BAYESIAN SEMIPARAMETRIC METHODS FOR LONGITUDINAL, MULTIVARIATE, AND SURVIVAL DATA

by
Michael Lindsey Pennell

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2006

Approved by:

Dr. David Dunson, Advisor
Dr. Lawrence Kupper, Committee Chair
Dr. Amy Herring, Committee Member
Dr. Jainwen Cai, Committee Member
Dr. Stephen Rappaport, Committee Member

ABSTRACT
MICHAEL LINDSEY PENNELL: BAYESIAN SEMIPARAMETRIC METHODS FOR LONGITUDINAL, MULTIVARIATE, AND SURVIVAL DATA.
(Under the direction of Dr. David Dunson.)

In many biomedical studies, the observed data may violate the assumptions of standard parametric methods. In these situations, Bayesian methods are appealing since nonparametric priors, such as the Dirichlet process (DP), can incorporate a priori knowledge regarding the shape or location of an unknown distribution and exact inferences are available using Markov chain Monte Carlo methods. Despite the promise of Bayesian nonparametric methods, computation can be difficult under large sample sizes. In addition, there is a paucity of methods for multiple event time data and for testing across multiple groups.

In this dissertation, we propose three methods which address important computational, modelling, and testing issues in Bayesian nonparametrics. Our first method is a computationally simple approach to fitting Bayesian semiparametric random effects models to large longitudinal data sets. Our approach involves fitting a model to a smaller set of pseudo-data, which is constructed using expert opinion. The research was motivated by data from the Collaborative Perinatal Project, which was a prospective epidemiology study consisting of over 30,000 children.

We next develop a dynamic frailty model which accounts for age-dependent changes in susceptibility to a repeated health event, such as the occurrence of new tumors. Our model generalizes the traditional shared frailty model for multiple event time data to accommodate smooth, time dependent trends in the frailty, baseline hazard, and covariate effects. We also relax our assumptions on the frailty using DP priors.

Lastly, we present a Bayesian nonparametric method for testing for changes in a response distribution with an ordinal predictor. The research was motivated by data from toxicology studies, in which dose may affect both the shape and location of the response distribution. Using a generalization of the dynamic mixture of DPs (Dunson, 2006, *Biostatistics*, to appear), we test for equivalence in the unknown distribution across dose groups and estimate threshold doses. Our method accommodates multivariate responses without complication.

ACKNOWLEDGMENTS

Many thanks to David Dunson for his ideas and support in completing this dissertation. I would like to thank Dr. Aimin Chen and Dr. Matthew Longnecker, National Institute of Environmental Health Sciences (NIEHS) for providing data from the Collaborative Perinatal Project, Dr. Clinton Grubbs, Univ. Alabama Birmingham, for providing the chemoprevention data analyzed in Chapter 4, and Dr. Jack Taylor, NIEHS, for providing the genotoxicity data that we used to illustrate our method in Chapter 5. I would like to recognize the efforts of my expert panel in Chapter 3: Walter Rogan, MD, Allen Wilcox, MD, NIEHS and Dr. Robert McMurray, Department of Exercise Physiology, Dr. Diane Holditch-Davis, School of Nursing, UNC-Chapel Hill. I would also like to acknowledge the Environmental Pre-doctoral Training Grant in Biostatics and the Intramural Research program of the National Institute of Environmental Health Sciences for their financial support of my graduate studies and research. Finally, I would also like to thank my committee for their time and effort.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In many analysis settings, the observed data do not possess the characteristics of a known distribution. For example, count data can often have a larger proportion of zeros than would be anticipated under the Poisson distribution (Carota and Parmigiani, 2002; Dunson, 2004). Time to event data can also have characteristics such as non-monotone hazards that contradict the behavior of parametric models such as the Weibull. When violations of assumptions are minor, fully parametric methods should be adequate, but in more extreme cases these methods can be overly restrictive making inferences questionable.

Such problems have motivated the vast literature on nonparametric statistical methods. Robust frequentist methods exist for conducting two-sample tests (e.g., the Wilcoxon and Kolmogorov-Smirnov tests) and estimating distributions (e.g., the Kaplan Meier and Nelson-Aelen estimates of the survival function and cumulative baseline hazard). Semiparametric regression models also exist which only require assumptions regarding the link function between covariates and the response. For example, Cox's partial likelihood (1975) can be used to estimate regression coefficients in the proportional hazards model without assuming a specific distribution for the event times.

Although it is common for one to doubt the exact distributional form of certain data, usually there is some a priori intuition regarding its behavior. Bayesian nonparametric methods are advantageous in this respect since they provide convenient methods of incorporating prior information regarding the shape of a distribution into inferences. Some examples of commonly used nonparametric priors include the Dirichlet process (Ferguson, 1973, 1974), Pólya tree (Ferguson, 1974; Lavine, 1992, 1994), and neutral to the right processes (Doksum, 1974; Ferguson and Phadia, 1979). These priors can often be centered on a known distribution making it possible to both use the parametric form

when appropriate and to move away from it when its fit is poor. Bayesian computation, often through the use of Markov chain Monte Carlo (MCMC) methods (Gelfand and Smith, 1990; Tierney, 1994), provides exact posterior estimates of the unknown distribution as well as parameters and other functionals of interest. Frequentist methods typically rely on asymptotic evaluations which can be particularly troublesome in nonparametric settings where the number of unknown parameters increases with sample size.

Despite the promise of Bayesian nonparametric methods, computation can be difficult. Advances in Gibbs sampling methodologies (e.g., MacEachern, 1994; West et al., 1994; MacEachern and Müller, 1998; Neal, 2000) have made Bayesian nonparametrics more feasible, but these approaches have some limitations. In particular, the commonly used Pólya urn sampler for Dirichlet process mixture models (MacEachern, 1994; West et al., 1994) is not computationally feasible for large data sets, such as those from multi-center longitudinal studies. Variational approaches (Blei and Jordan, 2006) are more efficient for large samples, but do not use true posterior distributions for inference.

Another limitation of Bayesian nonparametrics is its lack of breadth. For instance, there are very few nonparametric or semiparametric Bayesian methods for multiple event time data. These data are common in biomedical studies in which the event of interest may be repeated infections, hospitializations, or recurrences of disease. Some examples include chemoprevention and cocarcinogenicity studies measuring the rate of appearance of palpable tumors of the skin and breast of animals exposed to a known carcinogen (Dunson, 2000; Gail et al., 1980; Forbes and Sambuco, 1998). Current methods for analyzing these of data, such as the shared frailty model (Vaupel et al., 1979; Clayton and Cuzick, 1985), account for correlations between tumors from the same animal using random effects. However since animal-specific susceptibility may unexpectedly change with age, generalizations of these models are needed to allow frailties to vary dynamically; ideally, such methods would be nonparametric and computationally efficient.

There has also been little consideration of Bayesian nonparametric testing in the k-sample setting. In particular, there are no Bayesian methods which test for changes in the shape of a response distribution across an ordinal predictor. Such methods would be appealing for toxicology data since differences amongst subjects in their response to treatment may cause changes in variance, skewness, or modality with dose. In these settings, methods which test for changes in the location parameter of a distribution (e.g., the method of Gopalan and Berry, 1998) may ignore important dosages.

In this dissertation, we provide an introduction to Bayesian nonparametric inference and propose three semiparametric methods which address some of the computational, modelling, and testing issues mentioned above.

In Chapter 2, we provide a literature review of Bayesian nonparametric priors with particular attention given to the Dirichlet process and other priors used in hierarchical models and survival analysis.

In Chapter 3, we describe an empirical Bayes method which makes fitting semiparametric random effects models feasible for large data sets. The method uses expert elicitation to construct a smaller set of pseudo-data which summarizes the scientifically important differences in the response and predictor values. We then fit a random effects model to the pseudo-data, assigning a nonparametric Dirichlet Process prior (DPP) to the random effects. This method was motivated by data from the Collaborative Perinatal Project (CPP), which was a prospective epidemiology study consisting of over 30,000 children.

In Chapter 4, we propose a semiparametric dynamic frailty model for multiple event time data, which is motivated by data from studies of tumorigenesis. The model presents many interesting innovations over current methods including a computationally simple method of introducing correlation amongst time dependent frailties, piecewise constant hazards, and dynamic regression coefficients. We also relax our assumptions on the frailty using DPPs. To illustrate our method, we analyze data from a cancer chemoprevention study.

In Chapter 5, we discuss a Bayesian nonparametric method for testing for changes in a response distribution across an ordinal predictor, such as dose. Using a dynamic mixture of Dirichlet processes (DMDP; Dunson, 2006), we allow the response distribution to change flexibly at each level of the predictor. In addition, we assign hierarchical priors to the mixture weights to obtain probabilities of no effect of the predictor and to identify thresholds in toxicology data, such as the lowest observed adverse effects level (LOAEL). The method also provides a natural framework for performing tests across multiple outcomes. We apply our method to simulated data and real data from a genotoxicity experiment.

Finally, we summarize our proposed methods in Chapter 6 and discuss some challenges for future research.

CHAPTER 2

BAYESIAN NONPARAMETRIC INFERENCE

Using the notation of Walker et al. (1999), let $Y_1, \ldots, Y_N \overset{iid}{\sim} F$ be defined on some space Ω. In the Bayesian parametric framework, F would be a known distribution function, P_Θ, and priors would be assumed for the unknown parameters Θ. Nonparametric inference presents a different line of thinking in that F is treated as an unknown function with prior P_Ω.

Seemingly the most popular nonparametric prior is the Dirichlet process (Ferguson 1973, 1974) due to its ease in implementation. Some important applications and theoretical work with the Dirichlet process include Antoniak (1974), Susarla and Van Ryzin (1976), Sethuraman (1994), Escobar (1994), MacEachern (1994), West et al. (1994), Escobar and West (1995), Bush and MacEachern (1996), Mukhopadhyay and Gelfand (1997), Kleinman and Ibrahim (1998), MacEachern and Müller (1998), Neal (2000), Kottas and Gelfand (2001), Carota and Parmigiani (2002), and Dunson (2004). Some more general classes of priors have also demonstrated good properties in certain applications such as the Pólya tree (Ferguson, 1974; Lavine, 1992, 1994; Muliere and Walker, 1997; Walker and Mallick, 1997; Hanson and Johnson, 2002) and neutral to the right processes (Doksum, 1974; Kalbfleisch, 1978; Ferguson and Phadia, 1979; Hjort, 1990). Reviews of previous work in Bayesian nonparametric inference can be found in Walker et al. (1999) and Müller and Quintana (2004). In addition, summaries of nonparametric priors used in survival analysis are provided by Ibrahim et al. (2001). We will begin this chapter with a detailed discussion of the Dirichlet process and later we will provide brief discussions of the Pólya tree (Section 2.2) and independent and dependent increments models, including neutral to the right processes (Section 2.3).

2.1 The Dirichlet Process

2.1.1 General Framework

The Dirichlet process was originally introduced by Ferguson (1973, 1974) as a convenient method of eliciting a nonparametric prior for F using the Dirichlet distribution, which we now define. For a $(k-1)$-dimensional vector of positive random variables, $(z_1, \ldots, z_{k-1})$, the $(k-1)$-variate Dirichlet distribution, Dirichlet$(\alpha_1, \ldots, \alpha_k)$ is defined by the joint density

$$f(z_1, \ldots, z_{k-1}) = \left(\frac{\Gamma(\alpha)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \right) \left(\prod_{j=1}^{k-1} z_j^{\alpha_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} z_j \right)^{\alpha_k - 1}, \tag{2.1}$$

where $\sum_{j=1}^{k-1} z_j \leq 1$, $\alpha = \sum_{j=1}^{k} \alpha_j$, and

$$\mathrm{E}(Z_j) = \frac{\alpha_j}{\alpha} \qquad \mathrm{Var}(Z_j) = \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha + 1)}. \tag{2.2}$$

Now consider the disjoint subsets of Ω, $B_1, \ldots, B_k$, where $\Omega = \bigcup_{j=1}^{k} B_j$. For an axis of values, $B_1, \ldots, B_k$ can be thought of as the set of non-overlapping intervals which comprise the axis. The Dirichlet process prior (DPP), $F \sim \mathrm{DP}(\alpha_0 F_0)$, assumes that $(F(B_1), \ldots, F(B_k))$ has a Dirichlet distribution with parameters $(\alpha_0 \mathrm{F}_0(B_1), \ldots, \alpha_0 \mathrm{F}_0(B_k))$, where F_0 is a known distribution, or base measure, and α_0 is a precision parameter which accounts for deviations from this parametric structure. As shown by Ferguson (1973), the posterior distribution of F is also a Dirichlet process with parameter $\alpha_0 F_0 + N \cdot F_N$, where F_N is the empirical cdf. Thus, by choosing a small α_0, one can obtain a diffuse prior, causing posterior estimates to be more data driven.

A couple of different authors have provided tractable representations of the DP. For instance, Blackwell and MacQueen (1973) developed a Pólya urn representation of the DP, which has proven useful in the development of Gibbs sampling algorithms (e.g., Escobar and West, 1995; MacEachern, 1994; West et al., 1994). This representation is described in greater detail in Section 2.1.3. Sethuraman (1994) developed an alternative

representation of the DP where $G \sim \text{DP}(\alpha_0 G_0)$ implies that

$$G = \sum_{i=1}^{\infty} w_i(\mathbf{v})\delta_{\theta_i}$$
$$\theta_i \overset{\text{iid}}{\sim} G_0 \text{ and } w_i(\mathbf{v}) = v_i \prod_{j<i}(1 - v_j) \text{ where } v_j \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_0), \tag{2.3}$$

δ_{θ_i} denotes a point mass at $\boldsymbol{\theta}_i$, and $\mathbf{v} = \{v_1, v_2, \ldots\}$. The mixing proportions $w_i(\mathbf{v})$ in the above representation are generated by successively breaking a "stick" of unit length into an infinite number of pieces, and thus (2.3) is often referred to as the *stick-breaking representation* of the DP. This representation makes it clear that G is discrete under a DPP, a result that was previously shown by Blackwell (1973).

2.1.2 Dirichlet Process Mixture (DPM)

The discrete nature of the DP is obviously problematic under a continuous Y. A simple solution to this problem is to use a Dirichlet process mixture or DPM (Antoniak, 1974). Consider a random vector $\mathbf{y}_i$ of length n_i whose distribution, F, is known and dependent upon a set of latent variables $\boldsymbol{\phi}_i$. In a Dirichlet process mixture, the DPP is shifted to $\boldsymbol{\phi}_i$ to ensure that $\mathbf{y}_i$ has a continuous distribution while still relaxing distributional assumptions. Thus, the model has the following hierarchical structure:

$$\begin{aligned} \mathbf{y}_i|\boldsymbol{\phi}_i &\sim F(\cdot; \boldsymbol{\phi}_i) \\ \boldsymbol{\phi}_i|G &\sim G \\ G|\alpha_0, \boldsymbol{\psi}_0 &\sim \text{DP}\Big(\alpha_0 G_0(\cdot; \boldsymbol{\psi}_0)\Big), \end{aligned} \tag{2.4}$$

where $\boldsymbol{\psi}_0$ are the parameters of the parametric base measure G_0. Using the stick-breaking representation of the DP, the DPM can also be described by the following process:

$$\begin{aligned} &\mathbf{y}_i|z_i \sim F(\cdot; \boldsymbol{\theta}_{z_i}) \quad z_i|\mathbf{v} \sim \text{Multinomial}(\mathbf{w}(\mathbf{v})) \quad i = 1, \ldots, N \\ &v_j|\alpha_0 \sim \text{Beta}(1, \alpha_0) \qquad \boldsymbol{\theta}_j|\boldsymbol{\psi}_0 \sim G_0(\cdot; \boldsymbol{\psi}_0) \qquad j = 1, 2, \ldots, \end{aligned} \tag{2.5}$$

where z_i indicates the mixture component with which $\mathbf{y}_i$ is associated and $\mathbf{w}(\mathbf{v}) = \{w_1(\mathbf{v}), w_2(\mathbf{v}), \ldots\}$.

A simple example of a DPM is the seminal work by Escobar (1994) whose objective

was to provide nonparametric estimates of normal means. The hierarchical structure of this model is simply $Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i \sim G$, and $G \sim \mathrm{DP}(\alpha_0 G_0)$. Other authors have used the DPM for density estimation (e.g., West et al., 1994; Escobar and West, 1995) and to construct semiparametric hierarchical models, as discussed in Section 2.1.4.

As noted by Neal (2000), sometimes a Dirichlet process mixture is referred to as a Mixture of Dirichlet processes (MDP) in the literature because the full conditional posterior of G is an MDP (Antoniak, 1974). However, this characterization will be avoided since models are usually characterized by their prior distributions and not their posteriors (Neal, 2000).

2.1.3 Computation under the DPM

Pólya Urn Gibbs Sampling

As mentioned earlier, the Pólya urn representation of the DP (Blackwell and MacQueen, 1973) has motivated Gibbs sampling methods for the DPM. Given $\mathbf{\Phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N\}$ constitute a set of exchangeable random variables, when G is integrated over its DPP it can be shown that $\mathbf{\Phi}$ may be generated according to the sequence (Escobar, 1994):

$$\begin{aligned} \boldsymbol{\phi}_1 &\sim G_0 \\ \boldsymbol{\phi}_2|\boldsymbol{\phi}_1 &\sim \frac{\alpha G_0 + \delta_{\boldsymbol{\phi}_1}}{\alpha_0 + 1} \\ &\vdots \\ \boldsymbol{\phi}_N|\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{N-1} &\sim \frac{\alpha G_0 + \sum_{j=1}^{N-1} \delta_{\boldsymbol{\phi}_j}}{\alpha_0 + N - 1}, \end{aligned} \tag{2.6}$$

As noted by West et al. (1994), any realization of $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N$ lies in a set of $K \leq N$ distinct values or clusters, with common values $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$. Thus, the conditional prior of $\boldsymbol{\phi}_i$ given $\boldsymbol{\phi}_i^{(i)} = \{\boldsymbol{\phi}_j : j \neq i\}$ is:

$$\left(\frac{\alpha_0}{\alpha_0 + N - 1}\right) G_0 + \left(\frac{1}{\alpha_0 + N - 1}\right) \sum_{j=1}^{K^{(i)}} n_j^{(i)} \delta_{\boldsymbol{\theta}_j^{(i)}}, \tag{2.7}$$

where $\boldsymbol{\phi}_i^{(i)}$ takes on $K^{(i)}$ distinct values, $n_j^{(i)}$ of which have the common value $\boldsymbol{\theta}_j^{(i)}$.

From (2.7), the full conditional posterior of $\boldsymbol{\phi}_i$ follows immediately

$$\pi(\boldsymbol{\phi}_i|\mathbf{Y}, \boldsymbol{\phi}_i^{(i)}) = q_{i0}G_{i0} + \sum_{j=1}^{K^{(i)}} q_{ij}\delta_{\boldsymbol{\theta}_j^{(i)}}, \tag{2.8}$$

where $\mathbf{Y}$ denotes the total data and each q_{ij} is a multinomial probability given by

$$q_{ij} = \begin{cases} c \cdot \alpha_0 h_i(\mathbf{y}_i) & j = 0 \\ c \cdot n_j^{(i)} f_j(\mathbf{y}_i|\theta_j) & j > 0, \end{cases}$$

where

- G_{i0} is the posterior obtained by updating the base measure G_0 with the likelihood, or equivalently
$$dG_{i0}(\boldsymbol{\phi}_i) \propto f_i(\mathbf{y}_i|\boldsymbol{\phi}_i)dG_0(\boldsymbol{\phi}_i),$$
- $h_i(\mathbf{y}_i)$ is the observed value of the marginal density of $\mathbf{y}_i$ under the base measure,
$$h_i(\mathbf{y}_i) = \int f_i(\mathbf{y}_i|\boldsymbol{\phi}_i)dG_0(\boldsymbol{\phi}_i)$$
- c is a normalization constant

The predictive distribution of a future $\boldsymbol{\phi}_i$, $\boldsymbol{\phi}_{N+1}$ is also easily obtained under the Pólya urn representation of the DP,

$$\pi(\boldsymbol{\phi}_{N+1}|\boldsymbol{\phi}) = \left(\frac{\alpha_0}{\alpha_0 + N}\right)G_0 + \left(\frac{1}{\alpha_0 + N}\right)\sum_{j=1}^{K} n_j\delta_{\boldsymbol{\theta}_j}. \tag{2.9}$$

For more details on these prior, posterior, and predictive distributions, please see MacEachern (1994) and West et al. (1994).

As demonstrated by Escobar (1994) and Escobar and West (1995), posterior computation may proceed using a Gibbs sampling algorithm which updates $\boldsymbol{\Phi}$ from (2.8). However, this approach is subject to slow mixing since $\boldsymbol{\theta}$ is rarely updated. Thus, based on the initial ideas of MacEachern (1994), West et al. (1994), propose a more efficient algorithm which updates the number of clusters (K), the configuration of subjects, and the unique values at each iteration of the MCMC. Let $\mathbf{S} = \{S_1, \ldots, S_N\}$ define a set of configuration indicators, where $S_i = j$ if $\boldsymbol{\phi}_i = \boldsymbol{\theta}_j$. The full conditional posterior of S_i

follows from (2.8),

$$P(S_i = j|\mathbf{Y}, \boldsymbol{\theta}^{(i)}) = q_{ij}. \tag{2.10}$$

Thus, the sampling algorithm proceeds as follows:

1. Given the values of $\boldsymbol{\theta}$ and $\mathbf{S}$ obtained from the previous iteration of a Gibbs sampler, sample $S_1, \ldots, S_N$ from (2.10) with a new $\boldsymbol{\phi}_i$ drawn from G_{i0} when $S_i = 0$.

2. Given the updated values of K and $\mathbf{S}$, update $\boldsymbol{\theta}$ using the posteriors,

$$\pi(\boldsymbol{\theta}_j|\mathbf{Y}, \mathbf{S}) \propto \left(\prod_{i:S_i=j} f_i(\mathbf{y}_i|\boldsymbol{\theta}_j) \right) dG_0(\boldsymbol{\theta}_j), \tag{2.11}$$

   for $j = 1, \ldots, K$.

The computational ease of DPMs depends upon whether or not the base measure is conjugate; i.e., does G_0 result in G_{i0} from the same family of distributions. When G_0 is conjugate, $h_i(\mathbf{y}_i)$ has a closed form and posterior computation is simple using the above sampling algorithm. However, when conjugacy does not hold, more complex sampling algorithms must be used. West et al. (1994) proposed estimating $h_i(\mathbf{y}_i)$ using either numerical quadrature or a Monte Carlo simulation which uses the average value of $f(\mathbf{y}_i|\boldsymbol{\phi}_i)$ over several values of $\boldsymbol{\phi}_i$ sampled from G_0. An alternative method was proposed by MacEachern and Müller (1998) which avoids numerical integration. Under their approach, N *candidate atoms* are sampled from G_0 at each iteration: $\boldsymbol{\theta}_{K+1}, \ldots, \boldsymbol{\theta}_{K+N}$. If $n_l^{(i)} > 0$ for subject i currently in cluster l, then $\boldsymbol{\phi}_i$ is sampled according to the probability function,

$$\pi(\boldsymbol{\phi}_i|\boldsymbol{\theta}^{(i)}, \mathbf{Y}) \propto \left(\frac{\alpha_0}{K+1} \right) f(\mathbf{y}_i|\boldsymbol{\theta}_{K+i}) \delta_{\boldsymbol{\theta}_{K+i}} + \sum_{j=1}^{K} q_{ij} \delta_{\boldsymbol{\theta}_j}, \tag{2.12}$$

else if $n_l^{(i)} = 0$, $\boldsymbol{\phi}_i$ is unchanged with probability $(K-1)/K$ and with probability $1/K$ it is sampled from (2.12) with the minor modification that K is replaced with $K-1$. More recently, Neal (2000) proposed a related algorithm which samples $m \geq 1$ candidate atoms for each subject, thus increasing the probability that each subject belongs to their own cluster. In the same article, Neal also describes a method in which cluster membership is updated using Metropolis steps.

Other Methods

Although Pólya urn sampling methods are easy to implement, in some situations they can be slow to converge and mix poorly, even when MacEachern's (1994) and West et al.'s (1994) algorithms are used. For instance, Jain and Neal (2004) note that when two or more mixture components have similar parameter values, the Gibbs sampler can become trapped in a local mode that corresponds to an incorrect clustering of data points. To address this issue, Jain and Neal proposed a Metropolis-Hastings procedure which increases the efficiency of cluster assignment by splitting and merging entire clusters at each step of the MCMC.

Another limitation of the Pólya urn sampling methods is that they do not provide direct samples from the DP. As a result, one cannot estimate credible intervals for the functional assigned the DPP. To address this issue, several authors have proposed computational algorithms which sample from a truncated version of Sethuramans' (1994) stick-breaking representation. For instance, Ishwaran and James (2001) proposed a blocked Gibbs sampler which updates the atoms, weights, and remaining hyperparameters from their joint distributions. For related approaches see Muliere and Tardella (1998), Ishwaran and James (2002), and Gelfand and Kottas (2002).

Some other computational methods for DPMs include alternatives to MCMC simulation. For example, importance sampling-type methods have been proposed by Liu (1996) and MacEachern et al. (1999). Newton and Zhang (1999) also proposed a predictive recursion method for estimating predictive distributions. Although these approaches are faster than Gibbs sampling, the importance sampling methods can produce large Monte Carlo error and predictive recursion tends to over-smooth (see, e.g., Quintana and Newton, 2000).

Recently, Blei and Jordan (2006) proposed a Variational Bayes (VB) approach to inference for the DPM. Under the stick-breaking representation (2.5), let $\mathbf{z} = (z_1, \ldots, z_N)'$ denote the mixture indicators for N subjects. The VB approach replaces the joint posterior of the stick-breaking parameters, $\pi(\mathbf{v}, \boldsymbol{\theta}, \mathbf{z}|\text{Data})$, with a *variational* distribution which truncates the stick-breaking process at M atoms:

$$q(\mathbf{v}, \boldsymbol{\theta}, \mathbf{z}) = \prod_{m=1}^{M-1} \text{Beta}(v_m; a_m, b_m) \prod_{m=1}^{M} q^*(\boldsymbol{\theta}_m; \boldsymbol{\eta}_m) \prod_{i=1}^{N} \text{Multinomial}(z_i; \boldsymbol{\pi}_m), \qquad (2.13)$$

where $q^*()$ denotes a distribution in the exponential family and $a_m, b_m, \boldsymbol{\eta}_m, \boldsymbol{\pi}_m$ $m = 1, \ldots, M$ are known as *variational* parameters, whose values are chosen to maximize a

lower bound on the log-marginal likelihood. As opposed to MCMC, VB has a single optimization criterion that can be used to assess convergence. In addition, Blei and Jordan have provided empirical evidence that VB is much faster than MacEachern's (1994) Gibbs sampler and Ishwaran and James' (2001) blocked Gibbs approach. However, a major disadvantage of this method relative to the MCMC approaches is that the estimates from VB are based on an approximation instead of the true posterior.

Hyperparameter estimation

As mentioned above, the base measure in the DPM, G_0, is usually fixed or assumed to come from a parametric family of distributions with a fixed hyperprior placed on its parameters, $\boldsymbol{\Psi}_0$ (see, e.g., West et al., 1994; Escobar and West, 1995; MacEachern and Müller, 1998). However, some authors have considered nonparametric estimation of G_0. For instance, some have assigned DPPs (e.g., Teh et al., 2006) or DPMs (e.g., Tomlinson and Escobar, 1999) to G_0. MacAuliffe et al. (2006) describe another approach in which G_0 is estimated every T^* iterations of the MCMC using kernel density estimates constructed from $\boldsymbol{\theta}$.

A few authors have also proposed methods for estimating α_0 from the data. For example, West (1992) assigned a gamma prior to α_0, while Carota and Parmigiani (2002) proposed a regression model. Liu (1996) developed sequential imputation and Gibbs sampling methods for approximating the MLE of α_0 (see also MacAuliffe et al., 2006).

2.1.4 Random Effects Modelling

The computational methods developed for DPMs have made nonparametric modelling of random effects feasible. For instance, when a DPP with a normal base measure is assigned to the unknown distribution of a random block effect (e.g., Bush and MacEachern, 1996) or a random coefficient (e.g., Kleinman and Ibrahim, 1998) in a linear model, conjugacy is achieved and computation may proceed using West et al.'s (1994) method. Hierarchical count data can also be modelled by a conjugate DPM by specifying a gamma base measure in the DPP for the random effect distribution (see, e.g., Dunson, 2004). However, as discussed in Mukhopadhyay and Gelfand (1997), a conjugate G_0 does not exist in all hierarchical generalized linear models, such as logistic regression. In these settings, posterior computation requires a more intensive approach, such as MacEachern and Müller's (1998) method.

Although DPMs improve the flexibility of random effects models, smoothness of the random effect distribution is compromised due to the almost surely discrete restriction of the DPP. However, this problem can be fixed by adding another level of hierarchy to the model; i.e. one may assume that G is fixed given the latent variables $\boldsymbol{\nu}_i$ and assign a DPP to the unknown distribution of $\boldsymbol{\nu}_i$. For example, Müller and Rosner (1997) use a DPM to model subject-specific coefficients within a nonlinear model for blood count data.

