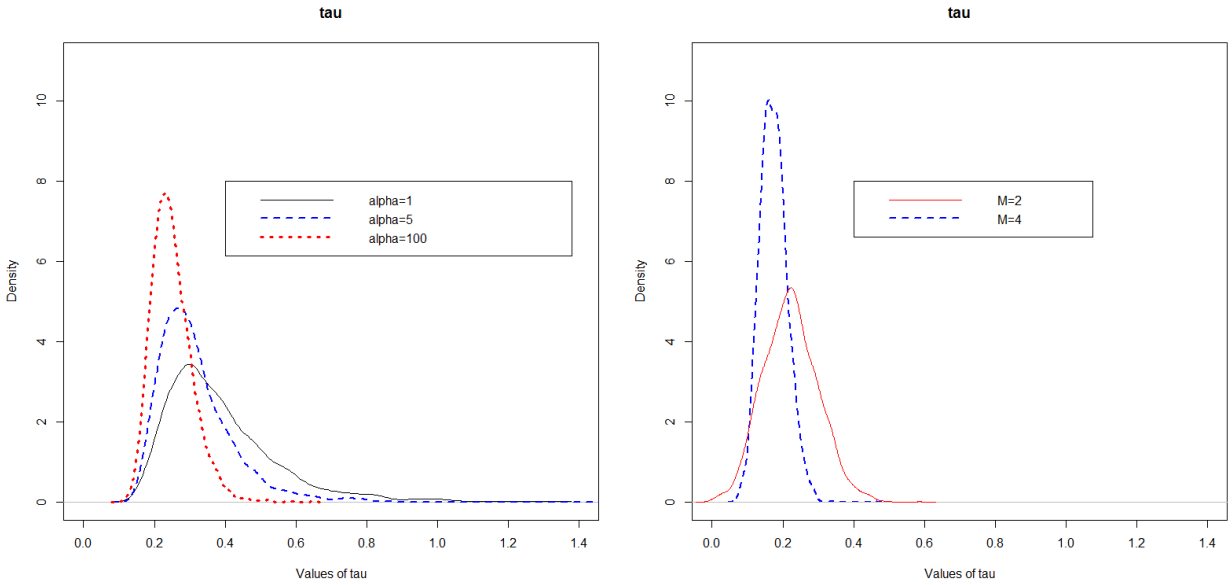

Figure 4–10: Data on possible risk-lowering effect of NSAIDS on breast cancer: The density of τ from conditional Dirichlet process and recursive Polya tree mixture model. The τ is just for the base distribution in `CDP`. In `RPTMM`, it is directly for F.

Burr (2012 [11]) considered 17 studies, and Table 4–13 lists the log risk ratios and the corresponding standard errors for these studies. Burr (2012 [11]) used the conditional Dirichlet process (Burr and Doss 2005 [12]), and concluded that "long-term use of NSAIDs appears to be associated with reduction of the risk of breast cancer at least at the study level". In next several section, I apply the recursive Polya tree mixture model to reanalyze it.

Task 1: Estimating heterogeneity

Table 4–14 lists the posterior mean of τ in the base distribution from conditional Dirichlet process (`CDP`) and ordinary Dirichlet process mixture models (first two

Model	τ	Model	τ	Model	τ
CDP($\alpha = 1$)	0.40 (0.19)	DP($\alpha = 1$)	0.52 (0.18)	RPTMM($M = 2$)	0.22 (0.08)
CDP($\alpha = 5$)	0.32 (0.11)	DP($\alpha = 5$)	0.47 (0.15)	RPTMM($M = 4$)	0.17 (0.04)
CDP($\alpha = 100$)	0.25 (0.06)	DP($\alpha = 100$)	0.40 (0.11)		

Table 4–14: The first column lists the estimate of τ with its standard error from conditional Dirichlet process, and α is the precision value. The second column contains the same estimate for ordinary Dirichlet process. The third column lists the estimates from recursive Polya tree mixture model, and M is the truncated level.

Model	μ	Model	μ	Model	μ
CDP($\alpha = 1$)	-0.33 (0.19)	DP($\alpha = 1$)	-0.27 (0.16)	RPTMM($M = 2$)	-0.24 (0.24)
CDP($\alpha = 5$)	-0.29 (0.12)	DP($\alpha = 5$)	-0.27 (0.19)	RPTMM($M = 4$)	-0.24 (0.18)
CDP($\alpha = 100$)	-0.26 (0.07)	DP($\alpha = 100$)	-0.27 (0.22)	.	.

Table 4–15: The first column lists the estimate of μ with its standard error from the conditional Dirichlet process, and α is the precision value. The second column is the same estimate from the ordinary Dirichlet process. The third column lists the estimates from recursive Polya tree mixture model, and M is the truncated level.

columns). In the recursive Polya tree mixture model, the posterior means of standard deviation τ are listed in the third column of Table 4–14.

Just as the previous discussion, the τ for the RPTMM has a more clear interpretation; while the τ in CDP or DPMM is just for the base distribution, and does not have clear interpretation.

Task 2: Estimating the mean effect

For the overall effect μ, Table 4–15 lists the posterior means from the conditional Dirichlet process, the ordinary Dirichlet process mixture model, and the recursive Polya tree mixture model. The posterior density plots are plotted in Figure 4–11.

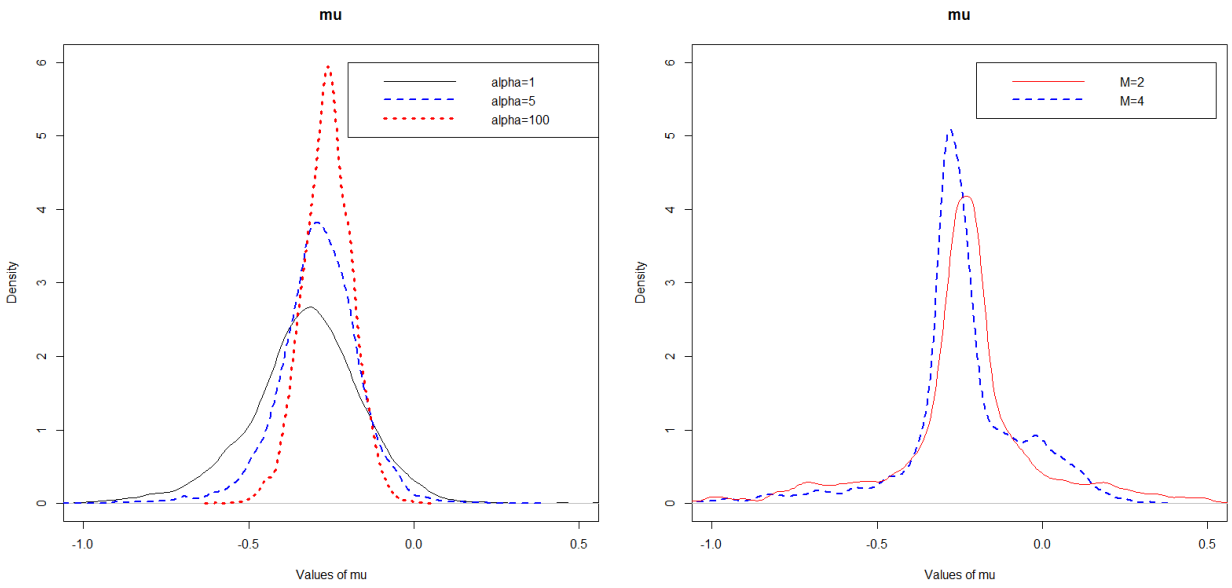

Figure 4–11: Data on possible risk-lowering effect of NSAIDS on breast cancer: Density of μ from conditional Dirichlet process and recursive Polya tree mixture model.

Model	Prob.	Model	Prob.
CDP($\alpha = 1$)	0.98	RPTMM($M = 2$)	0.90
CDP($\alpha = 5$)	0.99	RPTMM($M = 4$)	0.90
CDP($\alpha = 100$)	1.00	.	.

Table 4–16: The probability that the risk ratio is smaller than 1 from conditional Dirichlet process and recursive Polya Tree mixture model.

Trial	CDP($\alpha = 1$)	CDP($\alpha = 5$)	CDP($\alpha = 100$)	RPTMM(M=2)	RPTMM($M = 4$)
1	0.90	0.90	0.90	0.88	0.86
2	0.84	0.84	0.84	0.83	0.82
3	0.73	0.73	0.73	0.74	0.76
4	0.96	0.96	0.95	0.98	0.93
5	0.69	0.69	0.68	0.69	0.74
6	0.72	0.72	0.72	0.76	0.75
7	0.77	0.78	0.79	0.79	0.78
8	0.68	0.68	0.68	0.68	0.74
9	0.79	0.79	0.79	0.80	0.79
10	0.73	0.73	0.73	0.74	0.76
11	0.76	0.78	0.78	0.80	0.79
12	0.73	0.73	0.73	0.74	0.76
13	0.75	0.76	0.76	0.78	0.77
14	0.73	0.73	0.73	0.77	0.76
15	1.00	1.03	1.05	1.10	1.04
16	0.96	0.95	0.95	0.96	0.93
17	0.47	0.47	0.48	0.53	0.51

Table 4–17: The first three columns list the study effects from conditional Dirichlet process when precision values are set to 1, 5 and 100. The last column is the estimate from recursive Polya tree mixture model at truncated level $M = 2$ or $M = 4$.

Table 4–16 summaries the probability that risk ratio is smaller than 1 overall. All the probabilities are close to 1, which confirms the conclusion that: long-term use of NSAIDs is associated with reduction of the risk of breast cancer.

Task 3: Estimating study effects

Table 4–17 lists the posterior mean of risk ratios for each study. I find the estimates for the RPTMM are close to those from conditional Dirichlet process. Table 4–18 summaries the probability that risk ratios smaller than 1 for each study. I also find that the two methods give similar results.

Trial	CDP($\alpha = 1$)	CDP($\alpha = 5$)	CDP($\alpha = 100$)	RPTMM(M=2)	RPTMM($M = 4$)
1	0.72	0.78	0.79	0.80	0.84
2	0.83	0.87	0.89	0.89	0.93
3	0.98	0.99	0.99	0.99	0.99
4	0.62	0.66	0.66	0.62	0.68
5	0.99	0.99	0.99	0.99	1.00
6	1.00	1.00	1.00	1.00	1.00
7	0.99	0.99	1.00	1.00	1.00
8	0.97	0.98	0.98	0.99	0.99
9	0.95	0.96	0.96	0.97	0.98
10	0.97	0.98	0.98	0.97	0.98
11	0.87	0.88	0.88	0.86	0.92
12	0.98	0.99	0.99	0.99	0.99
13	1.00	1.00	1.00	1.00	1.00
14	1.00	1.00	1.00	1.00	1.00
15	0.47	0.40	0.30	0.37	0.34
16	0.62	0.67	0.68	0.67	0.70
17	1.00	1.00	1.00	1.00	1.00

Table 4–18: The probability that the risk ratio is smaller than 1 for each individual study.

In this chapter, I discuss the application of `RPTMM` to some examples related to Bayesian nonparametric hierarchical models. In next chapter, I build a Bayesian semiparametric accelerated failure time model based on `RPTMM`, and also discuss the extension to recurrent data analysis.

