Nonparametric Bayesian Inference in Biostatistics

by

Ying Yang

(Under the direction of Gauri Datta)

Abstract

Traditional parametric linear models are subject to several limiting constraints. In biomedical data analysis, parametric assumptions are often inappropriate because of multimodality and skewness arising from patient heterogeneity, presence of outliers, lack of important covariates, etc. For these reasons it is desirable to relax the parametric assumption, leading to a nonparametric approach to statistical modelling that accommodates these non-standard relationships in data.

This dissertation is a first step to understand the suitability of Polya tree priors and other nonparametric models for modeling biomedical data. In particular, the Polya tree prior is applied to repeated fractional data, cell line data, and microarray data. For repeated fractional data with a range of possible values from the unit interval and positive probability masses on 0 and 1, a latent variable is introduced to address probability point masses at 0 and 1. Posterior simulations for Polya tree priors on residual distributions, random effects distributions, and gene expression distributions are discussed. We propose new models and introduce Polya tree priors in these applications and develop novel algorithms to facilitate posterior inference.

Three case studies highlight aspects of inference with Polya trees. In one of the case studies we develop a nonparametric approach to inference about differential gene expression in microarray group comparison experiments. Future directions for research are discussed.

INDEX WORDS: Nonparametric Bayesian, Polya Tree, Dirichlet Process,Posterior Simulation, Fractional Data,Gene Expression

Nonparametric Bayesian Inference in Biostatistics

by

Ying Yang

M.S., The University of Georgia, 2000

M.S., Fudan University, 1996

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2004

Nonparametric Bayesian Inference in Biostatistics

by

Ying Yang

Approved:

Major Professor: Gauri Datta

Committee: Peter Müller
Lynne Seymour
William P. McCormick
Jaxk Reeves

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2004

Acknowledgments

Firstly, I would like to thank the Department of Statistics in the University of Georgia for allowing me to finish my Ph.D study while I work as a full time employee in Houston, Texas. In particular, I would deeply appreciate Dr. Peter Müller's guidance and encouragements that he has given to me over the past two years. No words can express my gratitude and respect to him. I would also like to thank Dr. Gauri Datta, Dr. Lynne Seymour, Dr. William P. McCormick, and Dr. Jaxk Reeves for their help during the completion of my dissertation research. I also had helpful conversations with Dr. Wesley O. Johnson and Dr. Timothy Hanson.

Above all, I want to thank those who were with me through the good and bad times: my parents; my husband, Ge Wen; my brothers and sister; and friends.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Linear models have been studied and applied intensively for continuous outcomes with normal errors. However, in many applications, this normality is not a reasonable assumption. This is particularly obvious in biomedical studies, where multimodality or skewness frequently arise due to patient heterogeneity, presence of outliers, lack of accounting for important covariates, alternative biologic mechanisms, etc.

Moreover, in linear modelling, the outcomes are often assumed to be independent. The independence assumption is not true for repeated measurement studies, where multiple measurements made on the same subject are naturally correlated. For this type of data, Laird and Ware (1982) present the random effects model, which is referred to as the normal linear random effects model. In this model, random effects are often assumed to be normally distributed. However, restricting the model to normally distributed random effects may be contrary to our prior beliefs. For the normal linear random effects model, Kleinman and Ibrahim (1998a) present an example where inference about the regression coefficients is sensitive to the assumption of normality about the random effects. Verbeke and Laesaffre (1996) show that the normal linear random effect can also perform poorly when the random effects have a mixture distribution.

Nonparametric approach can overcome the above limitations by assuming an unknown distribution for residuals or random effects. By doing so, we avoid to assume a specific form for the distribution. The nonparametric prior is robust with respect to misspecification of the

model. In particular, the nonparametric Bayesian approach for residuals or random effects is to specify a prior distribution on the space of all possible distribution functions. This prior is applied to the general prior distribution for residual or random effects. This can be accomplished with a Dirichlet Process prior distribution or Polya tree prior distribution.

1.2 Nonparametric Bayesian inference

A commonly used technical definition of nonparametric Bayesian models is probability models with infinitely many parameters (Bernardo and Smith 1994). Equivalently, nonparametric Bayesian models are probability models on function spaces.

The earliest priors for nonparametric problems seem to have been described by Freedman (1963) who introduced tail-free and Dirichlet random measures. Ferguson (1973, 1974) formalized and explored Dirichlet process (DP) prior. Escobar (1988), MacEachern (1994), Escobar and West (1995) extended DP to DP mixtures (MDP) in order to remove the constraint to discrete measures for DP. Posterior inference in MDP model is based on MCMC posterior simulation. Efficient MCMC simulation algorithm was discussed by MacEachern and Müller (1998). Ishwaran and James (2001) discussed sequential importance sampling-based methods for MDP models. Posterior consistency is discussed in Ghosal, Ghosh, and Ramamoorthi (1999). Nonparametric models based on Dirichlet process mixture are reviewed by MacEachern and Müller (2000).

Another type of tailfree process is Polya tree. Lavine (1992, 1994) proposed Polya tree as a generalization of the DP. Paddock et al (2003) and Hanson and Johnson (2002) introduced randomized Polya trees. Polya trees as priors in an accelerated failure time model were described in Walker and Mallick (1999) and Hanson and Johnson (2002). Walker and Mallick (1997) considered Polya tree as prior of random effects in generalized linear model.

Recent surveys of nonparametric Bayesian models appear in Walker et al. (1999), Dey et al. (1998), and Müller and Quintana (2004). A review of nonparametric Bayesian inference in survival analysis can be found in Sinha and Dey (1997).

However, nonparametric Bayesian methods are not appropriate in some applications. For example, when the data set is very small, posterior inference will be dominated by the prior if using nonparametric methods. It is also not proper to use nonparametric methods in experimental designs or when point estimation is of the interest.

1.3 Outline of the dissertation

The dissertation is organized as follows. In Chapter 2, we describe a model for repeated fractional data in section 2.2. In section 2.3 we discuss posterior simulation based on a normal random effects distribution in the proposed model. In section 2.4, we review the Dirichlet process prior and develop a posterior simulation scheme for inference using a Dirichlet process as a prior of the unknown random effect distribution in the fractional data model. A simulated data set is used to illustrate our proposed fractional data model in section 2.5. Section 2.6 summarizes the advantages and limitations of Dirichlet process.

In Chapter 3, we consider Polya tree (PT) priors for the distributions of residuals and random effects in linear models. In section 3.1, we provide the definition of PT models and discuss the properties and limitations. Posterior inference for PT models under independently and identically sampling is discussed in section 3.2. We provide some general suggestions for the choice of PT parameters. In section 3.3, we separately consider PT models as priors for residual and random effects distributions in linear models. In section 3.4, three examples are used to demonstrate the aforementioned methods.

In Chapter 4, we use three case studies with medical data sets to illustrate the application of nonparametric Bayesian methods in biomedical research. In our case study we develop a nonparametric Bayesian alternative to the popular empirical Bayes method (Efron

et al. 2001) for inference about differentially expressed genes in microarray group comparison experiments. Finally, we conclude the dissertation with summary and the extension of future research in Chapter 5.

In summary, nonparametric Bayesian methods provide alternative approaches to model common biostatistical data sets. Nonparametric Bayesian inference works fairly well when assumptions for the traditional parametric statistical methods are not satisfied because of heterogeneity of patients, presence of outliers, absence of accounting for important covariates, etc. It allows us to address scientific research questions without excessive dependence on technically convenient assumptions.

Chapter 2

A repeated fractional data model

Random-effects models are a traditional choice for the analysis of longitudinal data. Random effects are used to model dependence of repeated measurements from the some patients or other experimental units. Random effects can be thought of as unmeasured covariates whose values can be considered randomly distributed amongst study individuals. It is important to take account of unmeasured covariates or latent factors because they induce dependence among responses within an individual.

For continuous outcomes with normal errors, Laird and Ware (1982) proposed a normal linear random effects model. In this model, random effects were assumed to be centered around the mean regression coefficients for the populations, also known as the fixed effects. Conditional on random effects, repeated observations on a subject were considered independent. Goldstein (1986) and Longford (1987) proposed a model that incorporates nested random effect, enabling nested group-specific as well as individual-specific sources of heterogeneity (uncontrolled variation) to be modelled. Gilks et al (1993) presented a linear multiple-random-effects model that simultaneously accommodates group-specific sources of heterogeneity for several groupings of individuals with estimation using Gibbs sampling (Geman and Geman, 1984). Kleinman and Ibrahim (1998a) described a semiparametric Bayesian version of the normal linear random effect model, where a nonparametric prior distribution is specified for the random effects. In this chapter we consider the practically important case when the outcome variable is fractional data which is continuous between 0

and 1 plus positive point masses at 0 and 1. Statistical modelling on fractional data has not been investigated so far for longitudinal or repeated measurement data.

A standard approach for constrained data is the use of transformations to remove the constraint, such as a logit or probit transformation. Albert and Chib (1993) propose a probit regression model for binary and polychotomous outcomes. They impose normal regression structure on latent continuous data. Values of the latent data are simulated from suitable truncated normal distributions. After the latent data have been generated, the posterior distributions of the parameters are computed using standard results from normal linear models. Draws from these posteriors are used to sample new latent data. The process is iterated leading to a Gibbs sampling scheme. However, for the fractional data we are considering, a complication arises from the fact that 0 and 1 are included in the range of possible values, with positive probabilities. The conventional logit or probit transformation will no longer be valid.

In this chapter, we propose a simulation-based approach for computing the exact posterior distribution for parameters of interest. The key idea is to introduce additional latent variables to represent the awkward point masses at 0 and 1. A mixed effect model is imposed on latent variables and simultaneously accommodates individual-specific sources of heterogeneity. We start with standard normal mixed linear model assumptions as usual for continuous data and then extend to a nonparametric Bayesian model to accommodate the heterogeneity from other sources.

2.1 Model Formulation on Fractional Data

Suppose that a factional outcome vector y_i with n_i repeated measurements is observed in individual i. Responses y_{ij}, $j = 1, ..., n_i$, can be 0 or 1 with positive probability. Latent variables, z_{ij} are introduced to address this by including point masses at 0 and 1 in the

model.

$$y_{ij} = \begin{cases} 0 & \text{if } z_{ij} \leq 0 \\ z_{ij} & \text{if } 0 < z_{ij} < 1 \\ 1 & \text{if } z_{ij} \geq 1 \end{cases} \tag{2.1}$$
$$i = 1, \cdots, n, j = 1, \cdots, n_i$$

The z_{ij} are unknown. The distribution of z_{ij} is unconstrained and is continuous at 0 and 1. We can therefore proceed with standard linear mixed model assumptions as usual for continuous data, including normal distribution assumptions. We construct the following model for individual i,

$$z_i = X_i\beta + U_i\theta_i + e_i \tag{2.2}$$

where β is a $p \times 1$ vector of regression coefficients, commonly called fixed effects. The matrix X_i is a known $n_i \times p$ design matrix of fixed covariates. U_i is an $n_i \times q$ matrix of covariates for the $q \times 1$ random effect vector θ_i, and e_i is an $n_i \times 1$ vector of errors. In implementations of this model, it is common to assume e_i and θ_i are independent and $e_i \sim N_{n_i}(0, \sigma^2 I_{n_i})$.

For the distribution of random effects, we start with a normal distribution. Later we will introduce alternative and generalized models, as and if indicated by model diagnostics and criticism. We will assign nonparametric priors to the distribution of random effects.

Model (2.1) includes a monotonicity assumption. We assume that $Pr(y_{ij} = 1)$ increases as the location of $p(y_{ij} \mid 0 < y_{ij} < 1)$ increases, and similarly for $Pr(y_{ij} = 0)$. We feel this is reasonable in most applications.

For example, in an application with y_{ij} being the fraction of stained cells in immunohistochemistry data, it is reasonable to assume that the probability of all cells being stained ($y_{ij} = 1$) increases with the average fraction of stained cells going up. If it were desired to decouple $Pr(y_{ij} = 1)$ and $E(y_{ij} \mid 0 < y_{ij} < 1)$, this could be achieved by introducing two additional sets of latent variables, say u_{ij} and v_{ij}, with $Pr(y_{ij} = 1) = Pr(u_{ij} \geq v_{ij}, u_{ij} \geq z_{ij})$,

$Pr(y_{ij} = 0) = Pr(v_{ij} = \max(u_{ij}, v_{ij}, z_{ij}))$, and $Pr(y_{ij} \mid 0 < y_{ij} < 1) = Pr(z_{ij} \mid z_{ij} = \max\{u_{ij}, v_{ij}, z_{ij}\})$. We will not further pursue this model variation.

2.2 Normal linear random effect model

In the normal linear random effect model, the conditional distribution of $z_i (= (z_{i1}, \cdots, z_{in_i}))$ conditional on β and θ_i is given by

$$z_i | \beta, \theta_i \sim N_{n_i}(X_i\beta + U_i\theta_i, \sigma^2 I_{n_i}) \tag{2.3}$$

independently across experimental units, $i = 1, \cdots, n$. The linear model (2.3) defines the top level sampling model. Without loss of generality we assume independence across repeated observations, that is, a diagonal variance-covariance matrix. Little would change in the following discuss if we were to assume a non-diagonal variance-covariance matrix.

2.2.1 Prior Specification

We complete the model with conjugate priors. We choose the priors as follows. Let $N_p(m, S)$ denote a p-dimensional normal probability density function with moments (m, S).

For the fixed effects, we assume a conjugate multivariate normal prior

$$\beta \sim N_p(\mu_0, \Sigma_0) \tag{2.4}$$

Random effects are assumed to arise from a normal random effects model:

$$\theta_i \overset{iid}{\sim} N_q(0, \Sigma_\theta) \tag{2.5}$$

The prior on the variance of residuals is specified as follows.

$$\tau = (\sigma^2)^{-1} \sim Ga(\frac{\gamma_0}{2}, \frac{\lambda_0}{2}) \tag{2.6}$$

where $Ga(a,b)$ denotes a gamma distribution with mean $\frac{a}{b}$ and variance $\frac{a}{b^2}$. The conjugate priors of (2.4) through (2.6) are chosen for the technical convenience. Substantial prior information might require different prior distributions. Finally, $\mu_0, \Sigma_0, \Sigma_\theta, \gamma_0$, and λ_0 are fixed hyperparameters.

2.2.2 Posterior Inference

We implement posterior simulation by Gibbs sampling, resampling each of the indicated parameters conditional on the currently imputed values of all other parameters and the data. We did not use analytic forms because:

- All priors are only conditionally conjugate. They are not conjugate for the joint posterior distribution.
- Sampling one parameter at a time avoids manipulating an excessively large design matrix.
- Additionally, data are truncated by 0 and 1, which breaks joint conjugacy. Moreover, the normal linear mixed model is assumed on z, not on y.

Resampling z. Conditional on other parameters, resampling the latent z requires truncated normal sampling. From (2.3) we find

$$z_{ij} \begin{cases} \sim N(X_{ij}\beta + U_{ij}\theta_i, \sigma^2) I(z_{ij} \leq 0) & \text{if } y_{ij} = 0 \\ = y_{ij} & \text{if } 0 < y_{ij} < 1 \\ \sim N(X_{ij}\beta + U_{ij}\theta_i, \sigma^2) I(z_{ij} \geq 1) & \text{if } y_{ij} = 1 \end{cases} \tag{2.7}$$

where X_{ij} and U_{ij} are the jth row of matrices X_i and U_i, respectively. Generating random samples from a truncated normal is trivial. Suppose $x \sim N(\mu, \sigma^2) I(c_1 \leq x \leq c_2)$. Then x can be generated by letting $x = \sigma\Phi^{-1}(u) + \mu$, where u is a random sample generated from uniform distribution on the interval $[\Phi(\frac{c_1-\mu}{\sigma}), \Phi(\frac{c_2-\mu}{\sigma})]$, Φ is the normal cumulative

distribution function.

Resampling β. We proceed as in a standard normal linear regression.

From (2.3) we find

$$\underbrace{z_i - U_i\theta_i}_{z_i^\star} = X_i\beta + e_i.$$

Then

$$z_i^\star \sim N(X_i\beta, \tau^{-1}).$$

combining with the conjugate prior in (2.4) we find

$$P(\beta|\cdots) = N(m, v)$$

with moments

$$v = (\Sigma_0^{-1} + \tau \sum_i X_i' X_i)^{-1} \tag{2.8}$$

$$m = v(\Sigma_0^{-1}\mu + \tau \sum_i X_i' z_i^\star)$$

A list of other complete conditional posterior distributions is in the Appendix.

Gibbs Sampler

We implement MCMC posterior simulation by iterating over the complete conditional posterior distributions given in the above expressions. Starting with initial values $\beta, \theta_i, i = 1, ..., n$, and σ^2.

1. Generate z using (2.7).
2. Draw $\beta \sim P(\beta|\cdots)$ using the multivariate normal distribution given in (2.8).
3. Draw $\tau \sim P(\tau|\cdots)$ using $(A.1)$ in the Appendix.
4. For $i = 1, \cdots, n$, generate $\theta_i \sim P(\theta_i|\cdots)$ using $(A.2)$ in the Appendix.

Ergodic averages over the simulated parameter values approximate posterior integrals, including posterior means, posterior predictive distributions, etc. To verify convergence

of the MCMC simulations, Geweke's test (1992) was used to perform the convergence diagnostics. The assessment is done by BOA (Bayesian Output Analysis Version 1.0.1) in R.

2.3 Dirichlet Process Priors in the Linear Random effects model

Clearly it is not always appropriate to assume that the random effects arise from some known parametric family. The normal assumption on random effects can be restrictive. It is possible for the distribution of random effects to be multimodal and/or with unpredictable types of skewness. It might be desirable to take the random effects from a sufficiently large class to capture such possibilities. In biomedical data, multimodality frequently arises from patient heterogeneity, presence of outliers, lack of accommodating for important covariates, alternative biologic mechanisms, etc.

Examples are adult vs. pediatric populations, patients who are resistant to a given treatment, varying pharmacokinetics, invasive vs. non-invasive tumors, strokes arising from bursting blood vessels vs. blocked vessels, and so forth. For lower accounting of uncertainties and for improved prediction it is clearly critical to account for such heterogeneity. Also, the nature of the heterogeneity might be of inherit in itself, as , for example, in the discovery of new (sub-)issues of cancer. Our approach to achieve such generalization is the use of nonparametric Bayesian models.

A commonly used technical definition of nonparametric Bayesian models are probability models with infinitely many parameters (Bernardo and Smith 1994). In other words, a nonparametric Bayesian model is a probability model on a function space. Nonparametric Bayesian models are used to avoid critical dependence on parametric assumptions, to robustify parametric models, and to define model diagnostics and sensitivity analysis for parametric models by embedding them in a larger encompassing nonparametric model (Müller and Quintana 2004).

Bayesian nonparametric and semiparametric approaches include mixture models (West 1992), mixtures of Dirichlet processes (Ferguson 1973, Antoniak 1974, Escobar and West 1995, MacEachern and Müller 1998), and Polya tree priors (Lavine 1992,1994). For a recent review of nonprametric Bayesian models, see also Walker et al. (1999).

2.3.1 Dirichlet Process(DP)

The Dirichlet process is by far the most popular nonparametric model in the literature (for a recent review, see MacEachern and Müller 2000). Ferguson (1973) first introduced the DP as a random probability measure (RPM).

Definition: Let $\mathcal{X}$ be a space and $\mathcal{A}$ a σ-field of subsets, then a stochastic process G is said to be a Dirichlet Process on $(\mathcal{X}, \mathcal{A})$ with parameter αG_0 if for every $k = 1, 2, \cdots$ and measurable partition $(B_1, , \cdots, B_k)$ of $\mathcal{X}$, the random vector $(G(B_1), \cdots, G(B_k))$ has a Dirichlet distribution with parameters $(\alpha G_0(B_1), \cdots, \alpha G_0(B_k))$. Here $\alpha > 0$ is a scale parameter, G_0 is a probability measure on $\mathcal{X}$. We denote this by $G \sim DP(\alpha, G_0)$.

G is a random probability measure on $(\mathcal{X}, \mathcal{A})$, that is, the DP model is a distribution on distributions. Two parameters need to be specified. The prespecified probability measure G_0, the base measure, is the prior guess and expectation of G: $E(G(B_i)) = G_0(B_i)$. The precision parameter, α, describes the degree of faith in the choice of G_0; α could be regarded as an "equivalent prior sample size" (Antoniak 1974). For large values of α, a sample G is very likely to be closed to G_0. For small α, a sampled G is likely to put most of its probability masses on just a few points.

Two limiting cases highlight the nature of the prior. Consider $x_i \overset{iid}{\sim} G, i = 1, \cdots, n$, with a DP prior, $G \sim DP(\alpha, G_0)$. As $\alpha \to \infty$, the random measure G becomes closed to G_0, and in the limit the model is equivalent to $x_i \overset{iid}{\sim} G_0$. As $\alpha \to 0$, the random measure G concentrates on a few point masses, leading to an increasing probability of ties. In the limit, $x_1 = x_2 = \ldots = x_n$, with $x_1 \sim G_0$.

An essential motivation for the DP construction is the simplicity of posterior updating. For $x \in \mathcal{X}$, let δ_x denote the measure on $(\mathcal{X}, \mathcal{A})$ giving mass one to the point x:

$$\delta_x(A) = \begin{cases} 1 & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \end{cases}$$

Theorem (Ferguson 1973): Let G be a Dirichlet Process on $(\mathcal{X}, \mathcal{A})$ with parameters α and G_0, and let $X_1, \cdots, X_n$ be a sample of size n from G. Then the conditional distribution of G given $X_1, \cdots, X_n$ is a Dirichlet Process $(\alpha + n, G_1)$ with $G_1 \propto G_0 + \sum_{i=1}^{n} \delta_{X_i}$.

A useful constructive definition of the DP was given by Sethuraman (1994). Any $G \sim DP(\alpha, G_0)$ can be represented as

$$\begin{aligned} G(\cdot) &= \sum_{i=1}^{\infty} w_i \delta_{\mu_i}(\cdot) \\ \mu_i &\overset{i.i.d}{\sim} G_0 \\ w_i = U_i \prod_{j<i} (1 - U_j) \quad &\text{with} \quad U_i \overset{i.i.d}{\sim} Beta(1, \alpha) \end{aligned} \tag{2.9}$$

with point masses generated from G; and weights arising as rescaled Beta draws.

In words, every realization of the DP can be represented as an infinite mixture of point masses. The locations μ_i of point masses are an independently and identically distributed sample from G_0, and random weights w_i are generated by a "stick-breaking" procedure. The Dirichlet process selects a discrete distribution G with probability 1 (Ferguson 1973). This discreteness is in many applications inappropriate. Antoniak (1974) introduced the mixture of Dirichlet process model.

The mixture of DP (MDP) model convolutes the random measure G with a continuous kernel to represent a continuous distribution. Assuming i.i.d. sampling from the random measure we have:

$$y_i \sim \int K(y_i \mid \theta, v) dG(\theta), \quad G \sim DP(\alpha, G_0). \tag{2.10}$$

The kernel might be, for example, a Gaussian kernel, $K(y_i \mid \theta, v) = N(y_i; \theta, v)$. The mixture $K(y_i \mid \theta, v)dG(\theta)$ defines a random continuous distribution. For computational purposes it is often convenient to rewrite (2.10) as an equivalent hierarchical model. The hierarchical model is written as

$$\begin{aligned} y_i &\overset{iid}{\sim} p_{\theta_i,\nu}(y_i) \\ \theta_i &\overset{iid}{\sim} G(\theta_i), \;\; G \sim DP(\alpha, G_0) \end{aligned} \tag{2.11}$$

This model is referred as the mixture of Dirichlet process model(MDP). An example of mixture of Dirichlet process is given below.

EXAMPLE 1. $y_i \overset{iid}{\sim} N(\theta, \sigma^2)$. $\theta \sim G$ and $G \sim DP(\alpha, G_0)$. Then the marginal distribution of y is obtained through marginalizing out the random probability measure G. It is a mixture of normal models mixed with respect to location.

2.3.2 Dirichlet Process Prior for Random Effects

In this section, we use the DP model as random effects model in (2.2). The DP model removes the assumption of a parametric prior in (2.5) and replaces it with a random distribution G. The distribution G then has a Dirichlet process prior. The DP model is completed by assigning a normal base measure. The conjugate nature of the base measure and the assumed model for the latent z_{ij} significantly simplifies the computation. As before, let the observed outcome be y_{ij} and assume the distribution of the latent variable z_{ij} for the jth measurement from individual i be:

$$y_{ij} = \begin{cases} 0 & \text{if } z_{ij} \leq 0 \\ z_{ij} & \text{if } 0 < z_{ij} < 1. \\ 1 & \text{if } z_{ij} \geq 1 \end{cases}$$

$$\begin{aligned} (z_{ij} \mid \beta, \theta_i, \sigma^2) &\sim N(X_{ij}\beta + U_{ij}\theta_i, \sigma^2) \\ i = 1, \cdots, n, j &= 1, \cdots, n_i \end{aligned} \tag{2.12}$$

where X_{ij} and U_{ij} are the jth row of matrices X_i and U_i, respectively, and n_i is the number of observations from individual i. The prior specification is a conjugate model for β and τ, as before:

$$\begin{aligned} \beta &\sim N_p(\mu_0, \Sigma_0) \\ \tau = \frac{1}{\sigma^2} &\sim Gamma(\frac{\gamma_0}{2}, \frac{\lambda_0}{2}) \end{aligned} \tag{2.13}$$

and a nonparametric random effect prior for θ_i:

$$\theta_i \overset{iid}{\sim} G \tag{2.14}$$

$$G \sim DP(M, G_0), \quad G_0 = N_q(0, \Sigma_\theta) \tag{2.15}$$

The above specification results in a semiparametric specification in that fully parametric distributions are given in (2.12) and (2.13) and a nonparametric distribution is given in (2.14) and (2.15). Selecting G_0 to be normal emulates the conjugate relationship between the sampling distribution of z and priors in the usual Bayesian hierarchy. Its advantage lies in its simplicity of computation. It is possible to assign priors on (M, G_0). Kleinman and Ibrahim (1998) consider the case that places a prior on G_0. Escobar and West (1995) provided an algorithm to update M based on knowledge from data. If desired, the a.s. discrete random effects under G could be replaced by a DP mixture model as in (2.10), by assuming $\theta_i \sim \int N(\theta_i|\mu, S)dG(\mu)$.

Conditional on β and θ_i's, z_{ij} can be generated by (2.7) as before. Similarly, β can be sampled from (2.8) given $z, \theta = (\theta_1, \cdots, \theta_n)$ with element $\theta_i = (\theta_{i1}, \cdots, \theta_{iq})$ and σ^2. The full conditional for τ is the same with (A.1) under prior (2.13) conditional on z, β and θ. Now we describe how to update θ. We start by assuming M and Σ_θ to be known. Later, we will relax this restriction and place priors on M and Σ_θ. Compared to the parametric model in section 2.2, the only change in the Gibbs sampler is in Step 4 of the Gibbs samplers introduced in section 2.2.2. Conditional on all other parameters the complete conditional posterior for θ_i is derived as follows.

The closed form of joint distribution of $\theta_1, \cdots, \theta_n$ is given by

$$dF(\theta_1, \cdots, \theta_n) = \prod_{i=1}^{n} \frac{MG_0(\theta_i) + \sum_{j=1}^{i-1} \delta_{\theta_i}(d\theta_j)}{M + i - 1}$$

By the Theorem 2 (Escobar 1994), the conditional distribution of θ_i given $\theta_j, j \neq i, z_i, \beta, \sigma^2$ has the following close form:

$$\begin{aligned} dF(\theta_i|\theta_j, j \neq i, z_i, \beta, \sigma^2) &= \frac{\phi(z_i|\theta_i, \beta, \sigma^2)MG_0(d\theta_i) + \sum_{j=1, j\neq i}^{n} \phi(z_i|\theta_j, \beta, \sigma^2)\delta_{\theta_i}(\theta_j)}{A(z_i) + \sum_{j=1, j\neq i}^{n} \phi(z_i|\theta_j, \beta, \sigma^2)} \\ &= \frac{A(z_i)}{A(z_i) + \sum_{j=1, j\neq i}^{n} \phi(z_i|\theta_j, \beta, \sigma^2)} \times \frac{M\phi(z_i|\theta_i, \beta, \sigma^2)g_0(\theta_i)}{A(z_i)} \\ &+ \sum_{j=1, j\neq i}^{n} \frac{\phi(z_i|\theta_j, \beta, \sigma^2)}{A(z_i) + \sum_{j=1, j\neq i}^{n} \phi(z_i|\theta_j, \beta, \sigma^2)} \delta_{\theta_i}(\theta_j) \end{aligned}$$

where ϕ is the standard multivariate normal density function with mean $X_i\beta + U_i\theta_i$ and variance-covariance matrix $\sigma^2 I_{n_i}$ and $A(z_i)$ is defined as $A(z_i) = M \int \phi(z_i|\theta_i, \beta, \sigma^2) dG_0(\theta_i)$.

Gibbs sampling uses the simple structure of the conditional posteriors for the elements of θ, resulting in the following conditional distribution. For each $i = 1, \cdots, n$,

$$(\theta_i|\{\theta_j, j \neq i\}, z_i, \beta, \sigma^2) \sim q_0 G_\theta(\theta_i \mid z_l, \beta, \sigma^2) + \sum_{j\neq i} q_j \delta_{\theta_i}(\theta_j) \tag{2.16}$$

with the following definitions:

- $G_\theta(\theta_i|z_i, \beta) = \frac{M\phi(z_i|\theta_i, \beta, \sigma^2)g_0(\theta_i)}{A(z_i)}$ is the posterior distribution of θ_i if G_0 is the prior for θ_i. g_0 is the density function of G_0.