Some authors have developed generalizations of the DP to allow the distributions of random effects or other latent variables to depend on covariate values. For example, dependent nonparametric processes have been proposed by Cifarelli and Regazzini (1978), MacEachern (1999, 2000), Müller et al. (2004), Dunson (2006), and Dunson and Pillai (2004). In Chapter 5, we will discuss these methods in more detail and propose a generalization of the Dunson (2006) approach.

2.2 The Pólya Tree

2.2.1 General Framework

Another common nonparametric prior is a generalization of the Dirichlet process known as the Pólya tree or PT for short(Ferguson, 1974; Lavine, 1992, 1994). The Pólya tree is defined by an infinite set of binary partitions of the space Ω. Let B_0 and B_1 be obtained by splitting Ω into two pieces. B_0 and B_1 are split into (B_{00}, B_{01}) and (B_{10}, B_{11}), respectively, and this process is repeated *ad infinitum*. For some m, let $\epsilon = \epsilon_1 \cdots \epsilon_m$ with $\epsilon_k \in \{0, 1\}$ for $k = 1, \ldots, m$ so that ϵ defines a unique set of partitions, B_ϵ. A random probability measure F on Ω is said to have a Pólya tree prior if there exists nonnegative numbers $\mathcal{A} = (\alpha_0, \alpha_1, \alpha_{00}, \ldots)$ and random variables $\mathcal{C} = (C_0, C_{00}, C_{10}, \ldots)$ such that

1. All random variables in $\mathcal{C}$ are independent
2. For every ϵ, $C_{\epsilon 0} \sim \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$
3. For every $m = 1, 2, \ldots$ and every $\epsilon = \epsilon_1 \cdots \epsilon_m$,

$$F(B_{\epsilon_1 \ldots \epsilon_m}) = \left(\prod_{j=1;\epsilon_j=0}^{m} C_{\epsilon_1 \cdots \epsilon_{j-1} 0} \right) \left(\prod_{j=1;\epsilon_j=1}^{m} (1 - C_{\epsilon_1 \cdots \epsilon_{j-1} 0}) \right), \tag{2.14}$$

where the first terms (for $j = 1$) are interpreted as C_0 and $1 - C_0$ (Lavine, 1992). The formal shorthand notation is $F \sim \text{PT}(\Pi, \mathcal{A})$, where Π is the set of partition probabilities.

An attractive feature of the Pólya tree prior is that it is conjugate, i.e. $F|\mathbf{Y} \sim \text{PT}(\Pi, \mathcal{A}|\mathbf{Y})$, where $\mathcal{A}|\mathbf{Y} = \alpha_\epsilon + n_\epsilon$ and n_ϵ is the number of observations from $\mathbf{Y}$ in B_ϵ (Ferguson, 1974; Lavine, 1994). Thus, when $y_1, \ldots, y_N$ are all uncensored, the posterior predictive distribution of a future observation, Y_{N+1}, follows immediately

$$P(Y_{N+1} \in B_\epsilon) = \prod_{k=1}^{m} \frac{\alpha_{\epsilon_1 \cdots \epsilon_k} + n_{\epsilon_1 \cdots \epsilon_k}}{\alpha_{\epsilon_1 \cdots \epsilon_{k-1} 0} + \alpha_{\epsilon_1 \cdots \epsilon_{k-1} 1} + n_{\epsilon_1 \cdots \epsilon_{k-1}}}. \tag{2.15}$$

2.2.2 Applications

A Pólya tree priors can be centered on a probability distribution, F_0, by taking the partitions to coincide with percentiles of F_0 and assuming that $\alpha_{\epsilon 0} = \alpha_{\epsilon 1} = \alpha_\epsilon$ for each ϵ (Lavine, 1992). At level m, this would correspond to the partitions $B_j = \{(F_0^{-1}\big((j-1)/2^m\big), F_0^{-1}\big(j/2^m\big)\}$ for $j = 1, \ldots, 2^m$, where $F_0^{-1}(0) = -\infty$ and $F_0^{-1}(1) = \infty$. To facilitate computation, $\mathcal{C}$ is terminated at some finite level M and attention is restricted to the $r = 2^M$ partitions given by $\pi_M = (B_1, \ldots, B_M)$ (Lavine, 1992).

Some additional attention must be given to the choice of α_ϵ. As seen in (2.15), as α_ϵ decreases for each m, the posterior predictive distribution of Y_{N+1} approaches the empirical cdf. Thus, in this sense, α_ϵ is similar to the parameter α_0 in the DP in that it quantifies the prior confidence in the base distribution. However, interpreting α_ϵ as a precision parameter is limiting since it also determines the smoothness of F (Lavine, 1992). To see this trait, consider the probability of two Y_i's taking on the same value,

$$P(Y_i = y|Y_j = y) = \prod_{k=1}^{\infty} \frac{\alpha_{\epsilon_1 \cdots \epsilon_k} + 1}{\alpha_{\epsilon_1 \cdots \epsilon_{k-1} 0} + \alpha_{\epsilon_1 \cdots \epsilon_{k-1} 1} + 1}. \tag{2.16}$$

According, to Mauldin et al. (1992), (2.16) equals 0 is a sufficient condition for a continuous F, which can be approached by allowing α_ϵ to increase with m. Ferguson (1974) claims that $\alpha_\epsilon = m^2$ implies that F is continuous with probability one, which Lavine (1992) calls a "sensible canonical choice" for α_ϵ. Alternatively, the Pólya tree may be specialized to the Dirichlet process by choosing $\alpha_\epsilon = \alpha/(2^m)$, but as mentioned previously, this ensures that F is discrete (Blackwell, 1973; Ferguson, 1973).

Under censored data, the partitions of the Pólya tree must coincide with observed

censoring times to maintain conjugacy (Muliere and Walker, 1997). Thus in order to center the tree on a given distribution, Muliere and Walker (1997) propose setting each $\alpha_\epsilon = \gamma_m F_0(B_\epsilon)$, where γ_m is a constant. This method not only ensures that $\mathrm{E}(F(B_\epsilon)) = F_0(B_\epsilon)$, but it also allows one to specify γ_m such that the continuity of F is ensured. In addition, the predictive probabilities retain a simple form,

$$P(Y_{N+1} \in B_\epsilon) = \prod_{k=1}^{m} \frac{\alpha_{\epsilon_1 \cdots \epsilon_k} + n_{\epsilon_1 \cdots \epsilon_k}}{\alpha_{\epsilon_1 \cdots \epsilon_{k-1} 0} + \alpha_{\epsilon_1 \cdots \epsilon_{k-1} 1} + n_{\epsilon_1 \cdots \epsilon_{k-1}} - s_{\epsilon_1 \cdots \epsilon_{k-1}}}, \tag{2.17}$$

where $s_{\epsilon_1 \cdots \epsilon_{k-1}}$ is the number of observations censored in B_ϵ.

In some recent applications, Pólya tree priors have been used to model random error or subject-specific random effects in semiparametric regression models. For example, Walker and Mallick (1997) assigned Pólya tree priors to random effects in hierarchical generalized linear models and frailty models. Although the full conditional posterior of F is tractable, the conditional posteriors of the random effects are not. Thus, computation requires either an indirect sampling method to obtain values of the random effects (e.g., Metropolis sampling) or constructing additional latent variables to achieve conjugacy. These methods are more computationally intensive than the Pólya urn sampling methods for the DPM.

Another disadvantage of the Pólya tree is its sensitivity to the choice of partitions. This latter problem may be alleviated by using a mixture of Pólya trees in which the parameters of the centering distribution, F_0, and/or $\mathcal{A}$ are random. This approach also improves computational efficiency since it does not require one to choose F_0 with a large enough variance to cover the support of F. Hanson and Johnson (2002) used a mixture of Pólya trees in an accelerated failure time model.

2.3 Independent and Dependent Increments Models

2.3.1 Neutral to the Right Processes

The Dirichlet process is also a special case of a general set of nonparametric priors known as neutral to the right (NTTR) processes. As defined by Doksum (1974), a distribution function, $F(t)$, is NTTR if it can written in the form $F(t) = 1 - e^{-Y(t)}$ where $Y(t)$ is a process with independent increments and

1. $Y(t)$ is nondecreasing a.s.

2. $Y(t)$ is right continuous a.s.

3. $\lim_{t\to-\infty} Y(t) = 0$ a.s.

4. $\lim_{t\to\infty} Y(t) = \infty$ a.s.

$Y(t)$ has at most a countably finite number of discontinuity points $t_1, t_2, \ldots$ with independent jumps $W_1, W_2, \ldots$ NTTR priors are generally specified in terms of the differences

$$Z(t) = Y(t) - \sum_j W_j 1(t_j \leq t < \infty),$$

which do not have any points of discontinuity. As is the case with PT priors, NTTR priors are always conjugate (Doksum, 1974).

Some special cases of NTTR processes have proven to be useful in Bayesian survival analysis. For example, Kalbfleisch (1978) used a special NTTR process known as the gamma process to develop a semiparametric proportional hazards model. In Kalbfleisch's model, $Y(t) = \Lambda_0(t)$, the cumulative baseline hazard function, and $Z(t)$ is the increment in $\Lambda_0(t)$ at time t. Using a finite set of partitions of the time axis, $0 < \tau_1 < \cdots < \tau_M < \infty$, $Z(\tau_j) \overset{iid}{\sim} \text{Ga}(c_0 z_0(\tau_j), c_0)$ where $z_0(\tau_j)$ is an initial guess at the true value of the increment and c_0 is a precision parameter for $j = 1, \ldots, M$. Hjort (1990) also developed discrete time and continuous time beta process priors for the increments in the cumulative baseline hazard. The gamma and beta processes are easy to implement (due to conjugacy) and provide a simple structure for incorporating a priori knowledge about the hazard function into analyses.

2.3.2 Dependent Increments Models

Although NTTR priors have some nice properties, the independent increments assumption may not be reasonable in some settings. In particular, one would typically expect the hazards from adjacent time intervals to be correlated a priori. To address this issue, some authors have proposed generalizations. Let λ_{0j} denote the baseline hazard over the jth partition of the time axis. To induce correlation, Aslanidou et al. (1998) modelled the baseline hazard using the discrete time martingale process of Arjas and Gasbarra (1994):

$$\lambda_{0j}|\lambda_{01}, \ldots, \lambda_{0(j-1)} \sim \text{Ga}\left(\nu, \frac{\nu}{\lambda_{0(j-1)}}\right). \tag{2.18}$$

In (2.18), the smoothness of the baseline hazard is controlled by the value of ν. Sinha (1998) proposed an alternate model for the baseline hazard which uses a correlated process described by Gamerman (1991):

$$\log(\lambda_{0(j+1)}) = \log(\lambda_{0j}) + e_{j+1} \qquad e_{j+1} \sim \mathrm{N}(0, \sigma^2). \tag{2.19}$$

Sinha's prior only requires assumptions on the level of smoothing between adjacent intervals. However, posterior computation requires the use of a posterior likelihood, and thus future predictions cannot be made. More recently, Nieto-Barajas and Walker (2002) developed a Markov gamma process for piecewise constant hazards. Their approach induces correlation through the use of latent Poisson and gamma random variables and has a convenient conjugacy property.

CHAPTER 3

EMPIRICAL BAYES FITTING OF SEMIPARAMETRIC RANDOM EFFECTS MODELS TO LARGE DATA SETS

3.1 Introduction

When multilevel data are compiled from a large study, multiple centers, or lengthy followups, the number of observations can become massive. In these situations, it can be difficult to fit random effect models (Laird and Ware, 1982) using standard frequentist (e.g., Wolfinger et al., 1994) and Bayesian (e.g., Zeger and Karim, 1991) methods due to convergence or memory problems. These difficulties are illustrated by data collected in the Collaborative Perinatal Project (CPP), a prospective epidemiologic study of pregnant women and their children in the U.S. from 1959-1974. Recently, Chen et al. (2006) examined the relationship between maternal smoking habits and childhood obesity within $N = 34,866$ children in the CPP using generalized estimating equations, or GEE (Liang and Zeger, 1986). Although GEE allowed the authors to perform inferences on population mean effects, it would have also been interesting to assess how smoking varied in its effect across the children. In addition, a random effects model would have relaxed assumptions on missingness by requiring only missing at random (MAR) instead of missing completely at random (MCAR). Unfortunately, investigators were unable to fit random effects models to the CPP data using frequentist or Bayesian methods due, in part, to the large sample size. For example, SAS PROC

MIXED failed to converge.

When a data set is large, as in the CPP, it would also be advantageous to use the abundant information to relax assumptions of models, such as normality of random effects. Bayesian nonparametric or semiparametric methods are attractive in these settings since the random effect distribution can be assigned a prior which reflects a priori knowledge about the shape or location. For instance, a Dirichlet process prior (DPP) may be assigned to the random effect distribution (see, for example, West et al., 1994, Bush and MacEachern, 1996, Mukhopadhyay and Gelfand, 1997, and Kleinman and Ibrahim, 1998), which reduces the number of random effects to a set of $K \leq N$ unique values. Each of these K clusters represent subjects with common latent traits which may include interesting genetic or environmental factors worthy of future study. Despite the promise of the DPP, K increases rapidly with N which can lead to a scientifically implausible and computationally impractical number of clusters when N is very large.

Unfortunately, few authors have considered adapting the computational methods for the DPP to handle large data sets. Although Blei and Jordan's (2006) variational inference method can substantially reduce computation time, especially for large N, the approach relies on replacing the true posterior density with a lower bound having unknown accuracy. Potentially, the particle filtering methods described by Chopin (2002), Ridgeway and Madigan (2003), and Balakrishnan and Madigan (2006) could be generalized to make Bayesian nonparametric inference feasible for large data sets. In this paper, we consider an alternate approach which involves scaling-down the size of the data prior to performing MCMC. Existing methods for *data squashing* include methods which fit models to both real and generated data, also known as *pseudo-data*, which are representative of the complete data. For example, DuMouchel et al. (1999) and Madigan et al. (2002) construct pseudo-data using a moment matching and likelihood-based approach, respectively, while Owen (2003) uses a random sample of the complete data. Huang et al. (2005) proposed a related Bayesian method for fitting hierarchical models, though their approach is parametric and involves combining posteriors from several sub-samples of the data.

Motivated by the CPP data, we propose a data squashing procedure for fitting semiparametric random effects models to large, longitudinal data sets. Our method consists of two stages. First, a multivariate clustering procedure is used to identify $G << N$ groups of *scientifically indistinguishable* subjects, meaning that differences between subjects in each group are so small that they would not be considered signifi-

cant by an expert of the subject matter. In the second stage, we use a DPP to model the G cluster means, further clustering the groups from the first stage. By applying the DPP to the cluster means instead of the complete data, we reduce both the computation time and the number of latent classes. In addition, our use of expert opinion improves the scientific justification of clustering. For discussion of the importance of expert elicitation, refer to Kadane and Wolfson (1998), Meyer and Booker (2001), and Garthwaite et al. (2005).

In Section 3.2, we discuss the CPP data and previous results. In Section 3.3, we propose the method. Section 3.4 contains a series of simulation examples, Section 3.5 applies the approach to the CPP data, and Section 3.6 discusses the results.

3.2 Maternal Smoking and Childhood Growth Data

As described by Broman (1984), the Collaborative Perinatal Project (CPP) was a large prospective study of pregnancy and childhood development. The study consisted of $55,043$ pregnancies enrolled at 12 study centers in the U.S. between 1959 to 1965 and included measurements obtained from children starting at birth and concluding at age 8. The investigators targeted 20 different outcomes in the study including the presence of mental and communicative disorders in the children and physical growth.

The CPP measured smoking during pregnancy and child height and weight at followup visits. Chen et al. (2006) used the measurements at birth and at years 1, 3, 4, 7, and 8 to determine the effects of maternal smoking on childhood growth amongst 34,866 children (17,348 boys and 17,518 girls). Categories of smoking exposure included (1) never smoked, (2) ex-smokers, and (3) currently smoking based on questionnaire data at registration or subsequent prenatal visits. Being unable to implement random effects models due to the large sample size, the authors used GEE to demonstrate that mothers who smoked during pregnancy had infants with lower birth weight, but by age 8, these children had a greater risk of being overweight.

As mentioned in Section 3.1, mixed effects models have several advantages including their ability to assess heterogeneity across subjects and relaxed assumptions on missingness. In exploratory analyses of the data, we found that the heavier children at age 4 were more likely to miss followups at ages 7 and 8. Thus, the MCAR assumption of GEE may be violated. In this paper, we wish to address these concerns by fitting a random effects model to the CPP data. We focus on the effects of smoking on weight in females to illustrate the approach as Chen et al. found the largest effect in this group.

3.3 Methods

3.3.1 General Motivation

For $i = 1, \ldots, N$, let $\mathbf{y}_i = (\mathrm{y}_{i1}, \ldots, \mathrm{y}_{in_i})'$ denote a set of n_i longitudinal measurements on subject i. Letting $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$ denote a set of predictors, we focus on the linear random effects model

$$\big[\mathbf{y}_i \,|\, \mathbf{b}_i, \mathbf{X}_i\big] \sim \mathrm{N}\big(\mathbf{X}_i\mathbf{b}_i, \tau^{-1}\mathbf{I}_{n_i}\big), \tag{3.1}$$

where $\mathbf{I}_{n_i}$ is an $n_i \times n_i$ identify matrix and $\mathbf{b}_i = (b_{i1}, \ldots, b_{ip})' \sim H$, an unknown distribution with mean $\boldsymbol{\beta}$ and covariance $\mathbf{V}$.

As N becomes very large and both n_i and p remain modest, many subjects have essentially identical values with $\mathbf{y}_i \approx \mathbf{y}_j$ and $\mathbf{X}_i \approx \mathbf{X}_j$ for many different pairs i, j. Outcomes, such as weight, that are treated as continuous are often truncated or rounded when recorded, limiting the number of unique values in the data. In addition, values which are so close that a subject matter expert would consider them *scientifically indistinguishable* can be grouped together without loss of important information. Under these circumstances, the data are adequately summarized by values for $G << N$ clusters. For an observation i in cluster g let

$$\begin{aligned} \mathbf{y}_{gi} &= \overline{\mathbf{y}}_g + \boldsymbol{\epsilon}_{gi} \\ \mathbf{X}_{gi} = \overline{\mathbf{X}}_g + \boldsymbol{\Delta}_{gi} \qquad & \mathbf{b}_{gi} = \overline{\mathbf{b}}_g + \boldsymbol{\phi}_{gi} \end{aligned} \tag{3.2}$$

where $\overline{\mathbf{y}}_g$, $\overline{\mathbf{X}}_g$, and $\overline{\mathbf{b}}_g$ are the cluster-specific means of the response, predictors, and random effects, $\boldsymbol{\epsilon}_{gi}$ and $\boldsymbol{\phi}_{gi}$ are random variables, and $\boldsymbol{\Delta}_{gi}$ is a matrix of constants. When the G clusters adequately represent the heterogeneity in the data, the observed values of $\boldsymbol{\epsilon}_{gi}$, $\boldsymbol{\phi}_{gi}$, and $\boldsymbol{\Delta}_{gi}$ are all approximately zero. Thus, $\boldsymbol{\beta} = \mathrm{E}(\mathbf{b}_i)$ can be reasonably estimated by

$$\widehat{\boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{b}_i = \frac{1}{N}\sum_{g=1}^{G}\sum_{i\in g}(\overline{\mathbf{b}}_g + \boldsymbol{\phi}_{gi}) \approx \frac{1}{N}\sum_{g=1}^{G} m_g\overline{\mathbf{b}}_g = \widetilde{\boldsymbol{\beta}}, \tag{3.3}$$

where m_g is the number of subjects in cluster g.

Instead of fitting models to all N subjects, we propose an alternative approach in which we fit our model to the pseudo-sample, $(\mathbf{y}_1^*, \mathbf{X}_1^*), \ldots, (\mathbf{y}_G^*, \mathbf{X}_G^*)$, where $(\mathbf{y}_g^*, \mathbf{X}_g^*)$ represents the typical subject in cluster g (i.e., $\mathbf{y}_g^* = \overline{\mathbf{y}}_g$ and $\mathbf{X}_g^* = \overline{\mathbf{X}}_g$). In Section

3.2, we recommend a strategy for initial clustering of the N subjects in G groups. In Section 3.3, we propose a flexible Stage 2 clustering procedure which uses a DPP to avoid restrictions on H. Section 3.4 describes the MCMC algorithm and in Section 3.5 we discuss our approach to inference.

3.3.2 Stage 1 Clustering

We propose a stratified methodology to generate the first stage clusters. Although related to the data-sphere method used by DuMouchel et al. (1999), our procedure is geared to the random effects problem and incorporates knowledge of subject matter experts. Subject-specific data are first divided into q strata based on categorical predictors. For example, if there are two categorical predictors, one dichotomous and one with three levels, q should equal 6. Within each stratum, we wish to develop clusters of scientifically indistinguishable subjects based on the values of the continuous variables, i.e., the longitudinal responses and continuous predictors. For subject i in stratum j, we denote the values of these variables as $\mathbf{w}_{ji} = (\mathrm{w}_{ji1}, \ldots, \mathrm{w}_{jip_{ji}})'$. For ease in exposition, we will temporarily assume that $p_{ji} = p_j$ for $i = 1, \ldots, M_j$, where p_j is the number of continuous variables for each subject in stratum j and M_j is the stratum frequency. Prior to clustering, we transform $\mathbf{w}_{ji}$ to $\mathbf{z}_{ji} = (\mathrm{z}_{ji1}, \ldots, \mathrm{z}_{jip_j})'$, where

$$\mathrm{z}_{jik} = \frac{(\mathrm{w}_{jik} - \overline{\mathrm{w}}_{jk})}{\mathrm{s}_{\mathrm{w}_{jk}}},$$

and $\overline{\mathrm{w}}_{jk}$ and $\mathrm{s}_{\mathrm{w}_{jk}}$ denote the mean and standard deviation, respectively, of the kth continuous variable in stratum j.

Let the z-scores in stratum j be divided into G_j clusters whose location in p_j-space are represented by a set of data points or *seeds*, $\mathbf{c}_{j1}, \ldots, \mathbf{c}_{jG_j}$, where $\mathbf{c}_{jl} = (\mathrm{c}_{jl1}, \ldots, \mathrm{c}_{jlp_j})'$ and c_{jlk} is the average value of the kth standardized variable in cluster l. We assume that both the number of clusters and locations are unknown a priori, but through expert elicitation, we define a threshold r such that

$$d(\mathbf{z}_{ji}, \mathbf{c}_{jl}) = \sqrt{\sum_{k=1}^{p_j} (\mathrm{z}_{jik} - \mathrm{c}_{jlk})^2} \leq r \tag{3.4}$$

for subject j, i in cluster j, l. Thus, in a cluster of scientifically indistinguishable subjects, r is the elicited maximum distance between the data of a single subject and the

cluster seed, or the maximum radius of a cluster.

To elicit r, we recommend performing a set of exploratory cluster analyses and presenting the results to one or more subject matter experts. These analyses may be performed using a set of historical data, or alternatively, one stratum of the current data. In the latter method, the data used to elicit r will also be used in the second stage of the analysis, thus creating a sort of an empirical Bayes approach. In our analysis of the CPP data, we treated the data on male children of never smokers as our historical data, and we used it to choose an appropriate r for the female subjects. In our exploratory analyses, we used a range of r values to cluster the longitudinal weight of males with complete data (i.e., with followups at ages 0, 1, 3, 4, 7, and 8). Following each analysis, we plotted the growth curves from subjects in the cluster with largest radius (see Figure 3.1). Using these plots, we asked a panel of experts on body weight research (2 MD's, a PhD in Nursing, and an Exercise Physiologist) to tell us which clusters (each indexed by a radius, r) contain curves with potentially significant differences. In our example, 3 out of 4 panel members agreed that when r $\leq$ 2.14, the growth curves in each cluster were not significantly different. Thus, $r = 2.14$ was the obvious choice for the CPP. In other applications where there is substantial disagreement across the experts, the average elicited value could be used instead. Our method for choosing r is similar to the use of *opinion pools* to combine probability distributions elicited by several experts; for an example see Cooke and Goossens (2000).

Once we have specified r, we apply the following three-step methodology to cluster the continuous data in stratum j:

Step 1. Initialize cluster seeds.

Initialize G_j at 1 and let $\mathbf{c}_{j1}^{(0)} = \mathbf{z}_{j1}$. For $i = 2, \ldots, M_j$, if $d_{ji}^* = \min_l d(\mathbf{z}_{ji}, \mathbf{c}_{jl}^{(0)}) > r$, then increment G_j by 1 and define a new seed, $\mathbf{c}_{jG_j}^{(0)} = \mathbf{z}_{ji}$.

Step 2. Iteratively update the seeds. Initialize an index variable, t, at 1 and perform the following steps:

2.1 For $i = 1, \ldots, M_j$, if $d_{ji}^* \leq r$ assign $\mathbf{z}_{ji}$ to the cluster with the closest seed.

2.2 For $l = 1 \ldots, G_j$ compute

$$\mathbf{c}_{jl}^{(t)} = \frac{1}{m_{jl}} \sum_{i \in j,l} \mathbf{z}_{ji},$$

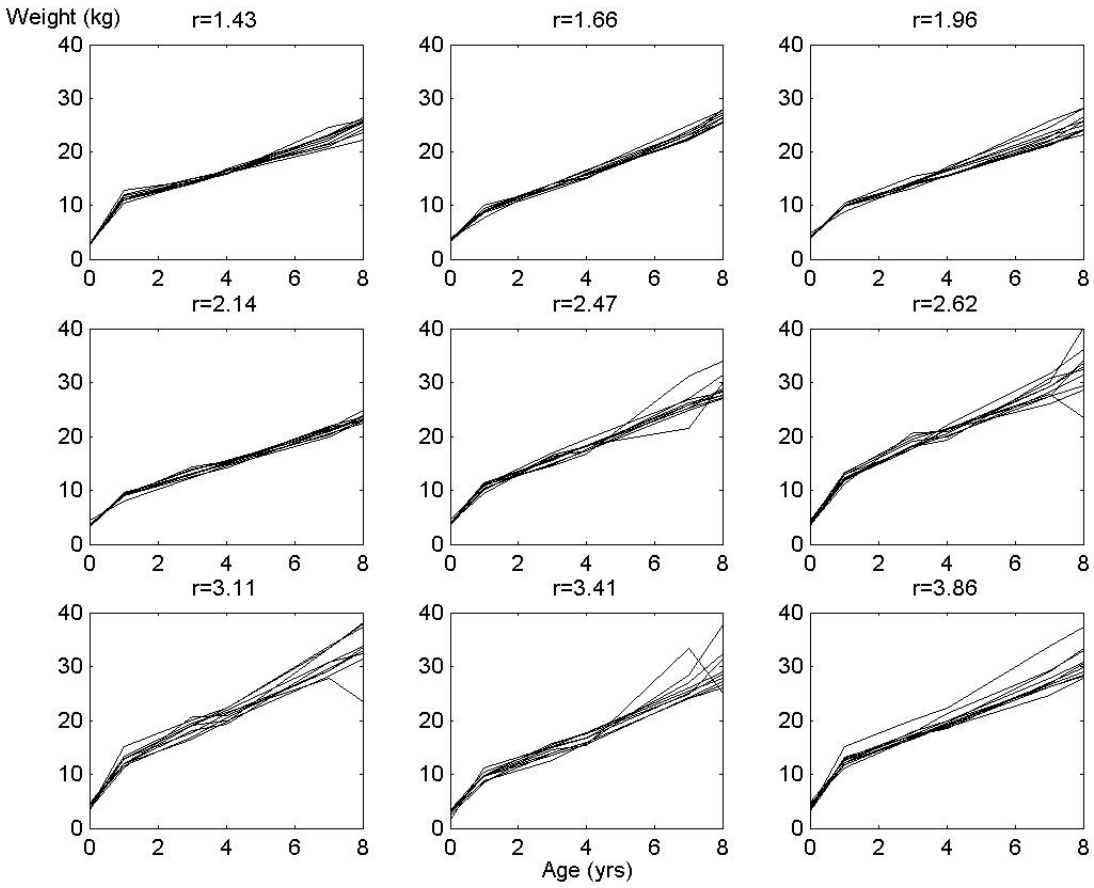

FIGURE 3.1: Plots used to elicit maximum radius, r, for CPP data. The subjects used in the analyses were male children of non-smoking mothers who were measured at each follow-up ($N = 1,115$). Each plot consists of 10 growth curves from the cluster with the largest radius, r. These curves correspond to the subjects furthest from and closest to the cluster seed, as well as 8 randomly chosen subjects.

where m_{jl} is the number of subjects currently in cluster j, l. Let $0 \leq \nu < 1$ denote a pre-specified convergence criterion such that changes in the cluster seeds less than or equal to $\nu \cdot d^*_{j0}$ are permissible, where d^*_{j0} denotes the minimum distance between the initial seeds. If $\max_l d(\mathbf{c}^{(t)}_{jl}, \mathbf{c}^{(t-1)}_{jl}) > \nu \cdot d^*_{j0}$, then increment t by 1 and repeat Steps 2.1 and 2.2, otherwise proceed to Step 3.

Step 3. Construct final clusters.