CHAPTER 5
Survival Analysis Models

5.1 A Bayesian semiparametric AFT model

In survival analysis, inference concerning the effect of covariates on the survival time T is of central importance. Cox's proportional hazards (PH) model (Cox 1972 [18]) is the most commonly used and popular model, and the associated partial likelihood theory of estimation is easily implemented using standard packages. However, the proportional hazard assumption may not be appropriate in a given analysis setting. In addition, it is also hard to interpret the regression coefficients, as Cox himself said (Reid 1994 [73]):

"Of course, another issue is the physical or substantive basis for the proportional hazards model. I think that's one of its weakness, that accelerated life models are in many ways more appealing because of their quite direct physical interpretation, particularly in an engineering context."

A popular alternative is the *accelerated failure time* (AFT) model, which assumes that for individual $i, i = 1, \ldots, N$,

$$T_i = \exp(x_i\beta)V_i \tag{5.1}$$

$$V_1, \ldots, V_N | G \sim G. \tag{5.2}$$

Here T_i denotes the survival time, x_i represents the covariates, β is a vector of regression coefficients, and G is a baseline survival distribution. A limitation of the parametric AFT model is that the baseline distribution G must be specified in advance. Kalbfleisch and Prentice (1980 [46]) introduced details of parametric AFT model: common choices for G are Weibull, gamma, log-normal, and log-logistic. The requirement to specify G in advance restricts the applicability of the AFT model.

Since Ferguson (1973 [28]) proposed the Dirichlet process, Bayesian nonparametric approaches have played important roles in developing flexible AFT models. Christensen and Johnson (1988 [16]) presented a semiparametric AFT model based on a simple Dirichlet process. However, Hanson and Johnson (2004 [38]) pointed out that "this approach does not allow a prior to be placed on β or credible intervals to be calculated for any parameter." Kuo and Mallick (1997 [52]) proposed an AFT model using a Dirichlet process mixture (DPM) model. Their approach is based on the algorithms from Escobar (1994 [26]) and Escobar and West (1995 [27]). Walker and Mallick (1999 [79]) introduced an AFT model which assumes that G is a random distribution with a simple Polya tree prior. Based on their work, Hanson and Johnson (2004 [38]) proposed a semiparametric AFT model using a stick-breaking process (Sethuraman 1994 [75]), and their algorithm "Chain Two" can be easily fitted to interval-censored data.

Although many flexible AFT models have been proposed, their popularity is still limited (Jara et al., 2011 [45]). One of the reasons is that these Bayesian semiparametric AFT models rely on complicated computing algorithms, and are hard to

be understood and implemented. In this Chapter, I develop a Bayesian semiparametric AFT model based on the recursive Polya tree mixture model and term this the `RPTMM-AFT` model, which not only enjoys the flexibility provided by Bayesian nonparametrics, but is also very easy to implement.

In Section 5.2, I introduce the detail of the `RPTMM-AFT` model. In Section 5.3, I study a simulated data set (Hanson and Johnson, 2004 [37]), in which the baseline distribution is assumed to have mixture form; a right-censored data set; and an interval-censored data set. In Section 5.4, I extend the semiparametric AFT model to the application of recurrent data analysis, and show that our approach can approximate the AFT model based on the reversible jump MCMC.

5.2 The model

I write the model of (5.1) in the regular regression form

$$Y_i = \log(T_i) = x_i\beta + \log(V_i) = x_i\beta + \epsilon_i,\ i = 1, \ldots, N,$$

$$V_1, \ldots, V_N | G \sim G \text{ i.i.d.}$$

The coefficients β are assigned a prior distribution, $\beta \sim \pi(\beta)$. Here I assume the prior for β is noninformative, but it is straightforward to implement the other kinds of prior.

In the following sections, I give details of the `RPTMM` tailored for survival data problems in general, and the AFT model in particular.

5.2.1 The algorithm

Step 1. Initialization step

I firstly initialize the parameters in the model. For each i, $\epsilon_i = \log(V_i)$ is assumed to arise from a mixture distribution

$$\epsilon_i|\theta_i, \sigma_i^2 \sim \mathrm{N}(\theta_i, \sigma_i^2) \qquad \theta_i \sim F$$

F is initialized as a uniform distribution on a broad support,

$$F^{(0)} \sim U(\min(Y) - |\max(Y) - \min(Y)|, \max(Y) + |\max(Y) - \min(Y)|).$$

Parameters θ_i and σ_i^2 are initialized as $\theta_1^{(0)} = \cdots = \theta_N^{(0)} = 0$ and $\sigma_1^{(0)} = \cdots = \sigma_N^{(0)} = 1$.

Step 2. Estimation of β

With θ_i and σ_i^2 fixed at their current values, the estimation process of β is very standard. Let $\widetilde{Y}_i = Y_i - \theta_i$. Then

$$\widetilde{Y}_i = x_i\beta + \epsilon_i - \theta_i = x_i\beta + \widetilde{\epsilon}_i,$$

where $\widetilde{\epsilon}_i|\theta_i, \sigma_i \sim \mathrm{N}(0, \sigma_i^2)$. Here, for a prior specification I have $\beta \sim \pi(\beta)$, where $f(\beta)$ is noninformative. From page 375 of Gelman et al. (2003 [33]),

$$\beta|\mathrm{Data}, \theta, \sigma^2 \sim \mathrm{N}(A, B), \tag{5.3}$$

where

$$A = (X^T W^{-1} X)^{-1} X^T W^{-1} \widetilde{Y} \qquad B = \tau^2 (X^T W^{-1} X)^{-1}.$$

Here W is an $N \times N$ diagonal matrix whose diagonal elements are $\sigma_1^2/\tau^2, \ldots, \sigma_N^2/\tau^2$. τ^2 is drawn from

$$\frac{1}{\tau^2} \sim \text{Gamma}\left(0.1 + \frac{N}{2}, 0.1 + \frac{1}{2}\sum_{i=1}^{N}(\epsilon_i - \overline{\epsilon})^2\right)$$

At each iteration υ ($\upsilon = 1, \ldots, \Upsilon$), I draw a value $\beta^{(\upsilon)}$ from (5.3), to obtain the final estimate of β as the average of all the drawn values from each step $\upsilon = 1, \ldots, \Upsilon$,

$$\widehat{\beta} = \frac{1}{\Upsilon}\sum_{\upsilon=1}^{\Upsilon}\beta^{(\upsilon)},$$

and the 95% credible intervals for β can be obtained from the quantiles of the drawn values $\beta^{(1)}, \ldots, \beta^{(\upsilon)}, \ldots, \beta^{(\Upsilon)}$ from each step.

Step 3. Estimation of θ_i

This is the key step in the `RPTMM-AFT` model, and in this step I assume the σ_i has been fixed at the current value. Consider the error term at iteration υ

$$\epsilon_i^{(\upsilon)} = Y_i - x_i\beta^{(\upsilon)}.$$

I model ϵ_i using the Bayesian hierarchical model as,

$$\epsilon_i|\theta_i, \sigma_i \sim \text{N}(\theta_i, \sigma_i^2), \text{ independently}$$

with

$$\theta_1, \ldots, \theta_N \sim F \text{ i.i.d.}$$

I assume that F has a Polya tree distribution, and use the recursive algorithm from Section 3.2 to perform Bayesian nonparametric analysis.

Estimation of σ_i^2

In the `RPTMM-AFT` model, $\epsilon_i|\theta_i \sim \mathtt{N}(\theta_i, \sigma_i^2)$ independently, and $\theta_1, \ldots, \theta_N \sim F, \mathtt{i.i.d}$. If σ_i is unknown to us and heterogeneity is believed to exist, each σ_i should be estimated. I could use the method described in Section 3.5 to estimate each individual σ_i at each step. Thus each individual σ_{Mj} corresponding to each component $j(j = 1, \ldots, 2^M)$ can be estimated as

$$\widehat{\sigma}_{Mj}^2 = \frac{1}{\gamma_j} \sum_{i=1}^{N} \omega_{ij} (\epsilon_i - \zeta_j)^2$$

where ζ_j is the midpoint of interval I_{Mj} and

$$\gamma_j = \sum_{i=1}^{N} \omega_{ij}.$$

The σ_i corresponding to each data ϵ_i is estimated as

$$\widehat{\sigma}_i^2 = \sum_{j=1}^{2^M} \omega_{ij} \widehat{\sigma}_{Mj}^2, \tag{5.4}$$

where $\omega_{ij} = P(\theta_i \in I_{Mj}|\epsilon_i)$.

5.2.2 Censored data

If Y_i is censored in $[C_1, C_2)$, it is easy to impute values through the truncated normal distribution. I sample a value for Y_i using

$$Y_i \sim \text{truncated N}(x_i\beta + \theta_i, \sigma_i^2) \text{ in } [C_1, C_2). \tag{5.5}$$

The sampling process for the truncated normal distribution can be done efficiently in `R` (see for example the function `ptnorm` in the package `msm`).

Summary of the algorithm

1. Initialize $F^{(0)}$, $\theta^{(0)}$, and $\sigma^{(0)}$. A Polya tree truncated at level M is built and centered on $F^{(0)}$.
2. For $l = 1, \ldots, \Upsilon$,
   - Sample β from equation (5.3);
   - Use `RPTMM` to estimate θ

     (a) For each data point ϵ_i, calculate $P(\theta_i \in I_{Mk}|\epsilon_i) = \omega_{ik}$ $(k = 1, \ldots, 2^M)$, and calculate $P_{Mj} = \sum_{i=1}^{N} \omega_{ij}/N$

     (b) Draw a large number of pseudo data $\theta_1^*, \ldots, \theta_d^*$

     (c) Built a new Polya tree using the empirical distribution of $\theta_1^*, \ldots, \theta_d^*$ as base distribution

     (d) Calculate θ_i using the `RPTMM` procedure.
   - Estimate σ_i^2 using equation (5.4).
   - Impute censored data according to equation (5.5).
3. Repeat Steps 1 and 2 many times.

Estimating the survival function

To estimate the probability $P(T > t|X, Y)$, I notice that

$$\begin{aligned} P(T > t|X, Y) &= P(\exp(X\beta + \epsilon) > t|X, Y) \\ &= P(\epsilon > -X\beta + \log(t)|X, Y). \end{aligned} \tag{5.6}$$

I could empirically estimate $P(\epsilon > -X\beta + \log(t)|X, Y)$ at iteration step υ through

$$\widehat{S}^{(\upsilon)}(t) = P(\epsilon > -X\beta + \log(t)|X, Y) = \frac{\sum_{i=1}^{N} I(\epsilon_i^{(\upsilon)} > -X\beta^{(\upsilon)} + \log(t))}{N},$$

where $I(.)$ is indicator variable. Let the estimate at the iteration υ be $\widehat{S}^{(\upsilon)}(t)$, so the final estimate will be

$$\widehat{S}(t) = \frac{\sum_{\upsilon=1}^{\Upsilon} \widehat{S}^{(\upsilon)}(t)}{\Upsilon}.$$

The samples $\widehat{S}(t)^{(1)}, \ldots, \widehat{S}^{(\Upsilon)}(t)$ can be used to build the credible intervals. For example, the 95% credible interval is $[\widehat{S}(t)_{0.025}, \widehat{S}(t)_{0.975}]$, where $\widehat{S}(t)_q$ is the q quantile of $\widehat{S}^{(1)}(t), \ldots, \widehat{S}^{(\Upsilon)}(t)$.