- $q_0 = cA(z_i) = cM \int \phi(z_i|\theta_i, \beta, \sigma^2) dG_0(\theta_i)$, just M times the density of the marginal distribution of $z_i = (z_{i1}, \cdots, z_{in_i})$ under the prior G_0.

- $q_j = c\phi(z_i|\theta_j, \beta, \sigma^2)$, the likelihood of z_i conditional on $\theta_i = \theta_j$.

- $c = \frac{1}{A(z_i) + \sum_{j=1, j\neq i}^{n} \phi(z_i|\theta_j, \beta, \sigma^2)}$ is the constant of normalization. The quantities q_j are standardized to unit sum, that is $q_0 + \sum_{j\neq i} q_j = 1$.

Following some algebra, $G_\theta(\theta_i|z_i,\beta)$ is a q-dimensional multivariate normal distribution with covariance matrix $Q_i = (\Sigma_\theta^{-1} + \tau U_i' U_i)^{-1}$ and mean $\tau Q_i U_i'(z_i - X_i\beta)$. In words, with probability q_0, we sample θ_i from $G_\theta(\theta_i|z_i,\beta)$, and we set $\theta_i = \theta_j$ with probability q_j.

When using the above conditional distributions in the MCMC algorithm, there may occur a problem if $\sum_{j\neq i} q_j$ becomes very large relative to q_0. The number of distinct θ_i's typically reduce to fewer than q due to the clustering of the θ_i's inherent in the Dirichlet process (Antoniak 1974). Then the MCMC chain is slow to converge. Escobar and West(1995) proposed a remixing algorithm to prevent the algorithm from getting stuck on a small set of θ_i's.

Some notations are now introduced to describe the remixing algorithm. Suppose $\theta = (\theta_1, \cdots, \theta_n)$ has $I \leq n$ distinct elements denoted by $b = \{b_1, \cdots, b_I\}$. Conditional on I, introduce indicators $s_l = j$ iff $\theta_l = b_j$. $S = \{s_l : l = 1, \cdots, n\}$ defines a configuration of θ. θ_{-l} denotes θ excluding the lth element. $S^{(l)}$ denotes the configuration of θ_{-l}. $I^{(l)}$ is the number of distinct values in θ_{-l}, with $n_j^{(l)}$ taking common value $b_j^{(l)}$.

Then we have

$$(\theta_l \mid z, \theta_{-l}, \beta, \tau) \sim q_0 G_\theta(\theta_l|z_l,\beta,\sigma^2) + \sum_{k=1}^{I^{(l)}} n_k^{(l)} q_k^{(l)} \delta_{\theta_l}(b_k^{(l)}) \tag{2.17}$$

with $q_k^{(l)} \propto \phi(z_l \mid \beta, b_k^{(l)}, \sigma^2 I_{n_l})$, and $q_0 + \sum_k n_k^{(l)} q_k^{(l)} = 1$.

After updating all the elements of θ, the newly generated θ implies a new configuration S. Once the set S is known, the posterior analysis of b_k's devolves into a collection of I independent analysis. Specially, the b_k's are conditionally independent with posterior densities

$$p(b_k|z,\beta,S,I,\Sigma_\theta,M) \equiv p(b_k|z_{(k)},\beta,S,I,\Sigma_\theta,M) \propto \prod_{j\in J_k} \phi(z_j|X_j\beta + U_j b_k, \sigma^2 I_{n_j}) dG_0(b_k,\Sigma_\theta) \tag{2.18}$$

for $k = 1, \cdots, I$. J_k is the set of indices of z in group k; i.e., $J_k = \{i : S_i = k\}$. $z_{(k)} = \{z_j : S_j = k\}$ is the observations in group k.

In summary, in order to sample from the conditional distribution of θ_l given all the other parameters in the model, one needs to do:

1. compute G_θ, the posterior distribution of θ_l given all the other parameters assuming that G_0 is the prior distribution for θ_l.

2. evaluate the marginal distribution for z_l assuming that G_0 is the prior distribution of θ_l.

3. sample θ_l from (2.17).

4. sample b_k from (2.18).

Typically, the covariance matrix Σ_θ in the base measure of the Dirichlet process in (2.15) is unknown and thus a suitable prior distribution can be placed on it. In the consequence, the base measure will no long marginally normal. For convenience, suppose

$$\Sigma_\theta^{-1} \sim Wishart(\nu, R_0)$$

where $\nu > 0$, and R_0 is a $q \times q$ positive definite matrix. Then

$$(\Sigma_\theta^{-1} \mid \nu, R_0) \propto |\Sigma_\theta^{-1}|^{\frac{\nu+q+1}{2}} exp\{-\frac{1}{2} tr((R_0\Sigma_\theta)^{-1})\},$$

Since $b_1, \cdots, b_I$ are I independent observations from $N_q(0, \Sigma_\theta)$, thus

$$\begin{aligned} p(\Sigma_\theta^{-1} \mid z, \beta, \theta, \sigma^2) &= p(\Sigma_\theta^{-1} \mid b_1, \cdots, b_I) \\ &\propto |\Sigma_\theta^{-1}|^{\frac{\nu+q+I+1}{2}} exp\{-\frac{1}{2} tr((R_0\Sigma_\theta)^{-1} + \sum_{k=1}^{I} b_k' \Sigma_\theta^{-1} b_k)\} \end{aligned}$$

so Σ_θ^{-1} can be updated by

$$(\Sigma_\theta^{-1} \mid z, \beta, \theta, \sigma^2) \sim Wishart(\nu + I, (R_0^{-1} + \sum_{k=1}^{I} b_k b_k')^{-1}) \tag{2.19}$$

The precision parameter, M, of Dirichlet process is extremely important for the model. Learning about M from the data may be addressed to incorporate M into the Gibbs sampling

analysis. If the prior for M is specified as $M \sim Ga(\alpha_M, \beta_M)$, then sampling M can be completed by the following steps.

- At each iteration, a latent variable η is sampled from beta distribution $(\eta|M, I) \sim Be(M+1, n)$ conditional on the most recent values of M and I.

- A new M is sampled from a mixture of two gamma distributions based on the same I and the newly generated η, that is,

$$(M|\eta, I) \sim \pi_\eta Ga(\alpha_M + I, \beta_M - log(\eta)) + (1-\pi_\eta)Ga(\alpha_M + I - 1, \beta_M - log(\eta)) \tag{2.20}$$

  where

$$\pi_\eta/(1-\pi_\eta) = (\alpha_M + I - 1)/(n(\beta_M - log(\eta))).$$

Finally, the complete Gibbs sampling schemes are summarized as follows:

1. Select starting values $\beta^{(0)}$, $\theta^{(0)}$, $\tau^{(0)}$, and $\Sigma_\theta^{(0)}$. Set $i = 0$.

2. Sample $z^{(i+1)}$ from $p(z|y, \beta^{(i)}, \theta^{(i)}, \tau^{(0)})$ according to (2.7).

3. Sample $\beta^{(i+1)}$ from $p(\beta \mid z^{(i+1)}, \theta^{(i)}, \tau^{(0)})$ according to (2.8).

4. Sample $\tau^{(i+1)}$ from $p(\tau \mid z^{(i+1)}, \beta^{(i+1)}, \theta^{(i)})$ according to (A.1).

5. Sample $\theta_l^{(i+1)}$ from $p(\theta_l \mid z^{(i+1)}, \beta^{(i+1)}, \tau^{(i+1)}, \theta_{-l}^{(i)}, \Sigma_\theta^{(i)})$ according to (2.17), for $l = 1, \cdots, n$.

6. Sample $b_k^{(i+1)}$ from $p(b_k \mid z^{(i+1)}, \beta^{(i+1)}, \tau^{(i+1)}, \Sigma_\theta^{(i)}, M^{(i)})$ according to (2.18), for $k = 1, \cdots, I$.

7. Sample $\Sigma_\theta^{-1(i+1)}$ from $p(\Sigma_\theta^{-1} \mid b^{(i+1)})$ according to (2.19).

8. Sample $M^{(i+1)}$ from $p(M \mid I, M^{(i+1)})$ according to (2.20).

2.3.3 Inference on the random distribution

In this subsection, we consider a generic Dirichlet process mixture model (MDP). Assume observations y_i are sampled independently and identically from a distribution represented as a mixture of kernels $p_\theta(y_i)$. Denote with $y = (y_1, \cdots, y_n)$ the observed data. A prior probability model is defined on the mixture by assuming a DP prior for a random mixing measure G.

$$y_i \overset{i.i.d}{\sim} \underbrace{\int p_{\theta,\nu}(y_i) dG(\theta)}_{H(y_i)}, \quad G \sim DP(MG_0(\cdot|\nu)) \tag{2.21}$$

The model is completed with a prior $p(\nu)$ on the hyperparameters ν. This is just model (2.11), marginalizing over θ_i. Model (2.12), (2.13), (2.14), and (2.15) is a special case of (2.21), using the normal linear regression, (2.12) as mixing kernel. In section 2.3.2, we explained the implementation of posterior inference by MCMC simulation. Note that posterior simulation in section 2.3.2 implicitly marginalized over G_0. That is, the random measure G itself never appeared in the posterior simulation.

The MCMC for MDP models is greatly simplified by marginalizing with respect to the unknown distributions G. Nevertheless, sometimes the final goal of an analysis is inference about G or H. In general, inference on the unknown distribution in DP mixture models is challenging. See Gelfand & Kottas (2002) for a discussion. However, some important simplifications are possible. The posterior means, $E(H|y)$ and $E(G|y)$, can be shown to be equal to the posterior predictive distributions $p(y_{n+1}|y)$ and $p(\theta_{n+1}|y)$ in the MDP model. They can be used to evaluate posterior estimates for G and H. Using full conditional posterior distributions that are already evaluated in the course of the MCMC simulation, the computation can be simplified by using an ergodic average of these conditional predictive distributions. This allows computationally efficient evaluation of $E(H|y)$ and $E(G|y)$. However, for the detailed full posterior inference more information is required. A computational algorithm is described as follows to allow easy (approximate) simulation.

First, we note that the posterior mean $E(H|y)$ is equal to the posterior predictive distribution. Let y_{n+1} denote a new , future observation, we find

$$p(y_{n+1}|y) = E[H(y_{n+1}|y, H)|y] = E[H(y_{n+1})|y]$$

Let γ denote the vector of all model parameters, and $\gamma^{(i)}$ denote the parameters imputed after i iterations of the MCMC simulation, we evaluate $p(y_{n+1}|y)$ as

$$p(y_{n+1}|y) = E[p(y_{n+1}|y,\gamma)|y] \approx \frac{1}{T}\sum_{i=1}^{T} p(y_{n+1}|\gamma^{(i)}, y) = \frac{1}{T}\sum_{i=1}^{T} p(y_{n+1}|\gamma^{(i)}). \tag{2.22}$$

Similarly, the posterior mean $E(G|y)$ is equal to the posterior predictive distribution. Let θ_{n+1} denote a new , future observation, we find

$$p(\theta_{n+1}|y) = E[G(\theta_{n+1}|y, G)|y] = E[G(\theta_{n+1})|y] \approx \frac{1}{T}\sum_{i=1}^{T} G(\theta_{n+1}|\gamma^{(i)}, y) = \frac{1}{T}\sum_{i=1}^{T} G(\theta_{n+1}|\gamma^{(i)})$$

Denote $\{\theta_j^{\star(i)}, j = 1, \cdots, k\}$ to be the distinct values of $\theta^{(i)} = (\theta_1^{(i)}, \cdots, \theta_n^{(i)})$, Write $n_j^{(i)}$ for the number of occurrences $\theta_l^{(i)} = \theta_j^{\star(i)}, l = 1, \cdots, n$ so that $n_1 + \cdots + n_k = n$. Then

$$p(\theta_{n+1}|y) = E[G(\theta_{n+1})|y] = M(M+n)^{-1}G_0(\theta_{n+1}) + (M+n)^{-1}\sum_{j=1}^{k} n_j \delta_{\theta_j^{\star(i)}}(\theta_{n+1}) \tag{2.23}$$

A simple example is used to illustrate the above algorithm. Assuming y_i are drawn from a normal distribution $N(\theta_i, \sigma^2), i = 1, \cdots, n$. $\theta = (\theta_1, \cdots, \theta_n)$ comes from some prior distribution G. A Dirichlet process prior is placed on G with a conjugate normal base measure $G_0 = N(b, \tau^2)$. Here $\gamma = (\theta, \sigma^2, b, \tau^2)$. We use a superindex (i) to identify the imputed parameter values after i iterations of the MCMC simulation. We find,

$$\begin{aligned}
p(y_{n+1}|\gamma^{(i)}) &= \int p(y_{n+1}|\theta_{n+1}, \sigma^{2(i)}) p(\theta_{n+1}|\theta, b^{(i)}, \tau^{2(i)}) d\theta_{n+1} \\
&= \frac{M}{M+n}\int f(y_{n+1}|\theta, \sigma^{2(i)}) dG_0(\theta) + \frac{1}{M+n}\sum_{j=1}^{k^{(i)}} n_j^{(i)} f(y_{n+1}|\theta_i^{\star(i)}, \sigma^{2(i)}) \\
&= \frac{M}{M+n} N(y_{n+1}; b^{(i)}, \sigma^{2(i)} + \tau^{2(i)}) + \frac{1}{M+n}\sum_{j=1}^{k^{(i)}} n_j^{(i)} N(y_{n+1}; \theta_j^{\star(i)}, \sigma^{2(i)}).
\end{aligned}$$

thus

$$p(y_{n+1} \mid y) = \frac{1}{T}\sum_{i=1}^{T} p(y_{n+1}|\gamma^{(i)})$$

Similarly,

$$p(\theta_{n+1}|\gamma^{(i)}) = \frac{M}{M+n}N(\theta_{n+1}; b^{(i)}, \tau^{2(i)}) + \frac{1}{M+n}\sum_{j=1}^{k^{(i)}} n_j^{(i)}\delta_{\theta_j^{\star(i)}}(\theta_{n+1})$$

where $\theta_j^\star, j = 1, \cdots, k$ are the unique values of θ_i, and n_j is the number of θ_i equal to $\theta_j^\star$.

Uncertainty in G is illustrated through posterior draws of G. Given $\theta^{(i)}$, the conditional posterior for G is a DP with updated parameters,

$$(G|\theta^{(i)}, Y) \sim DP(G_1, M+n) \quad \text{with} \quad G_1 = \frac{M}{M+n}G_0 + \frac{1}{M+n}\sum_{j=1}^{k^{(i)}} n_j^{(i)}\delta_{\theta_j^{\star(i)}} \tag{2.24}$$

The large total mass parameter $M + n$ implies that the random measure G is close to the conditional expectation G_1, the DP base measure in (2.24). We exploit this to approximate a posterior draw $G \sim p(G|\theta^{(i)}, y)$ as $G \approx G_1$.

2.4 A Simulation Study

In this subsection, we fit a mixed fractional data model to a simulated fractional date set to demonstrate the aforementioned methods in the estimation of random effects. Normal and Dirichlet process priors are placed as priors of the distributions of random effects. We simulated $n = 100$ data points for random effect from the follow normal mixture

$$\theta_1, \cdots, \theta_n \overset{iid}{\sim} 0.5N(-0.2, 0.15^2) + 0.5N(0.2, 0.15^2)$$

One covariate was generated by taking $X_{ij} \overset{iid}{\sim} N(0.5, 0.3^2), i = 1, \cdots, 100, j = 1, 2$. The true regression coefficient was set at $\beta = (\beta_0, \beta_1) = (0.8, -0.6)$, and the residual ϵ was generated from $N(0, 0.1^2)$. Let $\tilde{Y}_{ij} = \beta_0 + X_{ij}\beta_1 + \theta_i + \epsilon_{ij}, i = 1, \cdots, n, j = 1, 2$. The observed fractional data are given by $Y_{ij} = 1$ if $\tilde{Y}_{ij} \geq 1$; $Y_{ij} = \tilde{Y}_{ij}$ if $0 < \tilde{Y}_{ij} < 1$; and $Y_{ij} = 0$ if $\tilde{Y}_{ij} \leq 0$.

The histogram of observed Y is displayed in Figure 2.1. A latent variable z was introduced to remove the constraint of positive probability masses on points 0 and 1. Therefore, the model and priors are specified as follows:

$$
\begin{aligned}
y_{ij} &= \begin{cases} 1 & \text{if} \quad z_{ij} \geq 1 \\ z_{ij} & \text{if} \quad 0 < z_{ij} < 1 \\ 0 & \text{if} \quad z_{ij} \leq 0 \end{cases} \\
z_{ij} &= \beta_0 + X_{ij}\beta_1 + \theta_i + \varepsilon_{ij} \\
\theta_i &\overset{iid}{\sim} G, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \\
i &= 1, \cdots, n, j = 1, 2
\end{aligned}
$$

A vague normal prior is placed on $\beta = (\beta_0, \beta_1)$ and set it as $N((0,0), 100I_2)$.The distribution of random effect θ is estimated by the following two different approaches.

2.4.1 Normal prior

The distribution of the random effect θ is firstly set as normal distribution, that is $G = N(0, \sigma_\theta^2)$ with the hyperparameter prior $1/\sigma_\theta^2 \sim Ga(0.01, 0.01)$. Following the discussion in Section 2.2.2, the simulation was repeated 20000 times. The first 5000 iterations were discarded as burn-in. The point estimations and 95% confidence intervals of β and selected random effects are presented in Table 2.1. The posterior density of random effect θ is illustrated in Figure 2.2. It shows bimodality and suggests that a normal prior for random effects is not appropriate in this example.

2.4.2 Dirichlet prior for random effect

Now we specify a Dirichlet process prior for the distribution of random effect θ. Priors for other parameters are the same.

Assume

$$\theta_i \overset{i.i.d}{\sim} G, \quad i = 1, \cdots, n$$

$$G \sim DP(\alpha, G_0) \quad \text{with} \quad G_0 = N(0, \sigma_\theta^2)$$

$$1/\sigma_\theta^2 \sim Ga(\alpha_\theta, \beta_\theta)$$

Selecting G_0 to be normal emulates the conjugate relationship between the sampling distribution of z and priors in the usual Bayesian hierarchy. Its advantage lies in its simplicity of computation. α, the precision parameter, can take a fixed value or come from a parametric distribution.

α is firstly chosen to be 1, which reflects a large departure from normality. The initial values of θ_i is the average of y for each i. A Gamma prior is then placed on α.

The Gibbs sampler was run for 11,300 iterations. The first 2,300 iterations were discarded as a burn-in period. Every 5th iteration was used and the rest discarded, making a total sample size of 1,800. Convergence of the Gibbs sampler was assessed via Geweke's (1992) method.

2.4.3 Results

Table 2.1 shows posterior medians and 95% confidence intervals for parameters in the normal parametric and MDP models with the two values of α. As expected the estimates of fixed effects for different methods are close to the true values although 95% CI's for β_1 are wider under MDP priors. The reason for this deviation is shown in the density estimates presented in Figure 2.3. The first column in the table is for the normal model, the second column is for the DP model with $\alpha = 1$, and the third column is for the MDP model with a Gamma prior on α.

From the posterior distribution of the selected random effects, different facts are revealed. Not only are the 95% CI's different, the median values are very different as well compared to

the normal model. This is to be expected, as the random effects are directly affected by the relaxation of the Normal assumption. σ_θ is not easy to interpret in the MDP model, since it plays a complex role in the marginal posterior distribution of the θ_i.

From Table 2.1, it is clear that the point estimates and 95% confidence intervals of σ and σ_θ using normal parametric and nonparametric methods are quite different. The posterior densities of σ and σ_θ are illustrated in Figures 2.4 and 2.5.

Recall that the parameter α is a measure of the strength in the belief that G is G_0. Although it may be hard to quantify, α is a positive scalar that is related to how "clumpy" the data are. Clumpy data occur when the different subjects are concentrated into a few clusters. Recall also that α determines the prior distribution of k, the number of normal components in the mixture.

A value of $\alpha = 1$ reflects a large departure from normality, the mean value of $k = 6.7$. In Figure 2.6, the histogram of k is illustrated for the MDP model when a vague prior is placed on α. The mean value of k is 21.9 and the posterior mean of $\alpha = 9.4$. As mentioned, large values of α favor the base measure G_0 as the prior, i.e. the normal prior case. For this example, α is small which is an indication that the nonparametric model will give better estimates than normal model. The posterior histogram of α is given in Figure 2.7.

2.5 Conclusion

We have introduced semiparametric Bayesian inference for repeated fractional measurements. The main advantage of the proposed model are computational simplicity, ease of interpretation and explicit accounting and inference for clusters and subpopulations of patients. Computational simplicity is achieved by using mixture of normal models. Posterior MCMC proceeds by considering complete conditional posterior distributions conditional on indicators breaking the mixture.Conditional on these indicators inference reduces to a traditional

Table 2.1: 2.5%-, 50%-, and 97.5%-iles for various parameters from Normal MDP and the parametric normal model. β_0 is the intercept, β_1 is the fixed effect, and σ is the error standard deviation. θ_i is the intercept for subject i. σ_θ is the standard deviation in the base measure. $\bar{k}$ is the average number of clusters observed in the course of sampling

Parameter	Normal prior	DP prior α=1	DP prior $\alpha \sim G(\gamma_\alpha, \lambda_\alpha)$
β_0	0.830(0.768, 0.890)	0.810(0.695, 0.933)	0.813(0.715, 0.908)
β_1	-0596(-0.659, -0.534)	-0.559(-0.690, -0.437)	-0.560(-0.687, -0.438)
σ	0.093(0.080, 0.108)	0.268(0.241, 0.305)	0.273(0.240, 0.312)
θ_{69}	0.272(0.133, 0.401)	0.0002(-0.131, 0.129)	-0.006(-0.149, 0.126)
θ_{99}	-0.200(-0.338, -0.051)	-0.011(-0.152, 0.116)	-0.021(-0.174, 0.117)
σ_θ	0.247(0.194, 0.350)	0.099(0.052, 0.251)	0.083(0.048, 0.179)
$\bar{k}$		6.7	21.9

normal linear mixed model, greatly facilitating computation and interpretation of model parameters.

Limitations of the proposed model arise from the assumptions made in the DP prior model. For example, clusters generated by the DP prior are a priori stochastically ordered by size. For example, we can not express prior information that there might be subpopulations of a priori equal size. Also, the DP implies a specific form for predictive inference. In particular, the relative weights given to the base measure vs. prior observations are as given in (2.16). To generalize these weights one could use, with minimal changes to the posterior simulation algorithm, a species sampling model (SSM, Pitman 1996); or stick-breaking priors (Ishwaran and James 2001).

Finally, the model does not allow closed form inference. Although reasonably straightforward, posterior simulation is required.