3.1 Repeat Step 2.1 using $\mathbf{c}^{(t)}_{j1}, \ldots, \mathbf{c}^{(t)}_{jG_j}$.

3.2 For all $i : d^*_{ji} > r$, assign $\mathbf{z}_{ji}$ to its own cluster and update the value of G_j accordingly.

Step 1 of our method is related to the leader algorithm (Hartigan, 1975), while Step 2 can be thought of as a form of k-means clustering (MacQueen, 1967) since the cluster seeds are the means of the observations assigned to each cluster when the algorithm is iterated until complete convergence (i.e., $\nu = 0$). A proof of convergence of our algorithm is provided in Appendix A. After completing Steps 1-3 for $j = 1, \ldots, q$, we compute the means of the untransformed variables in each cluster, $\overline{\mathbf{w}}_{jl} = \sum_{i \in (j,l)} \mathbf{w}_{ji}$. As mentioned in Section 3.1, these data (plus the values of any categorical predictors) will constitute our $G = \sum_{j=1}^{q} G_j$ pseudo-subjects.

The above method is attractive for many large data sets since it leads to the quick formulation of first stage clusters chosen to have minimal scientifically-important distances between them. By choosing r based on expert elicitation, we induce a prior on the clustering process. Our initialization method then uses this prior to identify the most important separations in the data. Another attractive feature of our method is that all three steps may be implemented using PROC FASTCLUS (SAS, version 9) and sample code is available upon request from the authors.

In many longitudinal studies, including the CPP, $p_{ji} \neq p_{ji'}$ for several pairs $(j, i), (j, i')$ due to missing followups. A simple solution is to stratify by missingness, but sometimes the number of patterns may be too numerous to make this feasible. For instance, there are 58 different missingness patterns in the CPP data. Thus, to resolve this problem, we recommend stratifying by the most common patterns and assigning the remaining subjects to the stratum for which they have the least number of missing variables. In each of these strata, the initial cluster seeds are chosen using subjects with complete data. Then, in Steps 2 and 3, subjects with missing observations are assigned to clusters

based on adjusted distances,

$$d_{adj}(\mathbf{z}_{ji}, \mathbf{c}_{jl}) = \sqrt{\frac{p_j}{p_{ji}} \sum (\mathrm{z}_{jik} - \mathrm{c}_{jlk})^2}, \tag{3.5}$$

where the sum is taken over the p_{ji} nonmissing variables for subject i in cluster j. As before, these subjects may still be assigned to their own cluster if $d^*_{ji} > r$ in Step 3, and thus, we do not ignore any important outliers.

3.3.3 Dirichlet process clustering

In the remaining sections of this chapter, we will drop the stratum index from the Stage 1 clusters and refer to the pseudo-data as $(\mathbf{y}^*_1, \mathbf{X}^*_1), \ldots, (\mathbf{y}^*_G, \mathbf{X}^*_G)$. For pseudo-subject $g = 1, \ldots, G$, we assume

$$\begin{aligned} [\mathbf{y}^*_g \,|\, \mathbf{X}^*_g, \mathbf{b}^*_g, \tau] &\sim \mathrm{N}(\mathbf{X}^*_g \mathbf{b}^*_g, \tau^{-1}\mathbf{I}_{n^*_g}) \\ \mathbf{b}^*_g \sim H \qquad H &\sim \mathrm{DP}(\alpha H_0), \end{aligned} \tag{3.6}$$

where n^*_g is the number of measurements on pseudo-subject g, H_0 is a known distribution, and α is a precision parameter. In all the examples we will consider, $H_0 = \mathrm{N}(\boldsymbol{\mu}, \mathbf{D})$.

As discussed in Section 2.1.3, if we marginalize over the DPP for H, the sequence of random effects, $\mathbf{b}^*_1, \ldots, \mathbf{b}^*_G$, follows a Polya urn scheme (Blackwell and MacQueen, 1973), i.e.,

$$\mathbf{b}^*_k | \mathbf{b}^*_1, \ldots, \mathbf{b}^*_{k-1} \begin{cases} = \mathbf{b}^*_j & \text{with probability } \frac{1}{\alpha+k-1} \\ \sim H_0 & \text{with probability } \frac{\alpha}{\alpha+k-1}, \end{cases} \tag{3.7}$$

for $j < k$ and $k = 2, \ldots, G$. Thus, under the DPP, the random effects are clustered into $K \leq G$ different groups whose random effects are $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, where $\boldsymbol{\theta}_l \sim H_0$ for $l = 1, \ldots, K$ (MacEachern, 1994).

Let $\mathcal{S}_{1,i} \in \{1, \ldots, G\}$ and $\mathcal{S}_{2,i} \in \{1, \ldots, K\}$ index the stage 1 and 2 clusters of subject i, respectively. Given the frequencies of our Stage 1 clusters, $m_1, \ldots, m_G$, the probability that two, randomly selected subjects are in the same Stage 1 cluster is

$$P_{i,i'} = \Pr(\mathcal{S}_{1,i} = \mathcal{S}_{1,i'}) = \sum_{g=1}^{G} \frac{\binom{m_g}{2}}{\binom{N}{2}}, \tag{3.8}$$

which follows from the multivariate hypergeometric distribution. Also, under the DPP, the probability that two pseudo-subjects are grouped together is $1/(\alpha+1)$ (Antoniak, 1974). Therefore, a priori,

$$\Pr(\mathcal{S}_{2,i} = \mathcal{S}_{2,i'}) = P_{i,i'} + \frac{1 - P_{i,i'}}{\alpha + 1} \geq \frac{1}{\alpha + 1}. \tag{3.9}$$

Thus, our method increases the prior probability that two subjects are clustered together, relative to a DPP applied to N subjects. As a result, our prior favors a smaller, but more scientifically justified, number of clusters.

3.3.4 Posterior Computation

Computation under the DPP proceeds using the West et al. (1994) Pólya urn sampler described in Section 2.1.3 and the details regarding both the full conditional posterior distributions of $\mathbf{b}_1^*, \ldots, \mathbf{b}_G^*$ and the sampling algorithm are provided in Appendix B. The major difference between our implementation and the standard use of the sampler is that our conditional posteriors in terms of G pseudo-subjects instead of N subjects. It is important to note that if we were to apply the DPP to N random effects, instead of G, the MCMC would iterate very slowly for large samples and computation may be infeasible. In addition, the large matrices needed to update values for N subjects can cause memory problems in certain software, such as Matlab. This latter difficulty prevented us from applying the DPP to each subject in the CPP data.

To reduce the sensitivity of the Stage 2 clustering to subjectively chosen hyperparameters, we recommend placing hyperpriors on $\boldsymbol{\mu}$, $\mathbf{D}$, τ, and α. For our models, we use the priors

$$\begin{aligned} \pi(\boldsymbol{\mu}) &= \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad \pi(\mathbf{D}^{-1}) = \mathrm{W}(d_0, \mathbf{D}_0) \\ \pi(\tau) &= \mathrm{Ga}(\psi\tau_0, \psi) \qquad \pi(\alpha) = \mathrm{Ga}(a, b), \end{aligned}$$

where $\mathrm{W}(\cdot, d_0, \mathbf{D}_0)$ is the Wishart density with degrees of freedom d_0 and mean $\mathbf{D}_0$. Please see Appendix B for the full conditional posterior densities of these hyperparameters. Since each of the above priors are conjugate, it is straightforward to update the values of the hyperparameters within the MCMC using Gibbs steps.

3.3.5 Methods for Inference

In Section 3.3.1 , we demonstrated that population-average effects, $\boldsymbol{\beta}$, can be estimated by a weighed mean of $\overline{\mathbf{b}}_1, \ldots, \overline{\mathbf{b}}_G$. Although the DPP is applied to cluster means, $\mathbf{b}_1^*, \ldots, \mathbf{b}_G^*$ are computed based on one pseudo-subject and, as a result,

$$\text{Cov}\left(\frac{1}{N}\sum_{g=1} m_g \mathbf{b}_g^*\right) > \text{Cov}(\widetilde{\boldsymbol{\beta}}) = \frac{1}{N}\mathbf{V}.$$

However, given $\boldsymbol{\beta}$, a transformation can be made,

$$\dot{\mathbf{b}}_g = m_g^{-1/2}\mathbf{b}_g^* + (1 - m_g^{-1/2})\boldsymbol{\beta},$$

which preserves the mean for $\mathbf{b}_g^*$, but changes the covariance to $\mathbf{V}/m_g$ so that

$$\text{Cov}\left(\frac{1}{N}\sum_{g=1}^{G} m_g \dot{\mathbf{b}}_g\right) = \frac{1}{N}\mathbf{V}.$$

Based on the above results, we make a similar, posterior transformation of $\mathbf{b}_1^*, \ldots, \mathbf{b}_G^*$, which ensures that the variance of the population effects is reflective of the cluster size. Following convergence, let $\mathbf{b}_g^{*(t)}$ denote the value of $\mathbf{b}_g^*$ observed at iteration t, $t = 1, \ldots, T$. Prior to calculating the population mean, we replace $\mathbf{b}_g^{*(t)}$ with

$$\widetilde{\mathbf{b}}_g^{(t)} = m_g^{-1/2}\mathbf{b}_g^{*(t)} + (1 - m_g^{-1/2})\overline{\mathbf{b}_g^*}, \tag{3.10}$$

where $\overline{\mathbf{b}_g^*} = \sum_{t=1}^{T} \mathbf{b}_g^{*(t)}/T$. Note that for large T, $\text{Cov}(\overline{\mathbf{b}_g^*})$ approaches $\mathbf{0}$ and, thus, we do not (significantly) inflate the variances of $\widetilde{\mathbf{b}}_g$ by estimating the posterior mean. In the special case where m_g=1, we simply have $\widetilde{\mathbf{b}}_g^{(t)} = \mathbf{b}_g^{*(t)}$, and as the first stage cluster sizes grow, we shrink back towards the mean of the samples. By doing shrinkage within the first stage clusters instead of across the clusters, we do not obscure or mask non-normal features in the random effect distribution.

Now that we have corrected our estimates of the cluster-specific means, population effects can be estimated at each iteration of the MCMC as

$$\widehat{\boldsymbol{\beta}}_{(*)}^{(t)} = \frac{1}{N}\sum_{g=1}^{G} m_g \widetilde{\mathbf{b}}_g^{(t)}. \tag{3.11}$$

Thus, linear combinations of $\widehat{\boldsymbol{\beta}}_{(*)}$ can be used to test hypotheses about the average

effects of the predictors, similar to what is done with fixed effects in mixed models.

Inferences about heterogeneity can be based on the posterior clustering of the random effects. As in Bigelow and Dunson (2005), the Dirichlet process clustering can be summarized by post-processing the results from the MCMC using a hierarchical clustering procedure such as single linkage, which is also known as nearest-neighbors (Sneath, 1957). In this paper, we define a new set of Stage 2 clusters $k = 1, \ldots, K^*$ where for each pseudo-subject g in cluster k, there exists some other pseudo-subject g^* such that $\Pr(S_g = S_{g^*}) \geq p^*$, where, under the West et. al (1994) sampling algorithm for the DPM, S_g indicates the cluster membership of pseudo-subject g. To ensure adequate separation between our clusters, we choose $p^* = 0.5$ in our analyses. This clustering procedure can be implemented using the linkage and cluster functions in MATLAB (version 6). As seen in our analysis of the CPP data, the cluster-specific longitudinal trajectories and the proportion of subjects per cluster are useful in identifying outliers in the data.

3.4 Simulation Studies

We applied the approach to three simulated data examples. In each case, the true model for $\mathbf{y}_i$ given $\mathbf{b}_i$ was $\mathbf{y}_i \sim \mathrm{N}(\mathbf{X}_i\mathbf{b}_i, \mathbf{I}_6)$ where $\mathbf{X}_i = (\mathbf{x}_{i0}, \mathbf{x}_{i1}, \mathbf{x}_{i2})$ with $\mathbf{x}_{i0} = \mathbf{1}_6$, $\mathbf{x}_{i1} = \mathrm{u}_i \cdot \mathbf{1}_6$, $\mathrm{u}_i \in \{0, 1\}$, and $\mathbf{x}_{i2} = (0, 1, 3, 4, 7, 8)'/8$ for $i = 1, \ldots, N$. The predictor $\mathbf{x}_{i2}$ can be thought of as the age at followup for subject i and u_i as an exposure indicator, where $\sum_{i=1}^{N} \mathrm{u}_i = N/2$ in cases 1-3.

3.4.1 Case 1: Latent Class Data

In the first case, we simulated a single data set of size $N = 2000$ using the discrete distribution

$$\mathbf{b}_i = \begin{cases} \boldsymbol{\theta}_1 = (2.26, 0.46, 20.35)' & \text{with probability } 0.0792 \\ \boldsymbol{\theta}_2 = (3.14, 1.34, 22.76)' & 0.2969 \\ \boldsymbol{\theta}_3 = (3.30, 1.50, 23.20)' & 0.3065 \\ \boldsymbol{\theta}_4 = (3.46, 1.66, 23.64)' & 0.2969 \\ \boldsymbol{\theta}_5 = (4.77, 2.97, 27.23)' & 0.0205, \end{cases}$$

which has mean $\boldsymbol{\beta} = (3.25, 1.45, 23.06)'$. We will refer to all $i : \mathbf{b}_i = \boldsymbol{\theta}_j$ as Class j.

We applied our approach for $r = 2.14$ (elicited value), $r = 1.66$, and $r = 0$ (complete

data). Diffuse priors were chosen for $\boldsymbol{\mu}$ and τ with $\boldsymbol{\mu}_0 = (15, 0, 0)'$, $\boldsymbol{\Sigma}_0 = 100 \cdot \mathbf{I}_3$, $\tau_0 = 1$, and $\psi = 0.1$. The prior for $\mathbf{D}$ was centered on the identity matrix, with $d_0 = 3$. We also let $\alpha \sim \text{Ga}(a, 1)$ where we let $a = 0.25$ for $r = 2.14$ and $r = 1.66$, but chose $a = 0.1$ for the complete data to induce a similar prior for K across G. The MCMC was run for 25,000 iterations in each analysis with the first 5,000 iterations discarded as a burn-in and with every 10th sample collected to thin the chain. To speed up computation, we sampled each S_g conditional on the random effect values at the previous iteration.

Table 3.1 provides estimates of K and the population effects from our MCMC. Both the number of clusters and the values of the regression parameters are similar across r. In addition, the elicited r reduced computation time by approximately 19 hours, relative to complete data, which demonstrates the efficiency of our method.

After post-processing the results of our MCMC using nearest neighbor clustering, we obtained 4 Stage 2 clusters. One cluster consisted of an outlier from Class 2 but, as seen in Table 3.2, the remaining clusters demonstrate good agreement with the subjects' true clusters: one cluster consists of mostly Class 1 subjects, another is primarily comprised of subjects from Classes 2-4, while the third only contains subjects from Class 3. Thus, under the elicited r, our method effectively separated the extreme outliers (Class 5) from the rest of the data and, although less successful, was able to isolate most of the moderate outliers (Class 1). In addition, the parameter estimates within each cluster, $\widehat{\boldsymbol{\beta}}_{(*1)}, \widehat{\boldsymbol{\beta}}_{(*2)}$, and $\widehat{\boldsymbol{\beta}}_{(*3)}$, are comparable to the true values within Classes 1, 2-4, and 5, respectively. Under $r = 1.66$ and $r = 0$, the parameter estimates of the three largest clusters were similar to those listed in Table 3.2. However, the number of singleton clusters increased as r decreased. This exemplifies the importance of the expert elicitation as the value of r will significantly impact the number of outliers in Stage 2.

3.4.2 Cases 2-3: Continuous Random Effects

In Case 2, $\mathbf{b}_i \sim \text{N}\big(\boldsymbol{\beta}, \text{diag}(\boldsymbol{\omega})\big)$, where $\boldsymbol{\beta} = (3.3, 1.5, 23.2)'$ and $\boldsymbol{\omega} = (0.4, 0.4, 3)'$, while in Case 3

$$\mathbf{b}_i \sim 0.65 \cdot \text{N}\big(\boldsymbol{\beta}_1, \text{diag}(\boldsymbol{\omega}_1)\big) + 0.35 \cdot \text{N}\big(\boldsymbol{\beta}_2, \text{diag}(\boldsymbol{\omega}_2)\big),$$

where $\boldsymbol{\beta}_1 = (2.9, 1.1, 22.2)'$, $\boldsymbol{\beta}_2 = (4, 2.25, 25)'$, $\boldsymbol{\omega}_1 = (0.075, 0.1, 1)'$, and $\boldsymbol{\omega}_2 = (0.175, 0.2, 2)'$. Since computation was more intensive than in Case 1, we reduced our sample sizes to 1000 in each study. The Stage 1 clustering and MCMC proceeded as in Case 1, but with different priors for α; in Case 2, $a = 1$ for $r > 0$, while in Case

TABLE 3.1: Means and 95% credible intervals for K and $\widehat{\boldsymbol{\beta}}_{(*)}$ from simulation case 1. The true values of the population effects were $\boldsymbol{\beta} = (3.25, 1.45, 23.06)'$.

r	G	K	$\widehat{\beta}_{(*)0}$	$\widehat{\beta}_{(*)1}$	$\widehat{\beta}_{(*)2}$
2.14	93	7.06	3.21	1.47	23.03
		(4, 14)	(3.18, 3.25)	(1.42, 1,52)	(22.98, 23.07)
1.66	215	8.2	3.25	1.44	23.03
		(4, 18)	(3.22, 3.29)	(1.39, 1.49)	(22.98, 23.08)
0	2000	8.3	3.25	1.45	23.02
		(4, 15)	(3.21, 3.29)	(1.40, 1.49)	(22.97, 23.07)

3, $a = 3$ for $r = 2.14$ and $a = 2$ for $r = 1.66$. In both cases, $a = 0.5$ when $r = 0$.

Under normal random effects, the parameter estimates under $r = 2.14$ were virtually identical to those provided by a random effects model fit to the complete data (see Table 3.3). However, when the random effects came from a mixture of normals, it appears as if $r = 2.14$ underestimates the variability in the population, resulting in population effects which are slightly biased. Note that in simulating data from a mixture of normals, we do not account for the expert opinion that there are no important differences within each cluster. Hence, these results demonstrate the robustness of our method to r. Also, even for a sample size of 1,000 choosing $r = 2.14$ instead of $r = 0$ reduced computation time from approximately 1.5 days to less than an hour and the computational gain will increase with sample size.

3.5 Analysis of the CPP Data

3.5.1 Methods

We now return to the CPP data discussed in Section 3.2. In our analysis, we considered modelling the longitudinal weight of girls by age and exposure category: child of never

TABLE 3.2: Summary of Stage 2 clusters from simulation case 1. Parameter estimates are the posterior means and 95% credible intervals within each cluster. The table omits one singleton cluster consisting of a subject from Class 2. Clusters are ordered by the magnitude of their parameter estimates.

		Class Frequencies			**Parameter Estimates**		
Cluster (k)	N_k	1	2-4	5	$\widehat{\beta}_{(*k)0}$	$\widehat{\beta}_{(*k)1}$	$\widehat{\beta}_{(*k)2}$
1	175	173	2	0	2.20	0.28	20.63
					(2.09, 2.30)	(0.15, 0.41)	(20.48, 20.78)
2	1785	1	1784	0	3.26	1.56	23.19
					(3.23, 3.30)	(1.50, 1.61)	(23.15, 23.24)
3	39	0	0	39	5.41	2.83	26.31
					(5.21, 5.62)	(2.60, 3.07)	(26.0, 26.1)

TABLE 3.3: Means and 95% credible intervals for K and $\widehat{\boldsymbol{\beta}}_{(*)}$ from simulation cases 2 and 3. The true values of the population effects were approximately $\boldsymbol{\beta} = (3.3, 1.5, 23.2)'$ in each case.

r	G	K	$\widehat{\beta}_{(*)0}$	$\widehat{\beta}_{(*)1}$	$\widehat{\beta}_{(*)2}$
Case 2					
2.14	81	54.5 (44, 65)	3.33 (3.26, 3.40)	1.48 (1.34, 1.61)	23.21 (23.14, 23.27)
1.66	163	94 (76.5, 111)	3.36 (3.29, 3.42)	1.44 (1.32, 1.57)	23.19 (23.12, 23.26)
0	1000	492.3 (439.5, 541)	3.34 (3.28, 3.41)	1.49 (1.38, 1.60)	23.21 (23.15, 23.29)
Case 3					
2.14	48	31.6 (23, 40)	3.23 (3.17, 3.28)	1.71 (1.57, 1.85)	23.27 (23.21, 23.33)
1.66	101	55.0 (41, 69)	3.30 (3.24, 3.36)	1.45 (1.32, 1.59)	23.32 (23.26, 23.39)
0	1000	301.1 (238.5, 357.5)	3.28 (3.22, 3.35)	1.53 (1.42, 1.64)	23.30 (23.23, 23.37)

smoker ($N_1 = 6,684$), ex-smoker ($N_2 = 1,849$), or current smoker ($N_3 = 8,985$). In Stage 1, we stratified by exposure and the four most common missingness patterns: no missing data, missing followup at year 8, missing followups at years 3 and 8, and lost to followup following year 1. Within each stratum, we clustered under $r = 2.14(p_j/6)^{1/2}$ where p_j is the number of followups under the missingness pattern in stratum j. Note that the correction, $(p_j/6)^{1/2}$, is the reciprocal of the correction used in (3.5). These Stage 1 analyses generated $G = 526$ clusters across the 12 strata.

In Stage 2, we modelled the weight of pseudo-subject g using an intercept, $\mathbf{x}^*_{g0}$, indicators of smoking exposure ($\mathbf{x}^*_{g1}$ for ex-smokers and $\mathbf{x}^*_{g2}$ for current smokers), mean age at each followup ($\mathbf{x}^*_{g3}$), and ex-smoker by age ($\mathbf{x}^*_{g4}$) and current smoker by age ($\mathbf{x}^*_{g5}$) interactions. Age was centered around the mean value amongst the pseudo-subjects (3.16) and was assumed to have a linear effect due to the relatively few ages at which measurements were collected.

We used the same priors for τ, $\boldsymbol{\mu}$, and $\mathbf{D}$ as in the simulations and assigned a Ga(0.5,1) to α to express an a priori belief in few second stage clusters. We ran our MCMC for 45,000 iterations following a burn-in of 10,000, otherwise implementing as in Section 4.

3.5.2 Results

As in Chen et al. (2006), our estimated population effects suggest that a mother's smoking habits during pregnancy had a significant impact on the growth of female children. As seen in Table 3.4, the 95% credible intervals for the smoking-age interactions (β_4 and β_5) obtained using our method (denoted G-DPP) are above 0, suggesting that the effects of smoking on child weight increased with age. To describe the smoking effect, we provide estimates of the ex-smoker and current smoker effects at birth (η_{E0} and η_{C0}) and age 8 (η_{E8} and η_{C8}). At birth, the children of ex-smokers and current smokers were leaner than the children of never smokers, with the decrease being highly significant, $\Pr(\eta_{C0} < 0)$ and $\Pr(\eta_{E0} < 0) > 0.99$, but similar across the two groups, $\Pr(\eta_{C0} < \eta_{E0})$=0.668. However, at age 8, children in both exposure groups were significantly heavier, $\Pr(\eta_{C0} > 8)$ and $\Pr(\eta_{E8} > 0) > 0.999$, with the increase in weight being greater in the children of ex-smokers, $\Pr(\eta_{E0} > \eta_{C8}) = 0.997$. It is likely that some or most of the ex-smoker effect is due to confounding as Chen et al. found that adjustment for covariates such as center and pre-pregnancy weight resulted in an insignificant ex-smoker effect. However, the authors found that a current smoker effect

did persist following adjustment for confounders.

Table 3.4 also presents smoking effect estimates obtained using GEE as in Chen et al.'s (2005) covariate adjusted models. Although the GEE estimates suggest a significant effect of smoking on child weight, there is no significant ex-smoker by age interaction (p = 0.141). It is not surprising that GEE provides a flatter slope for the ex-smoker effect since, under the assumption of MCAR, it does not allow a child's observed weight to be related to her missingness pattern, which, as discussed in Section 2, appears to be the case in the CPP.

Another common method for large data sets is to fit a model to a random sub-sample of the data. Thus, we compared our population effect estimates to those obtained from fitting a semiparametric random effects model to two random samples of size 1752 (denoted RS1-DPP and RS2-DPP in Table 3.4). In each case, the ex-smoker effects had wide credible intervals and were insignificant. However, the results for the current smoker effects were not consistent across the random samples; in one sample the effect increased with age, while in the other sample, the effect was insignificant. These results demonstrate two key weaknesses of fitting a model to a random sample: a loss of power to detect an effect of a rare exposure and, since the method is sensitive to outliers in the data, dependence on the sample chosen. Our method does not suffer from either weakness since we preserve all scientifically important differences in Stage 1 and, by weighting our population effects by cluster size, we ensure that our estimates are reflective of the complete data. The two-stage methodology is also more computationally efficient; in this example it took approximately 30 more hours to complete the MCMC for RS1- and RS2-DPP.

Figure 3.2 summarizes the Dirichlet process clustering of the pseudo-subjects in Stage 2. Although the posterior mean and 95% credible interval for K were 10.2 (6, 17), the clustering probabilities, i.e. $\Pr(S_g = S_{g^*})$, indicate that many of these Stage 2 clusters are not well separated. However, we could identify outliers in the data when we post-processed the Dirichlet process clustering. We found that 15,740 subjects belong to a sub-population with "normal" traits, labelled "(1)" in Figure 3.2, and that 40 subjects (20 non-smokers, 2 ex-smokers, and 18 current smokers) belong to a small outlier cluster, labelled "(2)." The children in Cluster 2 are substantially heavier than the normal subjects and have steeper growth curves: Cluster 2 subjects averaged 3.5 kg at birth and 53 kg at age 8, while normal subjects averaged 3.1 kg at birth and 26.1 kg at age 8. The remaining 1,738 subjects in the CPP data were represented by pseudo-subjects who were not grouped with another pseudo-subject in at least half of

the iterations. Although some of these subjects appear to be outliers with unusual growth patterns, most (1,722) were lost to follow-up following birth or year 1 and the DPP could not accurately classify them due to their limited data. Had we not stratified by missingness in Stage 1, it is likely that many of these subjects would be grouped with the normal subjects. However, we discourage this practice as it increases the amount of imputation in the Stage 1 clusters.

Figure 3.3 provides the posterior mean of the ex-smoker and current smoker effects within the normal sub-population and Cluster 2 as well as the mean effect values for the remaining pseudo-subjects. As expected, the posterior means for normal subjects are similar to the population estimates. Other subjects have larger effect values. In particular, the average ex-smoker effect in Cluster 2 is 7.8 kg at age 7 and the average current smoker effect is 2.7 kg at age 8. In addition to exhibiting unusual growth, the children in Cluster 2 also had mothers who were, on average, 17.2 kg heavier prior to pregnancy than the mothers of normal children. This is an important result as Chen et al. (2006) found that pre-pregnancy weight is one of the strongest confounders of the association between smoking and child growth.

TABLE 3.4: Population effects of smoking in CPP analysis. The DPP estimates listed are means and 95% credible intervals; 95% confidence intervals are listed for the GEE results.

	Ex-smoker effects			**Current smoker effects**		
Method	β_4	η_{E0}	η_{E8}	β_5	η_{C0}	η_{C8}
G-DPP	0.11 (0.08, 0.14)	-0.08 (-0.15, -0.02)	0.82 (0.61, 1.02)	0.07 (0.05, 0.09)	-0.10 (-0.15, -0.05)	0.45 (0.29,0.60)
GEE	0.03 (-0.01, 0.06)	-0.004 (-0.11, 0.10)	0.40 (0.23, 0.58)	0.05 (0.03, 0.07)	-0.14 (-0.17, -0.11)	0.27 (0.09, 0.44)
RS1-DPP	-0.08 (-0.17, 0.03)	0.16 (-0.07, 0.31)	-0.45 (-1.14, 0.36)	0.07 (-0.001, 0.15)	-0.14 (-0.27, -0.01)	0.39 (-0.11, 1.00)
RS2-DPP	-0.09 (-0.18, 0.01)	0.12 (-0.12, 0.35)	-0.60 (-1.27, 0.14)	-0.01 (-0.07, 0.06)	-0.13 (-0.26, 0.01)	-0.21 (-0.66, 0.29)

3.6 Discussion

We have proposed a two-stage clustering procedure for fitting Bayesian semiparametric random effects models to large data sets. Our method uses expert elicitation to generate a smaller, biologically meaningful, pseudo-sample of data that summarize the important differences in the complete data. Then, by applying the DPP to these data, we substantially decrease the computational burden and generate scientifically interesting clusters in the posterior. Simulation studies have shown that our method can detect true trends in the data under discrete and continuous random effects, though there may be a small bias for multimodal, continuous distributions.

In applying our method to the CPP data, we have provided the first random effects analysis of the smoking data. Although our overall conclusions on the effect of maternal smoking during pregnancy are similar to those in Chen et al. (2006), we have also shown that their GEE methodology may have underestimated the effects of smoking on child weight. Our semiparametric method also allows inferences on heterogeneity in the smoking effects as well as the identification of clusters of subjects with large regression coefficients. Some of these outliers could be explained by confounders that were omitted from our model, such as maternal weight. Others likely reflect data entry or recording errors, and thus, an attractive feature of our approach is that inferences on subjects in the larger clusters are not sensitive to these outliers.