Estimating the intercept

For the model

$$Y = \log(T) = \beta_0 + X\beta + \epsilon,$$

$\epsilon \sim F$ is not assumed to have the mean zero. It will cause confounding of the intercept, but not influence the estimate of β. I rewrite the model as

$$Y = \log(T) = \beta_0 + \overline{\theta} + X\beta + \epsilon - \overline{\theta} = \widetilde{\beta}_0 + X\beta + \widetilde{\epsilon},$$

where $\overline{\theta}$ is the mean of pseudo $\theta_1^*, \ldots, \theta_d^*$ at each iteration υ. It is easy to verify that $E(\widetilde{\epsilon}|\theta, \sigma) = 0$. With the constraint, the $\widetilde{\beta}_0$ is the intercept I need to estimate.

5.3 Examples

Example 12. Simulated data

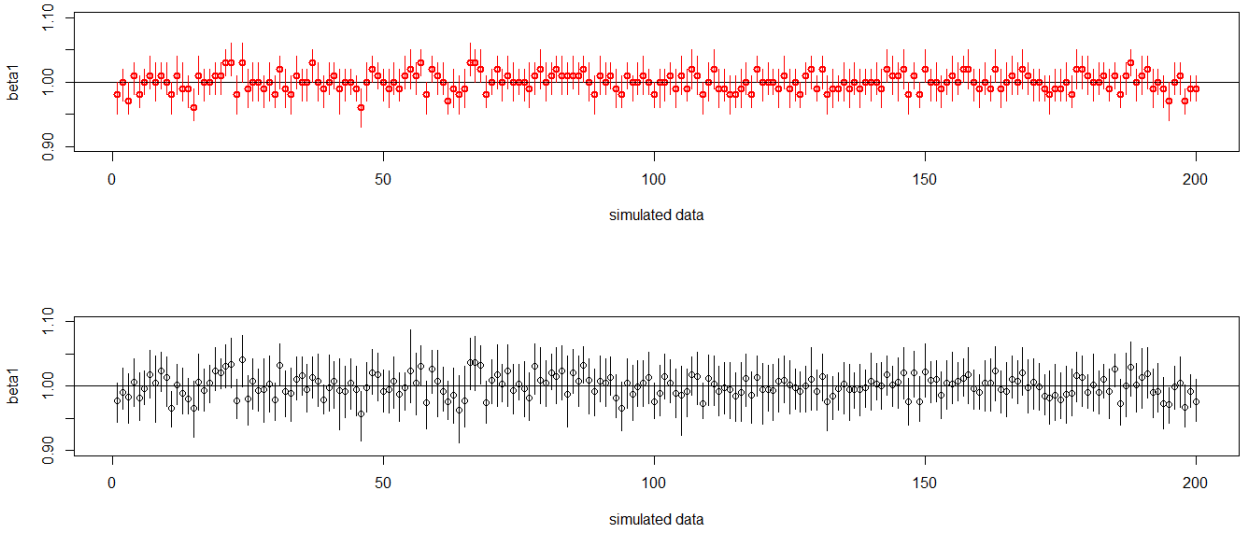

Figure 5–1: Simulated data: Estimating the true $\beta_1 = 1$. The top panel represents the estimates from `RPTMM-AFT` model, and the bottom panel represents the estimates from the `MPT-AFT` model. The 95% credible interval for each dataset is plotted.

I simulated 100 data points $V_1, \ldots, V_{100}$ from a mixture of normal distributions (from Hanson and Johnson 2002 [37]):

$$V_1, \ldots, V_{100} \sim 0.5\,\mathrm{N}(1, 0.15^2) + 0.5\,\mathrm{N}(3, 0.15^2)$$

Two covariates were generated by taking $x_{i,1} \sim 0.5\delta_0 + 0.5\delta_1$, where δ_a is a point mass at a; the second covariate $x_{i,2} \sim \mathtt{N}(0, 1)$. The true vector of coefficients was set at $\beta = (1, -1)$ and the survival times are calculated as

$$T_i = \exp(x_i\beta)V_i.$$

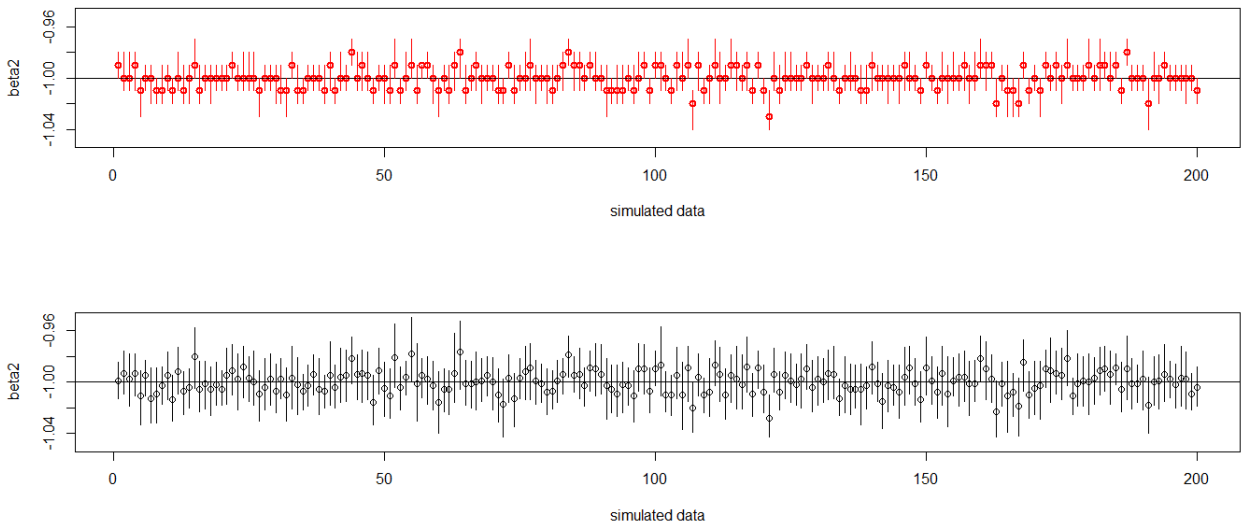

Figure 5–2: Simulated data: Estimating the true $\beta_2 = -1$. The top panel represents the estimates from `RPTMM-AFT` model, and the bottom panel represents the estimates from the `MPT-AFT` model. The 95% credible interval for each dataset is plotted.

The baseline survival distribution is a mixture distribution, which is very possible to occur in real applications. Hanson and Johnson (2002 [37]) fit this kind of simulated data using a mixture of Polya tree (`MPT`) model. They found that a parametric model cannot recover the accurate estimates of β, while by using the nonparametric baseline distribution (mixture of Polya trees), the true β can be accurately estimated.

200 data sets have been simulated and implemented using the proposed approach (`RPTMM-AFT`) and the AFT model based on mixture of Polya tree (`MPT-AFT`, using the R command `PTlm` in R package `DPpackage`). I plot the posterior mean and 95% credible intervals for the estimates of $\beta = (1, -1)$ for the 200 simulated datasets (in Figure 5–1 and 5–2), and find that both of the two Bayesian semiparametric

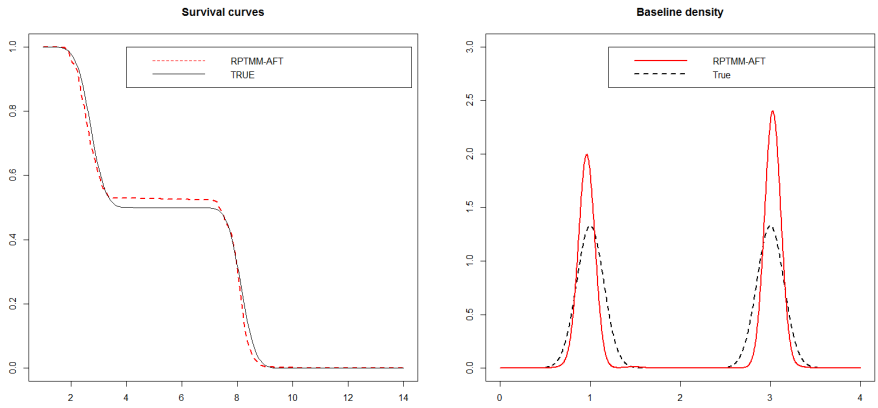

Figure 5–3: Simulated data: Survival curves and baseline density estimates for covariates $x = (1, 0)$.

AFT models give similarly accurate results. The average errors for the 200 data sets $\frac{1}{200}\sum_{i=1}^{200}(\widehat{\beta} - \beta)^2$ for the two Bayesian semiparametric AFT models are calculated. Both models give the average errors smaller than 0.001, and confirm their similar performance

In Figure 5–3, for one of the simulated data sets, I plot the survival curves and baseline density estimates for covariate combination $x = (1, 0)$. I find that they are very close to the true curves. Figure 1 of Hanson and Johnson (2002 [37]) also gives similar curves. I provide an alternative way which can approximate the semiparametric AFT model based on a mixture of Polya trees.

Example 13. Right censored case: Small cell lung cancer data

I demonstrate the proposed method for the small cell lung cancer data (Ying et al 1995 [83]). The standard therapy for patients with small cell lung cancer is to use the combination of drugs etoposide and cisplatin. However, whether to use etoposide

Models	β_0	β_1	β_2
Ying (1995)	3.03	-0.16 (-0.34,0.01)	-0.004 (-0.014,0.006)
Yang M1 (1999)	3.16 (2.86,3.17)	-0.17 (-0.24,-0.09)	-0.006 (-0.009,0.001)
Yang M2 (1999)	3.09 (2.95,3.20)	-0.16 (-0.23,-0.10)	-0.005 (-0.008,-0.001)
Huang et al (2007)	2.70 (2.37,3.01)	-0.15 (-0.24,-0.05)	0.001 (-0.005,0.007)
`RPTMM-AFT`	**3.24 (2.94,3.54)**	**-0.17 (-0.27,-0.08)**	**-0.006 (-0.011,-0.002)**

Table 5–1: Small cell lung cancer data. The first four rows list the estimates of coefficient and the 95% confidence intervals from the literature. The fifth row lists the posterior mean and 95% credible interval from `RPTMM-AFT` model

or cisplatin first is not clear. The data are from a clinical study, where in Group 1 cisplatin followed by etoposide, and in Group 2 etoposide followed by cisplatin. In this study, 121 patients with small cell lung cancer were randomly assigned to these two groups (62 patients to Group 1 and 59 patients to Group 2).

There was no loss to follow-up in this study, and each terminal event was either an observed death or administratively censored. Let $X_1 = 0$ if the patient is in Group 1 and 1 otherwise. Let X_2 denote the patient's entry age. Instead of using natural log, $Y = \log_{10}(T)$, so the model is $Y = \log_{10}(T) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

The estimates of coefficients are listed in Table 5–1. Ying et al. (1995 [83]) proposed a survival analysis with median regression models. Yang and Prentice (1999 [82]) used weighted empirical survival and hazard functions for this right censored data. Recently, Huang (2007 [42]) presented a least absolute deviations estimate for AFT model. Here, I fit these data using the Bayesian semiparametric `RPTMM-AFT` model.