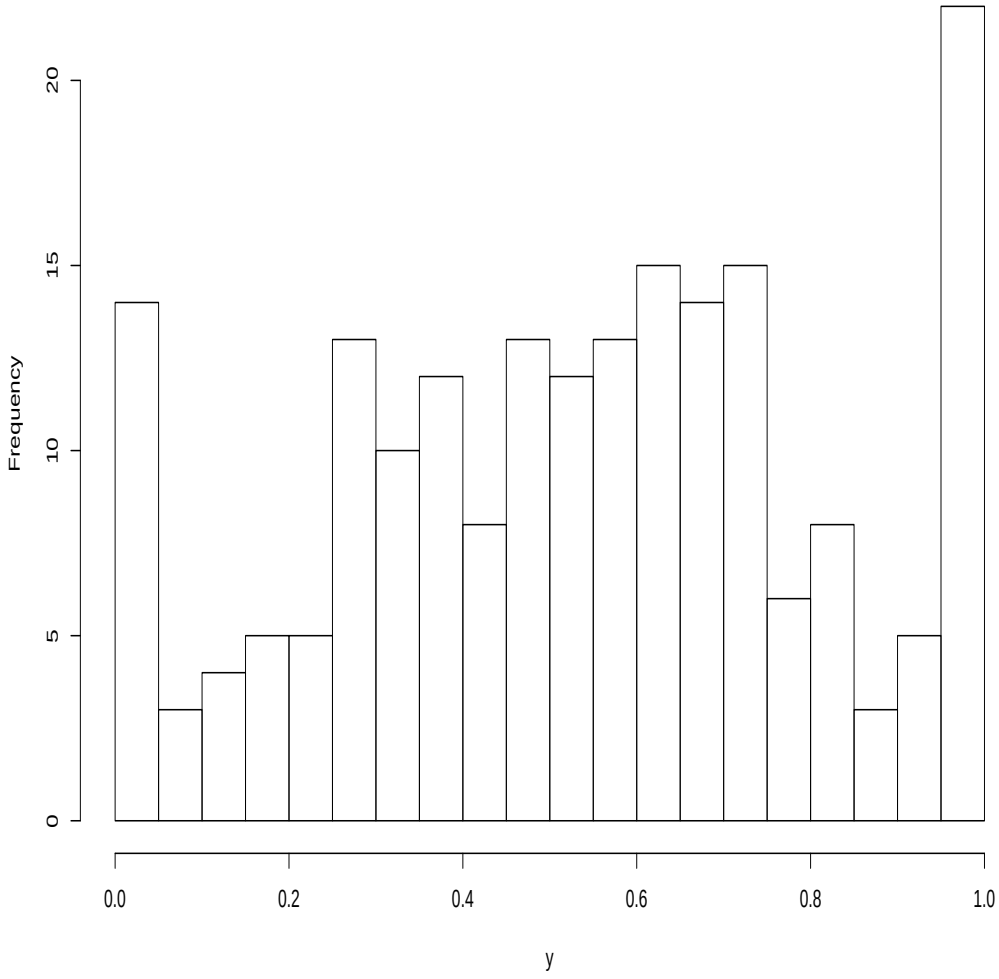

Figure 2.1: Histogram of y

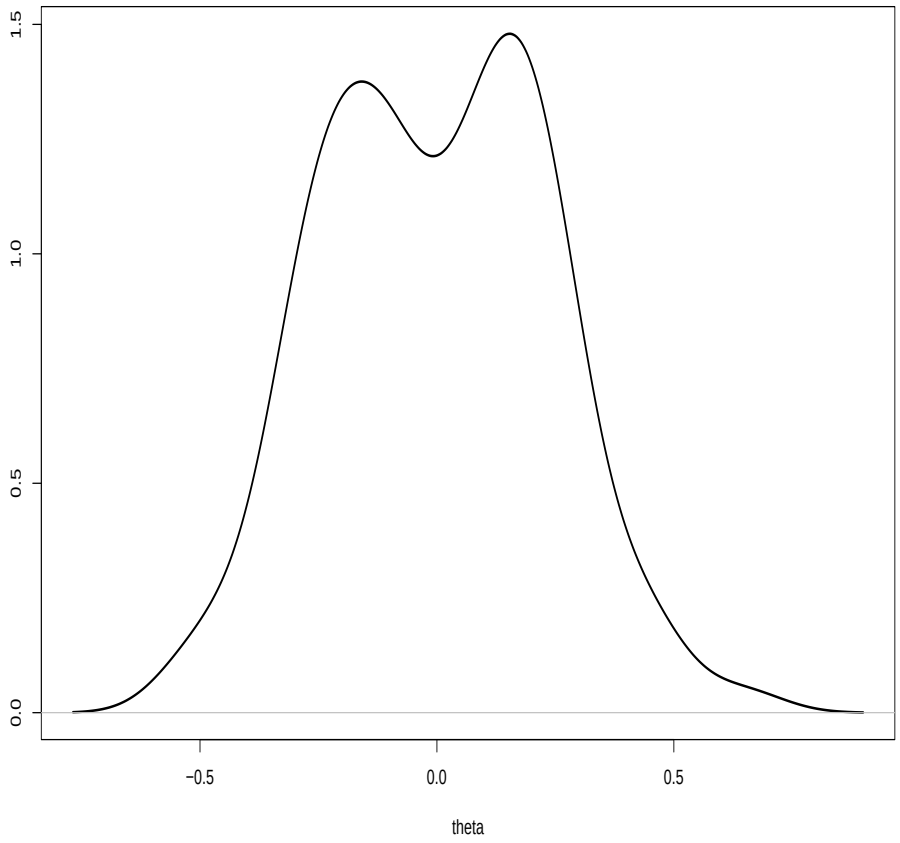

Figure 2.2: Density estimate of θ

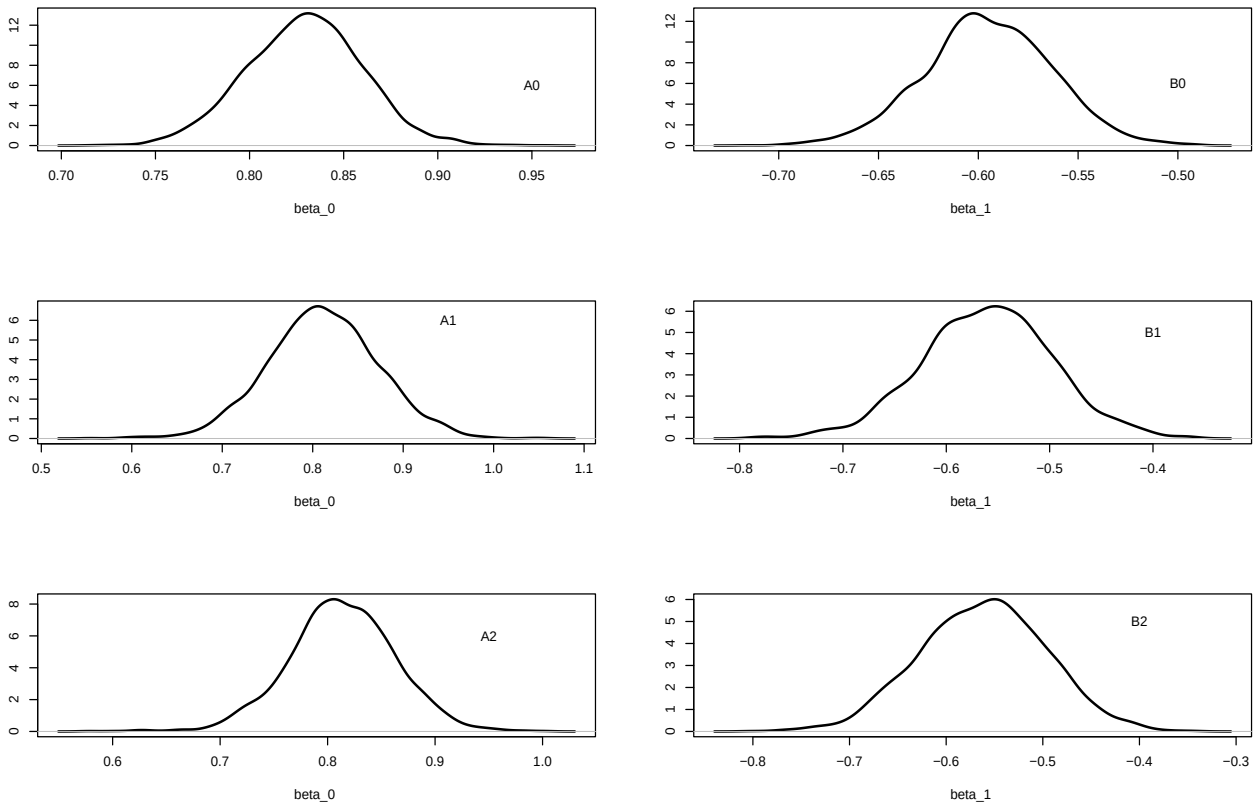

Figure 2.3: Density estimate of β

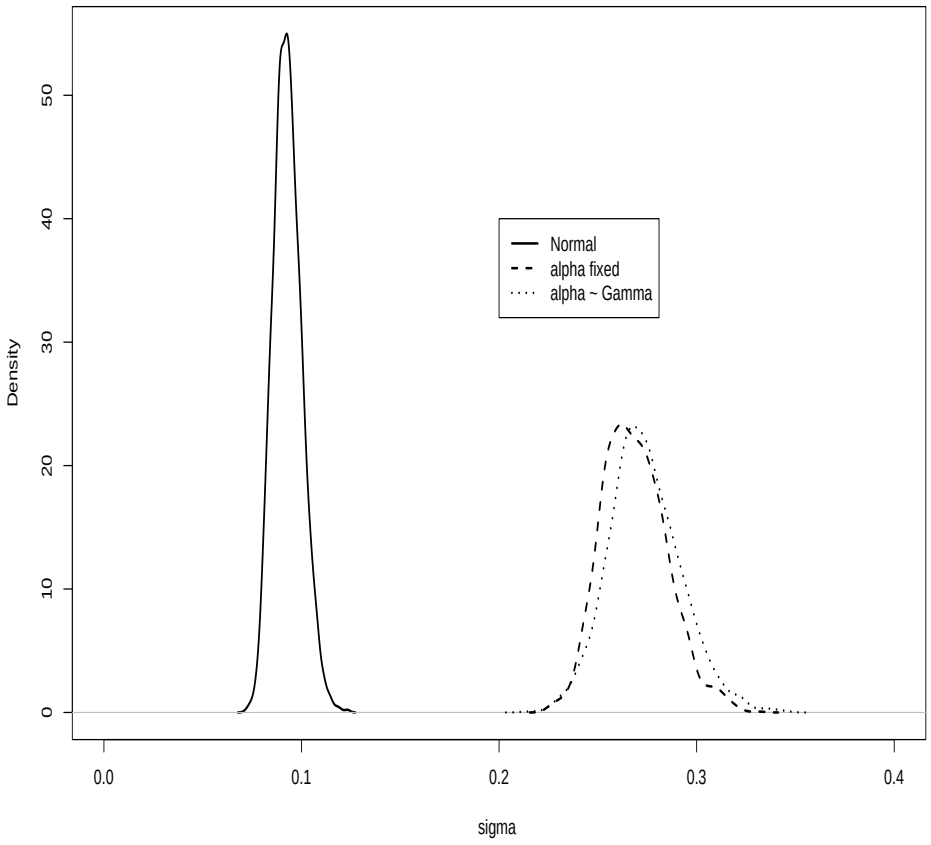

Figure 2.4: Estimated density estimate of σ under three models

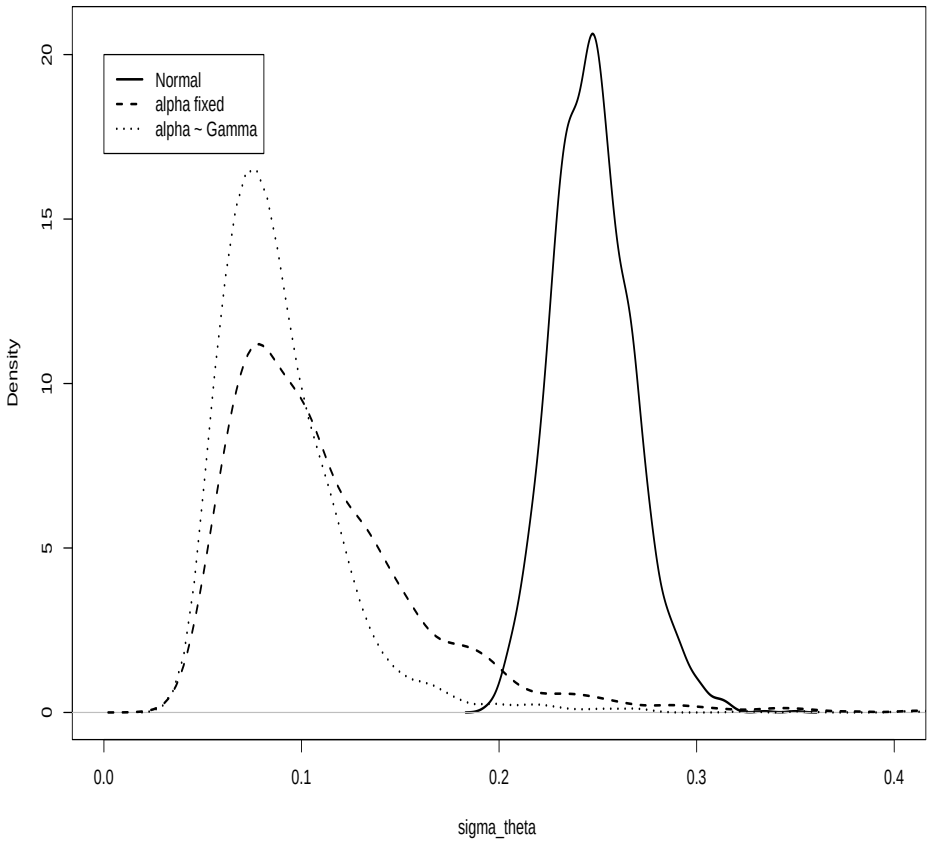

Figure 2.5: Estimated density estimate of σ_θ under three models

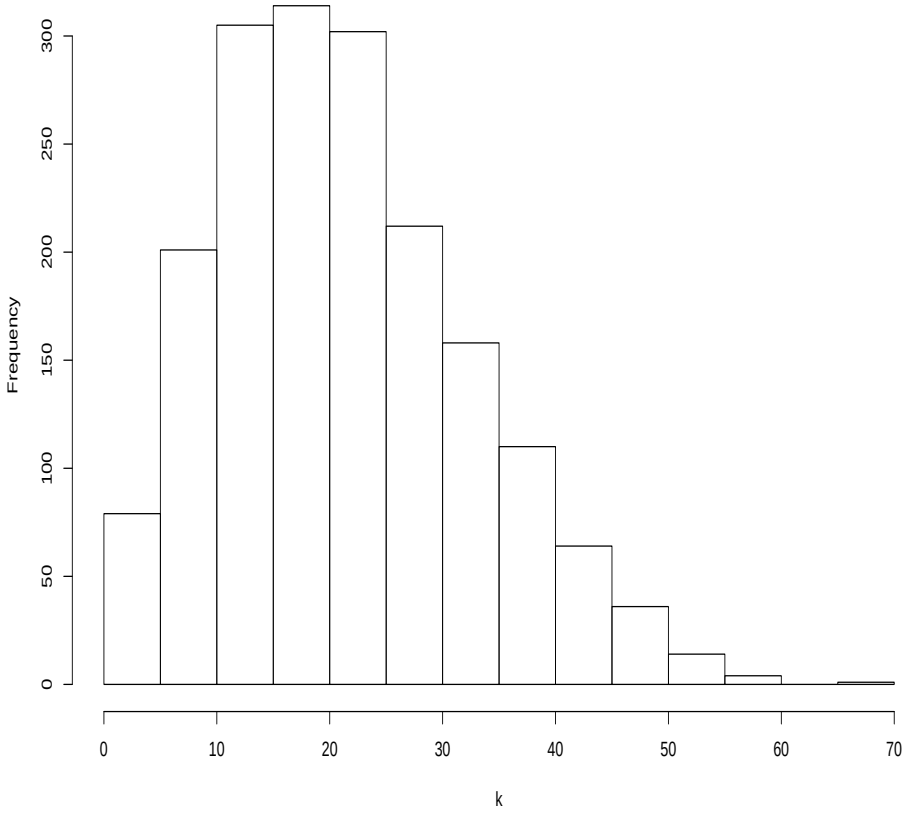

Figure 2.6: Observed histogram for k, number of clusters when $\alpha \sim Ga(\gamma_0, \lambda_0)$. $\bar{k} = 21.9$

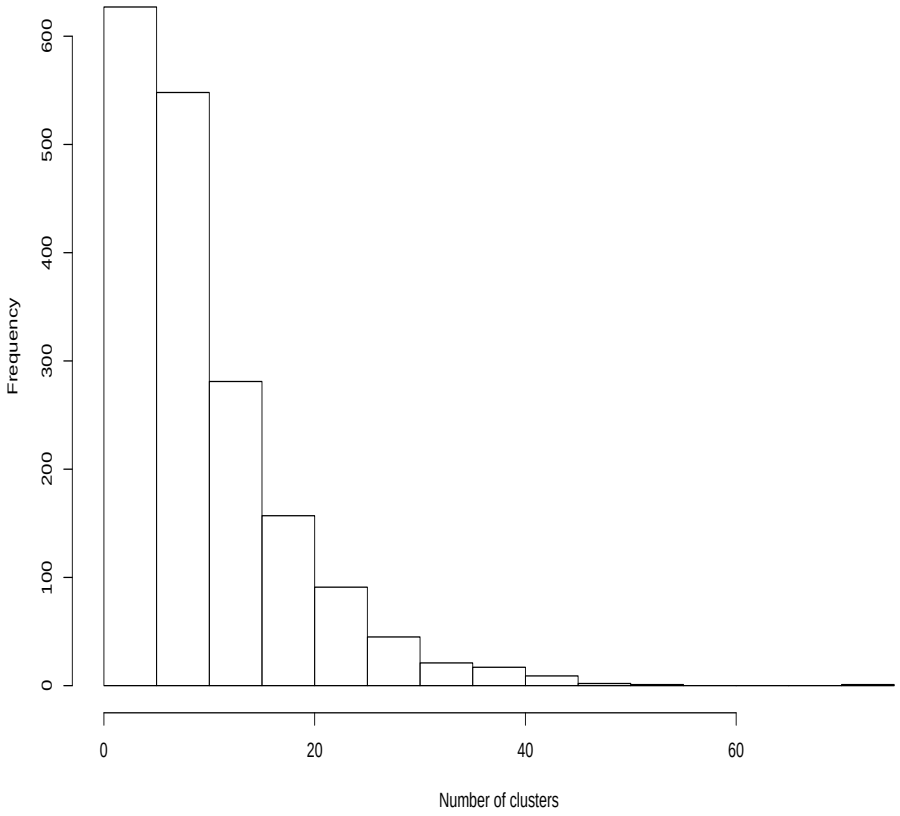

Figure 2.7: Histogram of α (precision parameter). $\bar{\alpha} = 9.4$

Chapter 3

Semiparametric models with Polya Tree priors

3.1 Introduction

The original description of the Polya tree (PT) model was introduced in Ferguson(1973). The Polya tree model provides an attractive generalization of the Dirichlet process (DP). Unlike the DP which assigns mass 1 to the set of all discrete distributions on $(\mathcal{X}, \mathcal{A})$, Polya tree distributions can be set up such that the random distribution G is continuous, or even an absolutely continuous distributions with probability one. Maudlin, Sudderth, and Williams (1992) show how to construct a Polya tree distribution using a Polya urn scheme. Lavine (1992, 1994) formally defines and develops the Polya tree model. He focuses on the binary tree construction used by Mauldin and Williams (1990) and Ferguson (1974) for constructing random distribution on the real line. He demonstrated how to construct a Polya tree prior with a given predictive density and how to use mixtures of PT's to model uncertainty about a parametric model.

Walker and Mallick (1997) illustrate the use of finite Polya trees as a Bayesian nonparametric prior for the random effects in a generalized linear model and frailty models. They show that it is possible to model the distribution using a Polya tree, whereas it is difficult using parametric models, for example, modeling the distribution about its median on the covariates. Muliere and Walker (1997) use a Polya tree prior distribution to determine a maximum tolerated dose (MTD) in a phase I trial. Through updating of the Polya tree, a predictive distribution for the critical dose level for the next individual is obtained, from which the maximum tolerated dose will follow. They show that Polya tree priors lead to good

estimation even with a modest number of patients and a limited number of doses. Walker and Mallick (1999) assigned a finite Polya tree prior to the error distribution in an accelerated failure time model. They consider two cases: one assumes error terms are exchangeable; the other assumes that error terms are partially exchangeable.

Polya tree models have limitations. Inferences for a PT depends on the sample space partitions for a specific tree structure. There are discontinuities in the predictive distribution at all partition boundaries. Paddock et al.(2003) proposed a randomized Polya tree to address the issue of partition dependence. They introduced observation specific parameters to jitter the tree partition. This smoothes out the discontinuities in the predictive distributions.

Berger and Guglielmi (2001) investigated the problem of testing the fit of data to a parametric model against a nonparametric alternative. They use a mixture of Polya trees centered at the hypothesized parametric model. A fixed sequence of partitions is used. The parameters that define the random probabilities of the nested partitions depend on unknown parameters in the null parametric model. On the other hand, Hanson and Johnson (2002), consider instead a mixture with respect to a hyperparameter that defines the partitioning tree. They discuss a median regression model in which the residual distribution is modelled as a mixture of Polya trees, centered at a parametric family of probability distributions with median 0. The mixture smooths out the effect of partitioning and thus the predictive error density is differentiable everywhere except 0.

3.1.1 Definition and Basic Properties

Polya trees were proposed as a generalization of Dirichlet Process (DP) in Lavine (1992, 1994), whose notation we follow.

Let $E = [0, 1]$,$E^0 = \emptyset$, E^m be the m-fold product $E \times E \times \cdots \times E$, $E^\star = \bigcup_{m=0}^{\infty} E^m$ and E^N be the set of infinite sequences of elements of E. Let Ω be a separable measurable space, $\pi_0 = \Omega$ and $\Pi = \{\pi_m; m = 0, 1, \cdots\}$ be a separating binary tree of partitions of Ω; that is,

let $\pi_0, \pi_1, \cdots$ be a sequence of partitions such that $\bigcup_{m=0}^{\infty} \pi_m$ generates the measurable sets and such that every $B \in \pi_{m+1}$ is obtained by splitting some $B' \in \pi_m$ into two subsets. Let $B_\emptyset = \Omega$ and, for all $\varepsilon = \varepsilon_1 \dots \varepsilon_m \in E^\star$, let $B_{\varepsilon 0}$ and $B_{\varepsilon 1}$ be the two subsets into which B_ε is split.

Definition (Lavine 1992): A random probability measure G on Ω is said to have a Polya tree distribution, or a Polya tree prior, with parameter $(\Pi, \mathcal{A})$, written as $G \sim PT(\Pi, \mathcal{A})$, if there exit non-negative numbers $\mathcal{A} = (\alpha_0, \alpha_1, \alpha_{00}, \cdots)$ and random variables $\mathcal{Y} = (Y_0, Y_1, Y_{00}, \cdots)$ such that

- all random variables in $\mathcal{Y}$ are independent;
- for every ε, $(Y_{\varepsilon 0}, Y_{\varepsilon 1}) \sim Beta(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$;
- for every $m = 1, 2, \cdots$, and every $\varepsilon = \varepsilon_1 \dots \varepsilon_m$,

$$G(B_{\varepsilon_1 \dots \varepsilon_m}) = \left(\prod_{j=1, \varepsilon_j = 0}^{m} Y_{\varepsilon_1 \dots \varepsilon_{j-1} 0} \right) \left(\prod_{j=1, \varepsilon_j = 1}^{m} (1 - Y_{\varepsilon_1 \dots \varepsilon_{j-1} 0}) \right)$$

  where the first term, i.e. for $j = 1$, are interpreted as Y_0 or $1 - Y_0$.

The random variable $Y_{\varepsilon 0}$ is the conditional probability of partition subset $B_{\varepsilon 0}$ given B_ε. For instance, for $m = 2$, $G(B_{00}) = Y_0 Y_{00}$, $G(B_{01}) = Y_0(1 - Y_{00})$, $G(B_{10}) = (1 - Y_0)Y_{10}$, and $G(B_{11}) = (1 - Y_0)(1 - Y_{10})$. The set Π determines the partition structure of the Polya tree. The parameters α_ε in $\mathcal{A}$ determine the smoothness of a realization of G and control how quickly the posterior predictive distribution moves from its prior mean to the empirical distribution.

Several properties facilitate the use of the Polya tree for nonparametric Bayesian inference.

Polya trees are conjugate. If $G|\Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A})$ and $x = (x_1, \cdots, x_n)|G \sim G$, then $G|x, \Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A}|x) = PT(\Pi, \mathcal{A}^\star)$, where $\mathcal{A}^\star = \{\alpha_\varepsilon^\star = \alpha_\varepsilon + n_\varepsilon : \varepsilon \in E^\star\}$ and $n_\varepsilon =$ the number of elements of x in B_ε. In words, the posterior distribution of G under i.i.d sampling

is also Polya tree with the same fixed partition sequence and the partitioning probability (Y_ε) is generated from Beta distribution with updated parameters $\alpha^\star_\varepsilon$, where $\alpha^\star_\varepsilon$ is equal to α_ε plus the number of $x_1, \cdots, x_n$ in subset B_ε.

The PT includes the Dirichlet process as a special case. A Polya tree is a Dirichlet process if $\alpha_\varepsilon = \alpha_{\varepsilon 0} + \alpha_{\varepsilon 1}$ for every $\varepsilon \in E^\star$ (Ferguson 1974).

Polya tree can choose $\mathcal{A}$ such that G is absolutely continuous with probability 1. In general, any $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = \rho(m)$ such that $\sum_{m=1}^{\infty} \rho(m)^{-1} < \infty$ guarantees G to be absolutely continuous. For example, Walker and Mallick (1999) and Paddock et al(2003) consider $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = cm^2$, where $c > 0$.

Let $x = (x_1, \cdots, x_n)$ denote the data of sample size n, i.e., $x_1, \cdots, x_n \overset{iid}{\sim} g_0$, where g_0 is the true density, with corresponding probability distribution G_0. Write $K(g_0, g) = \int g_0 log(\frac{g_0}{g})$ for the Kullback-Leibler divergence from g_0 to g. Let the Kullback-Leibler neighborhood $\{g : K(g_0, g) < \epsilon\}$ be denoted by $K_\epsilon(g_0)$. By Theorem 2 in Lavine (1994), the posterior is weakly consistent. If $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = 8^m$, Barron, Schervish, and Wasserman (1996) show that the predictive density is strongly consistent.

The joint marginal density generated from a random distribution with a PT prior has a closed form. The random probability measure G can analytically be integrated out. Suppose $x = (x_1, \cdots, x_n)|G \sim G$ and $G|\Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A})$, then the marginal joint density of $x_1, \cdots, x_n$ is given by

$$\begin{aligned} m(x_1, \cdots, x_n) &= \int p(x_1, \cdots, x_n \mid G) dp(G) \\ &= \int \prod_{i=1}^{n} g(x_i) dp(G) \\ &= f(x_1) \prod_{i=2}^{n} f(x_i \mid x_1, \cdots, x_{i-1}) \end{aligned}$$

where $f(x_1) = g_0(x_1)$ and

$$f(x_i \mid x_1, \cdots, x_{i-1}) = g_0(x_i) \lim_{M \to \infty} \prod_{m=2}^{M} \frac{\alpha'_{\varepsilon_m}(\alpha_{\varepsilon_{m-1}0} + \alpha_{\varepsilon_{m-1}1})}{\alpha_{\varepsilon_m}(\alpha'_{\varepsilon_{m-1}0} + \alpha'_{\varepsilon_{m-1}1})} \tag{3.1}$$

$\underline{\varepsilon}_m = \varepsilon_1 \cdots \varepsilon_m$ and $x_i \in B_{\varepsilon_1 \cdots \varepsilon_m}$. g_0 is the density function of centering distribution G_0 such that $E(G) = G_0$. $\alpha'_{\underline{\varepsilon}_m}$ is equal to $\alpha_{\underline{\varepsilon}_m}$ plus the number of observations among $x_1, \cdots, x_{i-1}$ that belong to $B_{\varepsilon_1 \cdots \varepsilon_m}$. In practice, we can always reduce the right side of (3.1) to a finite product. There are two reasons for doing so. The right side of (3.1) can be approximated by a finite product when m is large. On the other hand, $\alpha'_{\underline{\varepsilon}_m}$ will be equal to $\alpha_{\underline{\varepsilon}_m}$ when the partitioning process goes down to the Mth level such that no observations will fall into the subset which x_i belongs to.

3.1.2 Construction of a Polya Tree

Two issues should be considered when a Polya tree prior is constructed on Ω, e.g., $\Omega = \mathcal{R}$. One is how to construct the partition Π. The other is the selection of$\mathcal{A}$.

Through choosing Π, we may center the PT prior around a particular continuous distribution G_0. To do so, we take the partition points to align with percentiles of G_0. For instance, if $B_0 = (-\infty, G_0^{-1}\left(\frac{1}{2}\right)]$ (hence $B_1 = (G_0^{-1}\left(\frac{1}{2}\right), \infty)$), $B_{00} = (-\infty, G_0^{-1}\left(\frac{1}{4}\right)]$, $\cdots$ and $\alpha_0 = \alpha_1$, $\alpha_{00} = \alpha_{01}$, $\cdots$, then since $G(B_0) = Y_0 \sim Be(\alpha_0, \alpha_1)$, $E(G(B_0)) = \frac{1}{2} = G_0(B_0)$. Also, e.g., $G(B_{00}) = Y_0 Y_{00}$ implies $E(G(B_{00})) = \frac{1}{4} = G_0(B_{00})$ and for any $B \in \Pi$, $E(G(B)) = G_0(B)$. We need not confine ourselves to quartiles of the form $G_0^{-1}\left(\frac{j}{2^r}\right)$.

Figure 3.1 shows an example of the construction of a Polya tree prior on $(0, 1] = \Omega$ (Ferguson 1974). At the top level of the tree, Ω is split in half at the dyadic rational, 0.5. Thus $B_0 = (0, 0.5]$, $B_1 = (0.5, 1]$ and $\Omega = B_0 \bigcup B_1$. At the second level, B_0 and B_1 are again split at 0.25 and 0.75, respectively, which result in subsets $B_{00} = (0, 0.25]$, $B_{01} = (0.25, 0.5]$, $B_{10} = (0.5, 0.75]$, and $B_{11} = (0.75, 1]$, so on.

Another issue is how to select the parameters in $\mathcal{A}$. The parameters α_ε in $\mathcal{A}$ control how quickly the updated predictive distribution moves from the centering distribution G_0 to the empirical distribution. If the α_ε's are large, then the distribution of $x_{n+1} \mid x_1, \cdots, x_n$ is close G_0. If the α_ε's are small, then the distribution of $x_{n+1} \mid x_1, \cdots, x_n$ is close to

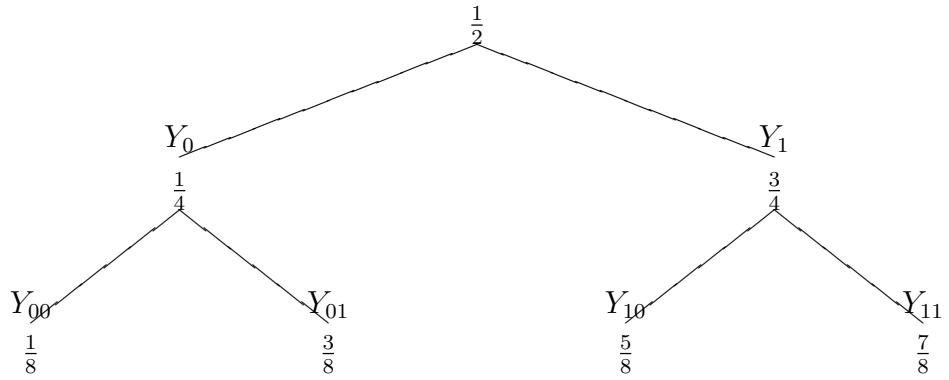

Figure 3.1: Construction of a Polya tree prior on (0,1] (Ferguson 1974)

the empirical distribution function. The parameters α_ε also express the belief about the smoothness of G. Ferguson (1974) provides conditions on $\mathcal{A}$ which yield discrete, continuous singular, and absolutely continuous distributions with probability one. For instance, for level $m = 1, 2, \cdots$, $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = 2^{-m}$ implies a Dirichlet process,$\alpha_\varepsilon = 1$ yields a random probability G of a type considered by Dubins and Freedman (1966) and shown to be continuous singular with probability one, and $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = m^2$ implies an absolutely continuous distribution with probability 1. Walker and Mallick (1999) and Paddock et al.(2003) considered $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = cm^2$, where $c > 0$.

Therefore, through selection of $\mathcal{A}$ and G_0, one can center the Polya tree prior around G_0 arbitrarily close, as determined by $\mathcal{A}$, in a manner analogous to the specification of baseline measure and precision parameter in the Dirichlet process. $\mathcal{A}$ can be thought of as a precision parameter and G_0 as a base measure.

3.1.3 Limitations of Polya Tree Models

Polya trees have some practical limitations. First, the resulting random probability measure is dependent on the specific partition sequence adopted. Second,using a fixed partitioning sequence Π results in discontinuities in the predictive distributions. Third, implementations for higher dimensional distributions require extensive housekeeping and are impractical. To mitigate problems related to the discontinuities Paddock et al. (2003) and Hanson and Johnson (2002) introduced randomized Polya trees. The idea is based on dyadic rational partitions, but instead of taking the nominal half-point Paddock et al.(2003) randomly choose a "close" cutoff. This construction is shown to reduce the effect of the binary tree partition on the first two points noted above. Alternatively, Hanson and Johnson (2002) consider instead a mixture with respect to a hyperparameter that defines the partitioning tree. The problem concerning high dimension persists though.

3.1.4 Finite Polya Tree

In practice, we can only carry out the above partitioning process to a finite level r. We then obtain a "partially specified Polya tree" (Lavine 1994), which is also called finite Polya tree, to approximate a realization from $PT(\Pi, \mathcal{A})$. A partially specified Polya tree will be defined and denoted by $G^r \sim PT(\Pi^r, \mathcal{A}^r)$. Let S^r be a finite subset of $E^\star$ such that, for every $\varepsilon = \varepsilon_1 \dots \varepsilon_m \in S^r$, $\varepsilon_1 \dots \varepsilon_j \in S^r (j < m)$ as well, and suppose that we have specified parameters $\{B_{\varepsilon 0}, B_{\varepsilon 1}, \alpha_{\varepsilon 0}, \alpha_{\varepsilon 1} : \varepsilon \in S^r\}$. Let $T_1 = \{G(B_{\varepsilon 1}) : \varepsilon \in S^r\}$ be the random probabilities assigned by the partially specified Polya tree, and let T_2 be the mass distribution of G conditional on T_1. Thus, $G = (T_1, T_2)$ and $L(G) = L(T_1) \times L(T_2|T_1)$.

Definition (Finite Polya Tree (Lavine 1994)) *The random variable T_1 has a finite Polya tree distribution with parameter $(\Pi^r, \mathcal{A}^r)$ if there exist sets $\Pi^r = \{B_{\varepsilon 1} : \varepsilon \in S^r\}$, $\mathcal{A}^r = \{\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1} : \varepsilon \in S^r\}$ and random variables $\mathcal{Y}^r = \{Y_\varepsilon : \varepsilon \in S^r\}$ such that:*

- *all the random variables in* $\mathcal{Y}^r$ *are independent;*

- *for every* $\varepsilon \in S^r$,Y_ε *has a Beta distribution with parameters* $(\alpha_{\varepsilon 0},\ \alpha_{\varepsilon 1})$*;*

- *for every* $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in S^r$,

$$G(B_{\varepsilon_1 \cdots \varepsilon_m}) = \left(\prod_{j=1,\varepsilon_j=0}^{m} Y_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} \right) \left(\prod_{j=1,\varepsilon_j=1}^{m} (1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1} 0}) \right)$$

The level r can be either fixed or we can assign it a prior distribution, e.g., a Poisson distribution. Walker and Mallick (1997, 1999) fixed $r = 8$. Hanson and Johnson (2002) recommend the rule of thumb $r \doteq log_2 n$, where n is the sample size. By allowing r to grow with the sample size, G accommodates finer and finer detail as more data are available. Or r can also be chosen such that there is at most one observation in each subset. It is also sensible to stop partition at the level under which the error of approximation is below a given predetermined value.