Although our method was motivated by a specific example, it can easily be extended to handle data with a slightly different form, or studies with different analysis objectives. For example, in studies where models are constructed for predictive purposes, one can use the pseudo-subjects to predict the random effects of future subjects. This methodology should work well for large data sets where the probability of a future outlier, dissimilar from previous outliers, is low. In some prospective epidemiology studies, there may be interest in fitting a model with many covariates, as was the case in Chen et al.'s analysis of the CPP. In these settings, it may be necessary to modify our first stage clustering to improve efficiency; for example, the clustering could be stratified based on propensity scores (Rosenbaum and Rubin, 1983) rather than across each covariate level. Finally, it would be interesting to modify our method to handle data with a large number of measurements on each subject, as in menstrual diary data (e.g., Harlow et al., 2000).

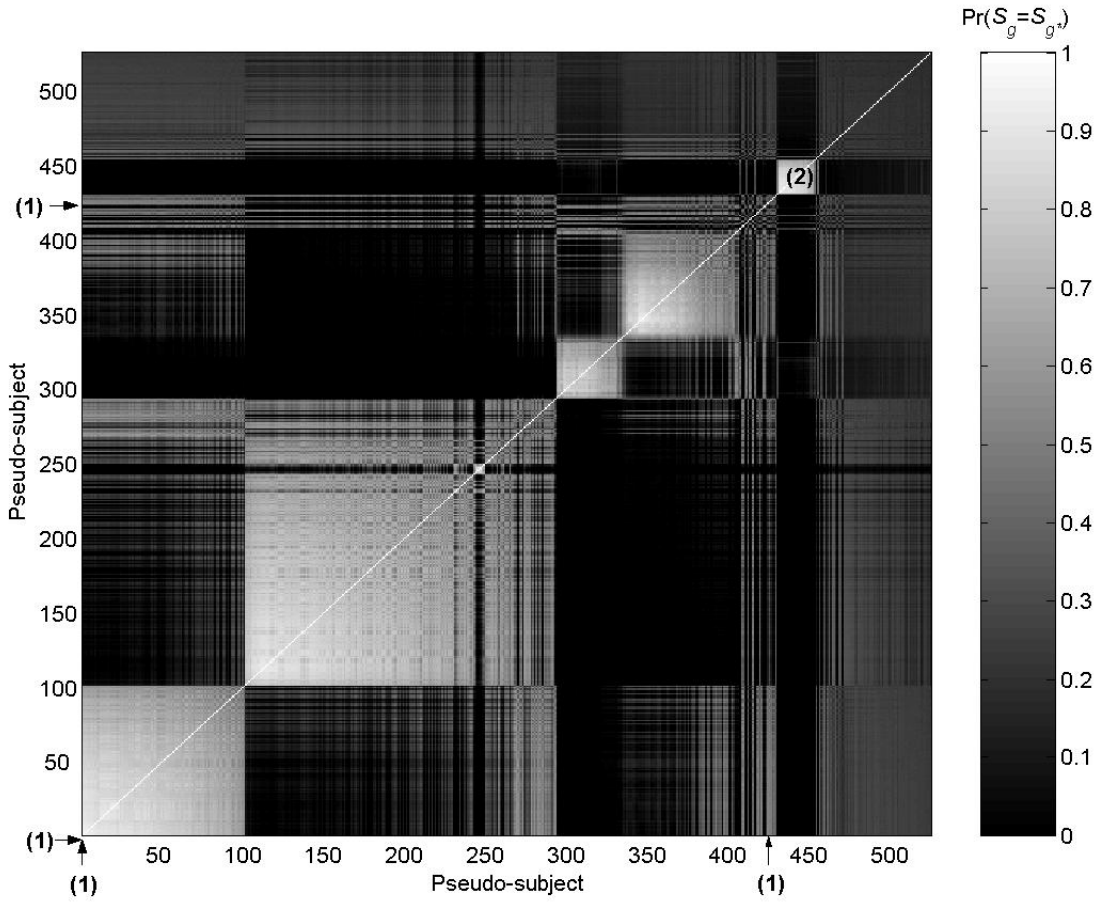

FIGURE 3.2: Dirichlet process clustering of CPP data. The order of the pseudo-subjects corresponds to the order of the singleton clusters in a dendrogram generated in Matlab (version 6). This dendrogram summarized nearest-neighbors clustering of the pseudo subjects using 1-Pr($S_g = S_{g^*}$) as the distance measure. The arrows denote subjects in the normal sub-population "(1)" (pseudo-subjects 1-425 in the figure). Cluster 2 (labelled "(2)") contains pseudo-subjects 430-453, while the remaining pseudo-subjects were not clustered with another pseudo-subject in at least half of the iterations.

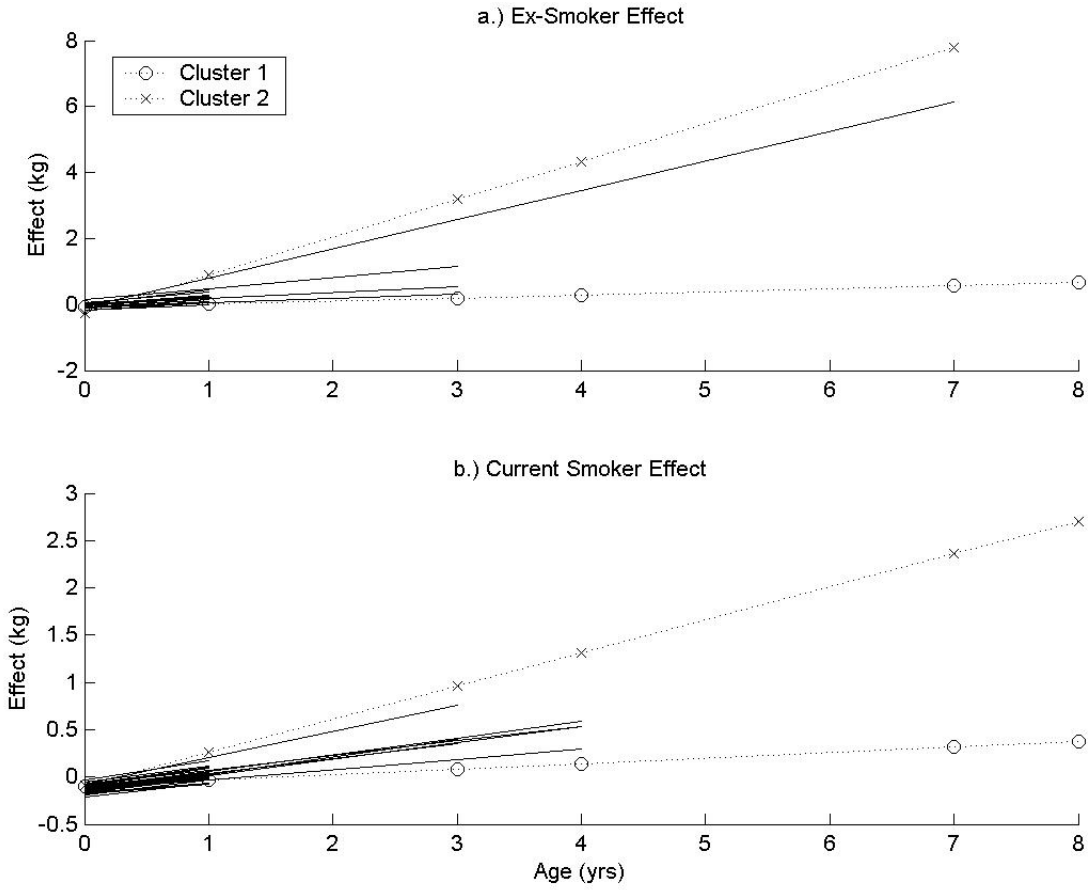

FIGURE 3.3: Mean smoking effects in the CPP data. Clusters 1 and 2 correspond to the groups of pseudo-subjects denoted in Figure 3.2. The solid, unlabelled lines correspond to the remaining pseudo-subjects. Effect estimates were computed up to the last followup of the exposed subjects. Estimates for unexposed pseudo-subjects are omitted.

CHAPTER 4

BAYESIAN SEMIPARAMETRIC DYNAMIC FRAILTY MODELS FOR MULTIPLE EVENT TIME DATA

4.1 Introduction

Many biomedical studies are designed to assess covariate effects on the time to recurrence of health-related outcomes, such as infections, hospitalizations, or recurrences of disease. For example, data of this type are collected in chemoprevention and carcinogenicity studies measuring the rate of appearance of palpable tumors of the skin and breast of mice (Gail et al., 1980; Forbes and Sambuco, 1998; Dunson, 2000). In these experiments, a rich set of data are available for each mouse, including times of appearance of each lesion, total number of tumors, and time of death.

A number of methods have been proposed to analyze tumor multiplicity data including recent work by Dunson and Dinse (2000) and Sinha et al. (2002). These articles relied on frailty-type models (Vaupel et al., 1979; Clayton and Cuzick, 1985) to accomodate baseline heterogeneity in risk of developing tumors. In these models, random effects (or frailties) measure an animal's risk relative to that for other individuals in the population, accounting for covariates. Standard shared and multivariate (Sargent, 1998) frailty models treat the frailties as time-independent factors, and hence do not allow a subject's risk to evolve dynamically as they age. Such formulations may be overly-restrictive when applied to tumorigenicity data, since biological changes occur-

ring with age result in complex and unanticipated trends in susceptibility to tumor development. A likely trend is that animals getting tumors relatively early may not be at higher risk later in life.

Recently, several authors have proposed more flexible, dynamic formulations. Yue and Chan (1997) and Yau and McGilchrist (1998) introduced a proportional hazard model for inter-recurrence times in which a subject's frailty changes following each event, and Lam et al. (2002) developed a related approach for the proportional odds model. In tumorigenicity studies, a time-varying frailty structure may be more realistic since it is more natural to model individual-specific risk as changing with age instead of according to previous occurrences of tumors. Relevant methods have been proposed by Henderson and Shimakura (2003), who developed a longitudinal Poisson regression model with gamma frailties which vary with time, and Paik et al. (1994), who proposed a proportional hazards frailty model with a time-specific random factor. Although promising, these methods involve complicated likelihoods and difficult computation, particularly when one considers generalizations (e.g., for joint modelling).

Bayesian approaches have several advantages for data collected from tumorigenicity studies, including ease of computation via MCMC, ability to incorporate prior information (e.g., from historical controls), and exact inferences on different aspects of the tumor response (time to first tumor, total tumor burden, etc). Unfortunately, in the Bayesian literature there has been limited consideration of dynamic frailty models and methods for multiple event time data in general. For recent Bayesian references on frailty models for multiple event time and multivariate survival data, refer to papers by Gustafson (1997), Sahu et al. (1997), Walker and Mallick (1997), Sargent (1998), Aslanidou et al. (1998), Sinha (1998), Chen and Sinha (2002), Dunson and Chen (2004), Sinha and Maiti (2004)), as well as a review in Ibrahim et al. (2001). Härkänen et al. (2003) proposed an innovative approach based on a model that allows subject-specific frailty trajectories to vary according to a latent class structure. In many settings, including animal tumorigenicity studies, it may be more natural to suppose that the age-specific risk trajectories vary according to a continuum, with each subject potentially having their own unique pattern.

Motivated by the tumor multiplicity application, we propose a Bayesian semiparametric dynamic frailty model. Our methodology generalizes the shared frailty model to allow time-varying frailties and regression coefficients. In addition, we use a multiplicative parameterization to introduce autocorrelation and smooth the time trajectories. To improve flexibility, we consider a nonparametric treatment of the frailty distribution

using a model which is related to the Dirichlet process (DP) mixture (Antoniak, 1974). For references on related approaches using DP mixtures in Bayesian analyses, refer to West et al. (1994), Bush and MacEachern (1996), Mukhopadhyay and Gelfand (1997), Müller and Rosner (1997), Kleinman and Ibrahim (1998), and Dominici and Parmigiani (2001). In addition, an alternative nonparametric approach for recurrent event time data was proposed by Ishwaran and James (2004). In this paper, we use DP priors to allow uncertainty in the distributions of a shared frailty and multiplicative innovations on this frailty. By centering the semiparametric model on a conditionally-conjugate dynamic gamma model, we facilitate posterior computation and lack of fit assessments of the parametric model using predictive distributions.

Section 4.2 proposes the model and prior structure. Section 4.3 outlines an MCMC algorithm for posterior computation. Section 4.4 applies the method to data from a cancer chemoprevention study, and Section 4.5 discusses the results.

4.2 Dynamic Frailty Model

4.2.1 Model Specification and Frailty Structure

Consider a study measuring the times of occurrence of repeated events within N subjects. The rate of event occurrence for subject i ($i = 1, \ldots, N$) at time t is denoted $\lambda_i(t)$. We partition the time axis into M finely-spaced intervals, $T_1, T_2, \ldots, T_M$, where $T_j = (\tau_{j-1}, \tau_j]$, $0 = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_M$, and τ_M is the maximum follow-up time in the study. The intervals are chosen to be sufficiently narrow so that it can be assumed that $\lambda_i(t) = \lambda_{ij}$ for all $t \in T_j$, $j = 1, \ldots, M$.

Suppose that subject i is followed for t_i^* time units, where $t_i^* \in T_{M_i}$ and $M_i \leq M$. Under these specifications, let $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iM_i})'$ denote a vector of time-varying frailties for subject i, and let $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ be a vector of p predictors for subject i and time interval T_j. We focus on models having the following structure:

$$\lambda_{ij} = \xi_{ij} \lambda_{0j} g(\mathbf{x}_{ij}; \boldsymbol{\beta}_j), \tag{4.1}$$

where λ_{0j} is the baseline hazard for the jth interval, $g(\cdot)$ is a known link function mapping from $\Re \to \Re^+$, and $\boldsymbol{\beta}_j$ are interval-specific regression parameters. We further express the baseline hazard as $\lambda_{0j} = \widehat{\lambda}_{0j} \Delta_j$, where $\widehat{\lambda}_{0j}$ is an initial guess at the baseline hazard (e.g., estimated from historical control data or based on expert elicitation) and

Δ_j is a multiplicative deviation from this guess.

Similar to what is done by Paik et al. (1994), the frailties are decomposed into time-independent and -dependent components:

$$\xi_{ij} = \phi_i \prod_{h=1}^{j} \phi_{ih}, \tag{4.2}$$

where ϕ_i is a subject-specific shared frailty and ϕ_{ih} is the multiplicative innovation over time interval h. This multiplicative structure provides a convenient framework for imposing autocorrelation amongst the frailties. To demonstrate this feature, consider a model in which $\phi_i \sim \text{Ga}(\psi_1, \psi_1)$ independently of $\phi_{ih} \overset{iid}{\sim} \text{Ga}(\psi_2, \psi_2)$, $h = 1, \ldots, M_i$. Given these distributions, the correlation between frailties from intervals j and $j + d$ $(d > 0)$ is

$$\text{Corr}(\xi_{ij}, \xi_{i,j+d}) = \sqrt{\frac{\psi_2^d\{(1+\psi_1)(1+\psi_2)^j - \psi_1\psi_2^j\}}{(1+\psi_1)(1+\psi_2)^{j+d} - \psi_1\psi_2^{j+d}}}. \tag{4.3}$$

Under the above model, ψ_1 provides an overall measure of between subject heterogeneity while ψ_2 can be thought of as both a smoothing parameter for the frailty trajectories and a measure of temporal heterogeneity within each subject.

Although we have observed that this dynamic gamma model performs well with some data, it can lead to spurious inferences when the actual distributions of the frailty components are distinctly non-gamma. For example, Walker and Mallick (1997) demonstrated that frailty distributions may differ across predictor-level resulting in multi-modal distributions when all the frailties are pooled together. With this in mind, we propose a semiparametric model which is flexible to unanticipated trends in the frailty.

For subject i and interval T_j, let r_{ij} denote the time at risk and let Z_{ij} denote the number of events experienced. Then, under expressions (4.1) and (4.2) with a DP specification for the distributions of ϕ_i and ϕ_{ij}, we have:

$$\begin{aligned} Z_{ij} &\overset{ind.}{\sim} \text{Poisson}\left(r_{ij}\phi_i\left(\prod_{h=1}^{j}\phi_{ih}\right)\lambda_{0j}g(\mathbf{x}_{ij}, \boldsymbol{\beta}_j)\right) \\ \phi_i &\overset{ind.}{\sim} G_1 \qquad\qquad \phi_{ij} \overset{ind.}{\sim} G_{j+1} \\ G_1 &\sim \text{DP}(\alpha_{01}G_{01}) \qquad\qquad G_{j+1} \sim \text{DP}(\alpha_{02}G_{02}), \end{aligned} \tag{4.4}$$

where $\text{DP}(\alpha_0 G_0)$ denotes the Dirichlet process centered on G_0 with precision α_0, and

we assume G_{01} is $\text{Ga}(\psi_1, \psi_1)$ and G_{02} is $\text{Ga}(\psi_2, \psi_2)$. This structure is centered on the dynamic gamma frailty model described above, but we allow the true frailty distribution to deviate from the parametric form. The amount of uncertainty in the gamma assumption for the two frailty components is controlled by the hyperparameters α_{01} and α_{02}, with small values of these parameters corresponding to little faith in the gamma forms.

Since $G_1, G_2, \ldots, G_{M+1}$ are modelled as discrete distributions under the Dirichlet process (Blackwell, 1973), there will be common values of the shared frailties and multiplicative innovations across subjects under (4.4). Thus, our method identifies clusters of subjects whose genetic traits convey a similar level of susceptibility at the outset of the study as well as clusters of subjects who experience similar increases in their susceptibility over each time interval. These latter clusters may identify subjects who experienced similar, unobserved events, such as subjects who experienced changes in gene expression that increased their susceptibility to developing new tumors.

4.2.2 Priors for Model Deviations and Regression Parameters

As with the frailties, the multiplicative deviations from the initial baseline hazard estimates are separated into time dependent and independent components:

$$\Delta_j = \nu_0 \prod_{h=1}^{j} \nu_h. \tag{4.5}$$

Assuming $\nu_0 \sim \text{Ga}(\kappa, \kappa)$ and $\nu_j \overset{iid}{\sim} \text{Ga}(\psi_3, \psi_3)$ for $j = 1, \ldots, M$, we have a convenient structure for introducing autocorrelation amongst the model deviations

$$\text{Corr}(\Delta_j, \Delta_{j+d}) = \sqrt{\frac{\psi_3^d\{(1+\kappa)(1+\psi_3)^j - \kappa\psi_3^j\}}{(1+\kappa)(1+\psi_3)^{j+d} - \kappa\psi_3^{j+d}}}\ , \tag{4.6}$$

where κ controls the degree of shrinkage of the posterior towards $\widehat{\boldsymbol{\lambda}}_0$, and ψ_3 measures smoothness in the deviations from the prior estimate. An appealing feature of this prior in the context of the tumor multiplicity application is that the prior variance increases with time. In many applications, events are known to be rare early and one can obtain a good prior estimate early on, but at later times there is much more uncertainty.

Although many other correlated prior processes have been proposed for the baseline hazard (cf. Arjas and Gasbarra, 1994; Gamerman, 1991; Gustafson et al., 2003), our

proposed structure is practically appealing for complex models since it retains the conditional conjugacy properties of the Gamma process (Kalbfleisch, 1978) without the need to introduce any additional latent variables (cf. Nieto-Barajas and Walker, 2002). These properties are highlighted in the following section on posterior computation.

A similar prior structure can be used to provide smooth trajectories for predictor effects. This technique is an alternative to dynamic covariate models based on random walks (cf. Gamerman, 1991; Sargent, 1997) and can simplify computation in certain applications. We consider the typical exponential link function, $g(\mathbf{x}_{ij}; \boldsymbol{\beta}_j) = e^{\mathbf{x}'_{ij}\boldsymbol{\beta}_j}$. However, our priors are developed in terms of the following re-parameterization of $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})'$:

$$\gamma^*_{jk} = e^{\beta_{jk}} = \prod_{l=1}^{j} \gamma_{lk},$$

where $\gamma^*_{jk} = \lambda_i(t; x_{ijk} = c+1)/\lambda_i(t; x_{ijk} = c)$ for $t \in T_j$ is the multiplicative change in hazard (i.e., hazard ratio) at time t attributable to a unit increase in the kth predictor, and $\gamma_{lk} = \gamma^*_{lk}/\gamma^*_{l-1,k}$ is the multiplicative innovation in this hazard ratio experienced over time interval T_l. In the absence of prior information about the effect of the kth predictor, it is reasonable to assume $\gamma_{lk} \overset{iid}{\sim} \text{Ga}(\psi_4, \psi_4)$ priors for the multiplicative increments on the hazard function. The prior is centered on no effect for a predictor with the degree of shrinkage and smoothness in the regression function controlled by the hyperparameter ψ_4, as is clear from the following expression:

$$\text{Corr}(\gamma^*_{jk}, \gamma^*_{j+d,k}) = \sqrt{\frac{\psi_4^d\{(1+\psi_4)^j - \psi_4^j\}}{(1+\psi_4)^{j+d} - \psi_4^{j+d}}} \quad . \tag{4.7}$$

In cases in which one has prior information about the regression function, the prior structure above could be used to model multiplicative deviations from these initial estimates. This approach is closely related to our priors for the baseline hazard. Another extension that may be useful in some applications, would be to allow the level of smoothing to vary across predictors. These generalizations are straightforward and we focus on the simple case for ease in exposition.

4.2.3 Elicitation of Hyperpriors

To reduce the sensitivity of analyses to subjectively chosen hyperparameters, we consider the use of hyperpriors in building our dynamic frailty model. A priori we assume that $\psi_1 \sim \text{Ga}(a_1, b_1)$ and $\psi_2 \sim \text{Ga}(a_2, b_2)$ to obtain some information from the data about the variance components of the frailties. Although a diffuse prior would be reasonable for ψ_1, allowing the level of between subject heterogeneity to be determined by the data, we recommend using an informative prior for ψ_2 centered on values which reflect the expected amount of correlation between Poisson counts from adjacent time intervals.

One could additionally specify gamma hyperpriors for ψ_3 and ψ_4, but preliminary results suggest that it can be difficult to specify a prior that ensures sufficient smoothing of the baseline hazard and predictor effects, since the likelihood tends to dominate the prior even for moderate sample sizes. Thus, to avoid over-fitting, we recommend choosing values of ψ_3 and ψ_4 which reflect both one's a priori intuition about the level of autocorrelation between time intervals and one's confidence in the initial guess of the baseline hazard. We also recommend performing a sensitivity analysis of this choice. Hyperpriors for α_{01} and α_{02} may also be elicited to determine the amount of deviation from the gamma frailty structure in the data, though we prefer to fix these parameters to avoid over-parameterization.

4.3 Posterior Computation

4.3.1 MCMC Methodology

In applications with moderate to large sample sizes and lengthy followups, our model will result in a large number of latent variables and unknown parameters. Fortunately, due to the conditionally conjugate structure of the priors, computation can proceed fairly easily using a hybrid MCMC algorithm consisting of Gibbs and Metropolis-Hastings steps (Gelfand and Smith, 1990; Tierney, 1994). For the complete details regarding the full conditional posterior distributions and updating algorithm of our MCMC, please see Appendix C.

For simplicity, our methods assume that data are right censored and that censoring is non-informative. To modify our MCMC algorithm to accommodate interval-censored data, one can simply include data augmentation steps for sampling the exact event times for interval-censored observations. In addition, dependent censoring may be

accommodated by incorporating the time-varying frailty as a predictor in a model for the censoring time. These models would generalize the shared frailty models of Liu et al. (2004).

Our proposed MCMC methodology performed well when applied to simulated data. Posterior estimates of the time-varying hazard ratios and future predictions of the hazard function tended to agree closely with the true values, even when the prior for the baseline hazard was poorly specified. As expected, results were somewhat sensitive to the prior when sample sizes were small. However, robustness improved with increasing sample sizes, likely reflecting the borrowing of information across time and hyperprior structure.

4.3.2 Identifiability and Computational Issues

As for most latent variable models, some identifiability restrictions are needed. In practice, ν_1 and ϕ_{i1} should be fixed at 1, since these parameters have the same contributions to likelihood as ν_0 and ϕ_i, $i = 1, \ldots, N$. We also recommend fixing ν_0 since there may be a tendency for weak identifiability between ν_0 and the mean of G_1, leading to slow mixing of the MCMC algorithm.

Another issue in implementation is the choice of knots for the piecewise constant hazard. The typical frequentist approach of choosing intervals at the unique failure times (see, e.g., Breslow, 1974) is inappropriate from a Bayesian perspective, since it involves using the data to choose priors. By choosing autocorrelated priors which borrow information across intervals, our approach allows for tightly spaced intervals. However, we recommend choosing knots so that intervals are at least as wide as the inter-exam times; data are not informative about changes on a finer scale, and the computational burden increases with the number of intervals.

4.4 Chemoprevention Application

4.4.1 Data Analysis

We illustrate our approach using data from a study of the effect of canthaxanthin, a carotenoid found in fruits and vegetables, on chemically induced mammary carcinogenesis (Grubbs et al., 1991). This data set has previously been used as an example by both Kokoska et al. (1993) and Dunson and Dinse (2000). The study consisted of 119 Sprague-Dawley rats administered one of four diets beginning at 34 days of age:

(1) 3390 mg/kg canthaxanthin, (2) 1130 mg/kg canthaxanthin, (3) 328 mg/kg retinyl acetate, or (4) vehicle control. There were 30 rats in treatment groups 1-3 and 29 in group 4. At age 55 days, each rat was administered 15 mg of a known carcinogen DMBA by gavage. Regular palpations began at 75 days of age and, in general, were performed twice a week until 235 days of age, or 180 days following administration of DMBA.

Our analysis focused on the tumor occurrence times, as measured from DMBA administration, of rats in treatment groups 1,2, and 4. We partitioned the study period into 24 time intervals, with each interval having a length of one week except for the first and last intervals (22 and 4 days respectively). When the animals died naturally or were sacrificed at the end of the study period, an extensive pathological examination was conducted to find tumors that may be undetectable by palpation. For this reason, we allowed for a multiplicative increase in the baseline hazard of tumor detection at the final examination by introducing an additional parameter, ω, into our model. Thus, the model we used has the following form:

$$\lambda_{ij} = \phi_i \nu_0 \widehat{\lambda}_{0j} \omega^{I_{ij}} \prod_{h=1}^{j} \phi_{ih} \nu_h \gamma_{h1}^{x_{i1}} \gamma_{h2}^{x_{i2}}, \tag{4.8}$$

where $I_{ij} = 1$ if $j = M_i$ and 0 otherwise and $x_{ik} = 1$ if subject i is in treatment group k and 0 otherwise, $k = 1, 2$. Assigning a $\text{Ga}(a_\omega, b_\omega)$ prior to ω results in the conditional posterior,

$$\pi(\omega|\cdot) = \text{Ga}\left(a_\omega + \sum_{i=1}^{N} Z_{iM_i}, b_\omega + \sum_{i=1}^{N} r_{iM_i} \phi_i \widehat{\lambda}_{0M_i} \nu_0 \prod_{h=1}^{M_i} \phi_{ih} \nu_h \gamma_{1h}^{x_{i1}} \gamma_{2h}^{x_{i2}}\right), \tag{4.9}$$

where the notation $\omega|\cdot$ denotes ω given all other variables.

In this study of cancer initiation, the retinyl acetate group (3) can be considered as a second control group, since it is known that this treatment is only effective in decreasing promotion (Grubbs et al., 1991). Thus, these data were used to choose the prior for the baseline hazard. In particular, we fit a Poisson regression model with a cubic polynomial in time to the group 3 data and used the resulting fitted curve as $\widehat{\boldsymbol{\lambda}}_0$. We chose Ga(5,0.5) and Ga(25,0.5) priors for ψ_1 and ψ_2, respectively, to express belief in low between and within subject heterogeneity, but high levels of autocorrelation amongst the frailties. Since tumor incidence should be negligible in the first three weeks of the study, we assumed $\gamma_{11} = \gamma_{12} = 1$. In addition, to express modest confidence in

the gamma frailty assumptions, we set $\alpha_{01} = \alpha_{02} = 5$. Finally, we let $\psi_3 = \psi_4 = 50$ to induce high levels of smoothing in both the baseline hazard and treatment effects, and set $a_\omega = 3$ and $b_\omega = 1$ to reflect our a priori belief that more tumors are observed in the exam following sacrifice.

Using our MCMC methodology, we ran a chain of 55,000 iterations with the first 5,000 discarded as a burn-in. To reduce storage requirements, every 10th observation was saved to thin the chain. It took approximately 4 hours to complete a Matlab (version 7.0) program run in batch on the statistical server at UNC-Chapel Hill (20 1.05 GHz processors).

Figure 4.1 provides the mean and pointwise 95% credible intervals for the hazard ratios comparing the canthaxanthin dose groups to control. Although the hazard ratios are initially near one, as time progresses they decrease toward a plateau between weeks 10-26. The posterior mean and 95% credible interval for the average hazard ratio over this interval is $\widehat{\gamma}^*_{[10,26]2} = 0.553$ $(0.343, 0.863)$ in the low dose group and $\widehat{\gamma}^*_{[10,26]1} = 0.489$ $(0.304, 0.758)$ in the high dose group. In addition, the estimated posterior probabilities of a chemopreventive effect in the two groups for this interval are $\Pr(\gamma^*_{[10,26]2} < 1 \,|\, \mathbf{Y}) = 0.995$ and $\Pr(\gamma^*_{[10,26]1} < 1 \,|\, \mathbf{Y}) = 0.999$, with $\Pr(\gamma^*_{[10,26]2} > \gamma^*_{[10,26]1} \,|\, \mathbf{Y}) = 0.676$, suggesting highly significant, but similar, effects overall in the two dose groups.