Estimates of the covariate effects from each model are similar. β_1 is significant, which indicates that Group 1 (cisplatin followed by etoposide) seems to perform

better. The 95% credible interval of β_2 is a little bit significant in `RPTMM-AFT` model.

Example 14. Interval censored case: Cosmetic effects of cancer therapy

Interval-censored survival data are very common in real applications. However, the literature to handle interval-censored cases using nonparametric approaches is limited. In frequentist statistics, Turnbull's EM algorithm (Turnbull 1976 [78]) seems to be a popular choice. In Bayesian nonparametric analysis, Doss and Huffer (2003 [23]) and Hanson and Johnson (2004 [38]) proposed two algorithms based on mixtures of Dirichlet processes. In this section, I discuss interval-censored data, and show that the proposed `RPTMM-AFT` model can provide accurately approximate Bayesian inference.

For women with early breast cancer, radiotherapy is an alternative treatment to mastectomy. An interesting question is whether a combination of radiotherapy and chemotherapy can reduce the cosmetic effect that induce breast retraction quickly. Beadle et al. (1984 [4]) presented a retrospective study of 46 radiation only patients, and 48 radiation plus chemotherapy patients. Patients went to a clinician typically every four to six months, and the level of breast retractions since the last visit was recorded as none, moderate or severe. If moderate or severe breast retraction occurred, the time of retraction was known only to lie between the present time and last visit, so is interval-censored. If no moderate or severe breast retractions have occurred at the last visit of a patient, the last visit time is considered as right censored.

I fit our `RPTMM-AFT` model to these data, under $Y = \log(T) = \beta_0 + \beta_1 X_1 + \epsilon$. The covariate of interest is $x_{1i} = 0$ if the ith patient had radiotherapy and chemotherapy, and $x_{1i} = 1$ if the ith patient had radiotherapy only. The posterior medians for β_1 are listed in the first row in Table 5–2. The posterior median of months for the radiotherapy only group is 38.8, with the 95% credible interval (30.5, 47.3). Adding chemotherapy to radiation treatment reduced the median time to retraction to 23.1 with 95% credible interval (20.0, 27.6). Both the algorithms based on mixture of Dirichlet process in Hanson and Johnson (2004 [38]) and the `RPTMM-AFT` model conclude that adding chemotherapy significantly reduces the time to deterioration.

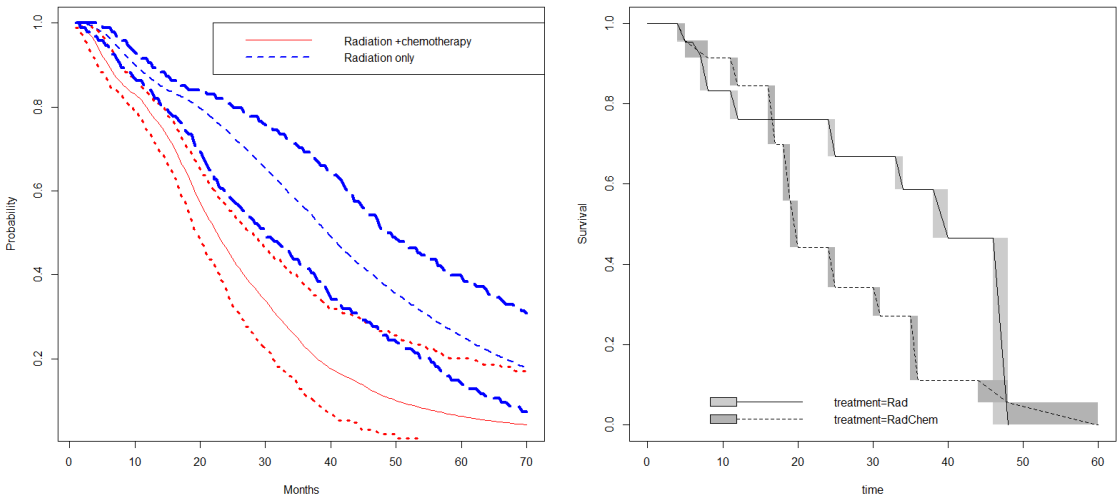

Figure 5–4: Breast cancer data set. Left panel: comparison of functions of retraction time for the two treatment groups and the 95% credible intervals from the `RPTMM-AFT` model. Right panel: Survival functions from EM algorithm (`R` package `interval`)

I also calculated the posterior medians and 95% credible intervals for the difference in survival across groups, and they are listed in Row 4-7 of Table 5–2. Before

Estimand	MDP ($\alpha = 20$)	RPTMM-AFT
Median of β_1	0.57 (0.16,0.91)	**0.53 (0.16,0.81)**
Median months for radiotherapy only	39.00 (28.00, 50.00)	**38.80 (30.50,47.30)**
Median months for radiotherapy+chemotherapy	22.00 (18.00, 29.00)	**23.1(20.00, 27.60)**
$S(t\|x = 1) - S(t\|x = 0)$ at month 10	0.08 (0.01,0.16)	**0.07 (0.01, 0.12)**
$S(t\|x = 1) - S(t\|x = 0)$ at month 20	0.25 (0.08,0.44)	**0.22 (0.07, 0.33)**
$S(t\|x = 1) - S(t\|x = 0)$ at month 30	0.30 (0.04, 0.51)	**0.30 (0.05,0.49)**
$S(t\|x = 1) - S(t\|x = 0)$ at month 40	0.29 (0.07, 0.52)	**0.30 (0.07,0.49)**

Table 5–2: Breast cancer data set: comparison of the Mixture of Dirichlet processes (Hanson and Johnson 2004 [38]) with the RPTMM-AFT model. The first row lists the posterior median of β_1. The second and third rows list the median months for each groups. The rest four rows list the differences in survival across groups at month 10, 20, 30, and 40. The "( )" lists the 95% credible interval.

10 months, there is little difference in survival between the two groups. After month 20, the difference becomes larger. All the estimates from RPTMM-AFT model are reasonably close to those from the AFT model based on the MDP algorithm (Hanson and Johnson 2004 [38]).

Klein and Moeschberger (1997 [48]) presented an analysis based on Turnbull's nonparametric maximum likelihood estimate (Turnbull 1976 [78]) of a survival function. Finkelstein and Wolfe (1985 [30]) used their semiparametric model and drew the estimated survival curves. Both their curves become constant after month 40.

I plot the comparison of distributions of retraction time for the two treatment groups in Figure 5–4, and I also plot the survival curves from EM algorithm using the R package interval. Our curves are very close to the Figure 4 of Hanson and Johnson (2004 [38]). There are two significant differences between the survival curves from Bayesian nonparametric models and EM algorithms:

- Before month 15, according to the EM algorithm, the survival curve from the "Radiation only" group is above the curve from the "Radiation plus Chemotherapy" group. While the Bayesian nonparametric models give different result. Through Table 5–2, I know that before month 10, the difference between the two groups is small.
- Bayesian nonparametric survival curves are smoother than those from EM algorithm.

5.4 Recurrent event data analysis

In many medical and scientific studies, patients can experience recurrent or repeated events during follow-up, such as recurrent infections or attacks in patients suffering from some disease within a longitudinal study. The challenge in recurrent data analysis is that the independence assumption between the event times is broken. Related failures of the same patient are dependent.

Lin et al. (1994 [56]) and Cook and Lawless (2006 [17]) give an excellent overview of the existing approaches to handle recurrent data, and most of them are based on the Cox's model. Komarek and Lesaffre (2007 [49]) proposed an accelerated failure time model for recurrent data, in which the error term is assumed to have a mixture of normal distributions. Reversible jump MCMC (Green 1995 [34]) is used to find the number of components and estimate parameters. In this section, I extend the `RPTMM-AFT` model proposed above to the recurrent data analysis. The error term is also assumed to arise from a nonparametrically specified distribution. Instead of complicated reversible jump MCMC, the recursive Polya tree mixture model is used.

For individual i, assume the failure time is $T_{i0} = 0, T_{i1}, \ldots, T_{in_i}$, I consider the log of gap time $Y_{ij} = \log(T_{ij} - T_{i(j-1)}), j = 1, \ldots, n_i$. Let the covariates involved be $X_{ij}, j = 1, \ldots, n_i$ for individual i. The model I consider is

$$Y_{ij} = X_{ij}\beta + \epsilon_{ij}, \tag{5.7}$$

where

$$\epsilon_{ij}|\theta_i, \sigma_i \sim \mathrm{N}(\theta_i, \sigma_i^2) \qquad \theta_i \sim F$$

Note that although an individual i can experience several events, ϵ_{ij} share the same θ_i and σ_i within the same individual for the normal kernel.

The estimating process is similar to the proposed algorithms, and the only change is that the ω_{ik} is estimated as

$$\omega_{ik} = P(\theta_i \in I_{Mk}|\epsilon_i) = \frac{P_{Mk} \prod_{j=1}^{n_i} \mathrm{N}(\epsilon_{ij}|\zeta_{Mk}, \sigma_i^2)}{\sum_{l=1}^{2^M} P_{Ml} \prod_{j=1}^{n_i} \mathrm{N}(\epsilon_{ij}|\zeta_{Ml}, \sigma_i^2)}.$$

I next discuss a well known data set (Manda et al. 1995 [61]), and fit it using Komarek and Lesaffre (2007 [49])'s nonparametric AFT model, and our `RPTMM-AFT` model. Although `RPTMM-AFT` model is extremely simple to implement, it can approximate the Komarek and Lesaffre (2007 [49])'s result.

Example 15. CGD data

The data set is from Fleming and Harrington (1991 [31]) and it is a double-blinded placebo controlled randomized trial of gamma interferon (γ-IFN) in chronic granulomatous disease (CGD). CGD is a group of rare inherited disorders of the

Covariates	Placebo	Gamma interferon
Number of patients	65	63
Pattern of inheritance 0: X-linked	41 (63.1%)	45 (71.4 %)
1: autosomal recessive		
Age (in years)	14.98 (9.64)	14.29 (10.12)
Corticosteroid use(0: No; 1: Yes)	2 (3.1%)	1 (1.6%)
Prophylactic antibiotic use (0: No; 1: Yes)	55 (84.6%)	56 (88.9%)
Gender (0: Male; 1: Female)	12 (18.5%)	12 (19.1 %)
Hospital region (0: USA; 1: Europe)	22 (33.8%)	17 (27 %)

Table 5–3: CGD data: baseline characteristics according to treatment group (from Manda et al. 1995 [61]).

immune function, with recurrent pyogenic infections. It is believed that the effect of gamma interferon can help to treat CGD patients.

A total 128 patients were followed, 65 patients on placebo, and 63 patients on gamma interferon treatment. 203 infections/censoring were observed, and the number of infections for per patient range from 1 to 8. Some of the potential confounder are involved, and they are listed in Table 5–3.

I fit the data using three models: a parametric AFT model whose frailty is assumed to be normal; a Bayesian semiparametric AFT model introduced in Komarek and Lesaffre (2007 [49])(R package `bayesSurv`); and the proposed `RPTMM-AFT` model

Table 5–4 lists the posterior mean and 95% credible intervals or confidence intervals for the three methods. I find

- The posterior means of the two Bayesian semiparametric models are very close. The parametric model gives the estimates closer to zero.