3.2 Posterior Inference under the i.i.d Sampling Model

3.2.1 Model

Assume observations x_i are independently and identically distributed, sampled from an unknown distribution G, with a Polya tree prior on G:

$$x_i \overset{iid}{\sim} G,\ i = 1, \ldots, n \text{ and } G \sim PT(\Pi, \mathcal{A}).$$

For the choice of parameter Π, we take the partition points to be quartiles of G_0. That is, $B_0 = (-\infty, G_0^{-1}(1/2))$, $B_1 = [G_0^{-1}(1/2), \infty)$, and, at level m, setting $B_j = [G_0^{-1}((j-1)/2^m), G_0^{-1}(j/2^m))$ for $j = 1, \cdots, 2^m$,with $G_0^{-1}(0) = -\infty$ and $G_0^{-1}(1) = +\infty$. Therefore $\{B_j; j = 1, \cdots, 2^m\}$ correspond to the 2^m partitions of level m. Under this parametrization $E(G) = G_0$. As for the choice of $\mathcal{A}$, it is convenient to take $\alpha_\varepsilon = cm^2$ for some $c > 0$ where $\varepsilon = \varepsilon_1 \cdots \varepsilon_m$ defines the index of a subset at level m. This implies that there exits a large

amount of variability for small m (higher levels). However, as we move down the levels (m large), $G(B_{\varepsilon 0})$ and $G(B_{\varepsilon 1})$ are close to reflect beliefs in the underlying continuity of G.

3.2.2 Posterior updating

According to the conjugacy property of the Polya tree, given an observation x_1 the posterior Polya tree distribution is easily obtained. We write the posterior as $G|x_1 \sim PT(\Pi, \mathcal{A}|x_1)$ with $\mathcal{A}|x_1$ given by

$$\alpha'_\varepsilon = \begin{cases} \alpha_\varepsilon & \text{if } x_1 \notin B_\varepsilon \\ \alpha_\varepsilon + 1 & \text{if } x_1 \in B_\varepsilon \end{cases} \tag{3.2}$$

For n independent observations, $x = (x_1, \cdots, x_n)$, $\mathcal{A}|x$ is given by $\alpha'_\varepsilon = \alpha_\varepsilon + n_\varepsilon$ where n_ε is the number of observations from $(x_1, \cdots, x_n)$ in B_ε.

3.2.3 Posterior Predictive Simulation

We exploit the structure in (3.2) to implement posterior predictive simulation. Formally, given $x_i \overset{iid}{\sim} G, i = 1, \cdots, n$ and $G \sim PT(\Pi, \mathcal{A})$, we consider posterior predictive simulation of $x_{n+1} \sim p(x_{n+1} \mid x_1 \dots x_n)$. Theoretically, the posterior predictive density $x_{n+1} \sim p(x_{n+1} \mid x_1 \dots x_n)$ can be evaluated by (3.1). Practically, the evaluation of the right side in (3.1) will be terminated at a finite level for two reasons. One reason is that the right side of (3.1) can always be approximated by the product of finite terms. The other reason is that starting at a certain level no observation will fall into the subinterval which x belongs to. Therefore, the product will not change from that level.

Based on (3.1), it is easy to sample the density $p(x_{n+1} \mid x_1 \dots x_n)$. In words, we "drop" a ball down (well, really up) the Polya tree. Starting with (B_0, B_1) at the root we generate the random probabilities $(Y_{\varepsilon 0}, Y_{\varepsilon 1})$ for picking the right and left partition in the respective next level. Recall that $Y_{\varepsilon 0} = G(B_{\varepsilon 0} \mid B_\varepsilon)$ and $Y_{\varepsilon 0} \sim Beta(\alpha'_{\varepsilon 0}, \alpha'_{\varepsilon 1})$. Going down the tree we run into some good luck. At some level m we will drop the ball into a subset B_ε, $\varepsilon = \varepsilon_1 \varepsilon_2 \cdots \varepsilon_m$,

that does not contain any data point. From level m onwards, dropping the ball proceeds as if we had no data observed. Thus we can generate the posterior predictive draw from the base measure G_0, restricted to B_ε. The posterior predictive draw can be generated by the following algorithm.

Algorithm 1

1. *Initialize:* $\varepsilon = \emptyset$.

2. *Iteration:* Loop over $m = 1, 2, \ldots$:

   (a) *Posterior PT Parameters:* Find $n_{\varepsilon 0} = \sum_{i=1}^{n} I(x_i \in B_{\varepsilon 0})$ and $n_{\varepsilon 1} = \sum_{i=1}^{n} I(x_i \in B_{\varepsilon 1})$, the number of x's in the two partitioning subsets for $B_\varepsilon = B_{\varepsilon 0} \cup B_{\varepsilon 1}$. Let $\alpha'_{\varepsilon 0} = \alpha_{\varepsilon 0} + n_{\varepsilon 0}$, and $\alpha'_{\varepsilon 1} = \alpha_{\varepsilon 1} + n_{\varepsilon 1}$.

   (b) *Generate Random Partitioning Probability:* Generate $Y_{\varepsilon 0} \sim Beta(\alpha'_{\varepsilon 0}, \alpha'_{\varepsilon 1})$, and set $\varepsilon_m \sim Bernoulli(1 - Y_{\varepsilon 0})$. If we want to keep the median of G unaffected by the centering distribution G_0, $\varepsilon_m \sim Bernoulli(0.5)$ for $m = 1$.

   (c) Set $\varepsilon = \varepsilon_1 \ldots \varepsilon_m$.

3. *Stop of the Recursion:* Stop the iteration over m for the smallest m^* such that $n_{\varepsilon_1 \ldots \varepsilon_{m^*}} = 0$ at $m = m^*$.

4. *Generate* x_{n+1}*:* Draw $x_{n+1} \sim G_0(x_{n+1}) \cdot I(x_{n+1} \in B_{\varepsilon_1 \ldots \varepsilon_{m^*}})$.

Artificial data is used to illustrated the above algorithm. We simulate $n = 400$ data points from a mixture of two normals with median 0.

$$x_1, \cdots, x_n \overset{iid}{\sim} 0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2) \tag{3.3}$$

For illustration purposes, we assume these observations are i.i.d. from an unknown distribution G. To estimate the predictive density, we assign a Polya tree prior on G. For the partition sequences of Polya tree, we try a simple centering distribution $G_0 = N(\mu, \sigma^2)$. It

is reasonable to fix μ at 0. For σ we consider values 1 and 10 to assess the effect of the standard deviation of the base measure on the posterior inference. Choose parameters α_ε in $\mathcal{A}$ to be $cm^2, m = 1, 2, \cdots$. For c we try values 0.1 and 5 in order to examine how c affects the posterior predictive density. We repeat the above algorithm 2000 times; thus 2000 x_{n+1}'s are generated.

The true density (solid line) of the mixture normal and histogram of the 400 observations are show in Figure 3.2. The data are fairly representative of the density. We compare the effects of the standard deviation σ and c through plots. The predictive density estimates are plotted in Figure 3.3. When c is small, the predictive density derived under the base measure with a small standard deviation is closer to the true density compared to the base measure with a large standard deviation. When c is large, the predictive density is less able to capture important features of the true distribution. The reason behind this phenomenon is that c plays a role as "precision parameter" in the Dirichlet process. For small values of c, the distribution of $x_{n+1} \mid x_1, \cdots, x_n$ is close to the empirical cdf. For large values of c, the distribution of $x_{n+1} \mid x_1, \cdots, x_n$ remains close to G_0.

In summary, we recommend to use a base measure with a scale (σ) comparable to the range of the data. For c we recommend small values, say $c = 0.1$ as a default for moderate size data.

3.2.4 Prior Predictive Sample

We can also use the above algorithm to generate a prior predictive sample,i.e., $x_i \overset{\text{iid}}{\sim} G$, with PT prior $G \sim PT(\Pi, \mathcal{A})$. The sample $(x_1, \ldots, x_n)$, also referred to as *marginal sample*, can be generated by first sampling x_1 from G_0 then sampling $x_2, \cdots, x_n$ using Algorithm 1. The procedure is as follows:

1. Generate $x_1 \sim G_0$

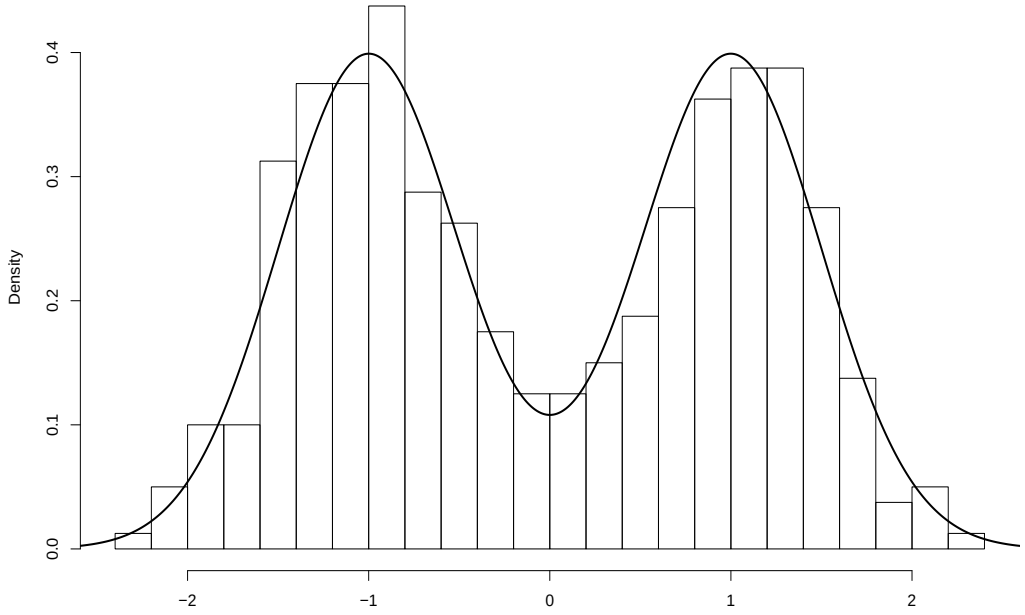

Figure 3.2: True density and a histogram of n=400 simulated observations.

2. Iterate over $i = 2, \ldots, n$:

   Use Algorithm 1 to generate $x_i \sim p(x_i \mid x_1, \ldots, x_{i-1})$.

3.2.5 Posterior Mean

Algorithm 1 provides a procedure to sample the posterior predictive density, which may be discontinuous at the countably infinite partition points. A minor variation of Algorithm 1 can be used to compute the posterior mean $E(G(B_{\varepsilon_1 \ldots \varepsilon_m}) \mid x_1, \ldots, x_n)$, i.e., $p(x_{n+1} \in B_{\varepsilon_1 \ldots \varepsilon_m} \mid x_1, \ldots, x_n)$. The proposed algorithm replaces the generation of random (conditional) probabilities for each of the nested partitioning intervals by the expected random probability,

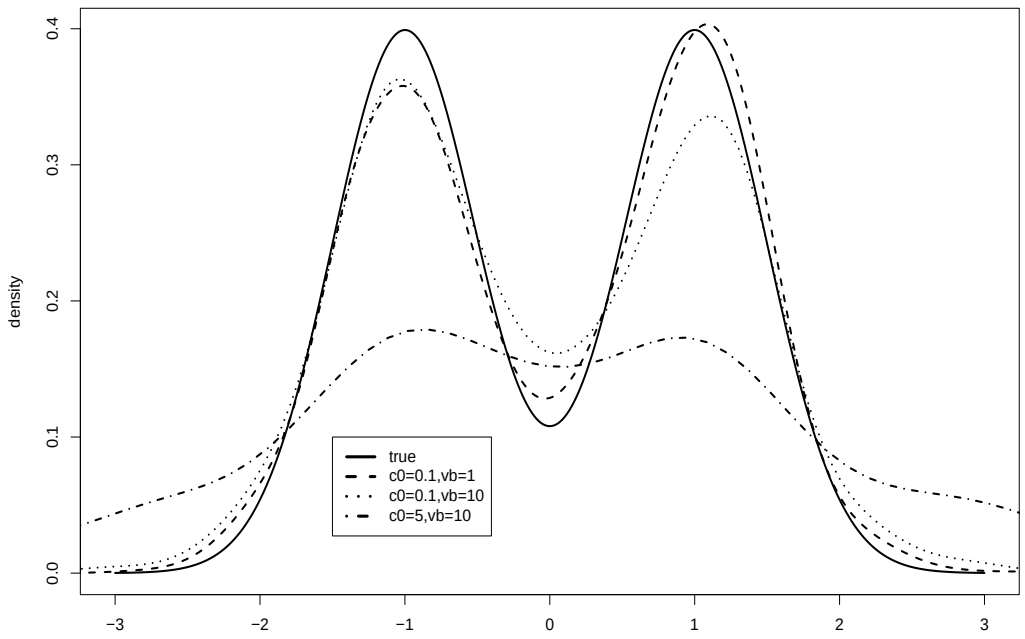

Figure 3.3: Predictive densities for different c and σ.

leading to the expected measure, $E(G \mid data)$, instead of a draw, $G \sim p(G \mid data)$. The procedure can be described as follows. Let $x = x_1, \ldots, x_n$ denote the observations from G. Consider some maximum level M, say $M = 10$. For all levels $m = 1, \ldots, M$ compute

$$\overline{Y}_{\varepsilon_1\varepsilon_2\cdots\varepsilon_{m-1}0} = E[Y_{\varepsilon_1\varepsilon_2\cdots\varepsilon_{m-1}0} \mid x] = \frac{\alpha'_{\varepsilon_1\varepsilon_2\cdots 0}}{\alpha'_{\varepsilon_1\varepsilon_2\cdots 0} + \alpha'_{\varepsilon_1\varepsilon_2\cdots 1}}.$$

where $\alpha'_{\varepsilon_1\cdots\varepsilon_m} = \alpha_{\varepsilon_1\cdots\varepsilon_m} + \sum_{i=1}^{n} I(x_i \in B_{\varepsilon_1\cdots\varepsilon_m})$. Recall that $\overline{Y}_{\varepsilon_1\varepsilon_2\cdots\varepsilon_{m-1}1} = 1 - \overline{Y}_{\varepsilon_1\varepsilon_2\cdots\varepsilon_{m-1}0}$ is the complement to one. Computing $\overline{Y}_{\varepsilon}$ is most elegantly implemented as a recursion. Let $\mathcal{S}_1, \cdots, \mathcal{S}_r$, $r = 2^M$ denote the subsets at level M. We find

$$E(G(\mathcal{S}_\varepsilon) \mid x_1, \ldots, x_n) \approx \prod_{m=1}^{M} \overline{Y}_{\varepsilon_1\epsilon_2\cdots\varepsilon_m} \equiv \hat{G}(\mathcal{S}_\varepsilon).$$

where $\varepsilon = \varepsilon_1 \cdots \varepsilon_M$. Here $G(\cdot)$ is the c.d.f. for the random measure G.

3.2.6 Simulation of A Posterior Draw

To generate random posterior draws $G \sim p(G \mid x_1, \ldots, x_n)$ can be proceeded as in section 3.2.5, replacing $\overline{Y}_\varepsilon$ by $Y_{\varepsilon 0} \sim Beta(\alpha'_{\varepsilon 0}, \alpha'_{\varepsilon 1})$, where $\alpha'_{\varepsilon 0} = \alpha_{\varepsilon 0} + \sum_i I(x_i \in B_{\varepsilon 0}$ and $\alpha'_{\varepsilon 1} = \alpha_{\varepsilon 1} + \sum_i I(x_i \in B_{\varepsilon 1}$. Let $G(\mathcal{S}_\varepsilon) = \prod_{m=1}^{M} Y_{\varepsilon_1 \varepsilon_2 \cdots \varepsilon_m}$. Plotting G against ε shows a random posterior draw of G. Plotting multiple draws G_i, $i = 1, 2, \ldots, I$ in the same figure illustrates uncertainty on the random measure.

Suppose that $x_1, \ldots, x_n$ come from (3.3). Assuming parameters $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = cm^2$, Figure 3.4 illustrates posterior draws of G under different c, where G is centering around $G_0 = N(0, \sigma^2)$. σ is fixed at 10. It is noted that posterior draws of G become smoother as c increases. This is because $G\{(-\infty, t]\} \mid x_1, \cdots x_n \rightarrow_d G_0(t)$ as $c \rightarrow \infty$, here G_0 is a smooth distribution.

3.3 Posterior Inference in Regression

In the above, we have discussed the posterior inference of PT models under i.i.d sampling model. In this section, we will discuss the posterior inference when PT is as a prior of the distribution of random residuals or the distribution of random effects in regression models.

Walker and Mallick (1997) assigned a Polya tree prior to the unknown distribution of random effects in a generalized linear model. Walker and Mallick (1999) and Hanson and Johnson (2002) considered an accelerated failure time model by specifying a Polya tree prior to the error distribution.

3.3.1 A PT Random Probability Measure As A Prior of Residual Distribution

Here we first discuss the model presented by Walker and Mallick (1999) who assigned a simple Polya tree prior to the distribution of regression error. Then we discuss how Hanson

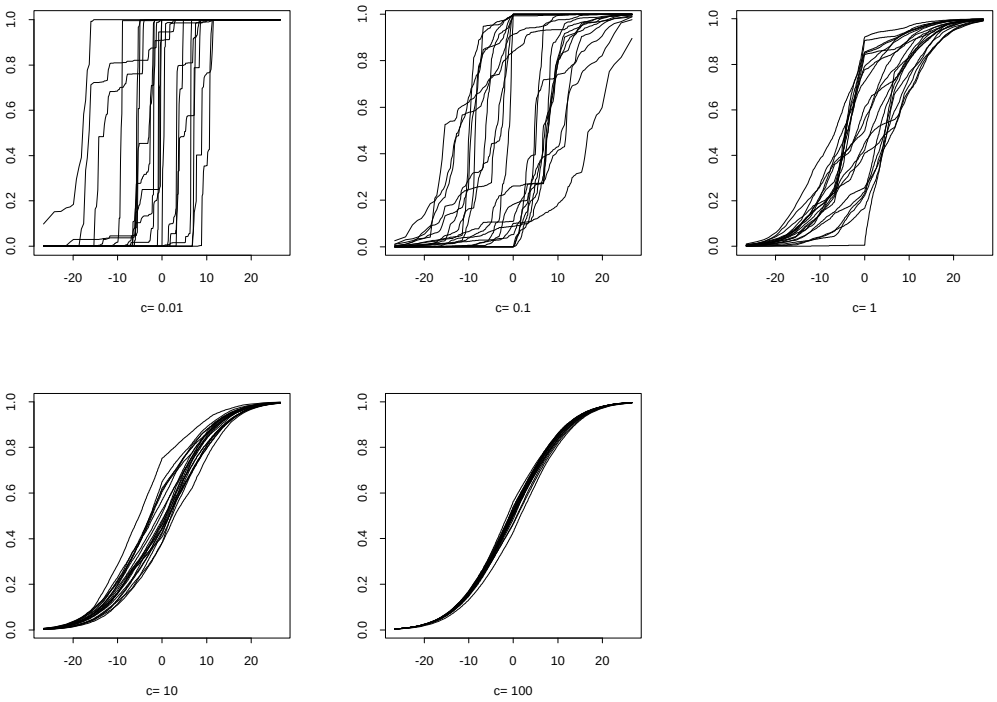

Figure 3.4: Posterior draw of $G \mid x_1, \ldots, x_n$.

and Johnson (2002) extended this model to mixture of Polya trees. The model is described as follows.

$$Y_i = -x_i'\beta + W_i, \quad \text{for} \quad i = 1, \cdots, n, \quad \beta \sim f_\beta(\beta)$$
$$W_1, \cdots, W_n \mid G \overset{iid}{\sim} G, \quad G \mid \Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A}) \tag{3.4}$$

Firstly, the distribution of error term is given a finite Polya tree prior, which is denoted as $G \mid \Pi^M, \mathcal{A}^M \sim PT(\Pi^M, \mathcal{A}^M)$. The Polya tree distribution is centered around a normal distribution G_0 with median (mean) 0 and a fixed large variance $\theta(i.e. G_0 = N(0, \theta))$. A fixed partition at level m is then generated by taking

$$B_{\varepsilon_1 \cdots \varepsilon_m} = (G_0^{-1}(\frac{j}{2^m}), G_0^{-1}(\frac{j+1}{2^m})], \tag{3.5}$$

where $j = 0, 1, \cdots, 2^m - 1$. $\mathcal{A}$ is assumed as

$$\mathcal{A} = \{\alpha_{\varepsilon_1 \cdots \varepsilon_m} = cm^2 : \varepsilon_1 \cdots \varepsilon_m \in \{0, 1\}^m\}, \tag{3.6}$$

with a fixed value of c. Walker and Mallick (1997, 1999) choose M to be 8. M can also be chosen such that at most one observation is in each partitioning subset at level M. For posterior simulation, we require the full conditional distributions $p(G|Y, \beta)$ and $p(\beta|Y, G)$.

Obtaining the posterior distribution of $p(G|Y, \beta)$ is straightforward. Since $W_i = Y_i + x_i'\beta$ are i.i.d. from G, i.e. $p(G|Y, \beta) = p(G \mid W_1, \cdots, W_n)$, we find that G can be generated by the algorithm described in Section 3.2.6. At level M, G is given as a sequence of probabilities, $\{V_k, k = 1, \cdots, 2^M\}$, on the sets at level M, say $S_1, \cdots, S_{2^M}$.

To sample β, the full conditional distribution is given by

$$p(\beta, \mid Y, G) \propto p(Y \mid G, \beta) f_\beta(\beta).$$

where $p(Y \mid G, \beta)$ is the likelihood for β given Y and G

$$p(Y \mid G, \beta) = \prod_{i=1}^{n} \left(\sum_{k=1}^{2^M} V_k I(Y_i + x_i'\beta \in S_k) \right)$$

A Metropolis-Hastings algorithm (Tierney 1994) can be used to sample β. A proposed sample for β can be taken from a multivariate normal with mean at current value and covariance matrix to be a scaled identity matrix.

Alternatively, in the traditional linear model with $W_1, \cdots, W_n \sim G_0$, the posterior distribution of β is proportional to $f_\beta(\beta) \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i + x_i'\beta)^2}{2\sigma})$. If β has a normal prior $N(\mu_\beta, \Sigma_\beta)$, then the posterior distribution of β is normal with mean $(\Sigma_\beta^{-1} + \sigma^{-2}x'x)[\Sigma_\beta^{-1}\mu_\beta - \sigma^{-2}x'y]$ and variance-covariance matrix $(\Sigma_\beta^{-1} + \sigma^{-2}x'x)^{-1}$, where $x = -(x_1, \cdots, x_n)'$.

In the above finite Polya tree, posterior simulation MCMC requires to sample G at each iteration. Now we consider another method for the posterior simulation. The random unknown probability measure G is integrated out and inference is based on the predictive distribution (Hanson and Johnson 2002). As a result, inference is exact up to MCMC error. To be specific, the joint conditional density is $p(W_1, \cdots, W_n \mid G) = \prod_{i=1}^n g(W_i)$, where g is the density function of G. After G is integrated out, we have

$$\begin{aligned} f(W_1, \cdots, W_n) &= \int p(W_1, \cdots, W_n \mid G) dp(G) \\ &= f(W_1) f(W_2 \mid W_1) \ldots f(W_n \mid W_1, \cdots W_{n+1}) \\ &= g_0(W_1) \prod_{i=2}^n f(W_i \mid W_1, \cdots, W_{i-1}) \end{aligned}$$

where g_0 is the density function of the centering distribution G_0. Using (3.1), we have

$$f(W_i \mid W_1, \cdots, W_{i-1}) = \left\{ \lim_{m\to\infty} \prod_{j=2}^m \frac{\alpha'_{\epsilon_1\cdots\epsilon_j} / (\alpha'_{\epsilon_1\cdots\epsilon_{j-1}0} + \alpha'_{\epsilon_1\cdots\epsilon_{j-1}1})}{\alpha_{\epsilon_1\cdots\epsilon_j} / (\alpha_{\epsilon_1\cdots\epsilon_{j-1}0} + \alpha_{\epsilon_1\cdots\epsilon_{j-1}1})} \right\} g_0(W_i) \tag{3.7}$$

where $\alpha'_{\epsilon_1\cdots\epsilon_j}$ is $\alpha_{\epsilon_1\cdots\epsilon_j}$ plus the number of elements of $W_1, \cdots, W_{i-1}$ in $B_{\epsilon_1\cdots\epsilon_j}$ such that $W_i \in B_{\epsilon_1\cdots\epsilon_j}$. As indicated before, the right side of (3.7) can be reduced to a limit product.

Under model (3.4) with Π and $\mathcal{A}$ defined in (3.5) and (3.6), respectively, the predictive density of error term $W_i \mid W_1, \cdots, W_{i-1}$ has the following close form

$$f_{W_i}(w \mid W_1, \cdots, W_{i-1}) = \lim_{m\to\infty} \left\{ \prod_{j=2}^m \frac{cj^2 + n_{\varepsilon(j,w)}}{2cj^2 + n_{\varepsilon(j-1,w)}} \right\} 2^{m-1} g_0(w) \tag{3.8}$$

where $n_{\varepsilon(j,w)}$ is the number of elements of $W_1, \ldots, W_{i-1}$ in the subset $B_{\varepsilon_1 \cdots \varepsilon_j}$ which contains w. g_0 is the density of base measure G_0 with respect to Lebesgue measure and has median 0.

The posterior distribution of $\beta \mid Y$ is given by

$$\begin{aligned} f_\beta(\beta \mid Y) &\propto f_\beta(\beta) f_Y(Y \mid \beta) \\ &= f_\beta(\beta) g_0(Y_1 + x_1'\beta) \prod_{j=2}^{n} f_{W_j \mid W_1, \cdots, W_{j-1}}(Y_j + x_j'\beta \mid Y_1 + x_1'\beta, \cdots, Y_{j-1} + x_{j-1}'\beta) \end{aligned} \tag{3.9}$$

Metropolis-Hastings algorithm can be used to obtain posterior inference. A suitable candidate generating density, $q(\beta^\star \mid \beta)$, where $q(\beta^\star \mid \beta) = q(\beta \mid \beta^\star)$, is chosen. Define the acceptance probability $\alpha(\beta^\star, \beta)$ as

$$\alpha(\beta^\star, \beta) = min \left\{ \frac{f_\beta(\beta^\star) f_{W_1, \cdots, W_n}(Y_i - x_i'\beta^\star, \cdots, Y_n - x_n'\beta^\star)}{f_\beta(\beta) f_{W_1, \cdots, W_n}(Y_i - x_i'\beta, \cdots, Y_n - x_n'\beta)}, 1 \right\} \tag{3.10}$$

At the kth iterate in the Markov chain we sample $(\beta^\star \mid \beta^k) \sim q(\beta^\star \mid \beta^k)$ and take

$$\beta^{k+1} = \begin{cases} \beta^\star & \text{with probability } \alpha(\beta^\star, \beta) \\ \beta^k & \text{with probability } 1 - \alpha(\beta^\star, \beta) \end{cases}$$

We construct a multivariate normal random walk candidate generating distribution for sampling $\beta^\star$ as follows. The covariance matrix is chosen to be as nearly as possible to the true posterior covariance matrix, perhaps by obtaining a crude estimate of covariance matrix from an initial sampling run (Müller 1991). Or the covariance matrix is chosen to provide a reasonable acceptance rate.

Alternatively, a guided walk Metropolis algorithm (Gustafson 1998) can be used to update β. A transition of the guided walk algorithm from (β^k, p^k) to (β^{k+1}, p^{k+1}) proceeds as

$$\beta^\star \leftarrow \beta^k + p^k |z|$$

where $z \sim N(0, s^2)$ and $p \in \{-1, +1\}$ satisfying $Pr(p = +1) = Pr(p = -1) = 0.5$. $\alpha(\beta, \beta^\star)$ is computed by (3.10). Then $(\beta^{k+1}, p^{k+1}) \leftarrow (\beta^\star, p^k)$ with probability $\alpha(\beta, \beta^\star)$, or $(\beta^{k+1}, p^{k+1}) \leftarrow (\beta^k, -p^k)$ with probability $1 - \alpha(\beta, \beta^\star)$.

In contrast to the random walk algorithm, the direction of the candidate $\beta^\star$ relative to the current state β^k is not random at each iteration. The chain moves consistently in the same direction until a candidate is rejected. Compared to the random walk Metropolis algorithm, the guided walk metropolis algorithm performs better in terms of efficiency and convergence time.

3.3.2 A PT Random Probability Measure as A Prior of Random Effect Distribution

For reasons of technical convenience, random effects are often assumed to come from the normal family of distributions (Zeger and Karim 1991). Kleinman and Ibriham (1998) extended the models to consider Dirichlet process prior for the random effects. Walker and Mallick (1997) assigned a Polya tree prior to the distribution of random effects in generalized linear model. Their approach involves centering the median of the random effects distribution at 0 and specifying the Polya tree to the finite level.