As discussed by Kokoska et al. (1993) and Dunson and Dinse (2000), in the presence of a significant effect on tumor incidence, there is typically interest in assessing which aspects of the tumor response profile are most affected. In particular, tumor biologists wish to distinguish effects on multiplicity (total number of tumors) and latency (time to tumor onset). In our model, the effects of treatment group k on multiplicity may be evaluated by computing the posterior probability $P_{Mk} = \Pr(\Lambda_{kM}/\Lambda_{0M} < 1|\mathbf{Y})$, where Λ_{kM} is the cumulative hazard at sacrifice for an animal in treatment group k $(k = 1, 2)$ with frailty $\boldsymbol{\xi}_i = \mathbf{1}$ (i.e., the prior mean in the dynamic gamma model) and Λ_{0M} is the cumulative baseline hazard. Also, for $k = 1, 2$, a beneficial effect of treatment k on latency may be evaluated by computing $P_{Lk} = \Pr(\mu_0 < \mu_k|\mathbf{Y})$ where

$$\mu_k = \sum_{j=1}^{M} j \left(\frac{\lambda_{0j} \omega^{1(j=M)} \gamma^*_{jk} (\tau_j - \tau_{j-1})}{\sum_{h=1}^{M} \lambda_{0h} \omega^{1(h=M)} \gamma^*_{hk} (\tau_h - \tau_{h-1})} \right), \tag{4.10}$$

which is the expected interval of onset of an animal in treatment group k, again assuming $\boldsymbol{\xi}_i = \mathbf{1}$, and μ_0 is the expected interval of onset for a control animal with the

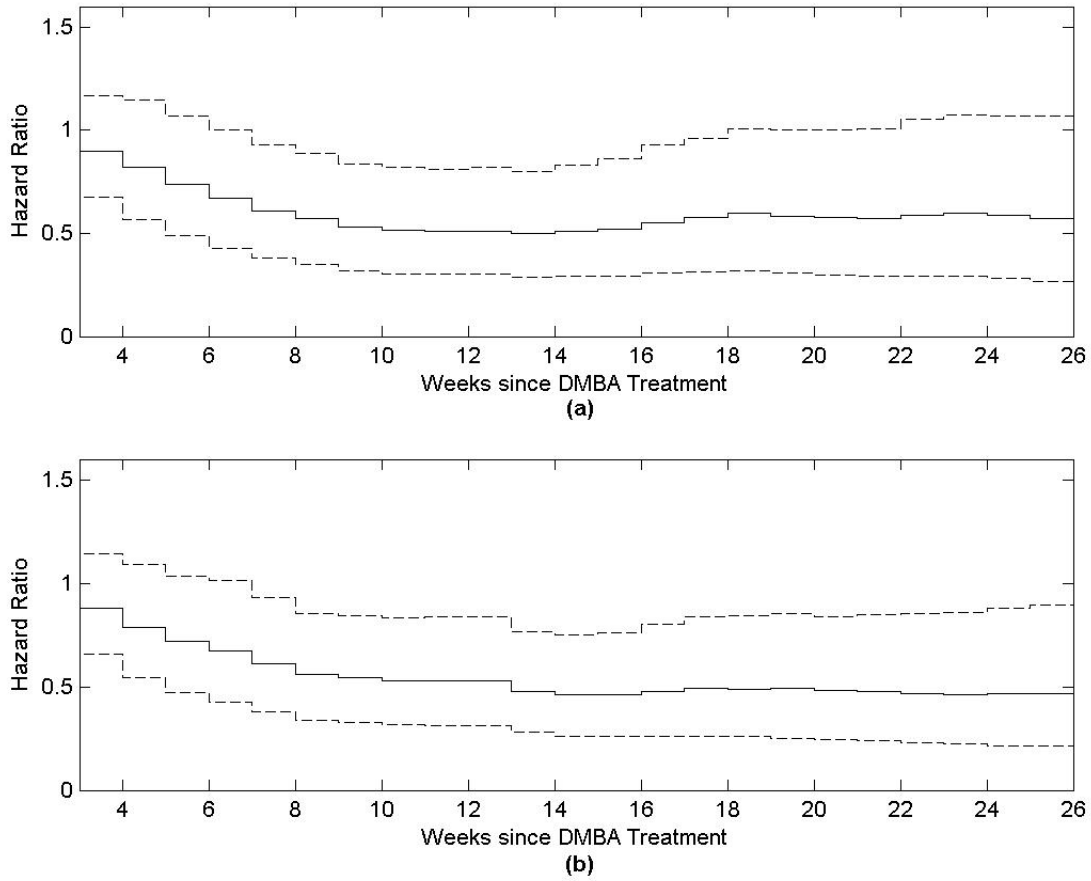

FIGURE 4.1: Posterior means (—) and pointwise 95% credible intervals (- - -) for hazard ratios comparing rats fed (a) 1130 mg/kg and (b) 3390 mg/kg canthaxanthin to mice administered a vehicle control diet.

same frailty. We found that both the low and high doses of canthaxanthin substantially decreased tumor burden, (P_{M1} and $P_{M2} > 0.99$), but there was no evidence of a beneficial effect on latency ($P_{L1} = 0.156, P_{L2} = 0.392$). Dunson and Dinse (2000) made similar conclusions regarding latency and multiplicity effect, however, they defined these phenomena in terms of both observed and induced, but unobserved tumors. By restricting our attention to observed tumors only, our method is not subject to biological assumptions on latent tumors. In addition, our dynamic frailty model is more biologically realistic than their shared frailty approach, and thus should provide a more realistic description of the canthaxanthin effect.

As seen in Figure 4.2, a substantial proportion of animals have frailties which evolve dynamically. In addition, two animals not depicted in the graphs have trajectories which increase toward values greater than four. These trends suggest that there are unmeasured factors which are affecting tumor incidence and a closer examination of the animals' genetic and physical traits should be made.

4.4.2 Goodness of Fit and Sensitivity Analyses

Goodness of fit can be assessed by comparing the predictive distribution of the weekly tumor counts with observed values in the data set. As seen in Figure 4.3, weekly predictions of tumor incidence prior to sacrifice agree with mortality adjusted means in the data. Assuming sacrifice at week 26, the posterior mean and 95% credible interval for the number of tumors discovered during the final exam are 0.520 (0.307, 0.808) for the vehicle control group, 0.285 (0.142, 0.509) for the low dose group, and 0.234 (0.111, 0.420) for the high dose group, which are in agreement with observed means (0.714, 0.167, and 0.222 tumors for the vehicle, low dose, and high dose groups, respectively).

We tested the sensitivity of our methodology to frailty assumptions by comparing the results using the priors from the previous section (denoted DPM1) against the fully parametric dynamic gamma model and the results obtained using an analysis with less confidence in the gamma assumption, expressed by letting $\alpha_{01} = \alpha_{02} = 2$ (DPM2). As seen in Figure 4.4, the predictive distributions of the frailties differ somewhat across each model. Most notably, low a priori confidence in the base model causes the distribution of ϕ_{N+1} to be more skewed and to have a fatter right tail. However, as seen in Table 4.1, parameter estimates and predictive probabilities are robust.

A second sensitivity analysis was performed for the smoothing parameters by re-

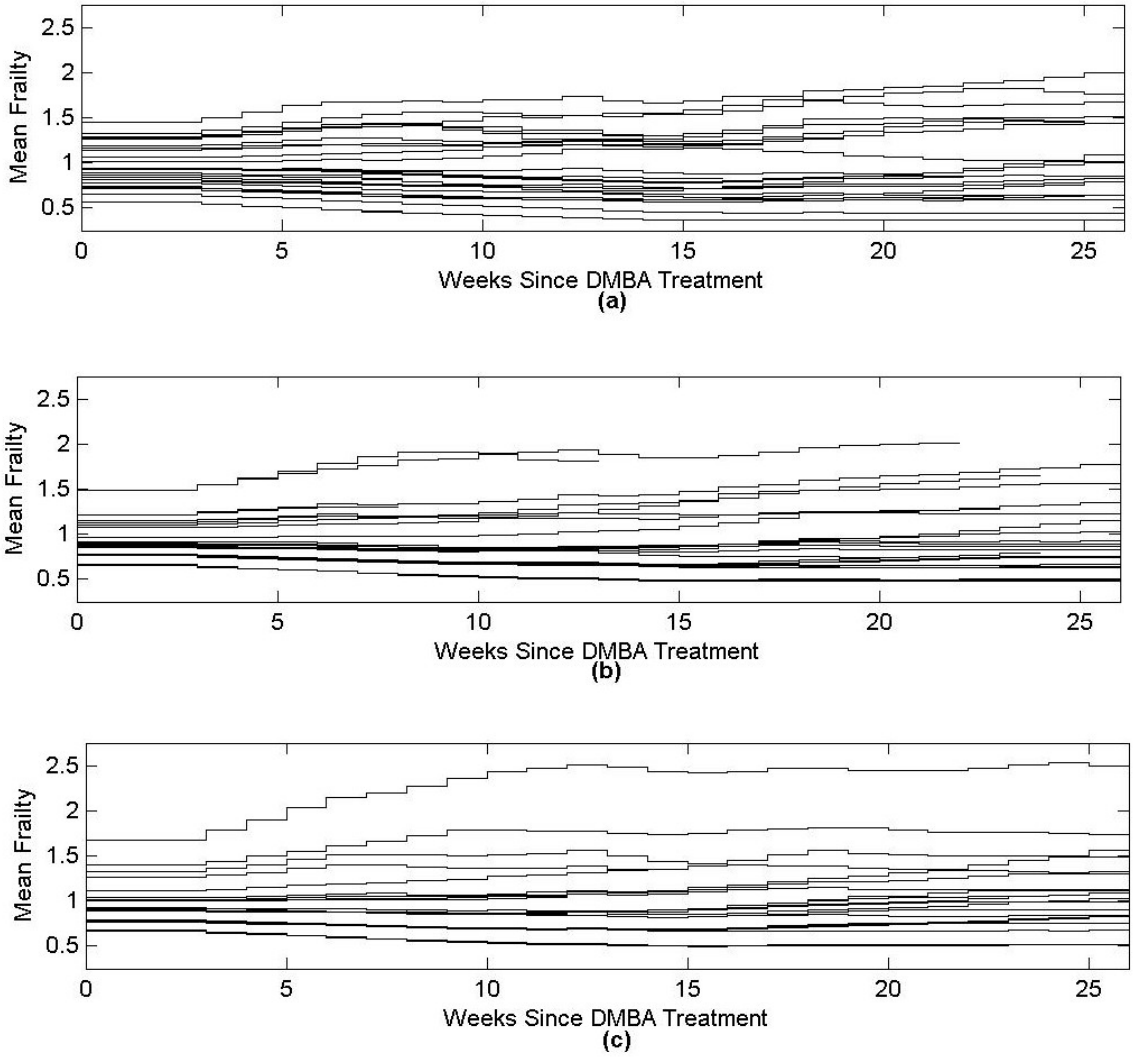

FIGURE 4.2: Posterior mean frailty trajectories for animals administered **(a)** vehicle, **(b)** 1130 mg/kg canthaxanthin, and **(c)** 3390 mg/kg canthaxanthin. Figure omits trajectories of animals 67 and 75 in the vehicle group due to their extreme values.

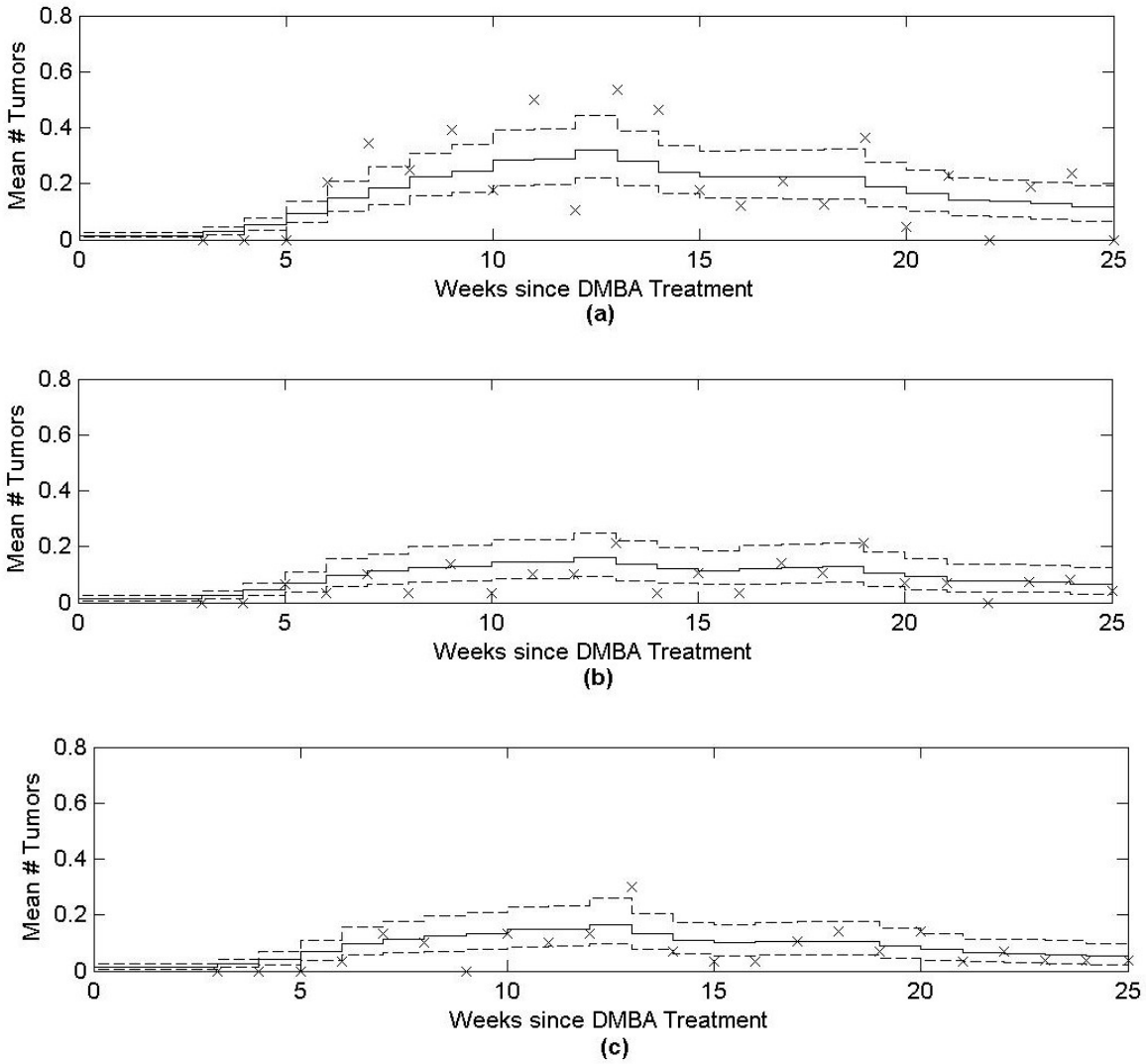

FIGURE 4.3: Observed (×) and predicted weekly tumor incidence prior to sacrifice for rats administered **(a)** a vehicle, **(b)** 1130 mg/kg canthaxanthin, and **(c)** 3390 mg/kg canthaxanthin. The pointwise 95% credible intervals (- - -) for the means (—) were calculated based on 1000 draws from the predictive distributions at each iteration of our MCMC algorithm.

TABLE 4.1: Sensitivity analysis of parameter estimates (mean and 95% credible intervals) and posterior probabilities from chemoprevention application.

	Model			
Parameter	DPM1	DPM2	DPM3	Dynamic Gamma
$\gamma^*_{[10,26]1}$	0.489 (0.304, 0.758)	0.485 (0.304, 0.748)	0.420 (0.254, 0.669)	0.482 (0.305, 0.735)
$\gamma^*_{[10,26]2}$	0.553 (0.343, 0.863)	0.548 (0.341, 0.855)	0.480 (0.289, 0.758)	0.546 (0.345, 0.845)
ω	7.95 (5.34, 11.19)	7.87 (5.23, 11.26)	7.60 (5.01, 10.82)	8.01 (5.41, 11.26)
ψ_1	5.94 (2.03, 12.80)	5.99 (2.04, 12.97)	5.97 (2.03, 13.02)	5.26 (2.03, 12.36)
ψ_2	48.8 (32.4, 69.9)	48.0 (31.0, 68.9)	49.5 (32.2, 70.5)	50.7 (34.0, 70.8)
Probability				
P_{M1}	0.999	> 0.999	> 0.999	0.999
P_{M2}	0.998	0.999	> 0.999	0.998
P_{L1}	0.156	0.139	0.167	0.167
P_{L2}	0.392	0.360	0.399	0.394

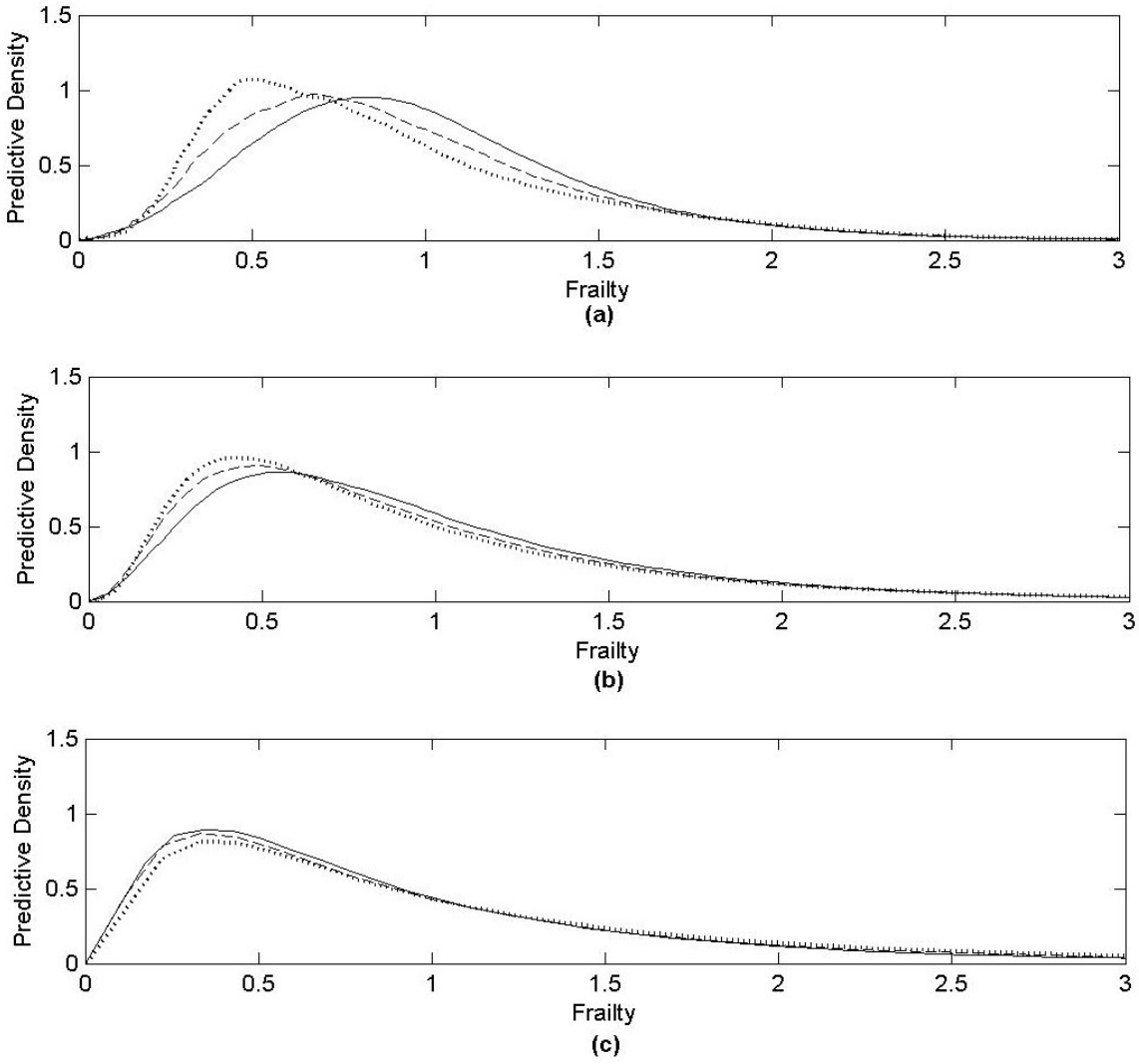

FIGURE 4.4: Comparison of predictive frailty distributions obtained using the dynamic gamma model (—), DPM1 (- - -), and DPM2 ($\cdots$). The densities of frailties from intervals **(a)** 1 (ϕ_{N+1}), **(b)** 11 ($\xi_{N+1,11}$), and **(c)** 24 ($\xi_{N+1,24}$) were approximated using a normal kernel smoother (width=0.05) applied to a posterior sample size of 500,000 (100 draws from the predictive distribution were taken at each iteration).

peating the analysis with $\psi_3 = \psi_4 = 25$ and $\alpha_{01} = \alpha_{02} = 5$ (denoted DPM3). Although Table 4.1 demonstrates that lower levels of smoothing decrease the estimated hazard ratios somewhat, conclusions regarding multiplicity and latency effects do not change. In addition, lowering ψ_3 and ψ_4 does not substantially improve goodness of fit.

4.5 Discussion

In this paper, we have developed a flexible method for inference in multiple event time data. Our method provides a convenient framework for smoothing hazard functions and time-dependent frailty trajectories and covariate effects. An appealing feature of our approach is the incorporation of dynamic frailties with nonparametric distributions. Even in a data set comprised of fairly homogeneous animals, our model is capable of identifying age-dependent shifts in susceptibility. By comparing the individuals' frailty trajectories, one can identify unusual individuals for genotyping and further examination. However, even in applications where the frailty is of no interest to the researcher, it would still be necessary to account for these trends to improve the accuracy of future predictions.

Although our proposed method accommodates a wide variety of time-dependent trends and distributions, the computational burden may not pay off in certain situations. In particular, studies of rare events or with few exam times may have minimal information in the data about time-varying frailty distributions. Although one could still apply our method in such a case, there may be minimal Bayesian learning and the prior would be utilized strongly in extrapolating over time. In these situations, a simple frequentist analysis may suffice.

By using Dirichlet process priors for a shared frailty and multiplicative innovations on the frailty, we have provided a less restrictive modelling framework than parametric alternatives. As mentioned previously, these priors ensure that the posterior distributions of the frailties are almost surely discrete (Blackwell, 1973; Ferguson, 1973). This restriction could be removed by using Pólya tree priors (Lavine, 1992, 1994) instead of the Dirichlet process, as was done by Walker and Mallick (1997) in the context of a shared frailty. However, a simpler alternative would be to model each ϕ_i and ϕ_{ij} using a Dirichlet process mixture. Although we did not consider such a methodology, this would be straightforward since it would just involve adding another level to the hierarchy.

Several extensions of our method would be interesting to pursue. For example, one

could consider priors with a mixture structure to allow selection of the frailty form. In particular, selection between no frailty, a shared frailty, and a dynamic frailty for each predictor. This approach would be a generalization of recent work by Dunson and Chen (2004). A formal method for testing for lack of fit of the dynamic gamma model would also be of interest. For instance, Berger and Guglielmi's (2001) method for comparing parametric and nonparametric models could potentially be adapted to our framework. The results from these lack of fit tests could be used to determine if the dynamic gamma model, or our nonparametric extension, should be used for data from similar studies. Another possible direction would be to apply our nonparametric framework to allow dynamic random effects in generalized linear models for longitudinal data and for joint modelling of longitudinal and survival data.

CHAPTER 5

NONPARAMETRIC BAYES TESTING OF CHANGES IN A RESPONSE DISTRIBUTION WITH AN ORDINAL PREDICTOR

5.1 Introduction

In many biomedical studies, there is interest in evaluating changes in a response distribution across an ordinal predictor. For example, toxicologists are often interested in assessing whether several biological responses vary in distribution with dose. In such settings, normality assumptions are typically not justified and there are biological reasons to expect changes in not only the location but also the variance and shape with dose. In particular, when there is substantial heterogeneity amongst subjects in their biological response to treatment due to variation in gene expression, timing of the cell cycle, and other factors, changes in variance, skewness, and even modality are natural.

Some frequentist nonparametric methods have been proposed for testing for trends across an ordinal predictor. In toxicology studies, Jonckheere's test (Terpstra, 1952; Jonckheere, 1954) is often used to test for stochastic order of a single outcome across dose groups. Related methods for multiple outcomes include Dietz's multivariate generalization of Jonckheere's Test (1989), Dietz and Killeen's (1981) test for a monotone trend in time, and O'Brien's (1984) rank-sum-type test (also refer to Huang et al.,

2005). Unfortunately, these methods are most sensitive to changes in the location of a distribution and may ignore important trends in shape (see, for example, footnote in Jonckheere, 1954). In contrast, k-sample tests based on empirical distribution functions (e.g, Ahmad, 1976; Kiefer, 1959) are sensitive to changes in distribution shape, but not changes in location and scale.

As mentioned in Chapter 1, Bayesian nonparametric approaches have several advantages over frequentist alternatives, including the ability to incorporate prior information (e.g., from historical controls) and exact inferences on unknown distributions. There is an extensive literature on nonparametric estimation (for a review see Chapter 2) and several methods are available for Bayesian nonparametric testing of goodness of fit (e.g., Berger and Guglielmi, 2001; Carota and Parmigiani, 1996; Verdinelli and Wasserman, 1998). However, there has been very little consideration of Bayesian nonparametric testing of differences in distributions across multiple groups. Gopalan and Berry (1998) describe an approach which uses a nonparametric Dirichlet process prior (DPP; Ferguson, 1973, 1974) to perform multiple comparisons, but the method only compares distribution means. Dunson and Taylor's (2005) quantile regression approach can assess the effects of the predictor on both the center of the distribution and the tails, but the method is not fully Bayesian and is not easily extended to multiple outcomes. Basu and Chib (2003) proposed an approach which compares semiparametric models using Bayes factors. This method may be promising when there are only 2-3 levels of an ordinal predictor. However, in toxicology studies with several dose groups, this approach would be unwieldy when one considers each possible contrast between dose groups.

In analyzing data from multiple groups, one would typically be interested not only in testing but also in estimation of group-specific distributions. Such estimation can potentially be accomplished by fitting a DP mixture (DPM; Antoniak, 1974) model separately to the different dose groups. For some references on applications of DPMs, refer to Sections 2.1.2 and 2.1.4. The disadvantage of such an approach is its inability to model trends and borrow information across the different dose groups, which is particularly important in applications having a modest number of subjects per group. Note that applications of the Basu and Chib (2003) procedure to test differences between groups would also have this disadvantage.

An alternative to using separate DPMs is to consider the dose group-specific distributions as a collection of dependent unknown distributions, which are assigned a dependent nonparametric prior. Dependent nonparametric priors have been the focus

of a growing body of literature. In the early work of Cifarelli and Regazzini (1978), dependence was induced through a regression model within the parametric base measure of the DP (for related work see Mira and Petrone, 1996; Tomlinson and Escobar, 1999; Carota and Parmigiani, 2002; Giudici et al., 2003). Although the method is straightforward, it has limited flexibility. Other authors proposed inducing nonparametric dependence by parameterizing random measures as products of DP distributed factors, though within a different framework than ours (see, e.g., Gelfand and Kottas, 2001 and our dynamic frailty model in Chapter 4). MacEachern (1999; 2000) proposed a dependent Dirichlet process (DDP) which characterizes dependency by defining a stochastic process for the atoms in the Sethuraman (1994) stick-breaking characterization of the DP. This approach has been successfully applied in ANOVA (De Iorio et al., 2004) and spatial (Gelfand et al., 2004) applications. A limitation is the assumption of fixed weights on the atoms, which does not allow new features to appear as dose increases. To solve this problem, Dunson (2006) proposed a dynamic mixture of DPs (DMDP), which is related to a mixture structure originally proposed by Müller et al. (2004) to combine inferences across related nonparametric models. More flexible dependent nonparametric priors for continuous predictors have been proposed by Griffin and Steel (2006), Dunson and Pillai (2004), and Duan et al. (2005).

Motivated by tractability in addressing the problem of nonparametric Bayes testing of changes with dose, we focus here on generalizing the Dunson (2006) approach. Our goal is to obtain posterior probabilities for local and global null hypotheses corresponding to equivalence in an unknown distribution between groups. Instead of requiring exact equivalence, we treat the distributions in adjacent groups as *effectively* equivalent if the total variation norm between their probability measures is less than ϵ. To borrow information across groups while assigning probabilities to the local null and alternative hypotheses, we use a hierarchical modification of the DMDP. Using an MCMC implementation, we obtain posterior model probabilities for the local and global hypotheses from a single run, while also producing posterior distributions for thresholds and estimates of group-specific distributions. These group-specific estimates will rely on borrowing of information from multiple groups, with an adaptive degree of shrinkage. Multiple response data can be accommodated without complication.

In Section 5.2, we describe our generalization of the DMDP and hyperprior structure. In Section 5.3, we outline a MCMC methodology for testing and estimation. In Section 5.4, we evaluate our approach using simulated data. In Section 5.5 we apply our method to a large data set from a genotoxicity experiment. In Section 5.6 we discuss

the results and future work.