Models		Parametric AFT	Komarek (2007)	`RPTMM-AFT`
Treatment	Estimate	1.10 (0.30)	1.25 (0.43)	**1.31 (0.33)**
	95% CI	(0.51,1.69)	(0.49,2.16)	**(0.69,1.96)**
Inheritance	Estimate	-0.62 (0.37)	-0.89 (0.44)	**-0.86 (0.37)**
	95% CI	(-1.35,0.10)	(-1.78,-0.04)	**(-1.54,-0.10)**
Age	Estimate	0.03 (0.02)	0.042 (0.02)	**0.041 (0.02)**
	95% CI	(-0.009,0.009)	(0.001,0.085)	**(0.007,0.08)**
Corticosteroids	Estimate	-1.57 (0.88)	-2.38 (1.22)	**-2.16 (0.97)**
	95% CI	(-3.30,0.15)	(-4.80,-0.03)	**(-3.82,-0.21)**
Prophylactic	Estimate	0.79 (0.42)	1.04 (0.48)	**1.03 (0.45)**
	95% CI	(-0.03,1.61)	(0.11,2.03)	**(0.21,1.98)**
Gender	Estimate	0.99 (0.51)	1.40 (0.72)	**1.44 (0.52)**
	95% CI	(-0.01,1.99)	(0.10,2.90)	**(0.43,2.52)**
Hospital	Estimate	0.74 (0.36)	1.11 (0.46)	**0.97 (0.45)**
	95% CI	(0.02,1.44)	(0.24,2.09)	**(0.02,1.82)**

Table 5–4: CGD data: Posterior means, 95% credible intervals of parameters for three methods: 1. parametric AFT model whose frailty is assumed to be normal; 2. a Bayesian semiparametric AFT model introduced in Komarek and Lesaffre (2007 [49])(`R` package `bayesSurv`); and 3. the proposed `RPTMM-AFT` model.

- The standard errors from the proposed `RPTMM-AFT` model are relatively larger than those from parametric model, but are smaller than those from Komarek's model.
- In parametric model, only treatment and hospital region are statistically significant. However, all the variables in the two Bayesian semiparametric models are significant.

All the covariates are significant for both of the Bayesian semiparametric models. The application of γ-IFN significantly reduces the rate of infection in CGD patients. Patients with X-linked inheritance pattern and with the presence of corticosteroids at the time of entry experienced higher rate of serious infection in CGD. Young patients and male patients also have a higher rate of infection. Patients in U.S.A hospitals seem to perform worse than those in European hospital.

I also did a "naive simulation". The errors $\epsilon_i = y_i - x_i\widehat{\beta}$ are calculated, and the $\widehat{\beta}$ are from Komarek's model. I ignore the dependence between these errors, and applied resampling techniques to produce 200 bootstrap samples. If y_i is right censored, it is still considered as censored in the bootstrap sample.

I fit the proposed `RPTMM-AFT` model to each of the bootstrap samples, and consider the $\widehat{\beta}$ as the "true" values. I record the proportion of the "true" values that are included in the credible intervals in the 200 bootstrap samples. For the seven covariates, these proportions are: Treatment (92%), Inheritance (96.5%), Age (96%), Corticosteroids (93.5%), Prophylactic (98%), Gender (96%), and Hospital Region

(93%).

Another "naive simulation" is to consider the $\widehat{\beta}$ from `RPTMM-AFT` to be the true values, simulate a $\theta_i^* \sim \widehat{F}$, and then simulate a $\epsilon_i^* \sim \mathtt{N}(\theta_i^*, \sigma_i)$. The simulated $Y^* = X\widehat{\beta} + \epsilon^*$. I also repeat the simulation 200 times to produce 200 simulated samples. The proportions to cover the "true" values for the seven covariates are: Treatment (92.5%), Inheritance (94%), Age (95%), Corticosteroids (94%), Prophylactic (93%), Gender (92.5%), and Hospital Region (94.5%). The proportions from the two naive simulations are around 95%, which indicate that the PRTMM-AFT model gives a reasonable fit.

The proposed `RPTMM-AFT` model can provide approximately similar results with Komarek and Lesaffre (2007 [49])'s nonparametric AFT model. In their models, the error terms are assumed to arise from a infinite mixture of normal distribution, and the reversible jump MCMC is used to estimate the number of components and parameters. The proposed `RPTMM-AFT` model is very easy to be implemented and understood.

5.5 Conclusions

In this section, I developed a fully Bayesian semiparametric accelerated failure time model, based on a recursive Polya tree mixture model. Our models are easy to implement and to be understood, compared with other semiparametric AFT models. When the baseline survival arise from some unexpected distribution (like a mixture distribution), our `RPTMM-AFT` model can still give accurate estimates of regression

coefficients.

The `RPTMM-AFT` model can also handle interval-censored data, and provide survival function estimates. In frequentist statistics, Turnbull (1976 [78])'s EM algorithm seems to be the only one to provide a nonparametric fit (see Doss and Huffer 2003 [23]). In the nonparametric Bayesian world, Doss and Huffer (2003 [23]) and Hanson and Johnson (2004 [38]) proposed the models based on mixture of Dirichlet process. The simple `RPTMM-AFT` model can approximate their approaches very well, and can easily implement fully Bayesian inference.

I also extend the proposed model to handle recurrent data, as an alternative to Komarek and Lesaffre (2007 [49])'s work. Both approaches can obtain similar estimates and provide fully Bayesian inference. Our approaches can avoid the requirement to specify the large number of prior parameters in advance.

CHAPTER 6
Conclusions and future work

In this thesis, I propose a recursive Polya tree mixture model (`RPTMM`). The motivation of our work is to provide a simple and computationally efficient Bayesian nonparametric approach. I model a random sample by assuming $Y_i|\theta_i \sim h(Y_i|\theta_i)$ independently, and the parameters are supposed to arise from some random distribution F. I model F using a Polya tree prior and use the empirical distribution of some pseudo parameters to be the base distribution. The proposed approach enjoys a number of advantages: (1) sampling from F is very straightforward, (2) it does not require any analytical computation, and (3) very fast.

I discuss the application of `RPTMM` to the baseball players' data in Brown (2008 [10]), and show that the proposed approach can compete with other Bayesian nonparametric and empirical Bayes methods. I also revisited the thumbtacks data in Liu (1996 [57]), and show that the "accurate" plots can also be obtained from `RPTMM`. I also discuss the density estimation problem, including bivariate density estimation. The proposed recursive Polya tree mixture model is shown to perform better than mixture of Polya tree model. In addition, I also show that the proposed approach can approximate a complicated conditional Dirichlet process model in meta-analysis. Since it is easy to sample from F, the mean or variance of F can obtained in a natural way.

In this thesis, I developed a semiparametric AFT model by assuming the error term arises from a random distribution F, and show that the `RPTMM-AFT` model can provide a full Bayesian inference. I revisit the simulated data proposed in Hanson and Johnson (2002 [37]), the data on a small cell lung cancer, and on the cosmetic effects of cancer therapy. The proposed approach is demonstrated to give similar performance to some more complicated approaches. I also extend the semiparametric AFT model to handle recurrent data.

In regression problems, I only consider the case where the response is a continuous variable. A future research problem is to extend the semiparametric regression model to the generalized linear model, and this may require more complicated computing algorithms. I investigate the performance of recursive Polya tree mixture models under a number of applications. However, many other situations could also be investigated, for example, large-scale multiple testing problems, such as gene expression data analysis.

CHAPTER 7
Epilogue

In this chapter, I discuss some questions raised by the examiners of my thesis. I discuss four issues related to the proposed `recursive Polya tree mixture model`.

1. Examples where we need this extra flexibility that the proposed model gives and other approaches lack.
2. Estimating the variance of F in meta-analysis
3. Assess whether the proposed `RPTMM` would produce too short credible intervals when sample size is small.
4. Compare the proposed `RPTMM` with the parametric AFT model.

7.1 Extra flexibility

In this section, I examine whether the proposed `RPTMM` can provide extra flexibility that other methods lack.

7.1.1 Comparison with parametric models

In my thesis, some examples have demonstrated the advantages of `RPTMM` by relaxing the parametric assumption concerning F.

One compelling example is the thumbtack data set discussed in Section 4. In this example

$$X_i \sim \text{Binomial}(Y_i, \theta_i),$$

and

$$\theta_1, \ldots, \theta_N \sim F.$$

By assuming F to arise from some random distribution, the `recursive Polya tree mixture model` captured the surprising "bimodal" structure of the data. As Liu (1996 [57]) mentioned, this surprising "bimodal" structure can not be captured through a parametric model.

Another example is the simulated data set (Example 12) discussed in Section 5.3, where the error term is assumed to arise from a mixture of log normal distributions. The least squares method can not accurately estimate the parameters in the simulated data set (Example 12), since the hidden assumption in the least squares method is that: the error term is a simple normal distribution.

By relaxing the parametric assumption of the error term, the proposed `recursive Polya tree mixture model` can successfully estimate the parameters, and also capture the mixture of log normal distributions.

The two examples demonstrate the extra flexibility and advantages of the `recursive Polya tree mixture model` compared with other parametric models.

7.1.2 Comparison with existing Bayesian nonparametric models

The `RPTMM` exhibits the same level of flexibility as many Bayesian nonparametric models. However, our proposed `recursive Polya tree mixture model` still enjoys a lot of advantages.

Case 1: General objectives in the Bayesian hierarchical model

In the Bayesian hierarchical model, suppose we observe a random sample $Y_1, \ldots, Y_N$,

$$Y_i|\theta_i \sim h(Y_i|\theta_i) \text{ independently,}$$

and $\theta_1, \ldots, \theta_N$ are assumed to arise from some random distribution F

$$\theta_1, \ldots, \theta_N \sim F \text{ i.i.d}$$

Generally speaking, there are two objectives in a Bayesian hierarchical model

- to estimate the individual $\theta_1, \ldots, \theta_N$
- to perform inference on F

In Chapter 4, I discuss a baseball players' dataset, which is related to estimating $\theta_1, \ldots, \theta_N$; and I also discuss a thumbtack data set, which is related to performing inference on F.

From the performance of different Bayesian methods to these two data sets, I compare the proposed `RPTMM` with Dirichlet process mixture model and empirical Bayes methods. Several points that show the advantages of the proposed `RPTMM` are emphasized here:

- The Dirichlet process mixture model (with the help of sequential imputation) successfully captures the bimodal structure in the thumbtack data set, but gives a poor fit to the baseball players' data set (see Table 4–3),
- The empirical Bayes approach is among one of the top models to estimate the batting averages in baseball players' data set, but it is not clear how to apply the empirical Bayes method to perform inference on F

- The proposed `recursive Polya tree mixture model` is not only one of the top models to estimate the batting averages in the baseball players' data, but also successfully captures the bimodal structure in the thumbtack data set.

The advantage of the proposed `recursive Polya tree mixture model` is that: it can not only provide a reasonable fit to estimate $\theta_1, \ldots, \theta_N$, but also perform inference on F.

Case 2: Density estimation

In Section 4, the density estimation problem is also discussed. Different approaches are compared based on some criteria (`ISE`, `WISE` and `LPML`).