In this section, we attempt to improve on certain aspects of the aforementioned methods in the random effect model setting. In particular, we present a random effect model in which the random effects are modelled as a Polya tree centered about a parametric probability distribution. With our approach, random effects can be directly sampled and inference will be based on the predictive density.

Without loss of generality, consider the fully specified random effect model

$$\begin{aligned} z_i &= x_i'\beta + \theta_i + w_i \\ \beta &\sim f_\beta(\beta) \\ w_i &\sim N(0, \sigma^2 I_k) \\ \theta_i &\overset{iid}{\sim} F, \quad F \sim PT(\Pi, \mathcal{A}) \\ i &= 1, \cdots, n, \; j = 1, \cdots, k \end{aligned} \tag{3.11}$$

with Π defined by (3.5) and $\mathcal{A}$ defined by (3.6). x_i is a $p \times k$ design matrix. β is a $p \times 1$ vector of fixed effects and θ_i is an individual random effect. A vague normal prior is taken for β and an inverse Gamma prior is assigned to σ^2, that is, $\beta \sim N_p(\mu_\beta, \Sigma_\beta)$ and $\sigma^2 \sim IG(a_1, a_2)$. Therefore the relevant full conditional distributions include $p(\beta \mid z, \theta, \sigma^2)$, $p(\sigma^2 \mid z, \beta, \theta)$, and $p(\theta_i \mid z, \theta_{-i}, \beta, \sigma^2)$, where $\theta_{-i} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_n), i = 1, \cdots, n$. Samples can be obtained using an MCMC algorithm and in particular a Metropolis-Hastings within Gibbs methods.

It is easy to show that the fully conditional distribution for β is normal with mean and covariance matrix given by $(\Sigma_\beta^{-1} + \sigma^2(x'x)^{-1})^{-1}(\Sigma_\beta^{-1}\mu_\beta + \sigma^2(x'x)^{-1}x'(z - \theta))$ and $(\Sigma_\beta^{-1} + \sigma^2(x'x)^{-1})^{-1}$, respectively. Here, x' is $nk \times p$ matrix with x_{ij} as the jth column of matrix x_i.

The full condition for σ^2 is an inverse gamma with updated parameters, $a_1 + n/2$ and $a_2 + \sum_{ij}(z_{ij} - x_{ij}'\beta - \theta_i)^2/2$.

We now describe how to update $p(\theta_i \mid z, \theta_{-i}, \beta, \sigma^2)$. The likelihood for θ_i, given z, σ^2, β, and θ_{-i}, is given by

$$l(z \mid \beta, \theta_i, \theta_{-i}, \sigma^2) = \prod_{i=1}^{n}\prod_{j=1}^{k} N(z_{ij} \mid \beta, \theta_i, \sigma^2)$$

Then the conditional distribution for θ_i can be written, up to a constant of proportionality. That is,

$$\begin{aligned} p(\theta_i|z, \theta_{-i}, \beta, \sigma^2) &\propto l(z \mid \beta, \theta_i, \sigma^2) p(\beta, \theta_i, \theta_{-i}, \sigma^2) \\ &\propto l(z \mid \beta, \theta_i, \sigma^2) p(\theta_i \mid \beta, \theta_{-i}, \sigma^2) p(\beta, \theta_{-i}, \sigma^2) \\ &\propto l(z \mid \beta, \theta_i, \sigma^2) p(\theta_i \mid \theta_{-i}) \end{aligned}$$

for each $i = 1, \cdots, n$. This can be done using a Metropolis-Hastings algorithm.

We construct Polya tree with mean measure F_0 and center the Polya tree prior at F_0. F_0 is the cumulative distribution function corresponding to a density f_0. It is reasonable to take F_0 as a normal distribution with median zero and a large variance. We consider the fixed partition defined by (3.5) and choose cm^2 for the parameters α's. At each level $m = 1, 2, \cdots$, we sample the set $B_{\varepsilon_1 \cdots \varepsilon_m}$ using Algorithm 1 described in section 3.2.3. $\theta_i^{(p)}$ is generated from f_0 constrained on $B_{\varepsilon_1 \cdots \varepsilon_m}$ once a level is reached such that no observation falls into $B_{\varepsilon_1 \cdots \varepsilon_m}$. The candidate $\theta_i^{(p)}$ is generated from the predictive distribution $p(\theta_i \mid \theta_{-i})$. The candidate is accepted with probability given by

$$min \left\{ 1, \frac{l(z_i \mid \beta, \theta_i^{(p)}, \sigma^2)}{l(z_i \mid \beta, \theta_i^{(c)}, \sigma^2)} \right\},$$

where $\theta_i^{(p)}$ is the proposed sample and $\theta_i^{(c)}$ is the current sample.

3.4 Examples

In this section, we use three examples to illustrate the aforementioned methods. The first example emphasizes the application of Polya tree as the prior of error distribution. The second example will focus on the implementation of Polya tree as prior of random effects distribution. In the third example, we used the simulated fractional dataset considered in Chapter 1 to demonstrate how a Polya tree distribution can be used as the prior of the distribution of random effects in the fractional data model.

3.4.1 Polya Tree Distribution as a Prior of Error Distribution

In this subsection, we will use two simulated datasets to examine the fit of model (3.4) and compare the estimates based on Polya tree prior with the results obtained based on the normal distribution. For simplicity, only one covariate (x_1) is considered and generated from $N(2,1)$. The true regression coefficient of this covariate was set at 1.

First, we let the random residual be sampled from $N(0, 0.15^2)$. Then we consider the random residuals sampled from a mixture of two normal distributions $0.5N(-0.2, 0.15^2) + 0.5N(0.2, 0.15^2)$. The median of residuals is set at 0, which is a standard practice in linear regression model. The size of the sample generated from both distributions is taken to be $n = 250$, large enough to provide an adequate representation of the distribution. Thus the observed Y is given by $Y_i = x_i + W_i$,where $W_i, i = 1, \cdots, n$ is simulated residual.

We fit both data sets under normal and Polya tree priors. Following the discussion in Section 3.2.3 for the choice of parameters $\alpha_{\varepsilon_1 \cdots \varepsilon_m}$ and G_0 in Polya tree prior, we take parameters $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = cm^2$ at the mth level with $c = 0.1$ and $G_0 = N(0, \sigma^2)$ where $\sigma = 5$. For the regression coefficients, we follow the standard approach, assuming $\beta \sim N_2(0, 100I_2)$, where I_2 is 2×2 identity matrix.

Table 3.1 summarized the estimated regression coefficients when true residuals come from a normal distribution. Results indicate that point estimates for the regression coefficients are fairly accurate under both normal and Polya tree prior distributions. The estimated G under Polya tree prior is plotted in Figure 3.5. The dashed line is the true density from which true residuals are generated. It is clearly demonstrated that nothing appears to be lost in the analysis using Polya tree even if the random residual distribution is a normal distribution.

The estimated regression coefficients are also presented in Table 3.1 when true residuals come from a mixture of normals. It is clearly to see that point estimates for the regression coefficients are fairly accurate under both normal and Polya tree prior distributions. The histogram of estimate residuals under normal assumption is displayed in Figure 3.6. It sug-

Table 3.1: Posterior regression estimates when residuals are sampled from normal and mixture of normals

	$W_i \sim N(0, 0.15^2)$		$W_i \sim 0.5N(-0.2, 0.15^2) + 0.5N(0.2, 0.1562)$	
Parameter	Normal	Polya tree	Normal	Polya tree
β_0	-0.02(-0.06, 0.03)	-0.03(-0.12,0.07)	0.03(-0.05, 0.10)	-0.02(-0.16, 0.19)
β_1	1.01(0.99, 1.03)	1.02(0.96, 1.06)	0.98(0.95, 1.02)	0.98(0.93, 1.05)

gests that normality assumption is not valid for this dataset. The estimated G under Polya tree prior is plotted in Figure 3.7. As can be seen the estimated G's are considerably close the the true density of residuals. Especially the estimated G can capture the multiple modes fairly well.

Although the point estimates for β's are accurate for all models, Polya tree prior adds more uncertainty to the model thus the high uncertainty occurs for β_0 and β_1 when the distribution of residual is assumed to have a Polya tree prior. If estimating regression coefficients is of interest, we will not loss much using traditional normal linear regression modeling when the true distribution of residuals is not normal. However, making prediction is also important in medical research, the prediction for a new patient will miss bimodal nature under normal assumption when the true distribution of residual is not normal.

3.4.2 Polya tree as prior of random effect distribution

We aim to estimate the distribution of random effects in the following example. The data set is constructed in which the actual distribution of random effects is much different than the family of base measures centering the Polya tree prior.

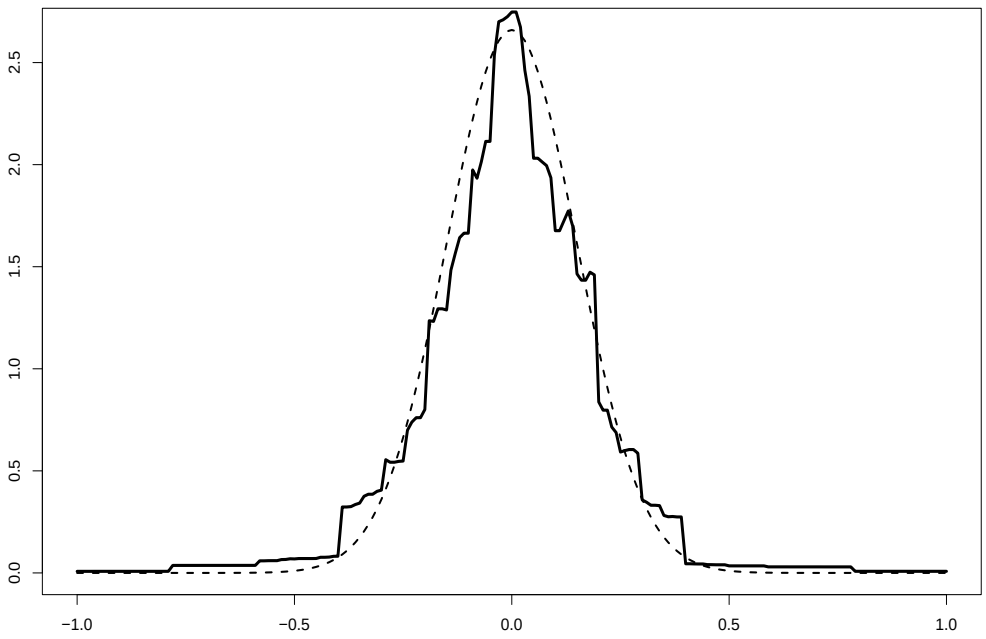

Figure 3.5: Posterior expectation of G when residuals are generated from a normal distribution. Dashed line is the true residual density. Solid line is the estimate of G under Polya tree prior

We simulate $n = 200$ data points from random effect model in which the baseline distribution for random effects is a normal mixture

$$\theta_1, \cdots, \theta_n \overset{iid}{\sim} 0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2).$$

Covariate $x_1, \cdots, x_n \overset{iid}{\sim} N(1, 1)$. The true vector of regression coefficients was set at $\beta = (0.8, 1)'$, the residuals (w's) were generated from $N(0, 0.5^2)$, and the observations are given by $z_i = (1, x_i)'\beta + \theta_i + w_i$.

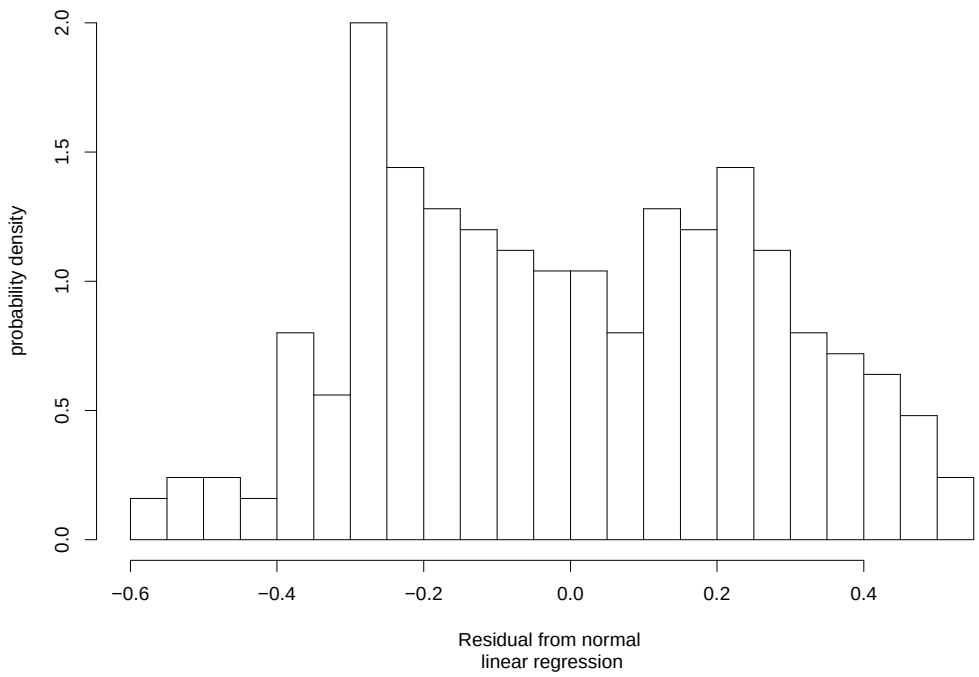

Figure 3.6: Histogram of residuals when residuals are generated from a mixture of normals.

As in section 3.4.1, we place a bivariate normal prior $N(0, 10I_2)$ for β. The residuals were assumed to follow a normal distribution with mean 0 and variance σ^2. An inverses gamma prior ($IG(0.1, 0.1)$) was assign to σ^2.

The base probability distribution for the Polya tree prior was taken to be normal with mean 0 and variance 25. The α-values in $\mathcal{A}$ were taken to be constant at level m and equal to cm^2. We fixed the values of c at 0.1 and 5 to assess the effect of c on the resulting inference.

The estimated regression coefficients are presented in Table 3.2. The point estimates of coefficients are accurate for all models. The posterior predictive density estimates for random effects are plotted in Figure 3.8. It suggests that small c do a better job at capturing the

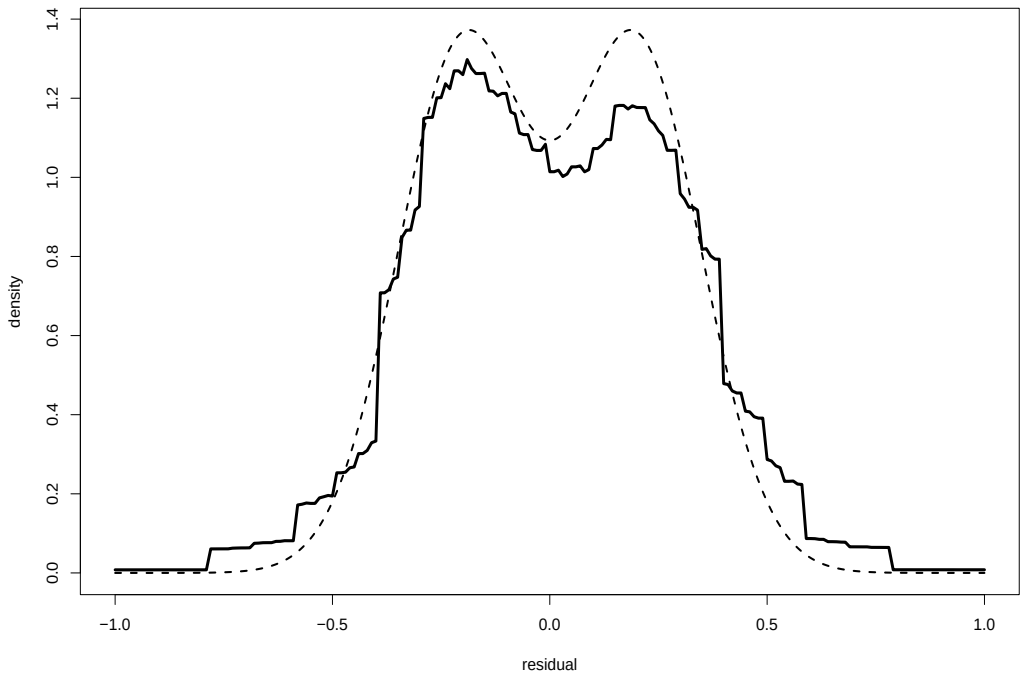

Figure 3.7: Posterior expectation of G when residuals are generated from a mixture of normals. Dashed line is the true residual density. Solid line is the estimate of G under Polya tree prior

bimodal of the true density than larger c because for large values of c we are performing a parametric analysis.

3.4.3 Polya tree as prior of random effect distribution in fractional data model

In this subsection we continue the fractional data example in Chapter 1, there the distribution of random effects was assumed to be normal or Dirichlet process. In the following, we place a Polya tree prior on the unknown distribution G, which is the distribution of random effects

Table 3.2: Posterior regression estimates for simulated dataset

	$c = 0.1$		$c = 5$	
Parameter	median	95% CI	median	95% CI
β_0	0.68	(0.34, 1.13)	0.85	(0.44, 1.40)
β_1	1.01	(0.85, 1.18)	0.92	(0.62, 1.22)
σ	0.56	(0.38, 0.74)	0.29	(0.15, 0.47)

Table 3.3: Posterior regression estimates for simulated fractional data

Parameter	median	95% CI
β_0	0.847	(0.720, 0.942)
β_1	-0.599	(-0.663, -0.534)
σ	0.093	(0.081, 0.109)

in fractional data model. The Polya tree distribution is assumed to center around $N(0, 5^2)$. At the mth level, $\alpha_{\varepsilon_1 \ldots \varepsilon_m}$ is taken to be cm^2. For simplicity, c is fixed to be 0.1. The estimated regression coefficients are presented in Table 3.3. The point estimates for β_0 and β_1 are fairly close to the true values. The posterior predictive density of θ_{n+1} is plotted in Figure 3.9. It shows that the estimated G captures the bimodal nature of random effect distribution very well.

3.5 Conclusion

We have introduced PT models for random sampling, for residual distributions in a regression model, and for random effects distributions in mixed effects models. We reviewed known algorithms for prior- and posterior simulation, and developed new algorithms for posterior

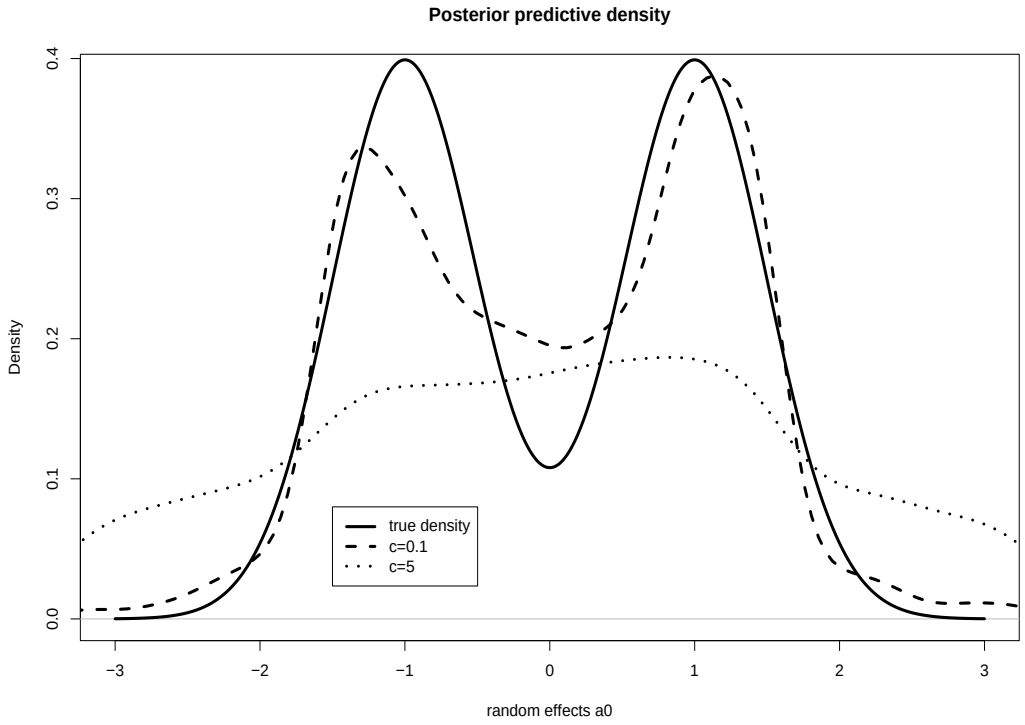

Figure 3.8: Posterior predictive density estimate for random effects.

simulation with analytically marginalized random measure and for evaluation of the posterior mean for the random distribution. Apart from implementations with finite PTs the use of PT priors on random effects distributions is new.

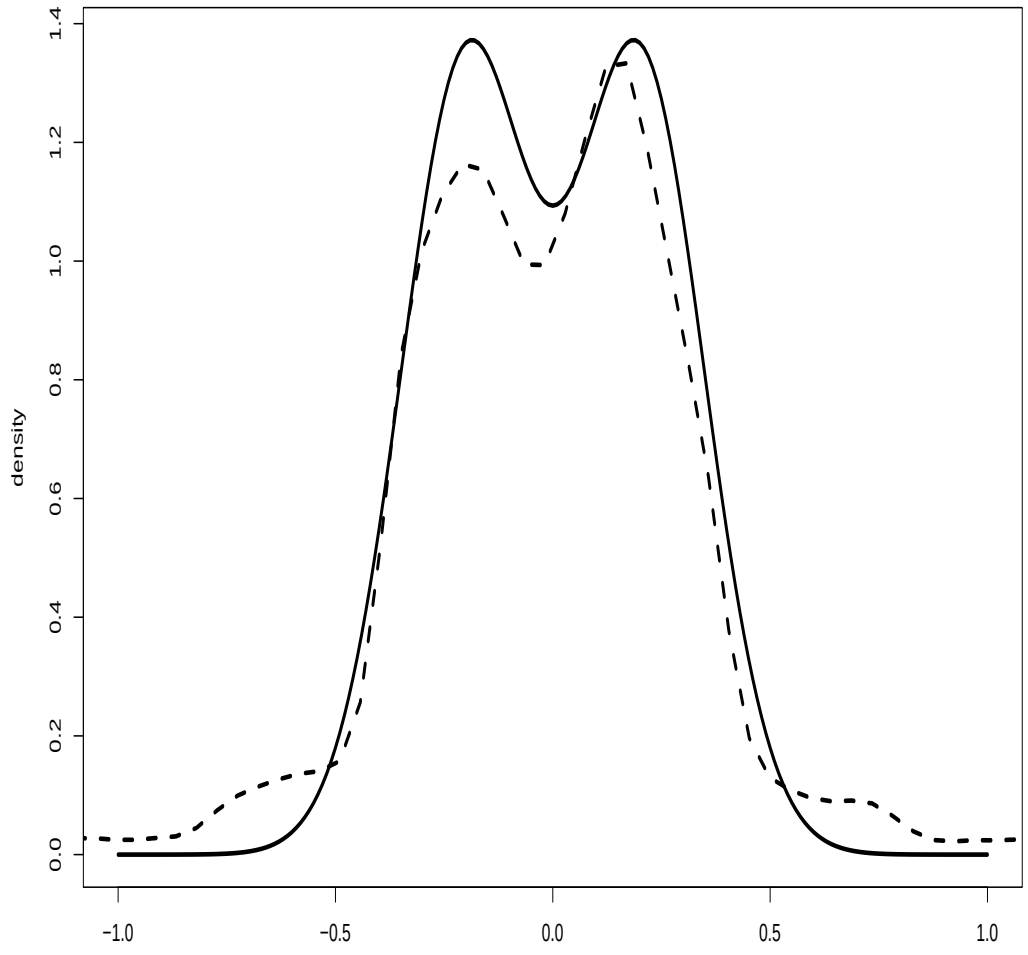

Figure 3.9: Posterior predictive densities of θ, solid line is the true density of G. The dashed line is the estimated G

Chapter 4

Bayesian Data Analysis for Some Biomedical Applications

4.1 Introduction

In this chapter we present three case studies to complete the development of the models introduced earlier. In section 4.1. we use a PT prior for the residual distribution in a regression model. We show how the models and algorithms developed in the earlier chapters are used in the data analysis and compare posterior inference under the proposed nonparametric model with a standard parametric model.

In section 4.2. we further develop the nonparametric random effects model introduced in section 3.3.2. In the context of a case study we show how the proposed nonparametric model for random effects with many corresponding experimental units is combined with a parametric random effects distribution for another set of random effects in the same model. We compare the resulting inference with a standard parametric model, highlighting the nature of the nonparametric extension.

Finally, in section 4.3. we develop a new nonparametric Bayesian approach to inference for differential gene expression in microarray group comparison experiments. The proposed inference is a natural nonparametric Bayesian extension of a popular empirical Bayes approach introduced in Efron et al. (2001). We define the model, explain appropriate posterior simulation algorithms and compare results with the empirical Bayes method. The results show how the nonparametric Bayesian model addresses some of the critical limitations of the empirical Bayes method. In particular, the nonparametric Bayes method explicitly acknowledges the

uncertainty in the unknown quantities and provides a full probabilistic description of the uncertainties.

4.2 Cell line data

HER2 represents an appealing target for humoral and cellular immunotherapy because HER2 is expressed at high levels in a variety of human cancers, such as breast cancer, ovarian cancer. Overexpression of antigenic proteins, such as HER2, is coupled with a high protein turnover, leading subsequently to a high number of major histocompatibility complex proteins (MHC) class I peptide complexes on the cell surface. Therefore, HER2 positive tumor cells are potentially good target cells for tumor-reactive cytotoxic T cells(CTLs) recognizing HER2-derived peptides in context with MHC class I molecules. Lytic activity of separated CTLs isolated from normal donors and patients with HER2 positive tumors was found to be low because HER2 is a self-antigen attributable to induction of tolerance. This study is to investigate whether the lytic potential of HER2-specific CTLs could be improved with the help of Herceptin, an inhibitory antibody against HER2, and monocyte, a white blood cell which can ingest dead cell or damaged cells.

The HER2 positive ovarian cancer cell line SKOV3 is obtained and treated by the combination of Herceptin and monocyte at different levels. Herceptin has 4 levels: 0,1,5,10($\nu g/ml$). Monocyte has 4 levels: 0, 10:1, 20:1, 40:1, which are coded as 0,1,2,and 3. The percentage of specific lysis was calculated as follows on the SKOV3 cell lines from 43 patients:

percent specific lysis=(experimental release-spontaneous release)/(maximum release - spontaneous release) $\times$ 100.

Here experimental release is the count from cell-free culture supernatant from MO/MA incubated with labelled target cells; spontaneous release is the count from cell-free culture supernatant from labelled target cells only, and maximum release is the count from lysed cells of labelled tumor cells only.

Let y_i denote the percentage of specific lysis from patient i. To constrain the percentage to be positive, We add 0.1 to y_i and then perform a logit transformation, that is, $z_i = logit(y_i + 0.1)$, where $logit(p) = log(\frac{p}{1-p})$. The histogram of z_i is presented in Figure 4.1. To illustrate the possible interaction between Herceptin and monocyte, an interaction plot of these two factors is displayed in Figure 4.2.

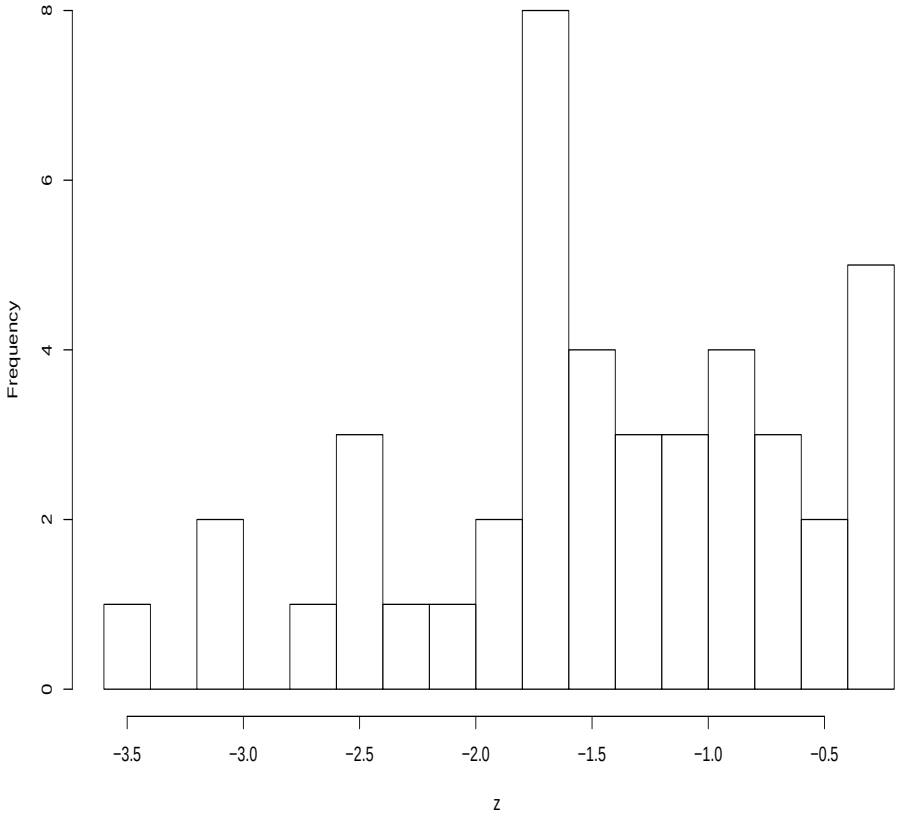

Figure 4.1: Histogram of cell line data.