5.2 Nonparametric Model and Prior Structure

5.2.1 General Framework

Let $\mathbf{y}_{hi}$ be a vector of q continuous outcomes measured on subject i $(i = 1, \ldots, n_h)$ in group h $(h = 1, \ldots, d)$ and X_h be a score for group h, where $X_{h-1} < X_h < X_{h+1}$. In a toxicological study, X_h is the dose administered to each animal in group h with $X_1 = 0$ representing the control group. Throughout Sections 5.2 and 5.3, we will assume that multiple outcomes are available, although in the case of a single outcome the simplification is straightforward. To relax distribution assumptions, we assume that $\mathbf{y}_{hi} \sim F_h$ which has the density function

$$f_h(\mathbf{y}_{hi}) = \int \mathrm{N}(\mathbf{y}_{hi}; \boldsymbol{\mu}_{hi}, \boldsymbol{\Sigma}_{hi}) dG_h(\boldsymbol{\mu}_{hi}, \boldsymbol{\Sigma}_{hi}), \tag{5.1}$$

where G_h is unknown. Hence, the density of $\mathbf{y}$ in group h is characterized as a nonparametric mixture of multivariate Gaussian densities, with the mixture distribution varying across groups. It is well known that mixtures of normals can accurately approximate any distribution.

It is clear that differences in the mixture distributions, G_h and G_{h+1}, imply differences in the outcome distributions, F_h and F_{h+1}. Hence, we focus on testing for changes in the mixture distributions. In particular, our focus is on local null hypotheses characterizing differences in adjacent groups and on global hypotheses representing intersections of these local hypotheses. As in the Kolmogorov-Smirnov test, we could specify a point null hypothesis which requires $G_1, \ldots, G_d$ to be *exactly* equivalent. However, many have argued that point nulls are artificial and would almost never be observed in practice (see, e.g., Berger and Delampady, 1987; Nickerson, 2000). In this paper, we use a more realistic null hypothesis under which the distributions are *effectively* eqivalent across groups. We formalize this hypothesis below.

Abusing notation, let $G_1, \ldots, G_d$ denote probability measures corresponding to the mixture distributions for the different groups. In addition, let B denote any Borel set such that $B \in \mathcal{B}$, with $B \subset \Re^q$ and $\mathcal{B}$ the Borel sigma-algebra of subsets of $\Re^q$, and let

the total variation norm of $G_{h+1} - G_h$ be defined as

$$||G_{h+1} - G_h||_{TV} = \max_{B \in \mathcal{B}} \Big| G_{h+1}(B) - G_h(B) \Big|.$$

We compare the local null hypothesis

$$H_{0h} : ||G_{h+1} - G_h||_{TV} \leq \epsilon \tag{5.2}$$

against the alternative

$$H_{1h} : ||G_{h+1} - G_h||_{TV} > \epsilon, \tag{5.3}$$

where ϵ is some small constant such that when H_{0h} holds, there is no appreciable difference in the mixture distributions across these two groups. The global null of no changes in the response distribution across groups corresponds to the intersection of these local nulls,

$$H_0 : \bigcap_{h=1}^{d-1} H_{0h}, \tag{5.4}$$

which we will test against the global alternative of at least one change in distribution across adjacent groups,

$$H_1 : \exists \ h : ||G_{h+1} - G_h||_{TV} > \epsilon. \tag{5.5}$$

5.2.2 DMDP Model

In addition to conducting hypothesis tests, we would like to estimate the density in each dose group. As mentioned in Section 5.1, a promising approach would be to assign a dependent nonparametric process to $G_1, \ldots, G_d$ thereby allowing us to borrow information across groups. Using the method of Müller et al. (2004), we could let $G_h = \pi G_0 + (1 - \pi) G_h^*$ for $h = 1, \ldots, d$, where G_0 is a global probability measure and G_h^* is an innovation measure specific to group h, with $G_0, G_1^*, \ldots, G_d^*$ assigned independent nonparametric priors. However, this approach results in an over-specified model in that $d + 1$ random measures are incorporated to characterize d unknown distributions. Teh et al. (2006) proposed an alternative model which assumes that the G_h are drawn from DPPs with a common DP-distributed base measure G_0. However, these hierarchical priors treat the groups as exchangeable. To incorporate information

on ordering in the groups, Dunson (2006) proposed a dynamic mixture of Dirichlet processes (DMDP), which is conceptually related to the Müller et al. (2004) approach, but avoids the over-specification problem.

We first consider the case of two groups ($d = 2$). To induce dependency, we model the mixture distributions using the following DMDP:

$$\begin{aligned} G_1 &\sim \mathrm{DP}(\alpha_0 G_0) \\ G_2 = (1-\pi_1)G_1 + \pi_1 G_1^* \quad G_1^* &\sim \mathrm{DP}(\alpha_1 G_{01}), \end{aligned} \tag{5.6}$$

where $0 \leq \pi_1 \leq 1$ and G_1^* is an innovation distribution that characterizes changes in the mixture distribution of $\mathbf{y}$ caused by increasing the group score from X_1 (e.g., dose $= 0$) to X_2 (e.g., dose > 0). Note that the DMDP implies the following densities for $\mathbf{y}$ in groups 1 and 2:

$$\begin{aligned} f_1(\mathbf{y}) &= \int \mathrm{N}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) dG_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ f_2(\mathbf{y}) &= (1-\pi_1) f_1(\mathbf{y}) + \pi_1 \int \mathrm{N}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) dG_1^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned} \tag{5.7}$$

Hence the distribution of $\mathbf{y}$ in group 2 is a mixture of the distribution in group 1 and a different DP mixture of normals characterized by mixture distribution G_1^*. This is a natural model for toxicology data since we would expect that the distribution in the treatment group, F_2, shares features with the distribution in the control group, F_1, but that innovations may have occurred due to stress induced by the test chemical. The amount of borrowing information from the control control group and the level of innovation are represented by the weights $(1 - \pi_1)$ and π_1 respectively.

In the above model, it is convenient to choose $G_0 \equiv G_{01}$, with these base measures having a conjugate normal-inverse Wishart form: $dG_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{N}(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \kappa\boldsymbol{\Sigma}) \cdot \mathrm{W}(\boldsymbol{\Sigma}^{-1}; \mathbf{V}_0, v_0)$, where $\mathrm{W}(\cdot; \mathbf{V}_0, v_0)$ denotes the Wishart density with mean $\mathbf{V}_0$ and degrees of freedom v_0. In this case, our nonparametric prior for F_1 is centered on a multivariate t-distribution, while the prior for F_2 is centered on a mixture of multivariate t-distributions. As α_0 and α_1 increase, we place more weight on this parametric base model for the unknown densities. By choosing hyperprior densities for α_0 and α_1 (as we describe in Section 5.2.4), the method can adapt flexibly to accommodate lack of fit in the multivariate t components.

Under the DMDP, the total variation norm of $G_2 - G_1$ is

$$\begin{aligned} ||G_2 - G_1||_{TV} &= \max_{B \in \mathcal{B}} \Big| G_2(B) - G_1(B) \Big| \\ &= \max_{B \in \mathcal{B}} \Big| (1-\pi_1) G_1(B) + \pi_1 G_1^*(B) - G_1(B) \Big| \\ &= \pi_1 ||G_1^* - G_1||_{TV}, \end{aligned} \tag{5.8}$$

which implies the following null hypothesis:

$$H_0 : ||G_1^* - G_1||_{TV} \leq \epsilon/\pi_1. \tag{5.9}$$

Thus, our null hypothesis can be expressed in terms a different total variation norm whose upper bound is defined by two parameters, ϵ and π_1.

Extending our model to d groups, we let

$$\begin{aligned} G_h &= (1-\pi_{h-1}) G_{h-1} + \pi_{h-1} G_{h-1}^* \\ &= \left\{ \prod_{l=1}^{h-1} (1-\pi_l) \right\} G_1 + \sum_{l=1}^{h-1} \left\{ \prod_{m=l+1}^{h-1} (1-\pi_m) \right\} \pi_l G_l^* \\ &= \omega_{h1} G_1 + \sum_{l=1}^{h-1} \omega_{h,l+1} G_l^* \\ G_l^* &\sim DP(\alpha_l G_0), \text{ for } l = 1 \ldots, h-1, \end{aligned} \tag{5.10}$$

where $\boldsymbol{\omega}_h = (\omega_{h1}, \ldots, \omega_{hh})'$ are probability weights on the different components in the mixture and $h = 2, \ldots, d$. We also have the local nulls

$$H_{0h} : ||G_h^* - G_h||_{TV} \leq \epsilon/\pi_h \tag{5.11}$$

with H_0 being the intersection of $H_{01}, \ldots, H_{0,d-1}$.

5.2.3 Model Space Prior

Note that in the two sample case, H_0 holds for any fixed ϵ if: 1.) π_1 is sufficiently small, regardless of G_1 and G_1^* 2.) if G_1 is sufficiently close to G_1^* in total variation distance, regardless of π_1. Hence, the value of π_1 and the independent DPPs on G_1 and G_1^* control the prior probability allocated to H_0. If probability close to zero is allocated to H_0, then we may expect the posterior probability of H_0 to be small unless the sample

size is large, clearly an unappealing property in most applications. Thus, we want to be able to control the prior probability allocated to H_0.

We first consider the possibility of controlling $\Pr(H_0)$ through the priors for G_1 and G_1^*. The DPPs $G_1 \sim \text{DP}(\alpha_0 G_0)$ and $G_1^* \sim \text{DP}(\alpha_1 G_0)$ imply that $G_1(B) \sim \text{Be}(\alpha_0 G_0(B), \alpha_0[1 - G_0(B)])$ and $G_1^*(B) \sim \text{Be}(\alpha_1 G_0(B), \alpha_1[1 - G_0(B)])$, respectively. It is easy to verify numerically, through simulating from the above beta distributions over a grid of values for $G_0(B)$, that $\Pr(||G_1^* - G_1||_{TV} < \epsilon) \approx 0$, for ϵ close to zero regardless of the values of α_0 and α_1. Thus zero probability is allocated by the total variation component, which implies that $\Pr(H_0)$ is entirely controlled by π_1.

Based on the above statements, one can effectively replace H_0 in (5.9), with $\pi_1 \leq \epsilon^*$, where ϵ^* is also close to zero. Also, using similar arguments, we can replace the local nulls in the d-group setting, equation (5.11), with

$$H_{0h} : \pi_h \leq \epsilon_h^*. \tag{5.12}$$

Note that under these new local null hypotheses, we have a new global alternative:

$$H_1 : \exists \ h : \pi_h > \epsilon_h^*. \tag{5.13}$$

To control the prior probability allocated to H_0, we propose a prior for π_h which is a mixture of its distribution under H_{0h} and H_{1h}. Our priors are related to those used by George and McCulloch (1993) and Ishwaran and Rao (2003) to perform Bayesian variable selection. Let ζ_h be a latent indicator variable that equals 1 when the local null H_{0h} is true and 0 otherwise and let $\Pr(\zeta_h = 1) = p_{0h}$. Given ζ_h, our prior for π_h has the following form:

$$\pi_h|\zeta_h \sim \zeta_h \text{Be}(\pi_h; a_{0h}, b_{0h}) + (1 - \zeta_h)\text{Be}(\pi_h; a_{1h}, b_{1h}) \text{ for } h = 1, \ldots, d-1. \tag{5.14}$$

The first component of the mixture, $\text{Be}(a_{0h}, b_{0h})$, is chosen to be essentially a spike at zero; a good default prior is $\text{Be}(1, 99)$ which is centered on a 1% chance of an outlier under H_{0h} and has a 99th percentile of 0.05, which is a reasonable choice for ϵ_h^*. A good default choice for $\text{Be}(\pi_h; a_{1h}, b_{1h})$ is the noninformative prior $\text{Be}(1, 1)$.

In the above mixture, $\mathbf{p}_0 = (p_{01}, \ldots, p_{0h})'$ is chosen to reflect prior knowledge about the chances of H_0 being true. For example, to assign equal prior weight to H_0 and H_1, $\mathbf{p}_0$ should be chosen to satisfy $\Pr(H_0) = \prod_{h=1}^{d-1} p_{0h} = 0.5$. Under the Bayesian Bonferroni method of Westfall et al. (1997), this may be achieved by setting $p_{01} = \cdots = p_{0,d-1} =$

$0.5^{1/d}$. However for large d, the probability of each local null hypothesis is nearly 1, which can make this approach overly conservative. We instead use a model space prior described by Hans and Dunson (2005). This approach induces dependency in the local null hypotheses by assuming $p_{0h} = p_0$ for $h = 1, \ldots, d-1$ where $p_0 \sim \text{Be}(a_p, b_p)$ and a_p and b_p are chosen so that on average $\Pr(H_0)$=0.5, i.e., $\text{E}(p_0^{d-1}) = 0.5$. This implies that a_p and b_p must satisfy the following:

$$\begin{aligned} 0.5 &= \int_0^1 p_0^{d-1} \frac{\Gamma(a_p + b_p)}{\Gamma(a_p)\Gamma(b_p)} p_0^{a_p - 1}(1 - p_0)^{b_p - 1} dp_0 \\ &= \frac{\Gamma(a_p + b_p)\Gamma(a_p + d - 1)}{\Gamma(a_p)\Gamma(a_p + b_p + d - 1)}. \end{aligned} \tag{5.15}$$

As in Hans and Dunson, we impose the constraint $a_p + b_p = 1$ to represent unit information in the prior and solve (5.15) numerically for a_p and b_p.

5.2.4 Hyperpriors for DP Parameters

To decrease the sensitivity of analyses to a subjectively chosen G_0, we assign the following hyperpriors to κ and $\boldsymbol{\mu}_0$: $\kappa^{-1} \sim \text{Ga}(a_\kappa, b_\kappa)$ and $\boldsymbol{\mu}_0 \sim \text{N}(\boldsymbol{\beta}, \boldsymbol{\Omega})$. Escobar and West (1995) use a similar hyperprior structure to fit a DP mixture of normals to an outcome from a single group of subjects. They demonstrated that the modality of $f(\mathbf{y})$ is most sensitive to small values of κ. Thus, we recommend a diffuse prior for κ centered on large values. We also recommend a diffuse prior for $\boldsymbol{\mu}_0$ in which $\boldsymbol{\beta}$ represents one's best guess at the overall mean of $\mathbf{y}$.

A conjugate Wishart prior may also be assigned to $\mathbf{V}_0$ (see West et al., 1994). However, under a diffuse prior for κ, our model may not converge unless the prior for $\mathbf{V}_0$ is very informative. Thus, similar to what has been done in applications of DP mixtures of normals to galaxy data (Escobar and West, 1995) and neurological data (West and Turner, 1994; Cao and West, 1996), we recommend fixing $\mathbf{V}_0$ using an estimate obtained from historical data, or alternatively, at values under which a $\text{N}(\boldsymbol{\beta}, \mathbf{V}_0)$ distribution sufficiently covers an expected range of values of $\mathbf{y}$. We also recommend fixing v_0 on low integer values, though larger values may be used if there is substantial prior information about $\mathbf{V}_0$.

The values of $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots \alpha_{d-1})'$ may also be fixed to represent one's intuition about the number of normal mixture components added in moving between two groups (for a related approach see Escobar and West, 1995). We instead choose to estimate

these quantities from the data and assume $\alpha_h \sim \text{Ga}(a_{\gamma_h}, b_{\gamma_h})$ a priori for $h = 0, 1, \ldots, d-1$. In practice, one may favor priors which assign high prior probability to low values of $\boldsymbol{\alpha}$ since this implies that the number of mixture components is small (Antoniak, 1974) and a very flexible density estimator can be produced with few mixture components. Hence, a reasonable choice is $a_{\gamma_h} = a_\gamma = 1$ and $b_{\gamma_h} = b_\gamma = 1$.

5.3 Posterior Computation

In this section, we provide the full conditional posteriors from our DMDP and discuss methods for hypothesis testing and density estimation. We have placed the details regarding our Gibbs sampling algorithm in Appendix D, since it is closely related to the West et al. (1994) method described in Section 2.1.3.

5.3.1 Full Conditional Posterior Distributions

To simplify notation, let $\boldsymbol{\phi}_{hi} = \{\boldsymbol{\mu}_{hi}, \boldsymbol{\Sigma}_{hi}\}$. As in Dunson (2006), the derivation of our full conditional posterior distributions involves re-writing the model in (5.10) as

$$\begin{aligned} \boldsymbol{\phi}_{hi} &= \sum_{l=1}^{h} 1(M_{hi} = l)\boldsymbol{\xi}_{hil} \\ M_{hi} &\sim \text{Multinomial}(1, \ldots, h; \omega_{h1}, \ldots, \omega_{hh}) \\ \boldsymbol{\xi}_{hil} &\sim G_l^{**}, G_l^{**} \sim DP(\alpha_l G_0), \text{ for } l = 1, \ldots, h, \end{aligned} \tag{5.16}$$

where $G_l^{**} = G_1$ for $l = 1$ and $G_l^{**} = G_{l-1}^*$ for $l = 2, \ldots, h$.

Let $\boldsymbol{\theta}_l = (\boldsymbol{\theta}_{l1}', \ldots, \boldsymbol{\theta}_{lK_l}')'$ denote the unique values of $\boldsymbol{\xi}_l = \{\boldsymbol{\phi}_{hi} : M_{hi} = l; h = l, \ldots, d; i = 1, \ldots, n_h\}$, where $\boldsymbol{\theta}_{lr} = \{\boldsymbol{\mu}_{lr}^*, \boldsymbol{\Sigma}_{lr}^*\}$ and $\boldsymbol{\mu}_{lr}^*$ and $\boldsymbol{\Sigma}_{lr}^*$ are the (l, r)th unique values of the mean and dispersion parameters, respectively. Let m_l and m_{lr} denote the total number of subjects having $M_{hi} = l$ and $\boldsymbol{\phi}_{hi} = \boldsymbol{\theta}_{lr}$, respectively. Also, let $\boldsymbol{\theta}_l^{(hi)}, K_l^{(hi)}, m_l^{(hi)}$, and $m_{lr}^{(hi)}$ denote the values obtained after excluding subject h, i. Under this notation, the full conditional prior of $\boldsymbol{\phi}_{hi}$ is

$$\sum_{l=1}^{h} \omega_{hl} \left\{ \left(\frac{\alpha_{l-1}}{\alpha_{l-1} + m_l^{(hi)}} \right) G_0 + \sum_{r=1}^{K_l^{(hi)}} \left(\frac{m_{lr}^{(hi)}}{\alpha_{l-1} + m_l^{(hi)}} \right) \delta_{\boldsymbol{\theta}_{lr}^{(hi)}} \right\}, \tag{5.17}$$

where δ_θ denotes a degenerate distribution with all its mass at θ. Letting

$$\omega_{hl0}^{(hi)} = \frac{\omega_{hl} \cdot \alpha_{l-1}}{\alpha_{l-1} + m_l^{(hi)}}$$

and $\omega_{hlr}^{(hi)}$ denote the multiplier on $\delta_{\boldsymbol{\theta}_{lr}^{(hi)}}$ in (5.17), the full conditional posterior distribution for $\boldsymbol{\phi}_{hi}$ is as follows:

$$\sum_{l=1}^{h} \left\{ \widetilde{\omega}_{hl0}^{(hi)} G_{0hi} + \sum_{r=1}^{K_l^{(hi)}} \widetilde{\omega}_{hlr}^{(hi)} \delta_{\boldsymbol{\theta}_{lr}^{(hi)}} \right\}, \tag{5.18}$$

where

$$\begin{aligned}
\widetilde{\omega}_{hl0}^{(hi)} &= c \cdot \omega_{hl0}^{(hi)} \cdot \mathrm{t}(\mathbf{y}_{hi}; v_0 - q + 1, \boldsymbol{\mu}_0, \mathbf{V}_0^*) \\
\mathbf{V}_0^* &= \frac{v_0(1+\kappa)}{v_0 - q + 1} \cdot \mathbf{V}_0 \\
\widetilde{\omega}_{hlr}^{(hi)} &= c \cdot \omega_{hlr}^{(hi)} \cdot \mathrm{N}(\mathbf{y}_{hi}; \boldsymbol{\mu}_{lr}^*, \boldsymbol{\Sigma}_{lr}^*) \\
dG_{0hi}(\boldsymbol{\phi}_{hi}) &= \mathrm{W}\left(\boldsymbol{\Sigma}_{hi}^{-1}; \mathbf{V}_{0hi}, v_0 + 1\right) \cdot \mathrm{N}\left(\boldsymbol{\mu}_{hi}; \frac{\kappa}{\kappa+1}(\mathbf{y}_{hi} + \kappa^{-1}\boldsymbol{\mu}_0), \frac{\kappa}{\kappa+1}\boldsymbol{\Sigma}_{hi}\right) \\
\mathbf{V}_{0hi} &= \frac{1}{v_0+1}\left(v_0 \cdot \mathbf{V}_0 + \frac{1}{\kappa+1}(\mathbf{y}_{hi} - \boldsymbol{\mu}_0)(\mathbf{y}_{hi} - \boldsymbol{\mu}_0)'\right)
\end{aligned}$$

and $\mathrm{t}(\cdot; v, \boldsymbol{\mu}, \mathbf{V})$ denotes the multivariate t density with v degrees of freedom, location parameter, $\boldsymbol{\mu}$, and scale $\mathbf{V}$. Gibbs sampling may proceed by sampling directly directly from (5.18). However to improve mixing, we apply the algorithm described in Appendix D.

Given p_0 and $\mathbf{M} = \{M_{hi}\ i = 1, \ldots, n_h; h = 1, \ldots, d\}$, the probability that $\zeta_h = 1$ (i.e., H_{0h} is true) is

$$p_{0h}^* = \frac{p_0}{p_0 + (1 - p_0) \cdot BF_h}, \tag{5.19}$$

where BF_h is the Bayes factor for H_{1h} versus H_{0h},

$$BF_h = \frac{C(a_{1h}, b_{1h}) \cdot C(a_{0h}^*, b_{0h}^*)}{C(a_{0h}, b_{0h}) \cdot C(a_{1h}^*, b_{1h}^*)},$$

with

$$\begin{aligned} a^*_{kh} &= a_{kh} + \sum_{j=h+1}^{d} \sum_{i=1}^{n_j} 1(M_{ji} = h+1), \\ b^*_{kh} &= b_{kh} + \sum_{j=h+1}^{d} \sum_{i=1}^{n_j} 1(M_{ji} < h+1), k = 0, 1 \end{aligned}$$

and $C(a,b) = \Gamma(a+b)/(\Gamma(a) \cdot \Gamma(b))$. Given each ζ_h, π_h and p_0 have the following posteriors:

$$\pi(\pi_h|\cdot) = \zeta_h \text{Be}(a^*_{0h}, b^*_{0h}) + (1-\zeta_h)\text{Be}(a^*_{1h}, b^*_{1h}) \;\; h = 1, \ldots, d-1 \tag{5.20}$$

$$\pi(p_0|\cdot) = \text{Be}\left(a_p + \sum_{h=1}^{d-1} \zeta_h, b_p + d - 1 - \sum_{h=1}^{d-1} \zeta_h\right), \tag{5.21}$$

where the notation $a|\cdot$ denotes a given all other variables. The full conditional posteriors for $\boldsymbol{\mu}_0$ and κ are similar to those provided in Escobar and West (1995), while the conditional posterior for α_h is as in West (1992), with the modification that the relevant number of clusters and sample size are K_{h+1} and m_{h+1}, respectively, $h = 0, 1, \ldots, d-1$. As seen in Appendix D, it is straightforward to sample from the full conditional posterior distributions of each of these variables using Gibbs steps.

5.3.2 Hypothesis Testing

The global null may be formally evaluated using Rao-Blackwellization (Gelfand and Smith, 1990). For T iterations of the MCMC with a burn-in of T_b iterations, we compute

$$\begin{aligned} \widehat{\Pr}(H_0|\text{Data}) &= \frac{1}{T-T_b} \sum_{t=T_b+1}^{T} \Pr(\zeta_1 = \cdots = \zeta_{d-1} = 1 | p_0^{(t)}, \mathbf{M}^{(t)}) \\ &= \frac{1}{T-T_b} \sum_{t=T_b+1}^{T} \prod_{h=1}^{d-1} p_{0h}^{*(t)}, \end{aligned} \tag{5.22}$$

where the superscript (t) denotes a value sampled at iteration t. Using the common convention of using posterior probabilities as Bayesian alternatives to p-values, one may reject H_0 if (5.22) is < 0.05. However, it is more appropriate to consider posterior probabilities as measures of evidence than measures of significance. Thus, we also

recommend reporting that there is evidence against H_0 when the posterior probabilities are only moderately small (i.e., between 0.05 and 0.1).

An attractive feature of our methodology is that we can test local hypotheses within the same Markov chain used to test H_0. Since we properly calibrate our prior for p_0 to give equal prior probability to H_0 and H_1, the posterior probabilities of these local nulls do not need to be adjusted (see Westfall et al., 1997). For instance, the posterior probability of a distribution change between groups h and $h+1$ would be the average value of p^*_{0h} over $T - T_b$ iterations. In toxicology studies, there is also interest in comparing each dose group to control (group 1). Under our methodology, the null hypothesis of no difference in the distributions of groups 1 and h is

$$H^*_{0,h-1} : \pi_j \leq \epsilon^*_j \;\; j = 1, \ldots, h-1 \tag{5.23}$$

and we estimate its posterior probability as

$$\widehat{\Pr}(H^*_{0,h-1}|\text{Data}) = \frac{1}{T - T_b} \sum_{t=T_b+1}^{T} \prod_{j=1}^{h-1} p^{*(t)}_{0j}. \tag{5.24}$$

5.3.3 Density Estimation

The predictive density of a future $\boldsymbol{\phi}_{h,n_h+1}$ can be easily obtained at each iteration of the MCMC:

$$\pi(\boldsymbol{\phi}_{h,n_h+1}) = \sum_{l=1}^{h} \left\{ \frac{\alpha_{l-1}}{\alpha_{l-1} + m_l} \cdot dG_0(\boldsymbol{\phi}_{h,n_h+1}) + \sum_{r=1}^{K_l} \left(\frac{m_{lr}}{\alpha_{l-1} + m_l} \right) \delta_{\boldsymbol{\theta}_{lr}} \right\}. \tag{5.25}$$

This implies that the predictive density of $\mathbf{y}_{h,n_h+1}$ is

$$f_{h,n_h+1}(\mathbf{y}) \;\; = \;\; \omega_0 \cdot \mathrm{t}(\mathbf{y}; v_0 - q + 1, \boldsymbol{\mu}_0, \mathbf{V}^*_0) + \sum_{l=1}^{h} \sum_{r=1}^{K_l} \omega_{hlr} \cdot \mathrm{N}(\mathbf{y}; \boldsymbol{\mu}^*_{lr}, \boldsymbol{\Sigma}^*_{lr}), \tag{5.26}$$

where the multivariate t-density results from integrating $\mathrm{N}(\cdot; \boldsymbol{\mu}_{h,n_h+1}, \boldsymbol{\Sigma}_{h,n_h+1})$ over G_0, and ω_0 and ω_{hlr} are the respective multipliers on $dG_0(\boldsymbol{\phi}_{h,n_h+1})$ and $\delta_{\boldsymbol{\theta}_{lr}}$ in (5.25). Following T iterations, the Rao-Blackwellized estimate,

$$\bar{f}_{h,n_h+1}(\mathbf{y}) = \frac{1}{T - T_b} \sum_{t=T_b+1}^{T} f^{(t)}_{h,n_h+1}(\mathbf{y}), \tag{5.27}$$

may be computed over a grid of values and compared to the observed data to evaluate goodness of fit.

In a toxicology study, one would also be interested in the posterior density of the lowest observed adverse effects level (LOAEL). Assuming that any change in the distribution relative to group 1 is an adverse effect, we can estimate the posterior density of the LOAEL as follows:

$$\begin{aligned}
\widehat{\Pr}(\text{LOAEL} = h|\text{Data}) &= \frac{1}{T-T_b}\sum_{t=T_b+1}^{T}(1-p_{0,h-1}^{*(t)})\prod_{j=1}^{h-2}p_{0j}^{*(t)} \quad h=2,\ldots,d \\
\widehat{\Pr}(\text{LOAEL} = d+1|\text{Data}) &= \frac{1}{T-T_b}\sum_{t=T_b+1}^{T}\prod_{j=1}^{d-1}p_{0j}^{*(t)}, \qquad (5.28)
\end{aligned}$$

where a LOAEL of $d+1$ implies that the LOAEL is greater than the largest dose considered in the study. Hans and Dunson (2005) use a similar approach to estimate the posterior density of the LOAEL under umbrella orderings.

5.4 Simulation Studies

5.4.1 Description of Data

We performed three simulation studies to evaluate the performance of our methodology. In each simulation case, we generated a vector of three responses for each subject, $\mathbf{y}_{hi} = (y_{hi1}, y_{hi2}, y_{hi3})'$, from a mixture of five multivariate normal distributions:

$$f_h(\mathbf{y}_{hi}) = \sum_{j=1}^{5} w_{hj}\mathrm{N}(\mathbf{y}_{hi}; \boldsymbol{\mu}_j, \tau_j \cdot \mathbf{I}_3),$$

where $\boldsymbol{\mu}_j = (\mu_{j1}, \mu_{j2}, \mu_{j3})'$ and $w_{hj} \propto \exp(\gamma_{j0} + (h-1)\gamma_{j1})$ for $h = 1,\ldots,6$ and $i = 1,\ldots,30$. For each mixture component, $\mu_{jk} = x_j + k - 1$ where x_j is a constant.