- The proposed `recursive Polya tree mixture model` performs better than mixture of Polya trees to estimate the density.
- Based on the criteria (`ISE`, `WISE` and `LPML`), the `recursive Polya tree mixture model` provides similar performance to the Dirichlet process mixture model.

Because of the discrete property of the Dirichlet process, the Dirichlet process mixture model would lead to a mixture of any parametric distributions (the number of components is not fixed) to estimate the density. For example, in the Ozone data set (in Section 4.3), the density plot in the Dirichlet process mixture model shows a two-cluster structure. However, the proposed `recursive Polya tree mixture model` would lead to a "smoothed histogram" and capture more information.

Another advantage of the proposed approach is that: it is straightforward to obtain the uncertainty of the estimates of density. For example, the density of a point x_i,

$$h(x_i) \sim \int h(x_i|\theta) dF(\theta),$$

can be estimated through the Monte Carlo method if the θ values can be sampled from F. The proposed `recursive Polya tree mixture model` is a recursive algorithm, and includes many iterative steps $1, 2, \ldots, \Upsilon$. With the help of Monte Carlo method, the pseudo data at each step can be plugged to the formula above to obtain the estimates

$$h(x_i)^{(1)}, \ldots, h(x_i)^{(\Upsilon)}.$$

Thus the 95% credible intervals can be easily obtained as the 0.025 and 0.975 quantiles of these $h(x_i)^{(1)}, \ldots, h(x_i)^{(\Upsilon)}$.

In the Dirichlet process mixture model, F is marginalized out. At each step the drawn θ's can not be considered as a sample from F. Thus, how to estimate uncertainty of density still needs further investigation in the Dirichlet process mixture model.

Conclusion

Compared with the parametric model, the `RPTMM` can capture more flexibility from data. Compared with the existing Bayesian nonparametric models, the proposed `RPTMM` can enjoy some computational advantages described above.

7.2 Estimating the variance of F in meta-analysis

In this section, I discuss a simulated data set to demonstrate that the proposed `RPTMM` performs well but the existing Bayesian nonparametric models do not give a reasonable fit.

In Section 4.6, I discussed the meta-analysis problem. The log odds ratios $Y_1, \ldots, Y_N$ are assumed to be from normal distributions

$$Y_i|\theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2),$$

and

$$\theta_1, \ldots, \theta_N \sim F \text{ i.i.d.}$$

One of the objectives in meta-analysis is to estimate the $\texttt{Var}(F)$. As discussed in Section 4.6, if the $\theta_1, \ldots, \theta_N$ are assumed to arise from a nonparametric random distribution F, the estimation of $\texttt{Var}(F)$ is not straightforward using existing Bayesian nonparametric models.

For example, in the mixture of Polya trees model, the random distribution F has a base distribution $\texttt{N}(\mu, \tau^2)$. The τ^2 in the base distribution does not equal to the $\texttt{Var}(F)$, and thus does not have a clear interpretation.

In Table (4–7) and (4–14), we observe big difference in estimating the $\texttt{Var}(F)$, especially when the precision values α for existing Bayesian nonparametric models are small. Note that when the precision values α for existing Bayesian nonparametric models are large, both the proposed `recursive Polya tree mixture model` and the existing Bayesian nonparametric models provide similar estimates.

An important question is: which one is more accurate to estimate $\texttt{Var}(F)$, the existing Bayesian nonparametric models (with small precision values) or the proposed `recursive Polya tree mixture model`?

A simulated meta analysis data set

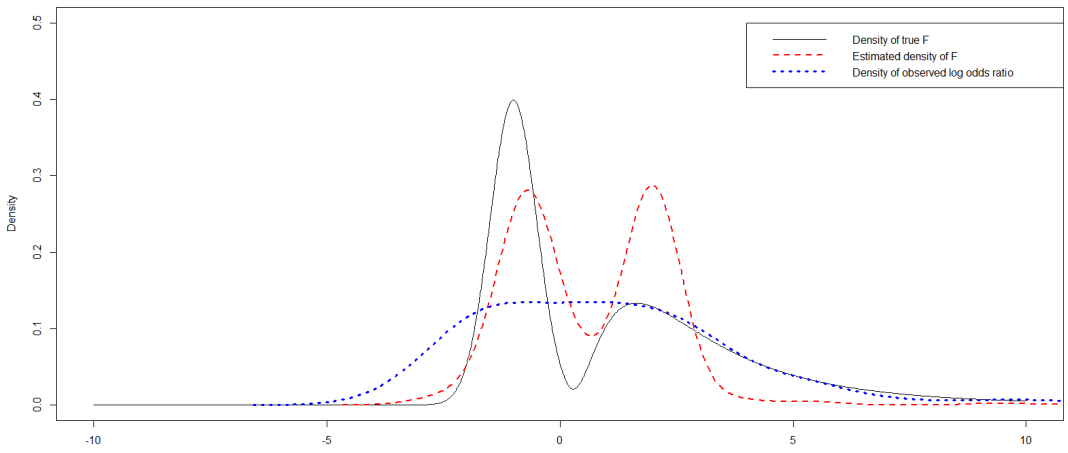

Figure 7–1: Simulated data set: The plot show the true density of $F \sim 0.5\text{Normal}(-1, 0.5^2) + 0.5\text{log-Normal}(1, \sqrt{0.5}^2)$, the estimated density of F from `RPTMM` (red line) and the density of observed log odds ratios (blue line)

To answer this, a new simulated data set was designed, where 100 θ values are generated from a mixture distribution

$$\theta_1, \ldots, \theta_{100} \sim 0.5\text{Normal}(-1, 0.5^2) + 0.5\text{log-Normal}(1, \sqrt{0.5}^2),$$

and log odds ratios $Y_1, \ldots, Y_{100}$ is assumed to arise from

$$Y_i \sim \text{N}(\theta_i, 1)\ i = 1, \ldots, 100.$$

The distribution of F is from a mixture distribution, and the true density plot of F is shown in Figure 7–1. Suppose the distribution is unknown to us, and only the simulated log odds ratios are available. For one simulated data set, the estimated

density plot of F from the proposed `RPTMM` is shown in Figure 7–1, and the two-cluster structure is successfully captured. The density plot of the observed log odds ratios (blue line) does not show any special structure.

In the mixture of Polya trees, the precision value α is set to be 1, and the hyperparameters are set according to the suggestions in the computing reference manual of the R package `DPpackage` (`http://cran.r-project.org/web/packages/DPpackage/DPpackage.pdf`). Compared with the mixture of Polya trees model, the proposed `RPTMM` is more "automatic", and does not require to set any hyperparameters.

The objective is to estimate the $\sqrt{\texttt{Var}(F)}$ (the true value is 3.0), and I simulate 100 data sets (each data set has 100 observations). I consider two methods to examine which method (`RPTMM` vs. mixture of Polya trees) give a more reasonable estimate of $\sqrt{\texttt{Var}(F)}$.

In the first method, for each of the simulate data set, the 95% credible intervals of both `RPTMM` and mixture of Polya tree are calculated, and check whether the true value 3.0 is included in the 95% credible intervals.

For an adequate model fit, we expect that around 95 out of 100 credible intervals would include the true value. For the proposed `RPTMM`, 93 credible intervals include the true value; however for the mixture of Polya trees, only 65 credible intervals include the true value. Thus, the mixture of Polya tree seems to provide incorrect inferences in this simulated case.

In the second method, I examine the average error. Suppose we define it as

$$\frac{1}{100}\sum_{i=1}^{100}(\sqrt{\widehat{\mathrm{Var}}(F_{(i)})}-3)^2,$$

where $\sqrt{\widehat{\texttt{Var}}(F_{(i)})}$ denotes the estimates from the `RPTMM` for simulated data set i. For mixture of Polya tree, the average error is 0.51, but for the `recursive Polya tree mixture model`, the average error is only 0.21. It is clear that the mixture of Polya trees fails to provide a reasonable estimate for the $\texttt{Var}(F)$.

The main reason for the failure of mixture of Polya tree is that it can only estimate the variance of base distribution, not the variance of F. However, the proposed `recursive Polya tree mixture model` directly estimates the variance of F.

In this simulated data set, I show that the `recursive Polya tree mixture model` is not only a computationally efficient method, but also gives reasonable answer where the existing Bayesian nonparametric models fail. In addition, to estimate the variance of F in meta analysis, the proposed `RPTMM` is also preferable.

7.3 Is the proposed `RPTMM` would produce too short credible intervals when sample size is small?

In this section, I discuss some problems related to the empirical Bayes idea, and discuss the credible intervals problem. Three topics are discussed in this section:

- I introduce the parametric empirical Bayes method, and discuss why sometimes it would exhibit the "over-shrinkage" problem,
- I explain why parametric empirical Bayes method produces a shorter confidence interval when the sample size is small,

- I discuss whether the `RPTMM` would provide reasonable credible intervals.

7.3.1 Empirical Bayes method in meta-analysis

The details of the empirical Bayes method to meta-analysis can be found from many standard textbooks or lecture notes. The introduction in this section is from a lecture note by Dr. David Draper, and the link of the lecture note is

`http://users.soe.ucsc.edu/~draper/San-Francisco-2011-notes-part-2.pdf`

In the meta-analysis problem, the observed log odds ratio $y_1, \ldots, y_N$ are from a normal distribution

$$y_i|\theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2),$$

and $\theta_1, \ldots, \theta_N$ are assumed to arise from a normal distribution

$$\theta_1, \ldots, \theta_N \sim \mathrm{N}(\mu, \tau^2).$$

The conditional distribution of θ_i, given data and parameters μ and τ^2 is

$$\theta_i|y_i, \mu, \tau^2 \sim \mathrm{N}(\theta_i^*, \sigma_i^2(1 - B_i)),$$

where

$$B_i = \frac{\sigma_i^2}{\sigma_i^2 + \tau^2},$$

and

$$\theta_i^* = (1 - B_i)y_i + B_i\mu.$$

Here the conditional mean of the effect for study i θ_i^* is a weighted average of the sample mean for that study (y_i) and the overall mean (μ).

To estimate the effect for study i, all the studies are used, which reflects the "borrowing information from others" in the estimation process. The B_i is the shrinkage effect.

Now the problem is how to estimate μ and τ. The likelihood function is

$$l(\mu, \tau^2|y) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{\sigma_i^2 + \tau^2}} \exp\{-\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{\sigma_i^2 + \tau^2}\}$$

The maximum likelihood estimate $\hat{\mu}$ and $\hat{\tau}^2$ is

$$\hat{\mu} = \frac{\sum_{i=1}^{N} W_i y_i}{\sum_{i=1}^{N} W_i},$$

and

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{N} W_i[(y_i - \hat{\mu})^2 - \sigma_i^2]}{\sum_{i=1}^{N} W_i},$$

where

$$W_i = \frac{1}{\sigma_i^2 + \hat{\tau}^2}$$

Thus the maximum likelihood estimate of θ_i is

$$\hat{\theta}_i = (1 - \hat{B}_i)y_i + \hat{B}_i\hat{\mu}, \tag{7.1}$$

where

$$\hat{B}_i = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + \hat{\tau}^2} = \frac{1}{1 + \hat{\tau}^2/\hat{\sigma}_i^2}, \tag{7.2}$$

Morris (1983 [65]) discussed the detail of the "over-shrinkage" problem. For small N, the uncertainty in τ^2 can not be accounted for fully, and therefore the $\hat{\tau}^2$ is underestimated. Morris (1983 [65]) pointed out that B is a convex nonlinear function of τ^2, and so substitution of a nearly unbiased estimator $\hat{\tau}^2$ of τ^2 into B

would produce an estimate of B that is biased to be too large. This conclusion can be proved by Jensen's inequality, and the details can be found in Morris (1983 [65]).