A linear regression model is used to fit the data. The model is given by

$$z_i = \beta_0 + \beta_1 H_i + \beta_2 M_i + \beta_{12} H_i * M_i + \varepsilon_i$$

$$i = 1, \cdots, n$$

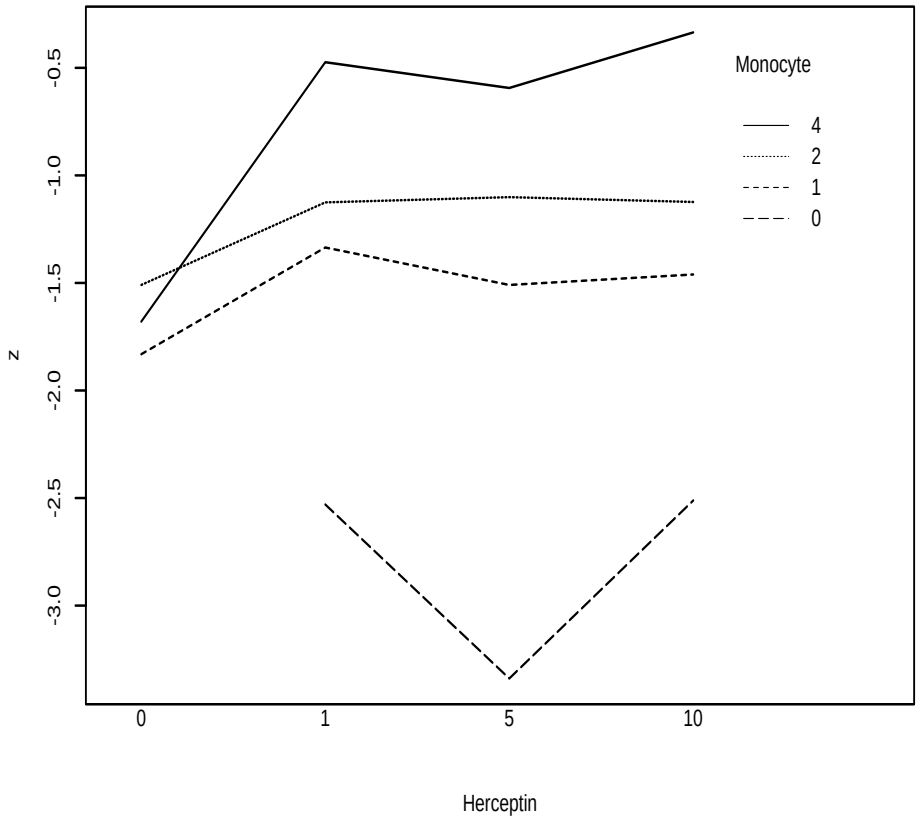

Figure 4.2: Interaction plot of Herceptin and monocyte.

Here H is the level of Herceptin. M is the level of monocyte, and $H * M$ is the interaction of Herceptin and monocyte.

PROC GLM procedure in SAS was firstly used to fit a conventional linear regression model. The estimated regression coefficients are summarized in Table 4.1. The histogram of estimated residuals is presented in Figure 4.3. The highly skewed nature of the histogram indicates that the normal assumption is not valid for this study. To test the normality

assumption, we display Q-Q plot of estimated residuals in Figure 4.4. Also, the Q-Q plot suggests possible skewness and fat tails of the distribution of residuals.

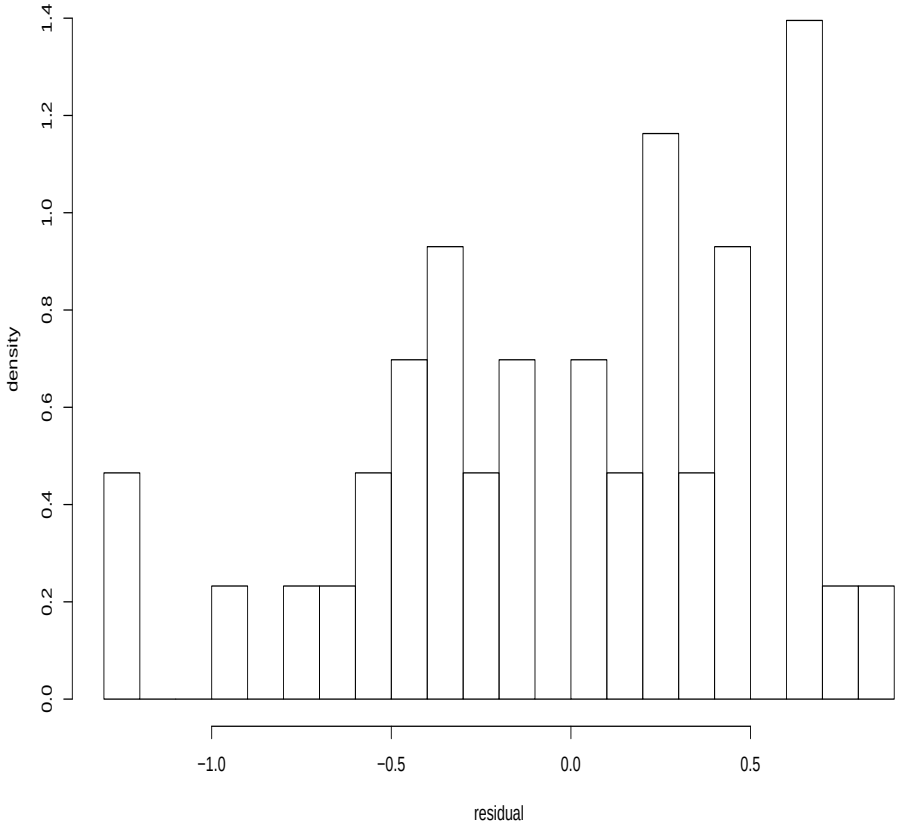

Figure 4.3: Histogram of estimated residuals under normal assumption.

Instead of assuming a normal distribution for residuals, we let residuals ε_i follow an unknown distribution, i.e., $\varepsilon_i \sim G$, and a Polya tree distribution is proposed to be the prior of the unknown distribution, i.e., $G \sim PT(\Pi, \mathcal{A})$. We take the default choices for the parameters (Π and $\mathcal{A}$) of the Polya tree. The partition sequence is chosen to center the Polya tree prior around a normal distribution with mean 0 and standard deviation of 5. The partition points coincide with percentiles of the centering distribution. For the assignment

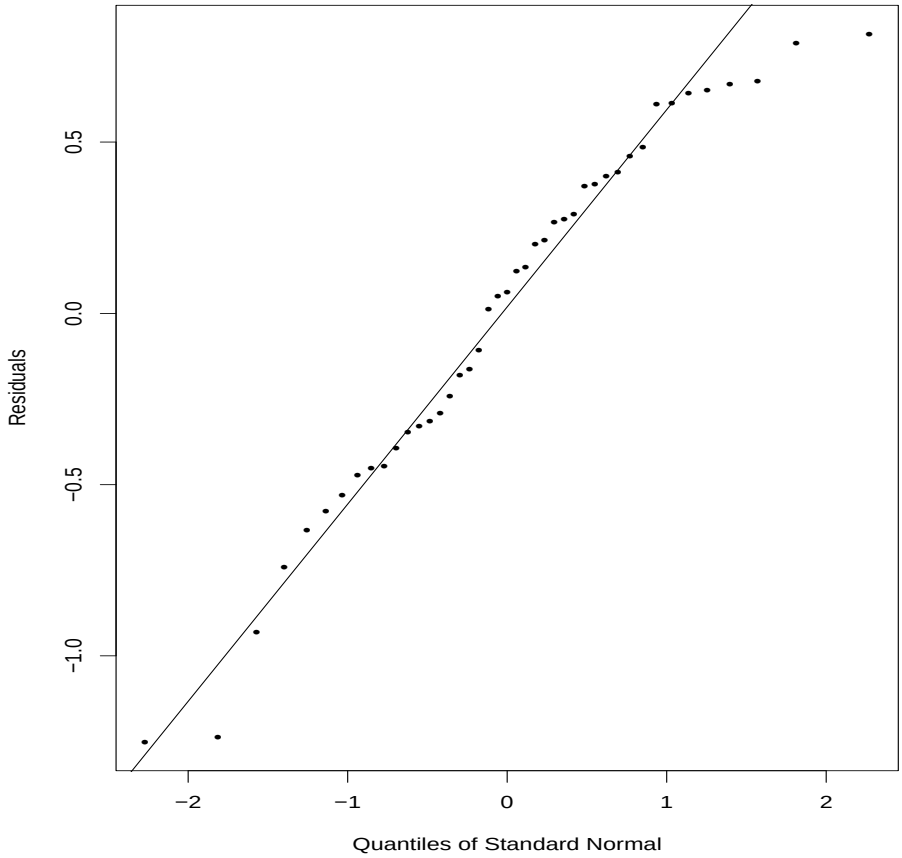

Figure 4.4: Q-Q plot of estimated residuals.

of $\mathcal{A}$, we use $\alpha_{\epsilon 0} = \alpha_{\epsilon 1} = cm^2$ for each ϵ at level m. The value of c is set at a fixed value of 10 in this example. For β, a normal prior with mean 0 and covariance matrix $20I_4$ is placed on $\beta = (\beta_0, \beta_1, \beta_2, \beta_{12})$, where I_4 is 4×4 identity matrix.

The MCMC algorithm is run for 30000 iterations. The first 3000 iterations are discarded as burn-in period, then samples are selected every 12th iteration, leaving us with 2250 Monte Carlo samples. Geweke's test (1992) was used to perform convergence diagnostics.

Table 4.1: Parameter estimates and 95% confidence interval

Parameter	Normal distribution median (95% CI)	Polya tree prior median (95% CI)
β_0	-1.448(-1.617, -1.269)	-1.612(-3.069, -0.056)
β_1	0.034(-0.009, 0.077)	0.034(-0.175, 0.256)
β_2	0.396(0.276, 0.513)	0.407(-0.165, 1.003)
β_{12}	0.032(0.004, 0.062)	0.031(-0.120, 0.183)

Posterior estimates for regression parameters are presented in Table 4.1. Noted the dramatically increased uncertainty for β_2 and β_{12} under the Polya tree prior. The estimated posterior densities of regression parameters are displayed in Figure 4.5.

The posterior expectation $E(G \mid z)$ of the unknown residual distribution is shown in Figure 4.6. The distribution in Figure 4.6 indicates the large variability of toxicities. As can be seen this is a distribution with fat tails as we expected. This large variability should be taken into account in any modeling procedure. The advantage of using Polya trees is that no specific form for the distribution is assumed.

The simulation-based implementation of posterior inference allows us to report inference on any event of function of the parameters of interest. In particular, we report posterior predictive inference for a future observation. Formally, this is $p(z_{n+1} \mid H_{n+1}, M_{n+1}, z_1, \cdots, z_n)$. Figure 4.7 displays predictive density curves for $H_{n+1} = 0$ and $M_{n+1} = 1$, i.e., Herceptin equal to 0 and monocyte equal to 1, and Herceptin equal to 1 and monocyte equal to 0. The predictive density curves of new observations are quite different from a normal density. In particular, we see skewness and multimodality are present.

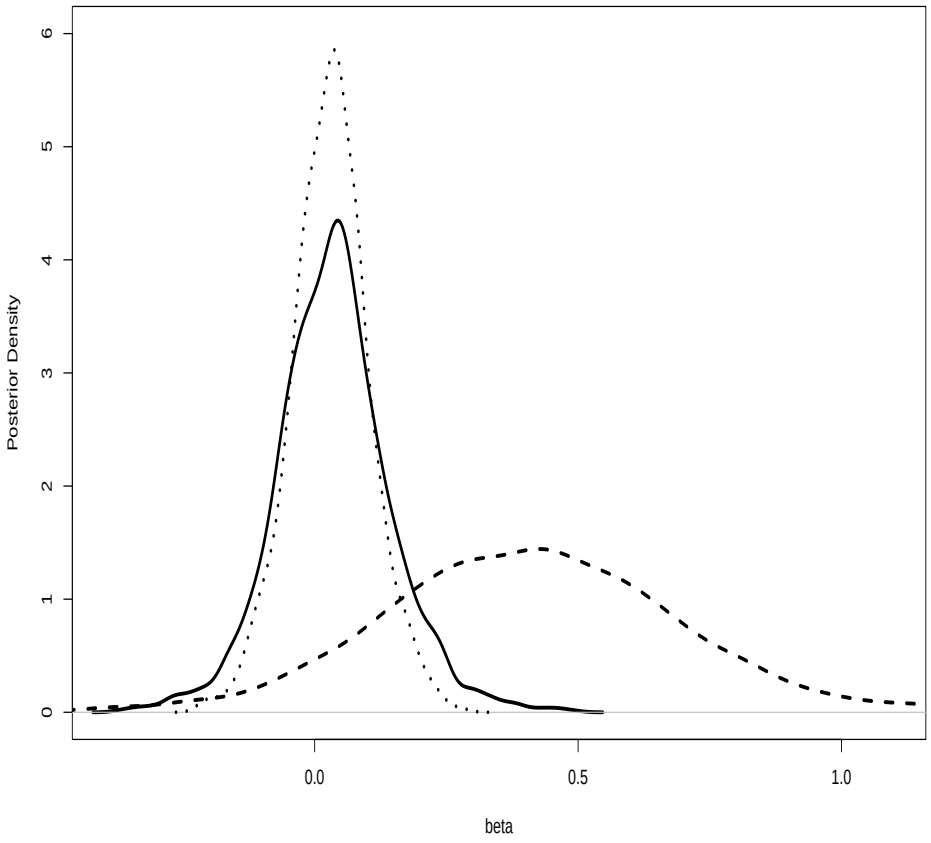

Figure 4.5: Posterior density estimates of βs, the solid line is the posterior density of β_1, the dashed line is the posterior density of β_2, and the dotted line is the posterior density of β_{12} .

4.3 Repeated Fractional measurement data

In cancer studies, some characteristics of the tumor microenvironment are often assessed by histologic evaluation of tumor biopsies. These include oxygenation and proliferation, two properties of the tumor environment which may influence the response to treatment. One approach to measure tumor oxygenation is the detection of bound nitroimidazoles based on

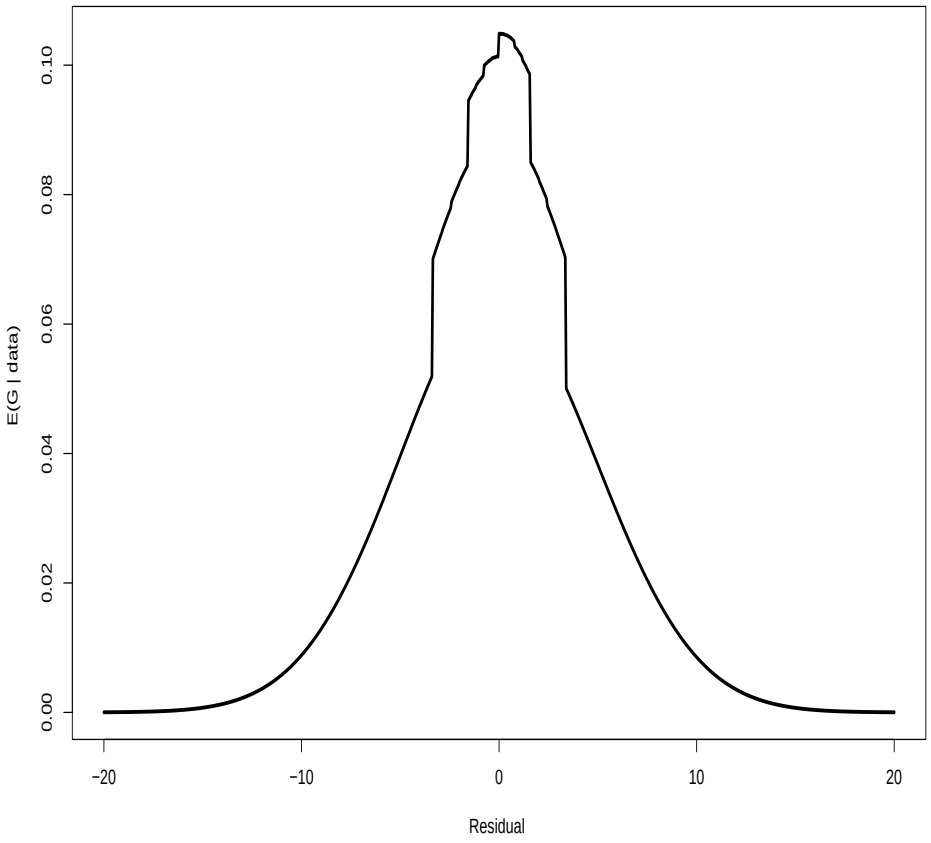

Figure 4.6: Posterior expectation of unknown distribution of residuals.

information from tumor biopsies. However, the accuracy of biopsy-based methods is related to how precisely the information derived from the biopsies represents the overall tumor microenvironment. Thrall et al (1997) studied binding of CCI-103F and pimonidazole, both 2-nitroimidazole compounds, in canine solid tumors for assessing pretreatment oxygenation and changes in oxygenation during irradiation.

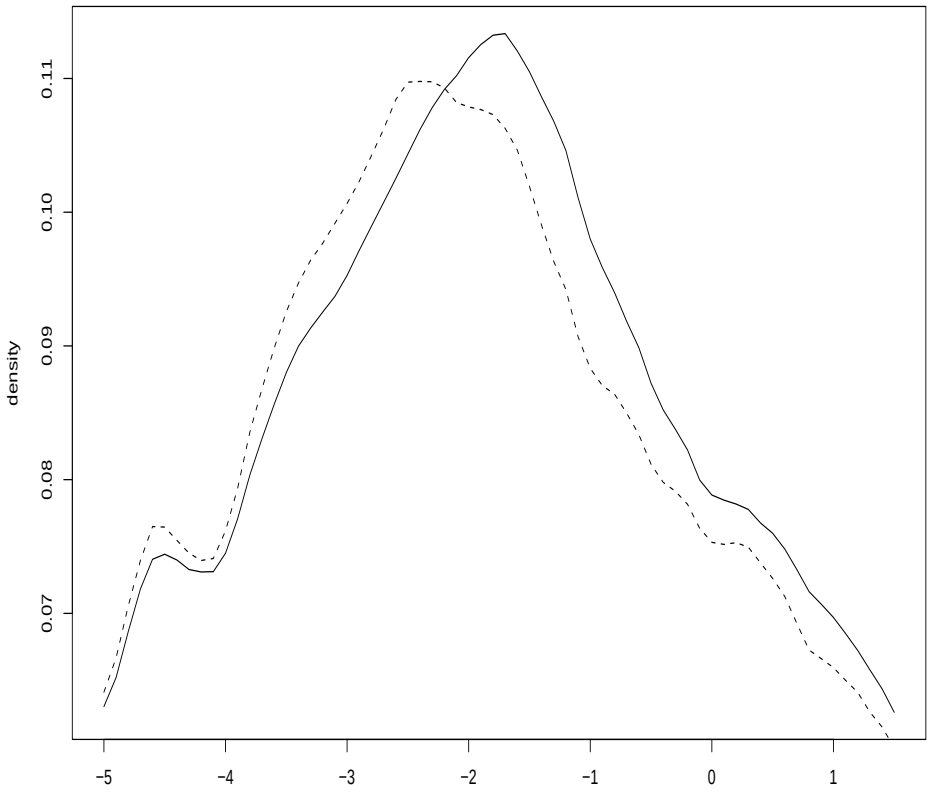

Figure 4.7: Predictive density for herceptin=0 and monocyte=1(solid line), and herceptin=1 and monocyte=0 (dashed line).

In this study, nine dogs, each with a primary solid tumor, were included. Twenty-four hours before treatment, CCI-103F was administered intravenously. Immediately prior to the treatment, a maximum of eight biopsies were obtained from different geographic regions of the tumor. One to four sections from each biopsy sample were placed on glass slides. The

volume fraction of hypoxic tumor tissue was reported by measuring the CCI-103F labelled area in each slide, which is calculated by

y =(CCI-103F labelled counts)/(CCI-103F labelled counts + unlabelled counts).

Slides from four out of eight biopsies were measured at 20 minutes after injecting the dye. Slides from the rest of biopsies in the same tumor were measured at 24 hours after injecting the dye. An important feature of the data is that the volume fraction of hypoxic tumor tissue can be 0 or 1 in some slides. The questions of interest include: "What is the average fraction of hypoxic cells in a tumor"; "How variable is the fraction within a tumor and between dogs?"; and "How is the measurements of hypoxic cells made in the same tumor affected by different measuring time?".

Figure 4.8 presents the average fraction of hypoxic cells in each biopsy for each dog. The x-axis denotes the biopsy label. Biopsies labelled as 1, 2, 3, and 4 were measured at 20 minutes. The fraction of hypoxic cells from biopsies labelled as 5, 6, 7, and 8 were measured at 24 hours.

We use conventional Bayesian and Bayesian semiparametric random effect models to fit the above fractional data. Some notations are introduced as follows. Let y_{ijk} denote the fraction of hypoxic cells in kth slide from jth biopsy of ith dog. Time t_{ij} is the measuring time for the jth biopsy from the ith dog, v_i is the volume of tumor from the ith dog, d_i denotes the random dog effect and $b_{j(i)}$ denotes the random biopsy effect within the ith dog. To include point mass probabilities for $y_{ijk} = 0$ and $y_{ijk} = 1$, we introduce a latent variable z_{ijk} in the model. Negative z_{ijk} is coupled into $y_{ijk} = 0$, and $z_{ijk} \geq 1$ is coupled into $y_{ijk} = 1$.

Thus we can proceed with a continuous probability model for z_{ijk}. The model is given by

$$y_{ijk} = \begin{cases} 0 & \text{if } z_{ijk} \leq 0 \\ z_{ijk} & \text{if } 0 < z_{ijk} < 1 \\ 1 & \text{if } z_{ijk} \geq 1 \end{cases}$$

$$z_{ijk} = \beta_0 + \beta_1 t_{ij} + \beta_2 v_i + d_i + b_{j(i)} + \varepsilon_{ijk},$$

$$\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$

$$i = 1, \cdots, 9, j = 1, \cdots, 8, k = 1, \cdots, 4$$

The model is completed by specifying the following priors:

$$\beta = (\beta_0, \beta_1, \beta_2) \sim N(\mu, \Sigma)$$

$$d_i \overset{iid}{\sim} N(0, \sigma_d^2)$$

$$b_{j(i)} \overset{iid}{\sim} F$$

The priors for hyperparameters are

$$\frac{1}{\sigma_d^2} \sim Ga(0.01, 0.01)$$

$$\frac{1}{\sigma^2} \sim Ga(0.01, 0.01)$$

We use a normal prior for β with mean 0 and covariance matrix $4I_3$, where I_3 is 3×3 identity matrix. Because of small number of dogs, the distribution of random dog effect is assumed to be normal. Two types of random effects distributions are considered separately for the distribution of biopsy effect. First, a normal prior, $F = N(0, \sigma_b^2)$ is specified. Then the parametric prior assumption is relaxed by specifying a nonparametric prior such as Polya tree prior, i.e., $F \sim PT(\Pi, \mathcal{A})$.

4.3.1 Normal prior for the biopsy effect

Assume $b_{j(i)} \overset{iid}{\sim} N(0, \sigma_b^2)$ and $\sigma_b^{-2} \sim Ga(0.01, 0.01)$.We run the Gibbs sampler for 35000 iterations, with the first 5000 being discarded as a burn-in period. In addition, due to high

autocorrelation, every 12th iteration was used and the rest discarded, making a total Monte Carlo sample size of 2500. Convergence of the Gibbs sampler was assessed via Geweke's (1992) method, using BOA (Bayesian Output Analysis Version 1.0.1) in R.

The histogram of mean biopsy effects is presented in the left panel of Figure 4.9. The right panel of Figure 4.9 displays the Q-Q normal plot of mean biopsy effect. It indicates that the distribution of biopsy effects takes on a multimodal appearance with a great deal of weight centered around 0. The non-normality may be due to heterogeneity across dogs, or biopsies, or other covariates, which is not presented here. Meanwhile, with a large dataset, we may consider to include outliers and sampling uncertainty arising from experiment layout into the model.

Histograms of β_0, β_1 and β_2 are illustrated in Figure 4.10. Note that the posterior distributions are reasonably symmetric. Posterior distributions of model parameters are summarized in Table 4.2. The results indicate that the proportion of hypoxic cell is strongly related to time but not significantly associated to volume of tumor.

4.3.2 Polya tree prior for the biopsy effect

Suppose the biopsy effect arises from a random probability measure F, and we assume a Polya tree prior for F. The prior on F was modelled as an infinite Polya tree with centering distribution $F_0 = N(0, 3^2)$. We place the same priors on fixed effects such as measuring time and volume of tumor and random dog effects as before. The parameters of the Polya tree prior are fixed as follows. The partitioning points are chosen to be the percentiles of F_0. c is set at a value of 0.1.

MCMC posterior simulation was run for 38000 iterations, with the first 4000 being discarded as burn-in period. Due to high autocorrelation, every 20th iteration was used and the rest discarded. Convergence of the posterior distributions of parameters was assessed via Geweke's (1992) method. The parameter estimates of β are presented in Table 4.2. Results

Table 4.2: Posterior median and central 95% posterior intervals for various parameters from the parametric normal model and Polya tree model. Here, β_0 is the intercept, β_1 is the slope over time, β_2 is the volume of tumor effect, σ is the error standard deviation, σ_d is the standard deviation of dog effect, and σ_b is the standard deviation of biopsy.

Parameter	Normal model	Polya tree model
β_0	0.303(0.127, 0.507)	0.347(0.173, 0.553)
β_1	0.113(0.019, 0.212)	0.054(-0.051, 0.169)
β_2	-0.001(-0.003, 0.003)	-0.001(-0.003, 0.002)
σ	0.093(0.085, 0.103)	0.093(0.085, 0.102)
σ_d	0.252(0.154, 0.527)	0.263(0.158, 0.510)
σ_b	0.197(0.163, 0.241)	

suggest that Polya tree prior introduces high uncertainty to the model, especially for the estimate of time effect.

The posterior distributions for β_0, β_1 and β_2 are shown in Figure 4.10. The posterior mean of F is shown in Figure 4.11, which indicates high between biopsy variability. It is noted that this is a fairly tight distribution about 0 as we would expect.

4.4 Microarray data

4.4.1 Introduction

In this section we develop a nonparametric Bayesian alternative to the popular empirical Bayes method (Efron et al. 2001) for inference about differential gene expression in microarray experiments. Subsection 4.4.2 provides some background on the data format and the nature of the experiments. Section 4.4.3 reviews some of the existing methods and how they relate to the proposed novel nonparametric Bayesian approach. In the following sections 4.4.4 and 4.4.5, we then proceed to introduce a nonparametric Bayesian model based on PT

priors and show how to implement posterior simulation in the proposed model. Finally, section 4.4.6. illustrates the proposed method in a simulation example and in a data set of colon cancer tissue samples.

Central to the empirical Bayes method is the notion of unknown distributions f_0 and f_1 of difference scores for non-differentially expressed genes and differentially expressed genes respectively. Details are discussed below, in section 4.4.4. The empirical Bayes method proceeds by using clever ad-hoc point estimates for f_0 and f_1 (actually for a ratio of unknown distributions). In contrast, a nonparametric Bayesian approach proceeds by explicitly acknowledging f_0 and f_1 as unknown quantities and assuming a prior probability model for them. The required model is a probability model on the unknown probability distributions f_0 and f_1, i.e., a nonparametric Bayesian model. We use a PT prior and proceed with posterior updating essentially as described in earlier chapters of this thesis. A minor complication arises from the fact that sampling involves an additional level of mixture. Genes are not known initially to be differentially or non-differentially expressed, leading to a sampling model that is a mixture of f_0 and f_1. The deconvolution of this mixture is the formalization of the scientific research question of identifying differentially expressed genes.

4.4.2 Background

The expression of thousands of genes are measured simultaneously using microarrays. This technology has been recognized by biomedical scientists as a powerful tool for gene discovery, expression profiling, as well as for the diagnosis and classification of cancers and other diseases. Microarrays measure mRNA concentrations by labelling the sample with a dye and then hybridizing them to spots on the array.

There are two main types of arrays: oligonucleotide arrays and cDNA arrays. In an oligonucleotide array, 14 to 20 probe pairs are used to interrogate each gene, each probe

pair has a Perfect Match (PM) and Mismatch (MM) signal, and the average of the PM-MM differences for all probe pairs in a probe set (called "average difference") is used as an expression index for the target gene.

In a cDNA microarray, the kind featured in this thesis, one base sequence matching all or part of a gene is printed on a glass slide. Two or more different fluorescent dyes label different samples. For example, the experimental sample is labelled with red dye and hybridized on the slide. A reference sample is labelled with green dye and hybridized on the same slide. By doing so, one can monitor multiple samples on the same array simultaneously. The log (red/green) intensity of RNA hybridization at each spot is measured.

In this section, we use the term "expression level" to refer to a summary measure of relative red to green channel intensities in a cDNA array or a summary difference of the PM and MM scores from an oligonucleotide array. See Wu (2001) for an introductory review of microarray technologies.