The parameter values used in the simulations are provided in Table 5.1. Cases 1 and 2 differed only in their values of the mixture probabilities; in Case 1 we let $\gamma_{j0} = 1$ and $\gamma_{j1} = 0$ for $j = 1\ldots,5$ to simulate under a null model, while in Case 2 the mixture probabilities varied with predictor level. As seen in Figure 5.1, the right tails of the marginal densities of y_1 become heavier with predictor level in Case 2 while, in Case 3, the modality of the distribution changes across h. The latter case may arise in toxicology experiments in which gene×dose interactions cause sub-populations to

TABLE 5.1: Parameter values for the mixture components in simulation cases 2 and 3.

			Case 2		
			j		
Parameter	**1**	**2**	**3**	**4**	**5**
x_j	-0.6	-0.5	-0.4	0.5	1
τ_j	0.16	0.25	0.36	0.903	1.323
γ_{j0}	1	1	1	-3	-3
γ_{j1}	-1	-0.25	-0.25	0.25	0.75
			Case 3		
x_j	-2	-1	0	1	2
τ_j	0.16	0.25	0.36	0.903	1.323
γ_{j0}	-3	-3	1	-3	-3
γ_{j1}	1	0.5	0	0.5	1

respond differently to treatment. The marginal densities of y_2 and y_3 follow a similar trend with dose, though their location parameters differ. Five data sets were simulated under each case.

5.4.2 Univariate Analyses

We first performed a univariate analysis on y_1 in order to compare our methodology to standard frequentist methods. In each case, we assumed $\mu_0 \sim \mathrm{N}(0, 100)$, $\kappa^{-1} \sim (0.5, 50)$, and $\alpha_h \sim \mathrm{Ga}(1, 1)$ for $h = 1, \ldots, d$ a priori and let $v_0 = 4$ and $V_0 = 1$. In our mixture priors for $\pi_1, \ldots, \pi_5$, we let $a_{0h} = a_{1h} = b_{1h} = 1$ and $b_{0h} = 99$ for $h = 1, \ldots, d-1$, as recommended in Section 5.2.3, and $p_0 \sim \mathrm{Be}(0.725, 0.275)$ to assign equal prior probability to H_0 and H_1. We ran our MCMC for a total $25,000$ iterations, discarding the first 5,000 as a burn-in, and saved every 10th iteration to thin the chain.

Our method correctly assigned high probability to H_0 in Case 1 (> 0.87 for each data set) and very low posterior probability to H_0 in Cases 2 and 3 (as seen in Table 5.2). These results were consistent with the p-values from a k-sample Kolmogorov-Smirnov test (Kiefer, 1959), while both Jonckheere's test and the Kruskal-Wallis test

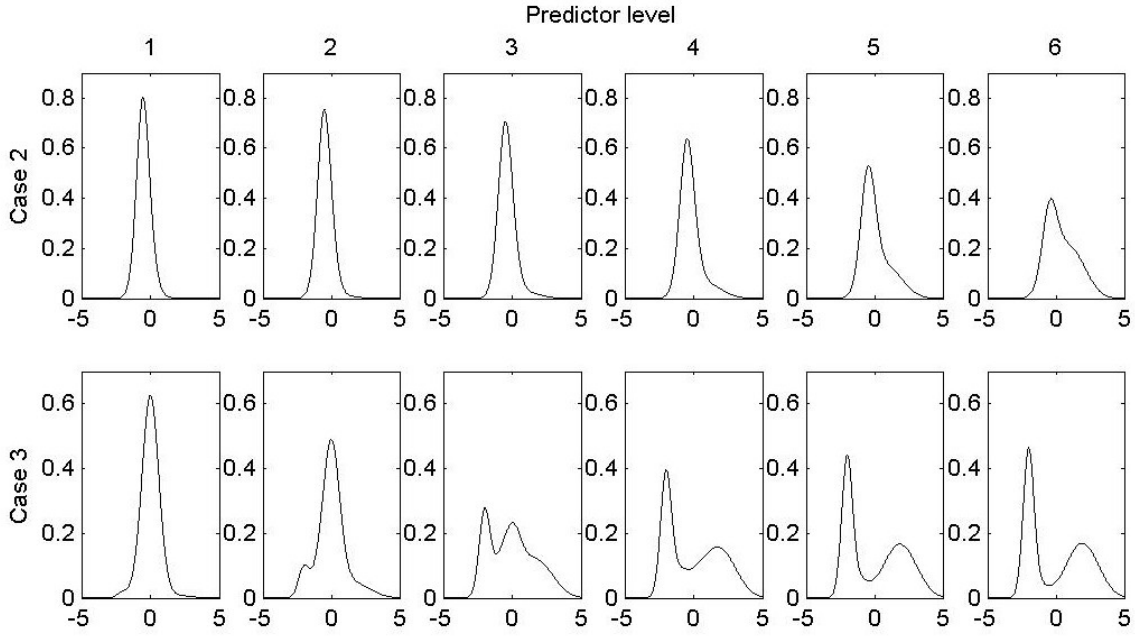

FIGURE 5.1: Marginal density of y_1 at each predictor level in simulation cases 2 and 3.

failed to reject the null in 1/5 Case 2 simulations and 5/5 Case 3 simulations. Since stochastic ordering was violated in Case 3, it is not surprising that Jonckheere's test was insignificant. Being sensitive to location but not shape changes, the Kruskal-Wallis test appeared to have lower power than our approach, which can detect any change in the distribution.

In addition to assessing evidence of any change across the groups, it is of interest to identify dose groups that differ from control. We compared our approach to methods commonly used for toxicology data, Dunn's method (1964) and Williams' (1986) modification of Shirley's method (1977). We also compared our method to $(d-1)$ 2-sample Kolmogorov-Smirnov tests performed under a Bonferroni correction. As seen in Table 5.2, our method found clear evidence that groups 5 and 6 differed from group 1 in 5/5 Case 2 simulations. In contrast, Shirley's method, which assumes a monotone trend across groups, found a significant difference in 4/5 simulations while Dunn's method and the Kolmogorov-Smirnov tests were more conservative. In Case 3, our method indicated that groups 3-6 were all clearly different from group 1. The Kolmogorov-Smirnov tests were more conservative and suggested group-related trends that were inconsistent with the simulated data; in two data sets a lower group was significantly different from

group 1, while a higher group was not found to differ from baseline.

5.4.3 Multivariate Analyses

We next repeated our analyses using a multivariate approach. In addition to testing for a predictor effect on the distribution of $\mathbf{y}$, in Cases 2 and 3 we also tested for an effect on the joint distribution of y_1, z_1, z_2 and y_1, y_2, z_2, where $(z_{hi1}, z_{hi2}) \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_2)$ for $h = 1, \ldots, 5$ and $i = 1 \ldots, 30$ independently of $\mathbf{y}_{hi}$. The purpose of these latter tests were to evaluate the performance of our multivariate method under varying numbers of affected outcomes. The implementation was very similar to that of univariate analyses, with $\boldsymbol{\mu}_0 \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_3)$ a priori and $\mathbf{V}_0 = \mathbf{I}_3$.

As in the univariate analyses, our method gave high posterior probability to H_0 in Case 1; the range of the posterior probabilities was 0.774-0.970 across the five data sets. Figure 5.2 summarizes the results for Cases 2 and 3. In Case 2, the posterior probabilities of H_0, H_{04}^*, and H_{05}^* were close to zero in each analysis. However, the posterior probabilities of H_{01}^*, H_{02}^* and H_{03}^* increased as we decreased the number of y's (the affected outcomes) in our model. This suggests that our method is conservative when only one outcome is moderately affected, which is a desirable feature for a multivariate method. In Case 3, the results were more consistent across the number of affected outcomes since the distribution changes in $\mathbf{y}$ were more substantial.

5.5 Genotoxicity Example

5.5.1 Data and Methods

We considered data from a genotoxicity study which used the comet assay (single-cell electrophoresis) to measure the effects of oxidative stress on DNA strand breaks. These data were previously analyzed in Dunson (2006), Dunson et al. (2003), and Dunson and Taylor (2005). In the comet assay, cells are embedded in agarose on a microscope slide and lysed to remove all cellular proteins. The cells are then subjected to electrophoresis to determine the extent of DNA damage; after being stained with ethidium bromide, intact DNA will appear as a single sphere under a fluorescent microscope while broken DNA fragments will migrate from the nucleus giving the image a comet-like appearance. Using imaging software, several surrogates of DNA strand breaks can be obtained

TABLE 5.2: Posterior probabilities of global and local null hypotheses for the univariate analysis of y_1 in simulation cases 2 and 3. The letters in superscript denote that the following frequentist tests were significant at the 0.05 level: J=Jonckheere, KW=Kruskal-Wallis, KS=Kolmogorov-Smirnov, S=Shirley, D=Dunn.

Data set	H_0	H_{01}^*	H_{02}^*	H_{03}^*	H_{04}^*	H_{05}^*
Case 2						
1	0.024^{KS}	0.823	0.753	0.700	0.029	0.024^{KS}
2	$< 0.001^{\text{J,KW,KS}}$	0.834	0.570	0.376	$< 0.001^{\text{S,D}}$	$< 0.001^{\text{S,D,KS}}$
3	$0.003^{\text{J,KW,KS}}$	0.802	0.655	0.016	0.003^{S}	$0.003^{\text{S,D,KS}}$
4	$< 0.001^{\text{J,KW,KS}}$	0.895	0.771	0.240	$< 0.001^{\text{S,D}}$	$< 0.001^{\text{S,D}}$
5	$< 0.001^{\text{J,KW,KS}}$	0.547	0.464	0.395	0.020^{S}	$< 0.001^{\text{S,D,KS}}$
Case 3						
1	$< 0.001^{\text{KS}}$	0.076	0.036	$< 0.001^{\text{KS}}$	$< 0.001^{\text{KS}}$	$< 0.001^{\text{KS}}$
2	$< 0.001^{\text{KS}}$	0.459	0.009	$< 0.001^{\text{KS}}$	$< 0.001^{\text{KS}}$	$< 0.001^{\text{KS}}$
3	$< 0.001^{\text{KS}}$	0.560	$< 0.001^{\text{KS}}$	< 0.001	$< 0.001^{\text{KS}}$	$< 0.001^{\text{KS}}$
4	$< 0.001^{\text{KS}}$	0.155	< 0.001	< 0.001	$< 0.001^{\text{KS}}$	< 0.001
5	$< 0.001^{\text{KS}}$	0.696	0.010	$< 0.001^{\text{KS}}$	$< 0.001^{\text{KS}}$	$< 0.001^{\text{KS}}$

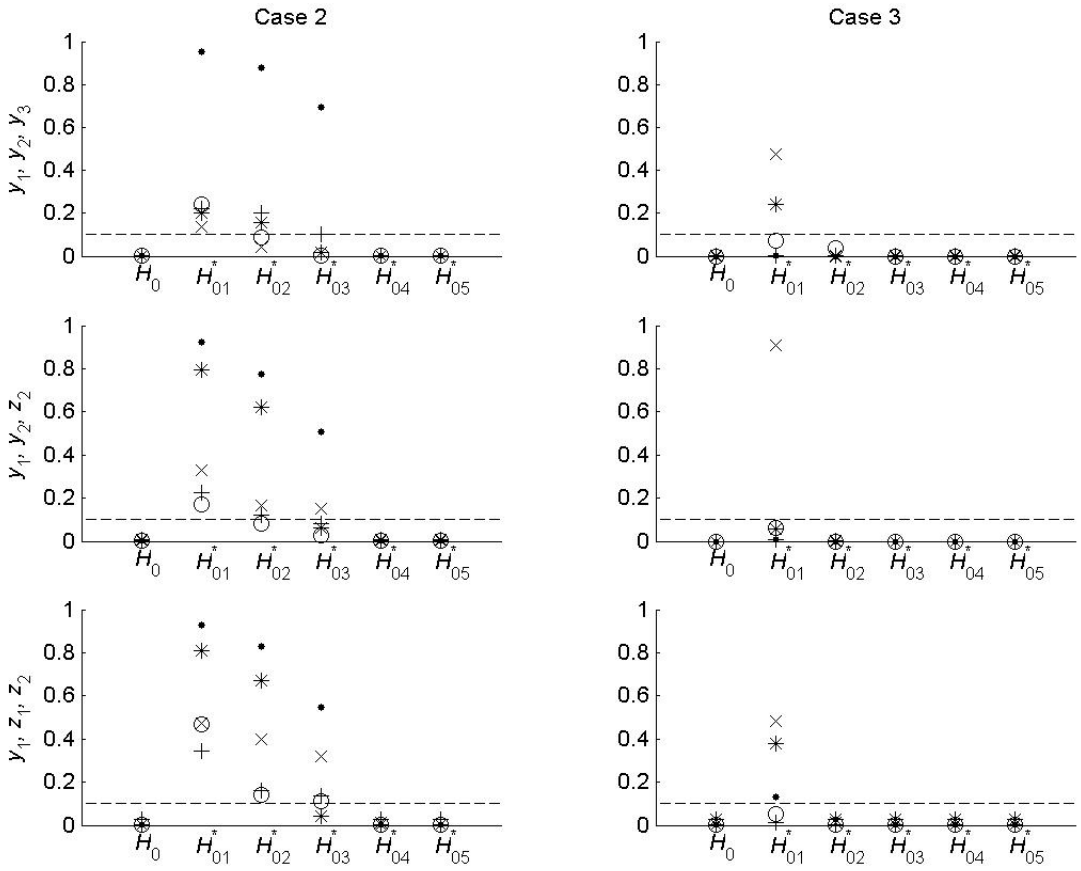

FIGURE 5.2: Posterior probabilities of global and local null hypotheses for the multivariate analyses in simulation cases 2 and 3. Labels for y-axes denote the different outcomes used in each analysis. Data points correspond to results from data sets 1 (+), 2 (o), 3 (*), 4 (·), and 5 (×). As a point of reference for evaluating each posterior probability, we have provided a dashed line at 0.1

including length of the comet tail and % DNA in the comet tail. Previous analyses of comet assay data have shown that these surrogates are non-Gaussian even after transformation (Duez et al., 2003; Dunson et al., 2003), and thus are a particularly interesting example for our methodology. For a discussion of statistical issues related to comet assay data see Lovell et al. (1999).

The data set we analyzed consisted of human lymphoblast cells exposed to 5 different concentrations of hydrogen peroxide (0, 5, 20, 50, and 100 μM H_2O_2). The goal of this experiment was to determine the sensitivity of the comet assay to detect DNA damage induced by H_2O_2, a known genotoxic agent. Several surrogates of the frequency of DNA strand breaks were measured for 100 cells in each dose group. However, we focus our analyses on the two best surrogates: $y_1 =$ % tail DNA and $y_2 =$ Olive tail moment (OTM), which is (% tail DNA)$\times$(distance between center of head and center of tail).

Our implementation was identical to that described in the simulation studies, with a few exceptions. Based on the range of surrogate values reported in Duez et al. (2003), we chose $\boldsymbol{\beta} = (50, 10)'$, $V_{011} = 180$, $V_{022} = 10$, and $V_{012} = V_{021} = 38.2$, where V_{0ij} denotes element (i, j) of $\mathbf{V}_0$. Note that we chose a large value for the covariance because % tail DNA and OTM are likely to be highly correlated. However, since we are uncertain about how well these mean and variance values represent the current data, we chose a relatively diffuse prior for $\boldsymbol{\mu}_0$, $\mathrm{N}(\boldsymbol{\beta}, 6 \cdot \mathbf{V}_0)$, and let $v_0 = 4$. Also, using the guidelines in Section 5.2.3, we assumed $p_0 \sim \mathrm{Be}(0.7, 0.3)$ a priori.

5.5.2 Results

Our method provided strong evidence in favor of an effect of H_2O_2 on the frequency of DNA strand breaks as $\widehat{\Pr}(H_0|\text{Data}) < 0.001$. As demonstrated by the posterior predictive densities in Figure 5.3, treatment with H_2O_2 changes the joint distribution of the % tail DNA and OTM from a unimodal distribution favoring small values to a multi-modal distribution supporting large levels of DNA damage. While genotoxic effects are evident at the smallest dose of H_2O_2 (the posterior probability that 5 μM is the LOAEL is > 0.999), they appear to level off at the higher doses since there is no appreciable difference in the distributions of the 20, 50, and 100 μM groups; $\widehat{\Pr}(\text{No difference}|\text{Data}) = 0.816$. Figure 5.3 demonstrates that these densities are in agreement with the observed data, supporting goodness of fit.

Although previous analyses of this data set have also demonstrated a relationship between H_2O_2 and the frequency of DNA strand breaks, we have provided several new

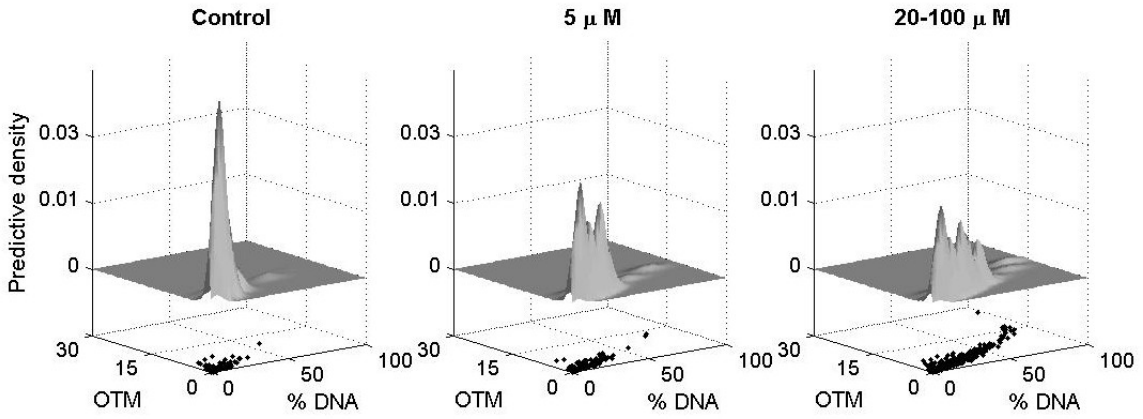

FIGURE 5.3: Posterior predictive density of % tail DNA and OTM in each dose group. The density presented for doses 20-100 μM H_2O_2 was averaged across the 20, 50 and 100 μM dose groups. '$\cdot$' denotes an observation.

pieces of information. For instance, neither the latent response models fit by Dunson et al. (2003) and Dunson (2006) nor the quantile regression approach of Dunson and Taylor (2005) provided a formal method for testing for changes across dose groups and for estimating threshold doses. In genetic epidemiology studies of DNA repair it is important to use minimal doses of the test agent to avoid high levels of cell death, which make the comet assay unreliable. Hence our finding that 20 μM is a threshold level for DNA damage has important implications for future experiments. Another advantage of our method is our ability to obtain smooth, nonparametric estimates of the joint density of the surrogates. Potentially, the information in these estimates (e.g., the number of normal mixture components in each dose group) could be used to elicit priors in future analyses of related experiments.

5.6 Discussion

In this chapter, we have proposed a nonparametric method which models distribution changes across an ordinal predictor and provides a formal approach for Bayesian testing of local and global changes across groups. Our method provides more informative results than many frequentist k-sample tests and, as demonstrated in simulation studies, may be more sensitive to changes in distributional shape. In addition, our method can

perform inferences on multivariate distributions and identify thresholds.

Although our method is relatively simple to program, the computational burden increases substantially with both sample size and the dimension of the response. For example, it took 23 hours to complete the MCMC for our analysis of the comet assay data using a Matlab (version 7.0) program run in batch on the statistical server at UNC-Chapel Hill (20 1.05 GHz processors); in comparison, it took approximately five hours to complete a single chain in our univariate analyses of the simulated data. Thus, a more efficient computational method would likely be necessary for massive, high dimensional data sets. One promising approach may be to develop a Variational Bayes (VB) implementation (see Blei and Jordan, 2006).

Some extensions of our method may also serve as exciting areas of future research. For example, a formal method of incorporating historical data would be useful in many applications, such as toxicology studies with extensive historical control databases. These data could potentially be used to generate an empirical estimate of G_0 for our model. It may also be of interest to incorporate order restrictions into our method which ensure that the the quantiles of the response are non-decreasing with dose. This extension would be related to the nonparametric approaches for stochastic ordering described by Gelfand and Kottas (2001) and Hoff (2003).

CHAPTER 6

CONCLUDING REMARKS

6.1 Overview

In this dissertation, we have proposed Bayesian semiparametric methods which address computational and substantive problems in biomedical data. Our work was motivated by real data examples from epidemiology and toxicology studies.

In Chapter 3, we developed a method which makes fitting semiparametric random effects models feasible for large data sets. The research was motivated by a computational problem that arose in collaborative research on the effect of maternal smoking on child growth in the Collaborative Perinatal Project (CPP). Our approach uses expert elicitation to generate a smaller data set which we model using a Dirichlet process mixture (DPM). Simulation studies demonstrated that our method is accurate and requires less computation time than DPMs fit to the complete data. We also presented the first random effects analysis of the CPP data and demonstrated that previous analyses using GEE may have generated biased estimates.

The problems we addressed in Chapters 4 and 5 were more substantive. In Chapter 4, we generalized the shared frailty model for multiple event time data to account for time-dependent changes in susceptibility across subjects, such as rats in a study of palpable tumors. Our method models smooth trends in the hazard, covariate effects and frailty, while relaxing distribution assumptions on the frailty using nonparametric priors. By accounting for changes in frailty in our chemoprevention example, we provided a more realistic description of the canthaxanthin effect than previous analyses which used shared frailty-type models.

Finally, in Chapter 5 we proposed a Bayesian method for testing for changes in an outcome distribution with an ordinal predictor. The approach is attractive for

toxicology data because it is sensitive to changes in distribution shape with dose (such as changes in skewness and modality) and can model trends across multiple outcomes. When we applied our method to genotoxicity data, we found that the modality of the bivariate distribution of % tail DNA and Olive tail moment (two surrogates of DNA damage) changes with the dose of H_2O_2. We also identified an interesting threshold that could have interesting implications for future experiments.

6.2 Future Research

At the end of Chapters 3-5, we discussed some possible extensions of our methods and areas of future research. We would like to briefly revisit a couple of issues that arose and discuss how they relate to more general problems in Bayesian inference.

In the discussion sections of Chapters 3 and 5, we mentioned that our methods may need to be modified to improve efficiency under high dimensional models (discussed in Chapters 3 and 5) or large data sets (discussed in Chapter 5). We anticipate that this will be an ongoing problem for Bayesian methods as many future biomedical studies, including large scale prospective studies and genetic epidemiology studies, will produce massive amounts of information that need to be analyzed quickly. In addition, high dimensional models will continue to play an integral role in the analysis of epidemiology data to ensure proper adjustment for confounders. Such issues have motivated the development of fast alternatives to traditional MCMC such as subsampling (Huang and Gelman, 2005), particle-filters (Chopin, 2002; Ridgeway and Madigan, 2003; Balakrishnan and Madigan, 2006), variational inference methods (e.g., Blei and Jordan, 2006), and our two-stage method in Chapter 3. However, we must continue to increase the number of computational tools available to ensure that Bayesian inference is viable in these settings.

The other issue issue we would like to address was noted at the conclusion of Chapter 4. In Section 4.5, we mentioned the possibility of removing the almost discrete restriction on our frailties using either a DPM or Pólya tree (Lavine, 1992, 1994). The DPM extension would be conceptually straightforward, though we would sacrifice parsimony in our model and may increase the computational burden. This is exemplifies a trade-off that exists in Bayesian nonparametrics. Ideally, one would like to choose priors that most accurately reflect their prior beliefs (e.g., that the frailty distribution is continuous) but since these models are more challenging to fit, people often adhere to simpler priors such as the DP. Thus, a continuing challenge will be to develop methods

which strike a balance between flexibility and computational convenience.

APPENDIX A

PROOF OF CONVERGENCE OF STAGE 1 CLUSTERING ALGORITHM IN CHAPTER 3

In Step 2 of our Stage 1 clustering algorithm we wish to minimize the *modified* squared error, $Q_j(\mathbf{z}_j, \mathbf{c}_j) = \sum_{i=1}^{M_j} Q_{ji}(\mathbf{z}_{ji}, \mathbf{c}_j)$, where $\mathbf{z}_j = (\mathbf{z}'_{j1}, \ldots, \mathbf{z}'_{jM_j})'$, $\mathbf{c}_j = (\mathbf{c}'_{j1}, \ldots, \mathbf{c}'_{jG_j})'$, and

$$Q_{ji}(\mathbf{z}_{ji}, \mathbf{c}_j) = \begin{cases} (d^*_{ji})^2 & d^*_{ji} \leq r \\ r^2 & d^*_{ji} > r. \end{cases}$$

Thus, the proof of convergence of the algorithm involves showing two conditions: 1.) changing the cluster assignment of a subject does not increase the modified square error, $Q_j(\mathbf{z}_j, \mathbf{c}_j)$, denoted Q_j hereafter 2.) updating the seed of a cluster does not increase Q_j.

1. Let Q_{ji} denote the contribution of subject j, i to Q_j prior to cluster assignment and Q^*_{ji} denote its value afterward. At iteration t, if subject j, i is:

   a.) moved from cluster j, l to cluster j, l' then

   $$Q^*_{ji} = d^2(\mathbf{z}_{ji}, \mathbf{c}^{(t-1)}_{jl'}) < d^2(\mathbf{z}_{ji}, \mathbf{c}^{(t-1)}_{jl}) = Q_{ji}.$$

   b.) assigned to cluster j, l' after not being assigned to a cluster at iteration $t-1$ then

   $$Q^*_{ji} = d^2(\mathbf{z}_{ji}, \mathbf{c}^{(t-1)}_{jl'}) \leq r^2 = Q_{ji}.$$

   c.) not assigned to a cluster after being in cluster l at iteration $t-1$ then

   $$Q^*_{ji} = r^2 \leq d^2(\mathbf{z}_{ji}, \mathbf{c}^{(t-1)}_{jl}) = Q_{ji}.$$

In each case, changing the cluster assignment of j, i does not increase its contribution to Q_j, thus demonstrating that Condition 1 holds.

2. Following the (t)th iteration of Step 2.1, the contribution of cluster j, l to Q_j is

$$\begin{aligned}
Q^*_{jl} &= \sum_{i \in jl} \sum_{k=1}^{p_j} (z_{jik} - \mathrm{c}_{jlk}^{(t-1)})^2 \\
&= \sum_{i \in jl} \sum_{k=1}^{p_j} (z_{jik} + \mathrm{c}_{jlk}^{(t)} - \mathrm{c}_{jlk}^{(t)} - \mathrm{c}_{jlk}^{(t-1)})^2 \\
&= \sum_{i \in jl} \sum_{k=1}^{p_j} (z_{jik} - \mathrm{c}_{jlk}^{(t)})^2 + m_{jl} \sum_{k=1}^{p_j} (\mathrm{c}_{jlk}^{(t)} - \mathrm{c}_{jlk}^{(t-1)})^2 \\
&\quad + 2 \sum_{i \in jl} \sum_{k=1}^{p_j} (z_{jik} - \mathrm{c}_{jlk}^{(t)})(\mathrm{c}_{jlk}^{(t)} - \mathrm{c}_{jlk}^{(t-1)}) \\
&= \sum_{i \in jl} \sum_{k=1}^{p_j} (z_{jik} - \mathrm{c}_{jlk}^{(t)})^2 + m_{jl} \sum_{k=1}^{p_j} (\mathrm{c}_{jlk}^{(t)} - \mathrm{c}_{jlk}^{(t-1)})^2 \\
&\geq \sum_{i \in jl} \sum_{k=1}^{p_j} (z_{jik} - \mathrm{c}_{jlk}^{(t)})^2 = Q_{jl},
\end{aligned}$$

where Q_{jl} is the contribution of cluster j, l to Q_j following Step 2.2. This demonstrates that updating the seed of a cluster does not increase its contribution to Q_j, thereby completing the proof of convergence. Similar arguments can be used to prove convergence of k-means clustering under squared-error loss (MacQueen, 1967).