Note that B_i is the shrinkage factor, and measures the "degree of shrinkage" for the estimation of θ_i. Larger values of B_i indicate more shrinkage, and vice versa. Thus, if $\hat{B}_i$ is overestimated, the maximum likelihood estimate of θ_i in (7.1) would "over-shrink" to the overall mean μ.

7.3.2 Why empirical Bayes would produce confidence intervals that are too short?

In this section, I discuss why the parametric empirical Bayes method would produce shorter confidence interval. The detail can be found in Morris (1983, [65]) and Carlin and Louis (2009, [14]).

Consider the Bayesian hierarchical model,

$$Y_i|\theta_i \sim h(Y_i|\theta_i) \texttt{ independently},$$

and

$$\theta_1, \ldots, \theta_N \sim F(\lambda).$$

The parametric empirical Bayes method would use the data to estimate the hyperparameter λ. However, Morris (1983, [65]) pointed out a problem of the "empirical Bayes" idea, that is, they would produce shorter confidence interval, especially when the sample size is small.

To see this, from elementary statistics, we have the iterated variance formula,

$$\mathrm{Var}(\theta_i|\lambda) = \mathrm{E}_{\lambda|Y}(\mathrm{Var}(\theta_i|Y_i, \lambda) + \mathrm{Var}_{\lambda|Y}(\mathrm{E}(\theta_i|y_i, \lambda)) \tag{7.3}$$

Methods	ACI(EB)	ACI (RPTMM)	Average error (EB))	Average error (RPTMM)
$N = 5$	71.0%	95.4%	0.43	0.41
$N = 10$	85.1%	97.0%	0.40	0.38
$N = 20$	92.1%	95.8%	0.35	0.40
$N = 30$	92.4%	95.2%	0.35	0.41
$N = 50$	94.3%	93.1%	0.34	0.37
$N = 100$	94.7%	93.2%	0.33	0.35

Table 7–1: Simulated case 7.3.1. The average coverage rates and average errors of the empirical Bayes methods and RPTMM are summarized.

An empirical Bayes procedure would estimate λ using the data, and ignore the posterior uncertainty about λ (the second term of the equation (7.3)). As a result, only the first term of equation (7.3) is considered (see Carlin and Louis, 2009, [14]).

Thus, the confidence interval of empirical Bayes would be too short, and especially when sample size is small (Morris, 1983 [65]). Morris (1983, [65]) introduced a method to produce reasonable confidence intervals for empirical Bayes method.

7.3.3 Will the proposed RPTMM produce too short confidence interval?

In this section, I examine whether the proposed RPTMM would produce too short credible intervals by examining two simple simulated cases.

Simulated case 7.3.1

In this simulated case, θ is generated from a normal distribution with mean 0 and variance 1,

$$\theta_1, \ldots, \theta_N \sim \text{Normal}(0, 1),$$

and the observed $Y_i (i = 1, \ldots, N)$ is also generated from a normal distribution with mean θ_i and variance 0.5

$$Y_i \sim \text{Normal}(0, 0.5).$$

The objective of generating θ from a normal distribution is to guarantee that the parametric empirical Bayes method can provide correct point estimation. Thus, if too short confidence intervals are produced by empirical Bayes methods, we could conclude that the problem is caused by underestimating $\mathrm{Var}(\theta_i|\lambda)$, not caused by point estimation.

Different sample sizes $N = 5, 10, 20, 30, 50, 100$ are considered, and we repeated sampling 100 data sets. To examine whether the confidence/credible intervals are reasonable, I summarize the number of data sets where the true θ values are included in the 95% confidence/credible intervals. The average coverage rate are defined as

$$\mathrm{ACR} = \frac{1}{100}\frac{1}{N}\sum_{i=1}^{100}\sum_{j=1}^{N} I(\theta_i \in 95\%\mathrm{CI}).$$

If empirical Bayes or `RPTMM` gives reasonable fit, the average coverage rate ACR is expected to be 95%. Table 7–1 summarize the average coverage rates and average errors of the empirical Bayes methods.

Note as expected, the average errors of parametric empirical Bayes method are a little bit smaller than `RPTMM`. Thus, from the point estimation perspective, the performance of empirical Bayes method is good.

However, just as Morris (1983, [65]) pointed out, when the sample size is small ($N = 5$ and $N = 10$), the average coverage rates of empirical Bayes method are only 71% and 85.1%. The values are significantly lower than the expected 95%, and indicates that when the sample size is small, the parametric empirical Bayes method produces too short confidence intervals. However, when the sample size is large

Methods	coverage rate	Average Error
Parametric AFT (Weibull)	97%	0.007
Parametric AFT (Exponential)	100%	0.012
Parametric AFT (Lognormal)	93.5%	0.015
Parametric AFT (Loglogistic)	88%	0.029
Mixture of Polya tree	98.5%	< 0.001
`RPTMM`	96%	< 0.001

Table 7–2: Simulated case 7.3.2. The coverage rate and average error for the β_1, and the true value is 1. $T_i = \exp(x_i\beta)V_i$, where $x_{i,1} \sim 0.5\delta_0 + 0.5\delta_1$, $x_{i,2} \sim \mathtt{N}(0,1)$, $\beta = (1,-1)$, and $V_1, \ldots, V_{100} \sim 0.5\,\mathrm{N}(1, 0.15^2) + 0.5\,\mathrm{N}(3, 0.15^2)$.

Methods	coverage rate	Average Error
Parametric AFT (Weibull)	100%	0.002
Parametric AFT (Exponential)	100%	0.003
Parametric AFT (Lognormal)	97%	0.003
Parametric AFT (Loglogistic)	94%	0.006
Mixture of Polya tree	97%	< 0.001
`RPTMM`	93%	< 0.001

Table 7–3: Simulated case 7.3.2. The coverage rate and average error for the β_2, and the true value is -1. $T_i = \exp(x_i\beta)V_i$, where $x_{i,1} \sim 0.5\delta_0 + 0.5\delta_1$, $x_{i,2} \sim \mathtt{N}(0,1)$, $\beta = (1,-1)$, and $V_1, \ldots, V_{100} \sim 0.5\,\mathrm{N}(1, 0.15^2) + 0.5\,\mathrm{N}(3, 0.15^2)$.

($N = 50$ and $N = 100$), the parametric empirical Bayes method produces reasonable confidence intervals.

As a comparison, the average coverage rates of the proposed `RPTMM` are close to 95%. Thus, the problem of the "too-short" confidence interval in empirical Bayes methods does not afflict the proposed `RPTMM`.

Simulated case 7.3.2

In this simulated case, I examine whether the `RPTMM-AFT` model can give reasonable credible intervals in a regression setting. Some 100 data points $V_1, \ldots, V_{100}$

from a mixture of normal distributions (from Hanson and Johnson 2002 [37]) were simulated:

$$V_1, \ldots, V_{100} \sim 0.5\,\mathrm{N}(1, 0.15^2) + 0.5\,\mathrm{N}(3, 0.15^2)$$

Two covariates were generated by taking $x_{i,1} \sim 0.5\delta_0 + 0.5\delta_1$, where δ_a is a point mass at a; the second covariate $x_{i,2} \sim \mathtt{N}(0, 1)$. The true vector of coefficients was set at $\beta = (1, -1)$ and the survival times are calculated as

$$T_i = \exp(x_i\beta)V_i.$$

200 data sets are simulated. The proposed `RPTMM`, and the parametric AFT models are fit to the 200 data sets. The 95% credible intervals are produced for each of the 200 data sets. If the true coefficients are included in the 95% data sets, the credible/confidence interval produced by that method is determined to be reasonable.

The coverage rate and average errors of the two coefficients for different methods are listed in Table 7–2 and 7–3.

First, the parametric AFT with different parametric distributions can not give accurate point estimation. Both of the average errors of the two Bayesian nonparametric methods are smaller than 0.001; while the average errors of the parametric AFT are much larger.

Second, by examining the coverage rates of each method, we conclude that

- the parametric AFT models with Weibull and exponential distributions produce too wide confidence intervals for the two parameters (note the coverage rates are close to 100%),

- the 95% confidence interval of β_1 in loglogistic AFT model is too short (only 88%),
- the 95% credible intervals of mixture of Polya trees seem to be a little bit wider (98.5% and 97%),
- both the lognormal AFT model and the proposed `RPTMM-AFT` give confidence/credible intervals close to 95%.

A fundamental difficulty in parametric AFT model is: how to choose the appropriate distribution? In addition, even the lognormal distribution is chosen (produces reasonable confidence interval), it's performance in point estimation is poor (note the average errors are 0.015 and 0.003).

The proposed `RPTMM` not only performs well in point estimation, but also gives reasonable credible intervals in this regression setting.

7.4 Compare the proposed `RPTMM` with the parametric AFT model

In Table 5–4 (Example 15), the regression coefficients in the parametric AFT model are found to be shrunk towards the null. The objective of this section is to discuss the mis-specification effects in the accelerated failure time model.

Robinson and Jewell (1991, [74]) discussed the problems with covariates adjustment in logistic regression models, and they summarized two conclusions

- In linear regression, omitting some covariates would not bias effect estimates, but decrease the precision of effect estimates,
- In logistic regression, omitting some covariates does bias the effect estimates (toward zero).

In Example 15, when the frailty term is assumed to arise from a parametric distribution, the regression coefficients are also driven toward zero. Note although in this example the regression coefficients are diluted, it does not indicate that in every data set, the parametric assumption of frailty would drive regression coefficients toward zero. It is an interesting research topic in the future to examine whether the parametric assumption of frailty would play some systematic role on coefficient estimates.

A simulated data was designed to examine: if heterogeneity exists in recurrent data, whether the parametric AFT model can obtain accurate coefficient estimates?

There are 75 patients, and each of them has four measurements. Two covariates are designed

$$x \sim \mathrm{N}(2, 0.5^2)$$

and

$$z \sim 0.5\delta_1 + 0.5\delta_2,$$

where δ_a is the point mass at a.

The response is designed to be (for $i = 1, \ldots, 75; j = 1, 2, 3, 4$)

$$y_{ij} = 2X_{ij} + 5Z_{ij} + b_i + \epsilon_{ij},$$

where

$$\epsilon_{ij} \sim \mathrm{N}(0, 1),$$

and b_i is assumed to arise from a mixture of log-normal distributions

$$b_i \sim \frac{35}{75}\text{log-Normal}(1, 0.1) + \frac{40}{75}\text{log-normal}(3, 0.1).$$

Suppose the response variables and two covariates are known to us, and the objective is to estimate the coefficients. The true coefficients are 2 and 5. I simulate 100 data sets and apply the parametric AFT model and the proposed `recursive Polya tree mixture model` to these 100 simulated data sets.