4.4.3 Methods

A common task in analyzing microarray data is to determine which genes are differentially expressed across two kinds of tissue samples or in general, samples obtained under any two experimental conditions. Specifically, it is of interest to detect genes with differential expression under the two conditions. Each gene has replicated measurements of expression levels under each condition.

A straightforward method is to use a traditional two-sample t-test for each gene (Devore and Peck 1997). Newton et al. (2001) proposed a parametric Bayesian modeling approach. Efron et al. (2001) discussed an empirical Bayes approach through the use of density estimates to approximate the distribution of expression levels for differentially and non-differentially expressed genes. They computed the posterior probability of differential expression by substituting estimates of relevant parameters and (ratios of) densities based

on the empirical distribution of expression levels. Pan et al. (2001a) suggested a mixture model approach, which follows the basic idea of Efron et al. (2001). They estimated the distribution of a t-type test statistic and its hull statistic using finite normal mixture models. A likelihood test is used to compare two distributions in order to identify genes with significantly changed expression.

Newton et al. (2004) proposed a semiparametric hierarchical mixture model and obtained gene-specific posterior probabilities of differential expression. Do et al. (2004) assume that the observed expression scores are generated from a mixture of two distributions that can be interpreted as distributions for affected and unaffected genes, respectively. They choose Dirichlet process models to represent the probability model for the unknown distributions.

Analogous to the above methods, we assume the unknown distributions of gene expression for differential and non-differential genes to have Polya tree priors. The marginal joint density of gene expression can be evaluated in a close form after integrating out the unknown distributions.

4.4.4 A mixture model for gene expression data

Suppose that Z_i is the difference score for gene i, summarizing the difference across samples observed under the biological conditions. See, e.g. Efron et al. (2001) for a possible difference of scores. In this example we will use 2-sample t-test statistics, but we do not make any assumption on this statistics. Gene i can be either differentially or non-differentially expressed under the condition of interest. We write the distribution of difference score Z_i as a mixture of two densities, f_0 and f_1, representing the density under differential and non-differential conditions, respectively. Thus, for $Z \in \{Z_i, i = 1, \cdots, n\}$, we assume $Z \sim f(Z)$ with

$$f(Z) = p_0 f_0(Z) + (1 - p_0) f_1(Z) \tag{4.1}$$

where p_0 is the probability that a gene is not differential in both experimental conditions.

By computing difference scores based on samples under the same biological condition, one could have additional samples of scores that are known to arise from f_0 by construction. This is essential to estimate f_0 with less uncertainty in the final inference.

Using Bayes' rule for given (f_0, f_1, p_0), the probability of differential expression

$$P_1(Z \mid f_0, f_1, p_0) = (1 - p_0)f_1(Z)/f(Z) \tag{4.2}$$

Equivalently, (4.1) can also written as

$$\begin{aligned} Z_i \mid r_i &\sim \begin{cases} f_0(Z_i) & \text{if } r_i = 0 \\ f_1(Z_i) & \text{if } r_i = 1 \end{cases} \\ Pr(r_i = 0) &= p_0, Pr(r_i = 1) = 1 - p_0, \quad i = 1, \cdots, n \end{aligned} \tag{4.3}$$

In the context of the hierarchical model (4.3) we can now recognize (4.2) as

$$P_1(Z \mid f_0, f_1, p_0) = Pr(r_i = 1 \mid Z, f_0, f_1, p_0) \tag{4.4}$$

Our goal here is to estimate the probability of differential expression, that is, $Pr(r_i = 1 \mid Z)$. This is the marginal of (4.4), marginalizing with respect to f_0, f_1 and p_0.

Efron et al. (2001) propose to estimate p_0 by an empirical Bayes approach, substituting point estimates for f_0/f_1 and p_0. To derive a point estimate for p_0 they observe that nonnegativity of P_1 implies $p_0 \leq min_Z f(Z)/f_0(Z)$, and propose to substitute the bound as point estimate $\hat{p_0} \equiv min_Z f(Z)/f_0(Z)$.

Do et al.(2004) proposed a fully model-based Bayesian approach that introduces a probability model on (f_0, f_1, p_0). They use Dirichlet process mixture models to define prior models for f_0 and f_1, that is

$$f_j(z) = \int N(z; \nu, \sigma^2) dG_j(\nu) \quad \text{and} \quad G_j \sim DP(M, G_j^\star), \quad \text{for} \quad j = 0, 1. \tag{4.5}$$

For the base measure $G_j^\star$ they use

$$G_0^\star = N(b, \sigma_0^2) \quad G_1^\star = 0.5N(-b_1, \sigma_1^2) + 0.5N(b_1, \sigma_1^2). \tag{4.6}$$

In words, the base measure for the non-differential gene scores is unimodal and centered at 0. The base measure for the differential gene scores is symmetric bimodal, reflecting the prior belief that differential expression (on the log scale) is equally likely on either direction.

Do et al. (2004)'s approach replaced the empirical estimates for f_0 and f_1 proposed by Efron et al. (2001) by posterior averages with respect to the posterior $p(f_j|Z), j = 0, 1$. Using the Dirichlet process mixture model in (4.5) and (4.6), a latent variable ν_i was introduced to break the DP mixtures assumed for f_0 and f_1. A stochastic size mixture of location problem arises in MCMC simulation. Another shortcoming of DP mixture is the computation intensive nature.

In this section, we follow Do et al. (2004)'s framework, but we will propose Polya tree priors for the unknown distributions f_0 and f_1. The posterior probability $Pr(r_i = 1 \mid Z)$ can easily be evaluated in closed form as discussed in Chapter 3.

Suppose that $f_0 \sim PT(\Pi_0, \mathcal{A})$ and $f_1 \sim PT(\Pi_1, \mathcal{A})$. Since the centering distribution plays a similar role as the base measure in Dirichlet process, for the construction of partition sequences for both Polya trees, we choose $F_0 = N(0, \sigma_0^2)$ and $F_1 = 0.5N(-b_1, \sigma_1^2) + 0.5N(b_1, \sigma_1^2)$ to be the centering distributions for the non-differential gene expressions and the differential gene expressions, respectively. The corresponding probability density functions are g_0 and g_1, respectively. For the choice of Polya tree parameters $\alpha_{\varepsilon_1 \ldots \varepsilon_m}$ in $\mathcal{A}$, we take $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = cm^2$ at the mth level with a fix value of c, e.g. $c = 0.1$. We assume a Beta prior $p_0 \sim Beta(a, b)$.

4.4.5 Posterior inference

Posterior simulation in the proposed model is carried out using MCMC simulation (Tierney 1994). The implementation is greatly simplified after marginalizing out the the unknown random measures f_0 and f_1. The algorithm is discussed as follows.

Given r_i, the posterior distribution of p_0 is a Beta distribution with parameters $a + n_0$ and $b + n - n_0$, where $n_0 = \sum_{i=1}^{n} I(r_i = 0)$, the number of genes which are not differential.

For $i = 1, \cdots, n$, we find (see below for the definition of $Z_{\Gamma_{0,i}}$):

$$p(r_i = 0 \mid r_{-i}, p_0, Z) \propto p(r_i = 0 \mid p_0) p(Z_i \mid r_i = 0, r_{-i}, p_0, Z_{-i})$$
$$\propto p_0 p(Z_i \mid r_i = 0, Z_{\Gamma_{0,i}}) = \pi_{i0}$$

Similarly, we have

$$p(r_i = 1 \mid r_{-i}, p_0, Z) \propto (1 - p_0) p(Z_i \mid r_i = 1, Z_{\Gamma_{1,i}}) = \pi_{i1}$$

where $Z_{\Gamma_{0,i}} = \{Z_j : j \neq i, r_j = 0\}$, $Z_{\Gamma_{1,i}} = \{Z_j : j \neq i, r_j = 1\}$, and r_{-i} is the vector $(r_1, \cdots, r_n)$ except the ith element. According to (3.1), we have

$$p(Z_i \mid r_i = k, Z_{\Gamma_{k,i}}) = g_k(Z_i) \lim_{M \to \infty} 2^{M-1} \prod_{m=1}^{M} \frac{cm^2 + n_{\varepsilon(m,Z_i)}(Z_{\Gamma_{k,i}})}{2cm^2 + n_{\varepsilon(m-1,Z_i)}(Z_{\Gamma_{k,i}})} \tag{4.7}$$

for $k = 0, 1$, where $n_{\varepsilon(m,Z_i)}(Z_{\Gamma_{k,i}})$ is the number of Z_j's in $Z_{\Gamma_{k,i}}$ falling into the same subinterval with Z_i at the mth level. As discussed in Chapter 3, the right side of (4.7) can be approximated by the limit of product.

Therefore, the posterior probability conditional on the currently imputed p_0 that gene i is differential given data and other parameters is

$$p(r_i = 1 \mid r_{-i}, p_0, Z) = \frac{\pi_{i1}}{\pi_{i0} + \pi_{i1}} \tag{4.8}$$

Averaging (4.8) across MCMC iterations we estimate the desired posterior probability of differentially expressed genes.

4.4.6 Simulation study and application

In this section, we first perform a small simulation study to demonstrate the proposed method. Results are compared with the know true parameter values in the simulation. Then we analyze the colon cancer data set reported in Alon et al (1999).

Simulation study

We simulate a sample of $n = 400$ gene difference scores $Z_i, i = 1, \cdots, n$ from $f = p_0 f_0 + (1 - p_0) f_1$ with $f_0 = N(0, 1)$ and $f_1 = 0.5N(-2, 1) + 0.5N(2, 1)$ and $p_0 = 0.5$. Following the above discussion, we assume f_0 and f_1 are unknown distributions and have PT priors. For the choices of PT parameters, we choose the centering distributions as $F_0 = N(0, 1)$ and $F_1 = 0.5N(-2, 1) + 0.5N(2, 1)$, and parameters $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = cm^2$ with $c = 0.1$.

Figure 4.12 shows the marginal posterior probability $\bar{P}_1(Z_i) = E(Pr(r_i = 1 \mid Z, f_0, f_1, p_0) \mid Z) = Pr(r_i = 1 \mid Z)$ for gene i, $i = 1, \cdots, n$. The structure of the proposed model implies that this marginal posterior probability of differential expression depends on the gene only through the observed score Z_i. It is meaningful to consider the marginal posterior probability as a function of Z.

Figure 4.13 shows the marginal posterior distribution $p(p_0 \mid Z)$. It is centered around 0.5, the true p_0.

Similar to Do et al. (2004), we can define the false discover rate (FDR) as

$$FDR = \frac{(1 - r_i)\delta_i}{\sum \delta_i}, \tag{4.9}$$

the fraction of false rejections, relative to the total number of rejections. Here δ_i is the indicator for rejecting the ith comparison, i.e., concluding that gene i is differentially expressed. The posterior expectation of FDR is easily computed as $\overline{FDR} = E(FDR \mid Z) = \{\sum(1 - \bar{P}_1(Z_i))\delta_i\} / \sum(\delta_i)$. If $\bar{P}_1(Z_i) > \gamma^\star$, we may classify gene i as differentially expressed. The threshold $\gamma^\star$ could then be selected to ensure that $\overline{FDR} \leq \tilde{\alpha}$ for some prespecified $\tilde{\alpha}$.

Gene expression profile of colon cancer

Gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array on over 6500 human genes in Alon et al. (1999). These samples were obtained from 40 patients while 20 patients provided both tumor and normal

tissue samples. Alon et al. (1999) focus on 2000 genes with highest minimal intensity across the samples. These 2000 genes comprise our data set. The microarray data set thus has $n = 2000$ rows and 62 columns. The data set was rearranged so that the tumors are labelled 1 to 40 and the normals 41 to 62. The first 11 columns report tumor samples collected under protocol P1 (using a poly detector), columns 12-40 are from tumor samples collected under protocol P2 (using total extraction of RNA), columns 41-51 are normal samples collected under P1 from the same patients as columns 1-11, and columns 52-62 are normal samples collected under protocol P2 from the same patients as columns 12-22.

Before we considered the posterior differential probability for each gene, we processed the data by taking the natural logarithm of each expression level. We then normalized to data to have mean zero and unit standard deviation. After normalization, we construct the gene difference score following the same procedure as Efron et al. (2001) and Do et al.(2004). The procedure is described as follows.

From the data matrix we construct a difference matrix, D, containing all the possible differences tumor and normal samples with the same protocol (P1 or P2) with the i-th row of D defined as the vector of all differences for the i-th gene. Meanwhile, in constructing D, we exclude differences of paired columns corresponding to the same patient. Including such differences would require the introduction of patient specific random effects to model the difference in variation between differences of paired and independent columns, respectively. Thus D includes possible effects due to differential expression in tumor versus normal samples. Specifically, it includes residual error plus a tumor versus normal effect for differentially expressed genes. The goal of the study is to identify those genes that are differentially expressed across tumor and normal samples.

Alternatively, one can use any other one-dimensional summary statistics of group difference for each gene. In our inference simulation, we use a traditional two-sample t-statistic.

However, it is important to note that we use t-statistics only as a convenience summary for each gene. We do not make any distribution assumptions on it.

In this example, we construct the difference score Z_i for gene i as follows:

$$Z_i = \frac{\bar{x}_{t,i} - \bar{x}_{n,i}}{S_i}, \quad i = 1, \cdots, 2000$$

where $\bar{x}_{t,i}$ and $\bar{x}_{n,i}$ are the average levels of expression for gene i in tumor and normal samples, respectively. S_i is the standard deviation of repeated expression measurements:

$$S_i = \sqrt{\frac{\sum_{i=1}^{40}(x_{t,i} - \bar{x}_{t,i})}{40} + \frac{\sum_{j=1}^{22}(x_{n,j} - \bar{x}_{n,j})}{22}}$$

Figure 4.14 shows the marginal posterior distribution $p(p_0 \mid Z)$. The posterior probabilities of differential expression $P_1(Z_i)$ for the 2000 genes ranges from 0.12 to 1.0, corresponding to $|Z|$ between 0.001 and 7.915, respectively. The first quartile, median, and the third quartile of the reported $P_1(Z_i)$ are 0.313, 0.511, and 0.842. Figure 4.15 shows the posterior probability $P_1(Z)$ for each Z.

Our approach provides a relatively easy and straightforward way to identify differentially expressed genes across the two conditions based on the posterior probabilities. In particular, there are 192 genes with posterior probabilities greater than 0.99. One hundred twenty-one genes have posterior probability greater than 0.998, where 119 genes have also been picked out by Significance Analysis of Microarrays (SAM) method. The smooth muscle gene cluster (J02854, T60155, M63391, D31885, X74295, X12369) has posterior probabilities of at least 0.998 for each individual gene. Alon et al identified 29 ribosomal protein genes (Table 1 in Alon et al. 1999) that appear to be related to cellular metabolism such as an ATP-synthase component and an elongation factor. Results using our approach show that only 10 of them have estimated posterior probabilities greater than 0.95.

4.4.7 Conclusion

We have developed a probability model for inference about differentially expressed genes in a microarray group comparison experiment. The advantages of the proposed nonparametric Bayesian approach is the introduction of the unknown distributions f_0 and f_1 of gene difference scores as random quantities. The implication of this choice is the opportunity to formally introduce prior information about the nature of f_0 and f_1. In particular, the introduced priors on f_0 and f_1 formalize the notion that non-differential expression should lead to difference scores centered around zero, and that differential expression should lead to difference scores that are either over-expressed, to the right of zero, or under-expressed to the left of zero. Posterior inference provides a full probabilistic description of all uncertainties. This makes it straightforward to report posterior probabilities of false positives and posterior expected false discovery ratios, allowing to calibrate decisions by the popular false discovery rate criterion. Implementation of posterior simulation is shown to be straightforward using the algorithms introduced in earlier sections.

The main limitation of the proposed approach is related to the assumed independence of the gene difference scores. This is a common technical convenience assumption that simplifies true prior information. Investigators are keenly aware of possible dependencies of gene expression in well established or hypothesized regulatory networks. Such prior information about dependence does not extend over all several thousand genes on the microarray, but is typically available for genes of specific interest. Another limitation in the proposed approach is the lack of decision theoretic explicit acknowledgement of the experimental goals. Microarray group comparison experiments might be carried out as a screening test to identify as many as possible promising candidates for further exploration, to identify a preferably small set of possible biomarkers, to provide a pilot data set to design a larger future experiment, and many other reasons. Depending on the goals different summaries of the posterior probabilities are appropriate. The use of false discovery ratios as defined in equation (4.9)

can be informally motivated by a 0-1 hypothesis testing loss. This is not always appropriate. Finally, an important limitation is the constraint to two groups, i.e., two biologic conditions of interest. Most investigators are interested in more than two biologic conditions. Typical examples are normal tissue versus different types of tumors, taken from different sites, different microenvironment, patients with different initial performance status, etc.

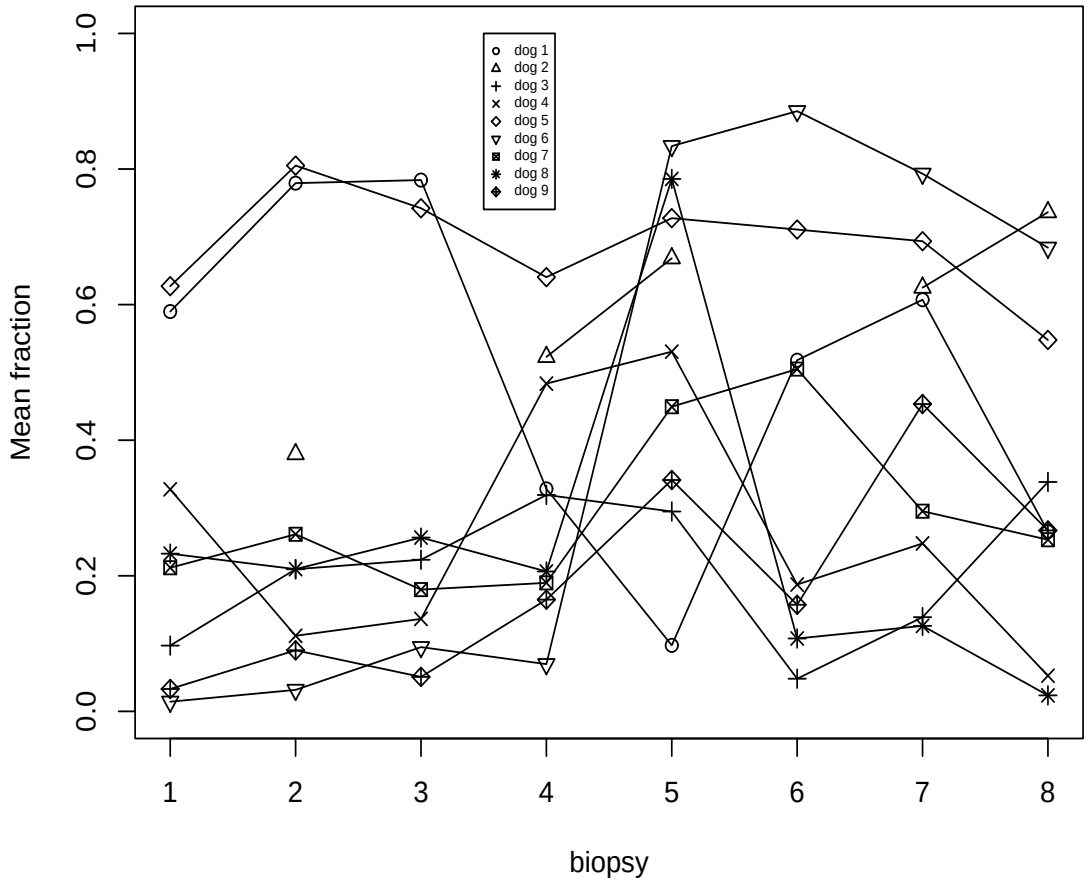

Figure 4.8: Average fraction of hypoxic cells in each biopsy for each dog.

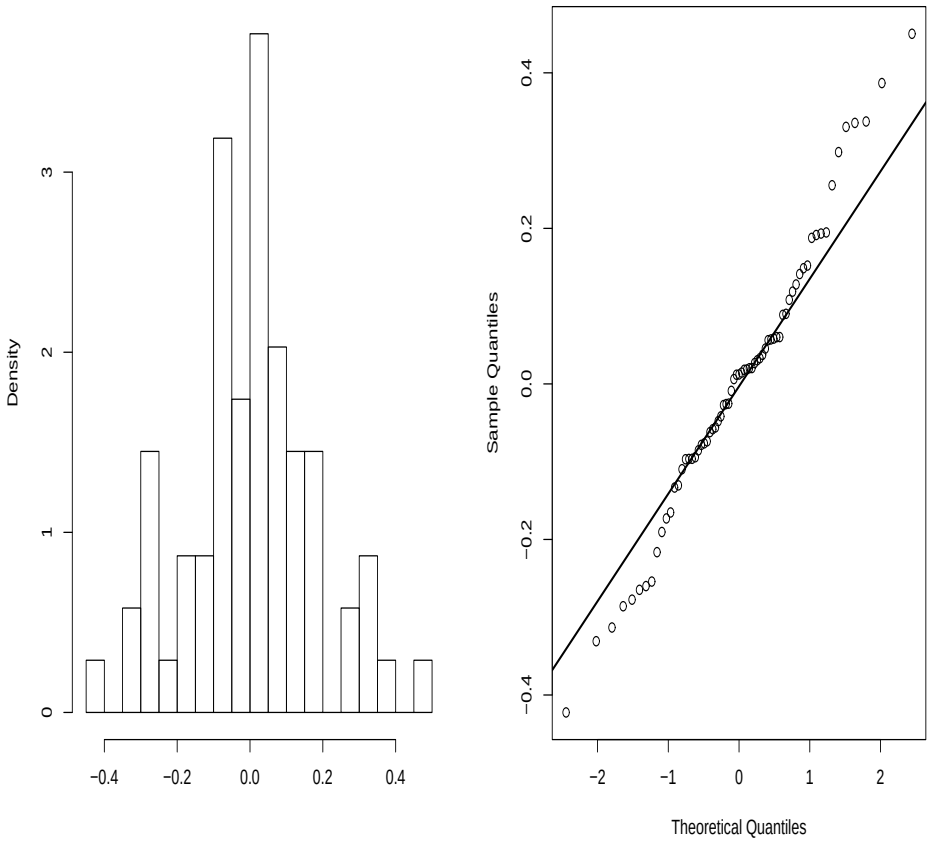

Figure 4.9: Histogram and Q-Q plot of mean biopsy effects under the parametric model.

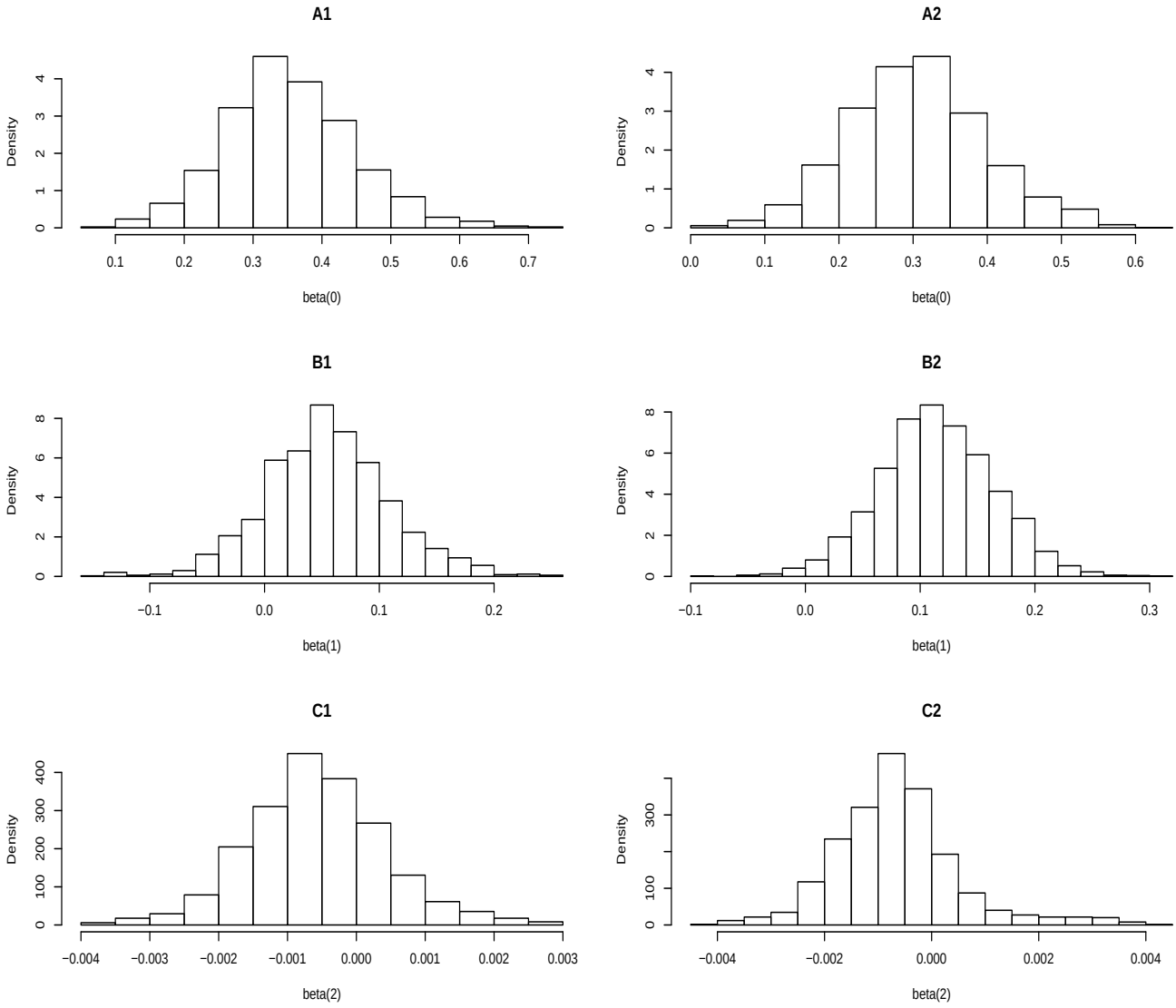

Figure 4.10: Histogram for β_0, β_1 and β_2 when biopsy effect has a Polya tree prior (A1, B1,C1) and normal prior(A2, B2,C2)

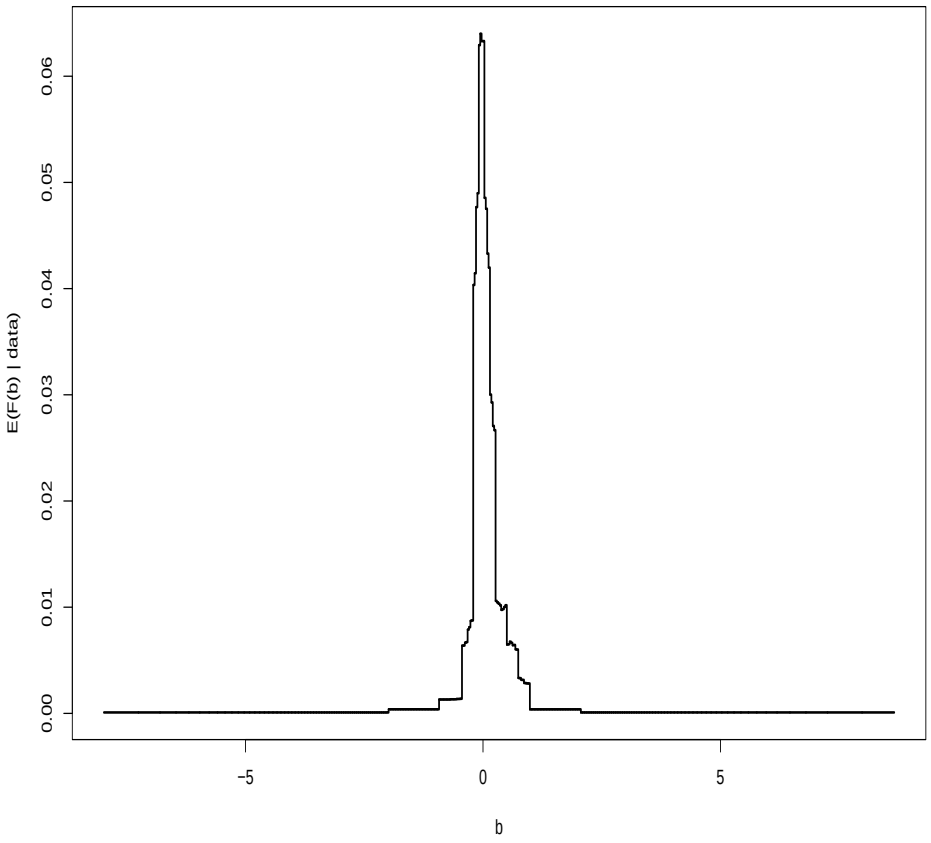

Figure 4.11: Posterior expectation of F.

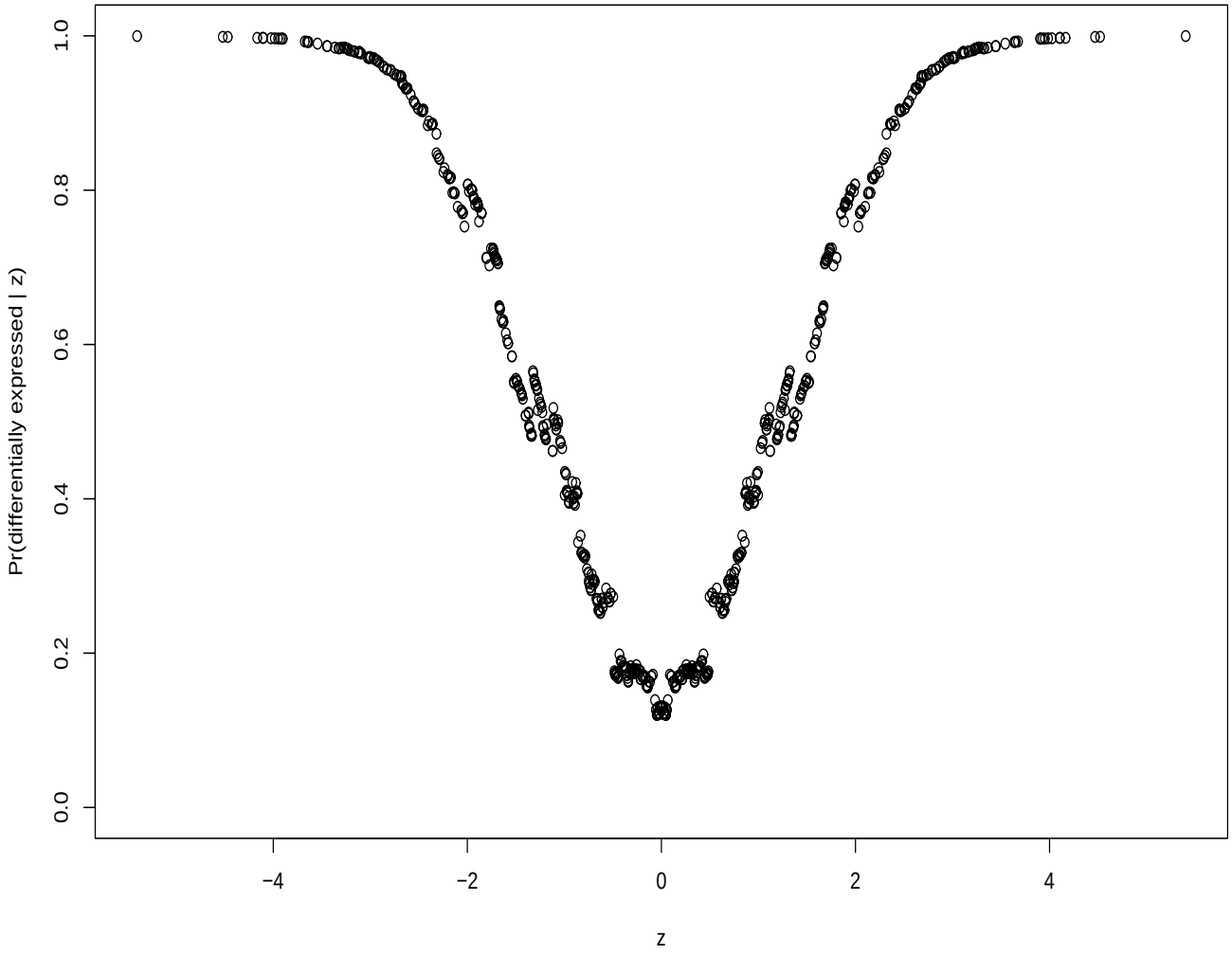

Figure 4.12: Posterior mean $P_1(Z_i) = E(Pr(r_i = 1 \mid Z, f_0, f_1, p_0) \mid Z)$.

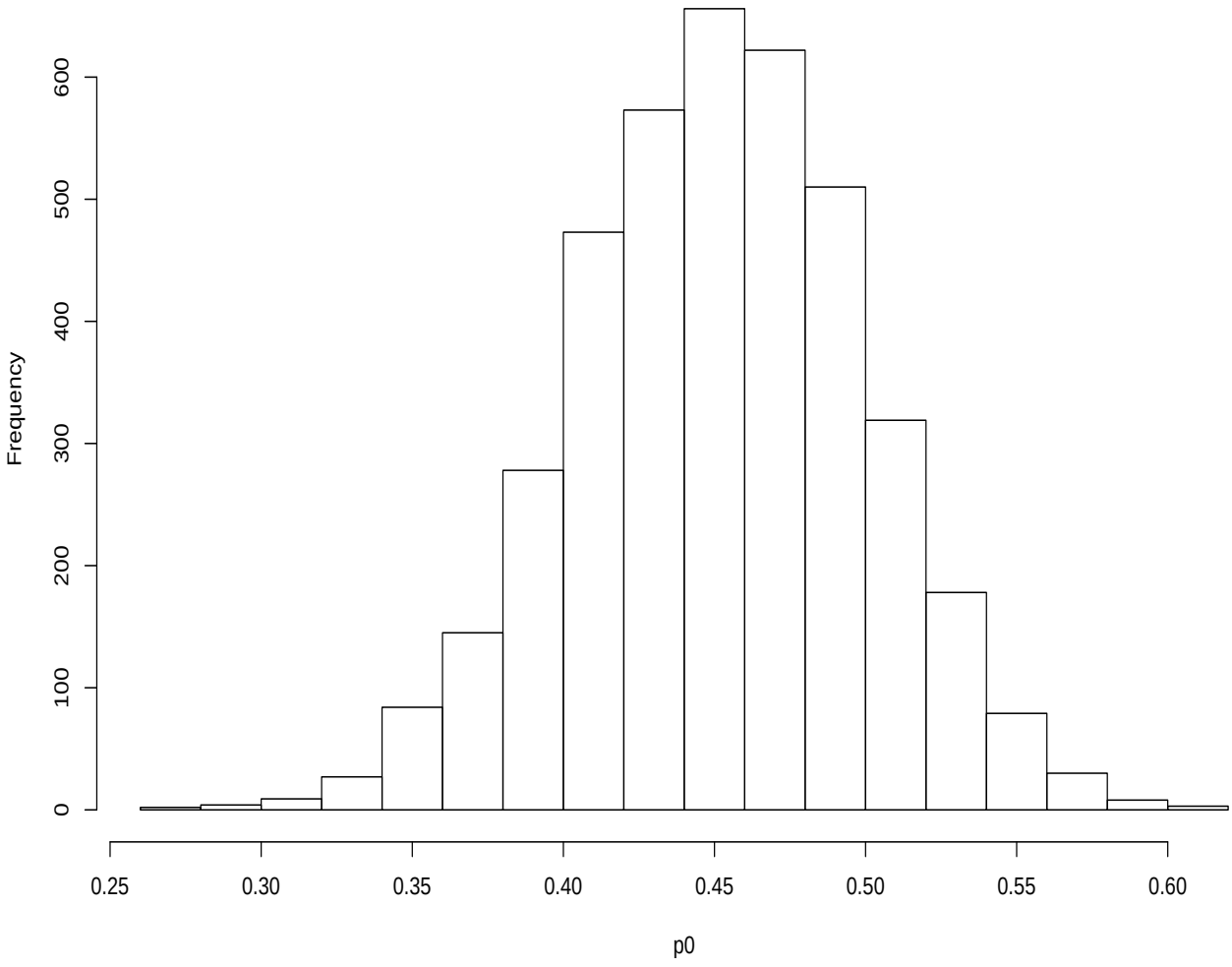

Figure 4.13: Histogram of the marginal posterior $p(p_0 \mid Z)$

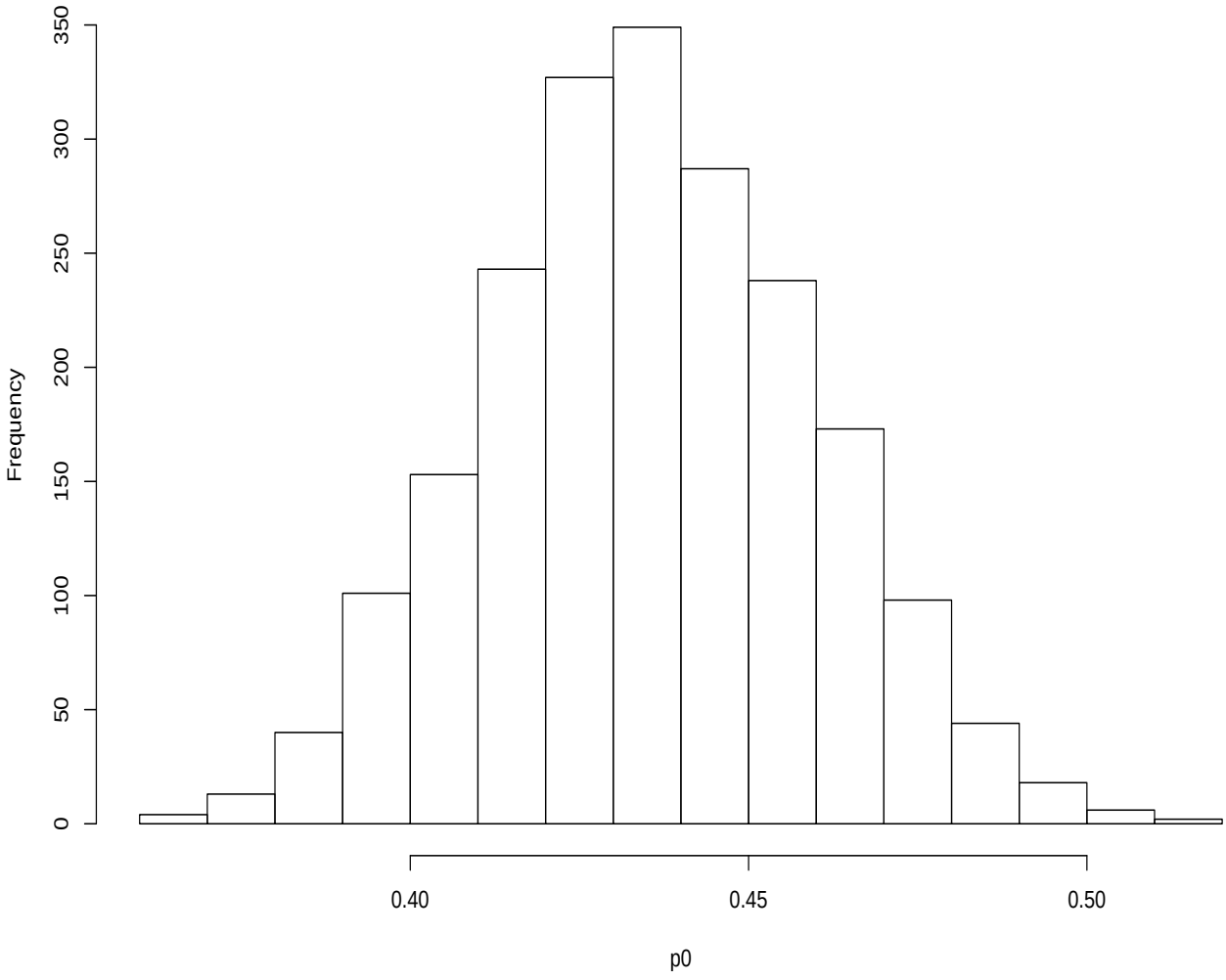

Figure 4.14: Histogram of the marginal posterior $p(p_0 \mid Z)$ (Alon colon cancer data)

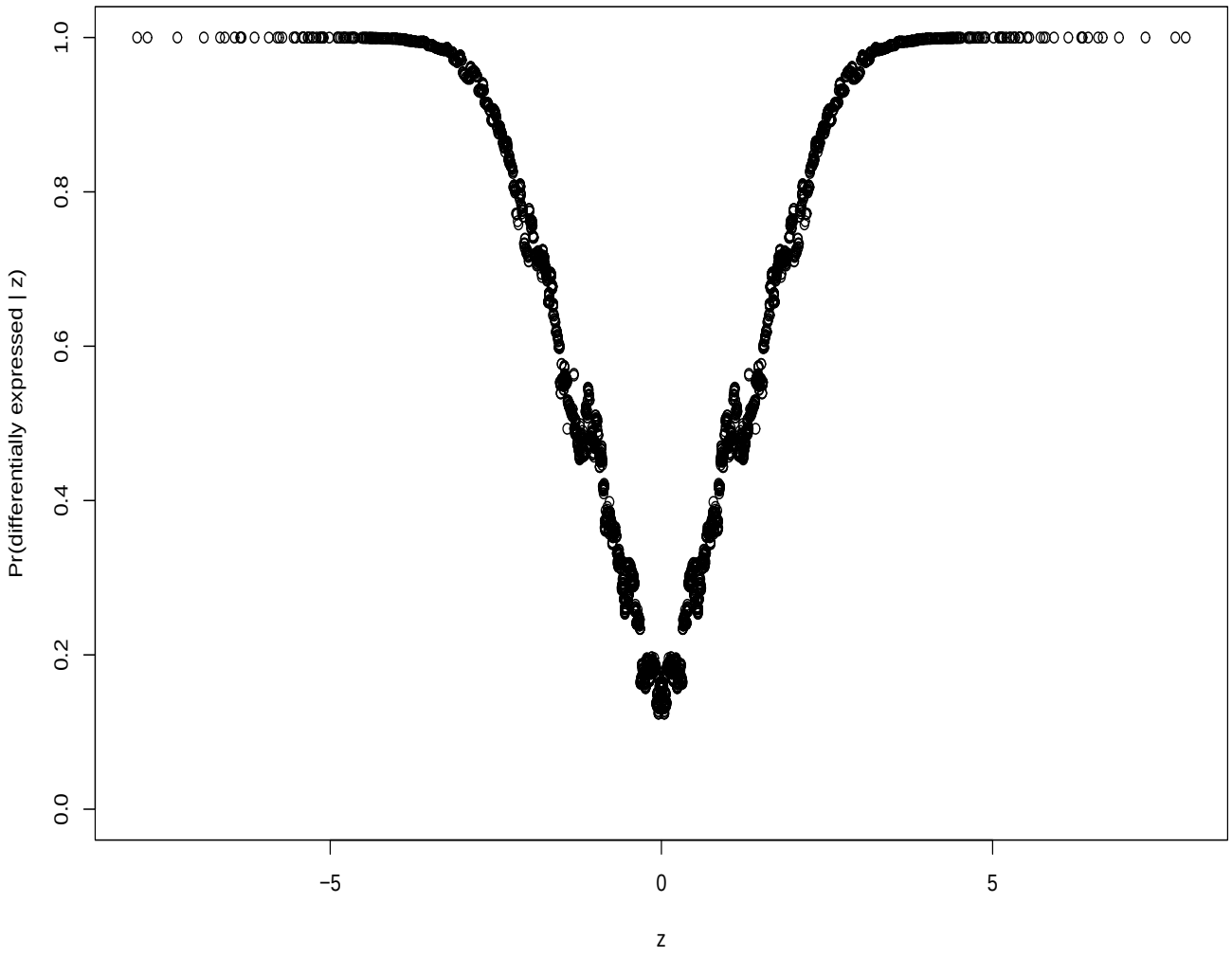

Figure 4.15: Posterior mean $P_1(Z_i) = E(Pr(r_i = 1 \mid Z, f_0, f_1, p_0) \mid Z)$ (Alon colon cancer data).

Chapter 5

Summary and Extensions

5.1 Summary

The goal of this research is to explore nonparametric Bayesian modelling with Polya trees in biomedical data. In this dissertation, we have proposed Bayesian nonparametric modelling for a class of important inference problems arising in biomedical data analysis. Proposed techniques included specifying a nonparametric prior for the distributions of random effects, residuals, and the unknown sampling distribution of gene expression. A Dirichlet process prior or Polya tree prior is specified as a nonparametric prior. We provide appropriate posterior simulation schemes.

The first important contribution of this dissertation is the development of a model and corresponding inference for repeated fractional data model. We introduced two models. First we use a fully parametric model. Then we generalize the model to semi-parametric model with a Dirichlet process prior for the random effect distribution.

The second important contribution is the use of Polya tree priors for the random effect distribution in the fractional data model. In the fractional data model, Polya tree priors avoid assuming a specific distribution for random effects. This allows us to estimate the distribution of random effects which provides insight into particular features of interest. For example, in the fractional data we obtained an estimate for the distribution of subject to subject variability.

The dissertation explores Polya tree as a prior for the residual distribution in linear models. It relaxes the normal assumption on residuals, allowing for features such as multimodality and skewness. A simulation study shows that the misspecification of the residual distribution can miss important details of the nature of data. It may lead to incorrect predictions.

The dissertation develops Polya tree models as priors for the unknown sampling distributions in microarray data. It provides a relatively easy and straightforward way to identify the differentially expressed genes across different biological conditions based on posterior probabilities.

The dissertation provides extensive discussion of computation issues for the different data settings. We provide a general guideline for the selection of Polya tree parameters. We discuss two posterior simulation schemes, posterior simulation for finite Polya trees and for Polya tree predictive densities. We developed new algorithms for posterior simulations with analytically marginalized random probability measure and for the evaluation of the posterior mean for the random distribution.

Naturally, many important inference problems and data formats remain that have not been addressed in this thesis. In particular, in this dissertation we did not consider Polya tree distribution in survival analysis, which is very common in medical studies. We did not consider Polya tree as a prior for dependent random distributions in meta analysis.

5.2 Future work

Application on survival data

Survival data is a very common format in medical research. Walker and Mallick (1999) described Polya tree as prior for the error distribution in an accelerated failure time. A finite Polya tree posterior simulation was used to obtain a predictive distribution for a future observation. Hanson and Johnson (2002) use a mixture of Polya trees to model regression

error in an accelerated failure time. We did not touch this field in this dissertation. However, we should implement the discussed methods to this common format of biostatistical data.

Microarray data

In section 4.3, we consider Polya tree as priors for the unknown distributions of gene expression under two conditions. In some microarray experiments, it is also of interest to detect whether genes are differentially expressed under more than two conditions. We will extend the proposed method to gene expressions measured under more than two conditions. Furthermore, we will generalize the method to the case that the number of conditions is unknown. The goal is to identify differentially expressed genes. The framework can be described as follows.

Suppose there are K conditions of interest, and there may be replicate measurements in each condition. We assume that some preprocessing technique has been used to adequately normalize the data so that the measurements can be viewed as approximations of relative gene expression in the samples. Let us initially consider three conditions ($K = 3$), with data $x_g = (x_{g,1}, \cdots, x_{g,n_1})$ from the n_1 replicate measurements in the first condition (C_1), $y_g = (y_{g,1}, \cdots, y_{g,n_2})$ from the n_2 replicate measurements in the second condition (C_2), and $z_g = (z_{g,1}, \cdots, z_{g,n_3})$ from the n_3 replicate measurements in the third condition (C_3).

Kendziorski et al (2002) proposed a hierarchical mixture model to account for replicate expression in multiple conditions. They derive the posterior probability of differential expression under two specific parametric formulations: a model based on Gamma distributed measurements and one based on log-normally distributed measurements. The method provides a way to infer patterns of differential expression among two or more conditions, but it relies on parametric model assumptions and the implementation of numerical optimization methods. In practice, the distribution of gene expression is usually unknown. A nonparametric Bayesian approach can be easily implemented in this setting.

Given three conditions, there includes 5 possible patterns. That is, equivalent expression across the three conditions, altered expression in just one condition, and distinct expression in each condition. Thus the marginal distribution od the data is

$$p_0 f_0(x_g, y_g, z_g) + p_1 f_1(x_g, y_g, z_g) + p_2 f_2(x_g, y_g, z_g) + p_3 f_3(x_g, y_g, z_g) + p_4 f_4(x_g, y_g, z_g)$$

where f_0 is the joint probability density (pdf) for equivalent expression across the three conditions, f_1, f_2, f_3 are the joint probability density for altered expression in the first, second, or third condition respectively, and f_4 is the joint pdf for genes distinctly expressed in each condition. p_k is the mixing proportion. By Bayes rule, the posterior probability of expression pattern k may be computed:

$$P(k \mid x_g, y_g, z_g) \propto p_k f_k(x_g, y_g, z_g)$$

To complete the model specification, we use Polya tree model to define prior models for f_k, that is, $f_k \sim PT(\Pi, \mathcal{A})$. We assume a Dirichlet prior $(p_0, \cdots, p_4) \sim Dir(\alpha_0, \cdots, \alpha_4)$. Let d_g denote the data vector for gene g, that is, $d_g = (x_g, y_g, z_g) = (d_{g,1}, \cdots, d_{g,N})$ where $N = n_1 + n_2 + n_3$. Then the joint marginal density for the equivalent expression is

$$f_0(d_g) = g_0(d_1) \prod_{i=1}^{N} f(d_i \mid d_1, \cdots, d_{i-1})$$

where g_0 is the density of the centering distribution and the predictive density $f(d_i \mid d_1, \cdots, d_{i-1})$ can be evaluated by equation (3.7)

Additional hierarchical level in repeated measurement model

In the fractional data discussed in the dissertation, the fractional response outcome was measured at two time points. It also can be the case that the fractional measurements will be obtained from the same biopsy or sample at more than two time points, or the measurement will be made at several time points per cycle for multiple cycles for each patient. The proposed

model will be generalized to incorporate the additional random effects, such as nested cycle effect.

In the immunobiology study, it is very important to understand the mechanism of direction action of immunobiologic therapies on the tumor or indirection modulation of the immune response. The proportions of specific lysis in blood samples are measured through flow cytometry. The proportions take values in a range between 0 and 1 including 0. The proposed model can also be applied to this type of data.

Bibliography

[1] Albert, J.H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data.*Journal of the American Statistical Association*, **88**, 669-680.

[2] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Lavine, A. J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In *Proceedings of the National Academy of Sciences*, **96**, 6745-6750.

[3] Antoniak, C.E. (1974) Mixtures of Dirichlet process with applications to nonparametric problems. *The Annals of Statistics*, **2**, 1152-1174.

[4] Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, **27(2)**, 536-561.

[5] Berger, J.O. and Guglielmi, A. (2001) Bayesian testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, **96**, 174-184.

[6] Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian theory*. New York: John Wiley & Son.

[7] Blackwell, D. and MacQueen, J.B. (1973) Ferguson distribution via Polya Urn schemes. *The Annals of Statistics*, **1**, 353-355.

[8] Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum likelihood from incomplete data via EM algorithm. *Journal of Royal Statistics Society, Ser. B*, **39**, 1-38.

[9] Devore, J. and Peck, R. (1997) *Statistics: the exploration and Analysis of Data*, 3rd ed, Duxbury Press, Pacific Grove, CA.

[10] Dey, D., Müller, P., and Sinha, D. (1998). *Practical nonparametric and semiparametric Bayesian statistics*. New York:Springer.

[11] Do, K. A., Müller, P., and Tang, F. (2004) A Bayesian mixture model for differential gene expression. Journal of Royal Statistics Society, Ser. B.

[12] Dubins, L. E. and Freedman, D. A. (1966) Random distribution functions. *Proceeding of Fifth Berkeley Symposium Mathematical Statistics and Probability*, **3**, 183-214. University of California Press.

[13] Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2001) Microarrays and their use in a comparative experiment. *Journal of the American Statistical Association*, **96**, 1151-1160.

[14] Escobar, M.D. (1988) Estimating the means of several normal populations by nonparametric estimation of the distributions of the means. *Unpublished PhD dissertation*, Yale University.

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

[16] Escobar, M.D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577-580.

[17] Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209-230.

[18] Ferguson T. S. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics*,**2** , 615-629.

[19] Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, **34** , 1386-1403.

[20] Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6(6)**, 721-740.

[21] Gelfand, A. E. and Kottas, A. (2002) A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **11(2)**, 289-305.

[22] Gelman A. and Rubin D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-511.

[23] Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (Disc: p189-193). *Bayesian Statistics* **4**. Proceedings of the Fourth Valencia International Meeting, 169-188.

[24] Ghosal, S., Ghosh, J.K., and Ramamoorthi, R. V. (1999) Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, **29**, 1264-1280.

[25] Gilks, W.R., Wang, C.C., Yvonnet, B., and Coursaget, P. (1993) Random-effects models for longitudinal data using Gibbs sampling. *Biometrics*, **49**, 441-453.

[26] Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized lease squares. *Biometrika*, **73**, 43-56.

[27] Gustafson, P. (1998) A guided walk Metroplois algorithm. *Statistics and Computing*, **8**, 357-364.

[28] Hanson, T. and Johnson, W.O. (2002) Modelling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, **97**, 1020-1033.

[29] Ishwaran, H. and James, L.F. (2001) Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, **96**, 161-173.

[30] Kendziorski C.M., Newton M.A., Lan H., and Gould M.N. (2002) On parametric empirical methods for comparing multiple groups using replicated gene expression profiles. *Technical Report ♯ 166*, **2002**, University of Wisconsin Department of Biostatistics and Medical Informatics.

[31] Kleinman K.P. and Ibrahim J.G. (1998a) A semiparametric Bayesian approach to the random effects model. *Biometrics*, **54**, 921-938.

[32] Laird N.M. and Ware J.M. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

[33] Lavine M. (1992) Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, **20**, 1222-1235.

[34] Lavine M. (1994) More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, **22**, 1161-1176.

[35] Longford, N.T. (1987) A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, **74**, 817-827.

[36] Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F.R., and Tsui, K. W. (2001) On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8**, 37-52.

[37] Newton, M. A., Noueriry, A., Sarkar, D., and Ahlquist, P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics*, **5**, 155-176.

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

[39] MacEachern, S.N. and Müller, P. (1998) Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**(2), 223-338.

[40] MacEachern, S.N. and Müller, P. (2000) Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models. In *Robust Bayesian Analysis*, F. Ruggeri and D. Ríos-Insua (eds), New York: Springer-Verlag, 295-315.

[41] Mauldin, R.D. and Williams, S.C. (1990) Reinforced random walks and random distributions.*Proceedings of the American Mathematical Society*, **110**, 251-258.

[42] Mauldin, R.D., Sudderth, W.D., and Williams, S.C. (1992). Polya trees and random distributions. *The Annals of Statistics*, **20**, 1023-1221.

[43] McLachlan, G. J., Bean, R. W., and Peel, D. (2001) A mixture model-based approach to the clustering of micrarray expression data. *Bioinformatics*, **18**, 413-422.

[44] Muliere, P. and Walker, S.(1997) A Bayesian nonparametric approach to determining a maximum tolerated dose. *Journal of Statistical Planning and Inference*, **61**, 339-353.

[45] Müller, P. (1991) Monte Carlo integration in general dynamic models. *Statistical Multiple Integration*. 145-163.

[46] Müller, P. and Quintana, F. (2004) Nonparametric Bayesian data analysis. *Statistical Science*, **19(1)**, 95-110.

[47] Paddock, S., Ruggeri, F., Lavine, M., and West, M. (2003) Randomised Polya Tree Models for Nonparametric Bayesian Inference. *Statistica Sinica*, **13(2)**, 443-460.

[48] Pan, W., Lin, J., and Le, C. (2001a) A mixture model approach to detecting differentially expressed genes with microarray data. *Technical Report*, **2001-011**, Division of Biostatistics, University of Minnesota.

[49] Pitman, J. (1996) Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, T.S. Ferguson, L. S. Shapeley and J. B. MacQueen (eds). Haywar, California: IMS Lecture Notes - Monograph Series, 245-268.

[50] Sethuraman, J (1994) A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, **4**, 639-650.

[51] Sinha, D. and Dey, D. K. (1997) Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*. **92**, 1195-1212.

[52] Thrall, D.E., Rosner, G.L., Azuma, C., McEntee, M.C., and Raleigh, J.A. (1997) Hypoxia marker labelling in tumor biopsies: quantification of labeling variation and criteria for biopsy sectioning. *Radiotherapy and Oncology*, **44**, 171-176.

[53] Tierney, L. (1994) Markov chains for exploring posterior distributions (Disc: p1728-1762). *The Annals of Statistics*, **22**, 1701-1728.

[54] Verbeke, G, and Lesaffre, E (1996) A linear mixed-effects model with heterogeneity in the random effects population. *Journal of the American Statistical Association*, **91**, 217-221.

[55] Walker, S.G. and Mallick, B.K. (1997) Hierarchical generalized linear models and faity models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society, Series B, Methodological*, **59(4)**, 845-860.

[56] Walker, S.G. and Mallick, B.K. (1999) Semiparametric accelerated life time model.*Biometrics*, **55**, 477-483.

[57] Walker, S.G., Damien, P., Laud, P. W. , and Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions (Disc: p510-527). *Journal of the Royal Statistical Society, Series B, Methodological*, **61**,485-509.

[58] West, M. (1992) Hyperparameter estimation in Dirichlet process mixture. *ISDS Discussion Paper 92-A03*, Duke University.

[59] Wu, T.D. (2001) Analyzing gene expression data from DNA microarrays to identify candidate genes. *Journal of Pathology*, **195(1)**, 53-65.

[60] Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*. **86**, 79-86.

APPENDIX

Resampling τ

From (3) we find

$$\underbrace{z_i - x_i\beta - U_i\theta_i}_{z_i^\star} = e_i$$

Then

$$z_i^\star \sim N(0, \sigma^2 I_{n_i})$$

independently across i. Together with (4),

$$\begin{aligned} P(\tau|\cdots) &= Gamma(a,b) \\ a &= \frac{\gamma_0 + \sum_{i=1}^n n_i}{2} \\ b &= \frac{\lambda_0 + \sum_{i=1}^n z_i^{\star\prime} z_i^\star}{2} \end{aligned} \tag{A.1}$$

Resampling $\theta_i, i = 1, \cdots, n$

From (3) we find

$$\underbrace{z_i - x_i\beta}_{z_i^\star} = U_i\theta_i + e_i$$

Then

$$z_i^\star \sim N(U_i\theta_i, \sigma^2 I_{n_i}).$$

Together with (3),

$$\begin{aligned} P(\theta_i|\cdots) &= N(m,v) \\ v &= (\Sigma_\theta + \tau U_i' U_i)^{-1} \\ m &= v\tau U_i' z_i^\star \end{aligned} \tag{A.2}$$