The average squared errors

$$\frac{1}{100}\sum_{i=1}^{100}(\hat{\beta}-\beta)^2$$

for the two parameters are used to compare the performance of the parametric AFT model and the `RPTMM`.

For the first covariate x, the average squared error for parametric AFT model is 0.81, while the value for `RPTMM` is only 0.02; for the second covariate z, the average squared error for parametric AFT is 0.79, while the value for `RPTMM` is also 0.02.

Note in this simulated, the parametric AFT model fails to estimate the regression coefficients accurately. When heterogeneity exists, the proposed `AFT` shows its advantage to accurately estimate the regression coefficients.

7.5 Summary

In this chapter, I discuss some questions raised by the examiners of my thesis.

In Section 7.1, the proposed `RPTMM` is shown to capture more flexibility from data compared with the parametric models. Compared with existing Bayesian nonparametric models, the proposed `RPTMM` can also enjoy some computational advantages.

In Section 7.2, a simulated data set is discussed to show that the `RPTMM` works but the existing Bayesian nonparametric models do not give reasonable fit.

In Section 7.3, I explain why the parametric empirical Bayes would produce too short confidence intervals, and show that the `RPTMM` does not have this problem.

In Section 7.4, a simulated data set in recurrent data is discussed. The parametric AFT fails to give a reasonable fit, whereas the proposed `RPTMM-AFT` can accurately estimate the coefficients.

References

[1] D.J. Aldous. *Exchangeability and related topics*. Springer, New York, 1983.

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

[3] A. Barron, M.J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

[4] G.F. Beadle, S. Come, I.C. Henderson, B. Silver, S. Hellman, and J.R. Harris. The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer. *International Journal of Radiation Oncology, Biology and Physics*, 10:2131–2137, 1984.

[5] L. Beckett and P. Diaconis. Spectrall analysis for discrete longitudinal data. *Advanced Applied Mathematics*, 103:107–128, 1994.

[6] C.M. Bishop. *Pattern Recognition and Machine Learning, New York*. Springer, 2006.

[7] D. Blackwell and J.B. MacQueen. Ferguson distributions via Polya Urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[8] D.M. Blei and M.I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[9] A.J. Branscum and T.E. Hanson. Bayesian nonparametric meta-analysis using Polya tree mixture model. *Biometrics*, 64(3):825–833, 2008.

[10] L.D. Brown. In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Annals of Applied Statistics*, 2(1):113–152, 2008.

[11] D. Burr. bspmma: an r package for bayesian semi-parametric models for meta-analysis. *Journal of Statistical Software*, 50(4), 2012.

[12] D. Burr and H. Doss. A Bayesian semiparametric model for random-effects meta-analysis. *Journal of American Statisticsl Association*, 100:242–251, 2005.

[13] C.A. Bush and S.N. MacEachern. A semiparametric Bayesian model for randomised block desigens. *Biometrika*, 83(2):275–285, 1996.

[14] Bradley Carlin and Thomas Louis. *Bayesian Methods for Data Analysis*. Chapman and Hall/CRC, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, 2009.

[15] C.H. Charles. A class of distribution function processes which have derivatives. *Journal of Applied Probability*, 1(2):pp. 385–388, 1964.

[16] R. Christensen and W. Johnson. Modeling accelerated failure time with dirichlet process. *Biometrika*, 75(4):693–704, 1988.

[17] R.J. Cook and J.F. Lawless. *The Statistical Analysis of Recurrent Events*. Springer, New York, 2006.

[18] D.R. Cox. Regression models and life-tables. *Journal of Royal Statistical Society. Series B*, 34(2):187–220, 1972.

[19] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

[20] D. Dey, P. Muller, and D. Sinha. *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer, New york, 1998.

[21] J. Diebolt and C.P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375, 1994.

[22] H. Doss. Bayesian Nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*, 22(4):1763–1786, 1994.

[23] H. Doss and F. Huffer. Monte Carlo methods for Bayesian analysis for survival data using mixtures of Dirichlet process priors. *Journal of Computational and Graphical Statistics*, 12(2):282–307, 2003.

[24] B. Efron. Bootstrap methods: Another look at jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[25] B. Efron and C. Morris. Data analysis using Stein's estimator and its generalizations. *Journal of American Statistical Association*, 70(350):311–319, 1975.

[26] M.D. Escobar. Estimating normal means with a Dirichlet process prior. *Journal of Americal Statistical Association*, 85(425):268–277, 1994.

[27] M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of Americian Statistical Association*, 90(430):577–588, 1995.

[28] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[29] T.S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4):615–629, 1974.

[30] D.M. Finkelstein and R.A. Wolfe. A semiparametric model for regression analysis of interval censored failure time data. *Biometrics*, 41:933–945, 1985.

[31] T.R. Fleming and D.P. Harrington. *Counting Processes and Survival Analysis*. Wiley, New York, 1991.

[32] A. Gelfand and A. Kottas. A computation approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11(2):289–305, 2002.

[33] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2003.

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

[35] P.J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Journal of Royal Statistical Society, Series B*, 59:731–758, 1997.

[36] T.E. Hanson. Inference for mixtures of finite Polya Tree models. *Journal of American Statistical Association*, 101(476):1548–1565, 2006.

[37] T.E. Hanson and W. Johnson. Modeling regression error with a mixture of Polya Trees. *Journal of American Statistical Association*, 87(460):1020–1033, 2002.

[38] T.E. Hanson and W. Johnson. A Bayesian semiparametric AFT model for interval-censored data. *Journal of Computational and Graphical Statistics*, 13(2):341–361, 2004.

[39] J. Higgins, S.G. Thompson, and D.J. Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of Royal Society of Statistics Series A*, 172(1):137–159, 2009.

[40] N.L. Hjort, C.C. Holmes, P. Muller, and S.G. Walker. *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, 2010.

[41] C.C. Holmes, F. Caron, J.E. Griffin, and D.A. Stephens. Two sample Bayesian nonparametric hypothesis testing, 2009.

[42] J. Huang, S. Ma, and H. Xie. Least absolute deviations estimation for the accelerated failure time model. *Statistica Sinica*, 17:1533–1548, 2007.

[43] H. Ishwaran and L.F. James. Gibbs sampling methods for Stick-Breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[44] H. Ishwaran and L.F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13:1211–1235, 2003.

[45] A. Jara, T.E. Hanson, F.A. Quintana, P. Muller, and G. Rosner. DPpackage: Bayesian semi and nonparametric modeling in R. *Journal of Statistical Software*, 40(5):1–30, 2011.

[46] J.D. Kalbfleisch and R.L. Prentice. *The statistical analysis of failure time data*. Wiley, New Jersey, 1980.

[47] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of American Statistican Association*, 53:457–481, 1958.

[48] J.P. Klein and M.L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 1997.

[49] A. Komarek and E. Lesaffre. Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution. *Statistica Sinica*, 17:549–569, 2007.

[50] A. Kottas and A.E. Gelfand. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96(456):1458–1468, 2001.

[51] M. Krnjajica, A. Kottas, and D. Draper. Parametric and nonparametric bayesian model specification: A case study involving models for count data. *Computational Statistics and Data Analysis*, 52(4):2110–2128, 2008.

[52] L. Kuo and B. Mallick. Bayesian semiparametric inference for the accelerated failure time model. *Canadian Journal of Statistics*, 25(4):457–472, 1997.

[53] P.W. Laud and J.G. Ibrahim. Predictive model selection. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):247–262, 1995.

[54] M. Lavine. Some aspects of Polya Tree distributions for statistical modelling. *The Annals of Statistics*, 20(3):1222–1235, 1992.

[55] M. Lavine. More aspects of Polya Tree distributions for statistical modelling. *Annals of Statistics*, 22:1161–1176, 1994.

[56] D.Y. Lin. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, 13:2233–2247, 1994.

[57] J.S. Liu. Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics*, 24(3):911–930, 1996.

[58] J.S. Liu. *Monte Carlo strategies in scientific computing*. Springer, New York, 1999.

[59] S.N. MacEachern, M. Clyde, and J.S. Liu. Sequential importance sampling for nonparametric Bayes models: The next generation. *The Canadian Journal of Statistics*, 27(2):251–267, 1999.

[60] S.N. MacEachern and P. Muller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.

[61] S.O.M. Manda and R. Meyer. Bayesian inference for recurrent events data using time-dependent frailty. *Statistics in Medicine*, 24:1263–1274, 1995.

[62] R. Martin and S.T. Tokdar. Semiparametric inference in mixture models with predictive recursion marginal likelihood. *Biometrika*, 98(3):567–582, 2011.

[63] J.D. McAuliffe, D.M. Blei, and M.I. Jordan. Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing*, 16(1):5–14, 2006.

[64] C.A. McGilchrist and C.W. Aisbett. Regression with frality in survival analysis. *Biometrics*, 47:221–225, 1991.

[65] C. Morris. Parametric empirical bayes confidence intervals. *Journal of American Statistical Association*, 78(381):47–55, 1983.

[66] P. Muliere and S.G. Walker. A Bayesian non-parametric approach to survival analysis using Polya trees. *Scandinavian Journal of Statistics*, 24(3):331–340, 1997.

[67] O. Muralidharan. An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, 4(1):422–438, 2010.

[68] R.M. Neal. Markov Chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[69] M.A. Newton. On a nonparametric recursive estimator of the mixing distribution. *The Indian Journal of Statistics, Series A*, 64(2):306–322, 2002.

[70] M.A. Newton and Y. Zhang. A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, 86(1):15–26, 1999.

[71] L.E. Nieto-Barajas and S.G. Walker. A Bayesian semi-parametric bivariate failure time model. *Computational Statistics and Data Analysis*, 51:6102–6113, 2007.

[72] J. Pitman. Exchangeable an partially exchangeable random partitions. *Probability Theory and related Fields*, 102(2):145–158, 1995.

[73] N. Reid. A conversation with Sir David Cox. *Statistical Science*, 9(3):439–455, 1994.

[74] L Robinson and N Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 2:227–240, 1991.

[75] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistical Sinica*, 4:639–650, 1994.

[76] W. Shen and T.A. Louis. Empirical Bayes estimation via the smoothing by roughening approach. *Journal of Computational and Graphical Statistics*, 8(4):800–823, 1999.

[77] T.C. Smith, D.J. Spiegelhalter, and A. Thomas. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, 14(24):2685–1699, 1995.

[78] B.W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295, 1976.

[79] S.G. Walker and B.K. Mallick. A Bayesian semiparametric accelerated failure time model. *Biometrics*, 55(2):477–483, 1999.

[80] S.G. Walker and D.A. Stephens. A multivariate family of distributions on $(0, \infty)^p$. *Biometrika*, 86(3):703–709, 1999.

[81] L. Wang and D.B. Dunson. Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011.

[82] S. Yang and R.L. Prentice. Semiparametric inference in the proportional odds regression model. *Journal of American Statistical Association*, 94(445):125–136, 1999.

[83] Z. Ying, S.H. Jung, and L.J. Wei. Survival analysis with median regression models. *Journal of American Statistical Association*, 90(429):178–184, 1995.

[84] T. Zhang and J.S. Liu. Nonparametric hierarchical Bayes analysis of binomial data via bernstein polynomial priors. *Canadian Journal of Statistics*, 40(2):328–344, 2012.