APPENDIX B

MCMC METHODOLOGY FOR CHAPTER 3

B.1 Methods for updating the random effects

Conditional on the other random effects, the prior for $\mathbf{b}_g^*$ is the mixture distribution

$$[\mathbf{b}_g^*|\boldsymbol{\theta}, \mathbf{n}^{(g)}, \alpha] \sim \left(\frac{\alpha}{\alpha + G - 1}\right) H_0 + \left(\frac{1}{\alpha + G - 1}\right) \sum_{k=1}^{K} \mathrm{n}_k^{(g)} \delta_{\boldsymbol{\theta}_k}, \tag{B.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_k')'$, $\mathbf{n}^{(g)} = (\mathrm{n}_1^{(g)}, \ldots, \mathrm{n}_K^{(g)})'$, $\mathrm{n}_k^{(g)}$ is the number of pseudo-subjects (other than g) with common random effect value $\boldsymbol{\theta}_k$, and $\delta_{\boldsymbol{\theta}_k}$ denotes a degenerate distribution at $\boldsymbol{\theta}_k$. After incorporating the likelihood for pseudo-subject g, $f(\mathbf{y}_g^*|\mathbf{b}_g^*)$, the full conditional posterior distribution of each $\mathbf{b}_g^*$ can be derived as

$$[\mathbf{b}_g^*|\mathbf{y}_g^*, \alpha, \boldsymbol{\theta}, \mathbf{n}^{(g)}] \sim q_{g0} H_{g0} + \sum_{k=1}^{K} q_{gk} \delta_{\boldsymbol{\theta}_k}, \tag{B.2}$$

where

$$q_{gk} = \begin{cases} c \cdot \alpha \cdot h(\mathbf{y}_g^*) & k = 0, \\ c \cdot \mathrm{n}_k^{(g)} \cdot f(\mathbf{y}_g^*|\boldsymbol{\theta}_k) & k > 0, \end{cases}$$

H_{g0} is a normal distribution with mean $\boldsymbol{\mu}_g = \mathbf{U}_g(\mathbf{D}^{-1}\boldsymbol{\mu} + \tau \mathbf{X}_g^{*'}\mathbf{y}_g^*)$ and covariance matrix $\mathbf{U}_g = (\mathbf{D}^{-1} + \tau \mathbf{X}_g^{*'}\mathbf{X}_g^*)^{-1}$,

$$h(\mathbf{y}_g^*) = \left(\frac{\tau}{2\pi}\right)^{\frac{n_g^*}{2}} |\mathbf{D}|^{-1/2} |\mathbf{U}_g|^{1/2} \cdot \exp\left\{ -\frac{1}{2}\left(\tau \mathbf{y}_g^{*'}\mathbf{y}_g^* + \boldsymbol{\mu}'\mathbf{D}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}_g'\mathbf{U}_g^{-1}\boldsymbol{\mu}_g\right)\right\},$$

and c is a normalization constant.

In order to update the random effect values from their full-conditional posteriors, we propose an MCMC algorithm which parallels that of MacEachern (1994) and West

et al. (1994):

1. For $g = 1, \ldots, G$, sample a random variable, $S_g \in \{0, 1, \ldots, K\}$, which equals k with probability q_{gk}. When $S_g = 0$, sample $\mathbf{b}_g^*$ from H_{g0} and increment K by one; for $S_g = k > 0$ set $\mathbf{b}_g^* = \boldsymbol{\theta}_k$.

2. For $k = 1, \ldots, K$ update $\boldsymbol{\theta}_k$ from its full conditional posterior, which is $\mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}_k}, \mathbf{R}_k)$, where

$$\boldsymbol{\mu}_{\boldsymbol{\theta}_k} = \mathbf{R}_k(\mathbf{D}^{-1}\boldsymbol{\mu} + \tau \sum_{g:S_g=k} \mathbf{X}_g^{*\prime}\mathbf{y}_g^*), \qquad \mathbf{R}_k = (\mathbf{D}^{-1} + \tau \sum_{g:S_g=k} \mathbf{X}_g^{*\prime}\mathbf{X}_g^*)^{-1}. \quad \text{(B.3)}$$

   Note that updating $\boldsymbol{\theta}_k$ changes the value of $\mathbf{b}_g^*$ for all g such that $S_g = k$.

B.2 Methods for updating the hyperparameters

Under the priors specified in Section 3.3.4 , the hyperparameter values may be updated by adding the following steps to our MCMC:

3. Sample $\boldsymbol{\mu}$ from

$$\pi(\boldsymbol{\mu}|\boldsymbol{\theta}, \mathbf{D}) = \mathrm{N}\bigg(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{D}^{-1}\sum_{k=1}^{K}\boldsymbol{\theta}_k), \boldsymbol{\Sigma}_{\boldsymbol{\mu}}\bigg), \quad \text{(B.4)}$$

   where $\boldsymbol{\Sigma}_{\boldsymbol{\mu}} = (\boldsymbol{\Sigma}_0^{-1} + K\mathbf{D}^{-1})^{-1}$.

4. Sample $\mathbf{D}^{-1}$ from

$$\pi(\mathbf{D}^{-1}|\boldsymbol{\theta}, \boldsymbol{\mu}) = \mathrm{W}\bigg(d_0 + K, \mathbf{D}_0 + \sum_{k=1}^{K}(\boldsymbol{\theta}_k - \boldsymbol{\mu})(\boldsymbol{\theta}_k - \boldsymbol{\mu})'\bigg). \quad \text{(B.5)}$$

5. Sample τ from

$$\pi(\tau|\mathbf{e}_1^*, \ldots, \mathbf{e}_G^*) = \mathrm{Ga}\bigg(\psi\tau_0 + \frac{n^*}{2}, \psi + \frac{1}{2}\sum_{g=1}^{G}\mathbf{e}_g^{*\prime}\mathbf{e}_g^*\bigg), \quad \text{(B.6)}$$

   where $n^* = \sum_{g=1}^{G} n_g^*$ and $\mathbf{e}_g^* = (\mathbf{y}_g^* - \mathbf{X}_g^*\mathbf{b}_g^*)$.

6. Sample α from

$$\pi(\alpha|z,K) = \pi_z \mathrm{Ga}\Big(a+K, b-\log(z)\Big) + (1-\pi_z)\mathrm{Ga}\Big(a+K-1, b-\log(z)\Big), \quad \text{(B.7)}$$

where

$$\frac{\pi_z}{(1-\pi_z)} = \frac{(a+K-1)}{G\big(b-log(z)\big)}$$

and $\pi(z|\alpha,K) = \mathrm{Be}(\alpha+1, G)$. As noted by West (1992) α may be updated by first sampling z from $\pi(z|\alpha,K)$ and then sampling α from (B.7).

The MCMC methodology proceeds by iterating over Steps 1-6 for a large number of iterations and discarding a burn-in period to allow convergence. However, to speed up computation, we recommend sampling each S_g conditional on the random effect values at the previous iteration. Although this modification may slow convergence down slightly, it is unlikely that it will affect posterior summaries obtained from long chains.

APPENDIX C

MCMC METHODOLOGY FOR CHAPTER 4

C.1 Full Conditional Posterior Distributions

Let $\mathbf{Y} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_N; t_1^*, \ldots, t_N^*; \mathbf{X}_1, \ldots, \mathbf{X}_N\}$ denote the observed data, where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iM_i})'$ and $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iM_i})'$. Also, let $\mathbf{\Phi} = \{\phi_i, \phi_{ij}, i = 1, \ldots, N; j = 1 \ldots, M_i\}$, $\boldsymbol{\nu} = (\nu_0, \nu_1, \ldots, \nu_M)'$, and $\mathbf{\Gamma} = \{\gamma_{hk}, h = 1 \ldots, M; k = 1, \ldots, p\}$. Given the above notation, the likelihood is proportional to

$$\begin{aligned} L(\mathbf{\Phi}, \boldsymbol{\nu}, \mathbf{\Gamma}|\mathbf{Y}) &= \prod_{i=1}^{N}\prod_{j=1}^{M_i}\left(\phi_i\nu_0\widehat{\lambda}_{0j}\prod_{h=1}^{j}\phi_{ih}\nu_h\prod_{k=1}^{p}\gamma_{hk}^{x_{ijk}}\right)^{Z_{ij}} \\ &\quad \times \exp\left\{-r_{ij}\phi_i\nu_0\widehat{\lambda}_{0j}\prod_{h=1}^{j}\phi_{ih}\nu_h\prod_{k=1}^{p}\gamma_{hk}^{x_{ijk}}\right\}. \end{aligned} \tag{C.1}$$

Under the dynamic gamma model, the full conditional posterior distributions of the ϕ_i and ϕ_{ij}'s are also gamma due to the Poisson form of (C.1):

$$\pi(\phi_i|\cdot) = G_i^* = \mathrm{Ga}\Big(\psi_1 + Z_i^*, \psi_1 + m_i^*\Big) \tag{C.2}$$

$$\pi(\phi_{is}|\cdot) = G_{is}^* = \mathrm{Ga}\Big(\psi_2 + Z_{is}^*, \psi_2 + m_{is}^*\Big), \tag{C.3}$$

where the notation $a|\cdot$ denotes a given all other variables and $Z_i^* = \sum_{j=1}^{M_i} Z_{ij}$, $Z_{is}^* = \sum_{j=s}^{M_i} Z_{ij}$

$$\begin{aligned} m_i^* &= \nu_0 \textstyle\sum_{j=1}^{M_i} r_{ij}\widehat{\lambda}_{0j}\prod_{h=1}^{j}\phi_{ih}\nu_h\prod_{k=1}^{p}\gamma_{hk}^{x_{ijk}} \\ m_{is}^* &= \phi_i\nu_0 \textstyle\sum_{j=s}^{M_i} r_{ij}\widehat{\lambda}_{0j}(\prod_{h=1}^{j}\nu_h\prod_{k=1}^{p}\gamma_{hk}^{x_{ijk}})\prod_{f\neq s}^{j}\phi_{if}, \end{aligned}$$

for $s = 1, \ldots, M_i$. The full conditional posterior densities of the elements of $\boldsymbol{\nu}$ and $\boldsymbol{\Gamma}$ are

$$\pi(\nu_0|\cdot) = \mathrm{Ga}\left(\kappa + \sum_{i=1}^{N} Z_i^*, \kappa + \sum_{i=1}^{N} \phi_i \sum_{j=1}^{M_i} r_{ij} \widehat{\lambda}_{0j} \prod_{h=1}^{j} \phi_{ih} \nu_h \prod_{k=1}^{p} \gamma_{hk}^{x_{ijk}}\right) \tag{C.4}$$

$$\pi(\nu_s|\cdot) = \mathrm{Ga}\left(\psi_3 + \sum_{i \in R_s} Z_{is}^*, \psi_3 + \nu_0 \sum_{i \in R_s} \phi_i \sum_{j=s}^{M_i} r_{ij} \widehat{\lambda}_{0j} \prod_{f \neq s}^{j} \nu_f \prod_{h=1}^{j} \phi_{ih} \prod_{k=1}^{p} \gamma_{hk}^{x_{ijk}}\right) \tag{C.5}$$

$$\begin{aligned} \pi(\gamma_{sk}|\cdot) \propto\ & \exp\left\{\left(\psi_4 + \sum_{i \in R_s} \sum_{j=s}^{M} Z_{ij} x_{ijk}\right) \log(\gamma_{sk}) - \gamma_{sk}\psi_4 \right. \\ & \left. -\nu_0 \sum_{i \in R_s} \sum_{j=s}^{M} r_{ij} \phi_i \widehat{\lambda}_{0j} \prod_{h=1}^{j} \phi_{ih} \nu_h \prod_{l=1}^{p} \gamma_{hl}^{x_{ijl}}\right\}, \end{aligned} \tag{C.6}$$

where $R_s = \{i : M_i \geq s\}$, for $s = 1, \ldots, M$.

When the Dirichlet process is used to model the frailty terms, the full conditionals of ϕ_i and ϕ_{is} are not G_i^* and G_{is}^*. Using the Pólya urn representation of the Dirichlet process (Blackwell and MacQueen, 1973; MacEachern, 1994; West, 1990), one can show that the prior distribution of ϕ_i given $\boldsymbol{\phi}^{(i)} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_N)'$ is the mixture

$$\left(\frac{\alpha_{01}}{\alpha_{01} + N - 1}\right) G_{01} + \left(\frac{1}{\alpha_{01} + N - 1}\right) \sum_{l=1}^{K^{(i)}} n_l^{(i)} \delta_{\theta_l^{(i)}}, \tag{C.7}$$

where δ_θ denotes the degenerate distribution with all its mass at θ, and the prior for ϕ_{is} given $\boldsymbol{\phi}_s^{(i)} = \{\phi_{i's} : i' \in R_s, i' \neq i\}$ and N_s total subjects in R_s is

$$\left(\frac{\alpha_{02}}{\alpha_{02} + N_s - 1}\right) G_{02} + \left(\frac{1}{\alpha_{02} + N_s - 1}\right) \sum_{l=1}^{K_s^{(i)}} n_{sl}^{(i)} \delta_{\theta_{sl}^{(i)}}, \tag{C.8}$$

where $\boldsymbol{\theta}^{(i)}$ and $\boldsymbol{\theta}_s^{(i)}$ denote the $K^{(i)}$ and $K_s^{(i)}$ unique values of $\boldsymbol{\phi}^{(i)}$ and $\boldsymbol{\phi}_s^{(i)}$, respectively, $n_l^{(i)}$ elements of $\boldsymbol{\phi}^{(i)}$ have value $\theta_l^{(i)}$, and $n_{sl}^{(i)}$ elements of $\boldsymbol{\phi}_s^{(i)}$ have value $\theta_{sl}^{(i)}$. After factoring in the likelihood, the full conditional posterior of ϕ_i is

$$q_{i0} G_i^* + \sum_{l=1}^{K^{(i)}} q_{il} \delta_{\theta_l^{(i)}} \tag{C.9}$$

where

$$q_{il} = \begin{cases} \frac{c_1 \alpha_{01} C(\psi_1, \psi_1)}{C(\psi_1 + Z_i^*, \psi_1 + m_i^*)} & l = 0 \\ \\ c_1 n_l^{(i)} (\theta_l^{(i)})^{Z_i^*} \exp\{-\theta_l^{(i)} m_i^*\} & l > 0, \end{cases}$$

$C(a, b) = b^a/\Gamma(a)$, and c_1 is a normalizing constant. Similarly, the full conditional posterior of ϕ_{is} $(s = 1, \ldots, M_i)$ is

$$q_{is0} G_{is}^* + \sum_{l=1}^{K_s^{(i)}} q_{isl} \delta_{\theta_{sl}^{(i)}} \tag{C.10}$$

where

$$q_{isl} = \begin{cases} \frac{c_{s+1} \alpha_{02} C(\psi_2, \psi_2)}{C(\psi_2 + Z_{is}^*, \psi_2 + m_{is}^*)} & l = 0 \\ \\ c_{s+1} n_{sl}^{(i)} (\theta_{sl}^{(i)})^{Z_{is}^*} \exp\{-\theta_{sl}^{(i)} m_{is}^*\} & l > 0. \end{cases}$$

Thus, the full conditional distributions of ϕ_i and $(\phi_{i1}, \ldots, \phi_{iM_i})'$ are mixtures of the gamma posteriors obtained under the dynamic gamma model and multinomial distributions with support on the unique values of each frailty component.

C.2 Updating Algorithm

In order to sample efficiently under the Dirichlet process mixture, we invoke a sampling scheme similar to that provided by MacEachern (1994) and West et al. (1994). Let there be K unique values in $(\phi_1, \ldots, \phi_N)'$ and K_s unique values in $\{\phi_{is} : i \in R_s\}$, which we denote by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$ and $\boldsymbol{\theta}_s = (\theta_{s1}, \ldots, \theta_{sK_s})'$, respectively, for $s = 1 \ldots, M$. We also define the discrete random variables S_i and S_{is} such that $S_i = k$ if $\phi_i = \theta_k$ and $S_{is} = l$ if $\phi_{is} = \theta_{sl}$, for $i \in R_s$. Our MCMC algorithm proceeds as follows:

1. Sample ν_0 from (C.4), given the current values of $\boldsymbol{\Phi}$, $\nu_1, \ldots, \nu_M$, and $\boldsymbol{\Gamma}$.

2. Sample S_i, for $i = 1, \ldots, N$ from a multinomial distribution with $\Pr(S_i = l) = q_{il}$, for $l = 0, 1, \ldots, K^{(i)}$, with a new ϕ_i drawn from G_i^* if $S_i = 0$.

3. Given the updated values of K and $S_1, \ldots, S_N$, generate a new $\boldsymbol{\theta}$ by sampling each θ_k from its full conditional posterior distribution, $\mathrm{Ga}\big(\psi_1 + \sum_{i:S_i=k} Z_i^*, \psi_1 + \sum_{i:S_i=k} m_i^*\big)$, for $k = 1, \ldots, K$. Assign the appropriate value of $\boldsymbol{\theta}^{(i)}$ to ϕ_i as indicated by S_i.

4. For s $= 1, \ldots, M$, perform the following steps:

   a.) For $i \in R_s$, sample S_{is} from the multinomial distribution with $\Pr(S_{is} = l) = q_{isl}$, for $l = 0, 1, \ldots, K_s^{(i)}$, with a new ϕ_{is} drawn from G_{is}^* if $S_{is} = 0$.

   b.) Update $\boldsymbol{\theta}_s$ by sampling each θ_{sl} from the full conditional posterior, $\text{Ga}\big(\psi_2 + \sum_{i:S_{is}=l} Z_{is}^*, \psi_2 + \sum_{i:S_{is}=l} m_{is}^*\big)$, for $l = 1, \ldots, K_s$. Assign the appropriate value of $\boldsymbol{\theta}_s^{(i)}$ to ϕ_{is}.

   c.) Sample ν_s from (C.5).

   d.) Sample γ_{sk} from (C.6) for $k = 1, \ldots, p$.

5. Update ψ_1 and ψ_2. Let $K^* = \sum_{s=1}^{M} K_s$. Under the DP model, the full conditional posterior densities of ψ_1 and ψ_2 depend only on $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$,

$$\pi(\psi_1|\cdot) \propto \left(\frac{\psi_1^{\psi_1}}{\boldsymbol{\Gamma}(\psi_1)}\right)^K \psi_1^{a_1-1} \exp\left\{-\psi_1\left(b_1 + \sum_{k=1}^{K}(\theta_k - \log\theta_k)\right)\right\} \tag{C.11}$$

$$\pi(\psi_2|\cdot) \propto \left(\frac{\psi_2^{\psi_2}}{\boldsymbol{\Gamma}(\psi_2)}\right)^{K^*} \psi_2^{a_2-1} \exp\left\{-\psi_2\left(b_2 + \sum_{s=1}^{M}\sum_{k=1}^{K_s}(\theta_{sk} - \log\theta_{sk})\right)\right\}, \tag{C.12}$$

   while under the dynamic gamma frailty model, the posteriors depend on each ϕ_i and ϕ_{ij}, respectively. Since (C.11) and (C.12) do not have closed forms, we recommend updating ψ_1 and ψ_2 using a Metropolis-Hastings random walk.

In our analysis of the chemoprevention data (Section 4.4), we modified the algorithm slightly to speed up computation. In steps 2 and 4a, we sampled each S_i and S_{is} using the cluster configuration at the previous iteration, which is similar to the method described in Appendix B.

It is also fairly straightforward to sample from predictive distributions at each iteration of the Gibbs sampler. Given $\boldsymbol{\phi}$ and $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_M$, the frailty of a future subject, $\boldsymbol{\xi}_{N+1}$, may be predicted by sampling from the distributions

$$\pi(\phi_{N+1}|\boldsymbol{\phi}) = \left(\frac{\alpha_{01}}{\alpha_{01}+N}\right)G_{01} + \left(\frac{1}{\alpha_{01}+N}\right)\sum_{l=1}^{K} n_l \delta_{\theta_l} \tag{C.13}$$

$$\pi(\phi_{N+1,s}|\boldsymbol{\phi}_s) = \left(\frac{\alpha_{02}}{\alpha_{02}+N_s}\right)G_{02} + \left(\frac{1}{\alpha_{02}+N_s}\right)\sum_{l=1}^{K_s} n_{sl} \delta_{\theta_{sl}} \tag{C.14}$$

for $s = 1, \ldots, M$. Count data for a future subject may then be simulated using the non-homogeneous Poisson process in equation (4.4).

APPENDIX D

MCMC METHODOLOGY FOR CHAPTER 5

Let $S_{hi} = (0, l)$ if $\boldsymbol{\phi}_{hi}$ is allocated to a new cluster in mixture component l and $S_{hi} = (r, l)$ if $\boldsymbol{\phi}_{hi} = \boldsymbol{\theta}_{rl}^{(hi)}$. At each iteration of our MCMC, we update the parameters in our model using the following steps:

1. Sample each S_{hi} from its full conditional posterior distribution which is multinomial with $\Pr(S_{hi} = (r, l)|\cdot) = \widetilde{\omega}_{hlr}$, for $l = 1, \ldots, h$, $r = 0, 1, \ldots, K_h$. Whenever $S_{hi} = (0, l)$ sample a new value for $\boldsymbol{\phi}_{hi}$ from G_{0hi} and assign subject h, i to their own cluster in component l.

2. Given the updated values of $\mathbf{S} = \{S_{hi}, h = 1, \ldots, d; i = 1, \ldots, n_h\}$, for $l = 1, \ldots, d$ and $r = 1, \ldots, K_l$ sample $\boldsymbol{\theta}_{lr}$ from

$$\mathrm{W}\left(\boldsymbol{\Sigma}_{lr}^{*-1}; v_0 + m_{lr}, v_0\mathbf{V}_0 + \mathbf{V}_{0lr}^*\right) \cdot \mathrm{N}\left(\boldsymbol{\mu}_{lr}^*; \kappa^*(\kappa^{-1}\boldsymbol{\mu}_0 + \sum_{h=l}^{d} \sum_{i:S_{hi}=(l,r)} \mathbf{y}_{hi}), \kappa^*\boldsymbol{\Sigma}_{lr}^*\right), \tag{D.1}$$

   where

$$\mathbf{V}_{0lr}^* = v_0\mathbf{V}_0 + \sum_{h=l}^{d} \sum_{i:S_{hi}=(l,r)} (\mathbf{y}_{hi} - \boldsymbol{\mu}_0)(\mathbf{y}_{hi} - \boldsymbol{\mu}_0)'$$

   and $\kappa^* = \kappa/(1 + \kappa \cdot m_{lr})$.

3. For $h = 1, \ldots, d-1$ sample: a.) ζ_h from (5.19) and b.) π_h from (5.20).

4. Sample p_0 from (5.21).

5. For $h = 0, 1, \ldots, d-1$ sample a latent variable, u_h, from $\pi(u_h|\alpha_h, m_{h+1}) = \mathrm{Be}(\alpha_h +$

$1, m_{h+1})$ and then sample α_h from its full conditional posterior

$$\begin{aligned}\pi(\alpha_h|u_h, K_{h+1}, m_{h+1}) &= p_{uh}\mathrm{Ga}(\alpha_h; a_{\gamma_h} + K_{h+1}, b_{\gamma_h} - log(u_h)) \\ &\quad +(1-p_{uh})\mathrm{Ga}(\alpha_h; a_{\gamma_h} + K_{h+1} - 1, b_{\gamma_h} - log(u_h)),\end{aligned} \tag{D.2}$$

where

$$\frac{p_{uh}}{1-p_{uh}} = \frac{a_{\gamma_h} + K_{h+1} - 1}{m_{h+1}(b_{\gamma_h} - log(u_h))}.$$

6. Sample $\boldsymbol{\mu}_0$ from $\pi(\boldsymbol{\mu}_0|\cdot) = \mathrm{N}(\boldsymbol{\mu}_0; \boldsymbol{\beta}^*, \boldsymbol{\Omega}^*)$ where

$$\boldsymbol{\beta}^* = \boldsymbol{\Omega}^*\left(\boldsymbol{\Omega}^{-1}\boldsymbol{\beta} + \frac{1}{\kappa}\sum_{h=1}^{d}\sum_{r=1}^{K_h}\boldsymbol{\Sigma}_{hr}^{*-1}\boldsymbol{\mu}_{hr}^*\right)$$

$$\boldsymbol{\Omega}^* = \left\{\boldsymbol{\Omega}^{-1} + \frac{1}{\kappa}\sum_{h=1}^{d}\sum_{r=1}^{K_h}\boldsymbol{\Sigma}_{hr}^{*-1}\right\}^{-1}.$$

7. Sample κ^{-1} from

$$\pi(\kappa^{-1}|\cdot) = \mathrm{Ga}\left(\kappa^{-1}; a_\kappa + \frac{1}{2}\sum_{h=1}^{d}K_h + 1, b_\kappa + \frac{1}{2}\sum_{h=1}^{d}\sum_{r=1}^{K_h}(\boldsymbol{\mu}_{hr}^* - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_{hr}^{*-1}(\boldsymbol{\mu}_{hr}^* - \boldsymbol{\mu}_0)\right). \tag{D.3}$$

REFERENCES

Ahmad, R. (1976). Multivariate k-sample problem and generalization of the Kolmogorov-Smirnov test. *Annals of Statistical Mathematics* **28**, 259–265.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

Arjas, E. and Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica* **4**, 505–524.

Aslanidou, H., Dey, D. and Sinha, D. (1998). Bayesian analysis of multivariate survival data using Monte Carlo methods. *Statistica Sinica* **26**, 33–48.

Balakrishnan, S. and Madigan, D. (2006). A one-pass sequential Monte Carlo method for Bayesian analysis of massive data sets. *Bayesian Analysis* **1**, 345–362.

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

Berger, J. and Delampady (1987). Testing precise hypotheses. *Statistical Science* **2**, 317–352.

Berger, J. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association* **96**, 174–184.

Bigelow, J. and Dunson, D. (2005). Semiparametric classification in hierarchical functional analysis. ISDS Discussion Paper 2005-18, Duke University.

Blackwell, D. (1973). The discreteness of ferguson selections. *Annals of Statistics* **1**, 356–358.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics* **1**, 353–355.

Blei, D. and Jordan, M. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis* **1**, 1–23.

Broman, S. (1984). The Collaborative Perinatal Project: an overview. In Medrick, S., Harway, M. and Finello, K., editors, *Handbook of Longitudinal Research*, pages 185–215, New York. Praeger.

Bush, C. and MacEachern, S. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika* **83**, 175–185.

Cao, G. and West, M. (1996). Practical Bayesian inference using mixtures of mixtures. *Biometrics* **52**, 1334–1341.

Carota, C. and Parmigiani, G. (1996). On Bayes factors for nonparametric alternatives. In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., editors, *Bayesian Statistics 5*, pages 507–511, London. Oxford University Press.

Carota, C. and Parmigiani, G. (2002). Semiparametric regression for count data. *Biometrika* **89**, 265–281.

Chen, A., Pennell, M., Klebanoff, M., Rogan, W. and Longnecker, M. (2006). Maternal smoking during pregnancy in relation to child overweight: follow-up to age 8 years. *International Journal of Epidemiology* **35**, 121–130.

Chen, M.H., I. J. and Sinha, D. (2002). Bayesian inference for multivariate survival data with a cure fraction. *Journal of Multivariate Analysis* **80**, 101–126.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika* **89**, 539–551.

Cifarelli, D. and Regazzini, E. (1978). Non parametrici in condizioni di scambiabilit a parziale e impiego di medie associative. Technical report, Quaderni Istituto Matematica Finanziaria, Turin, Italy.

Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A* **148**, 82–117.

Cooke, R. and Goossens, L. (2000). Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection Dosimetry* **90**, 303–309.

Cox, D. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

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

Dietz, E. (1989). Multivariate generalizations of Jonckheere's test for ordered alternatives. *Communications in Statistics, Part A* **18**, 3763–3783.

Dietz, E. and Killeen, T. (1981). A nonparametric multivariate test for monotone trend with pharmaceutical applications. *Journal of the American Statistical Association* **76**, 169–174.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability* **2**, 183–201.

Duan, J., Guindani, M. and Gelfand, A. (2005). Generalized spatial Dirichlet process

models. ISDS Discussion Paper 2005-23, Duke University.

Duez, P., Dehon, G., Kumps, A. and Dubois, J. (2003). Statistics of the comet assay: a key to discriminate between genotoxic effects. *Journal of the American Statistical Association* **76**, 169–174.

DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C. and Pregibon (1999). Squashing flat files flatter. In *Proceedings of the Fifth ACM Conference on Knowledge Discovery and Data Mining*, pages 6–15.

Dunson, D. (2000). Modeling of changes in tumor burden. *Journal of Agricultural Biological and Environmental Statistics* **6**, 38–48.

Dunson, D. (2004). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association* **100**, 618–627.

Dunson, D. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics* to appear.

Dunson, D. and Chen, Z. (2004). Selecting factors predictive of heterogeneity in multivariate event time data. *Biometrics* **60**, 352–358.

Dunson, D. and Dinse, G. (2000). Distinguishing effects on tumor multiplicity and growth rate in chemoprevention experiments. *Biometrics* **56**, 1068–1075.

Dunson, D. and Pillai, D. (2004). Bayesian density regression. ISDS Discussion Paper 04-33, Duke University.

Dunson, D. and Taylor, J. (2005). Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics* **17**, 385–400.

Dunson, D., Watson, M. and Taylor, J. (2003). Bayesian latent variable models for median regression on multiple outcomes. *Biometrics* **59**, 296–304.

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

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 578–588.

Ferguson, T. (1973). A Bayesian analysis of some non-parametric problems. *Annals of Statistics* **1**, 209–230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.

Ferguson, T. and Phadia, E. (1979). Bayesian nonparametric estimation based on censored data. *Annals of Statistics* **7**, 163–186.

Forbes, P. and Sambuco, C. (1998). Assays for photocarcinogenesis: relevance of animal models. *International Journal of Toxicology* **17**, 577–588.

Gail, M., Santner, T. and Brown, C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics* **36**, 255–266.

Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Applied Statistics* **40**, 63–79.

Garthwaite, P., Kadane, J. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Applied Statistics* **100**, 680–700.

Gelfand, A. and Kottas, A. (2001). Nonparametric Bayesian modeling for stochastic order. *Annals of the Institute of Statistical Mathematics* **53**, 865–876.

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

Gelfand, A., Kottas, A. and MacEachern, S. (2004). Bayesian nonparametric spatial modeling with Dirichlet process mixing. Technical Report AMS 2004-5, Department of Applied Math and Statistics, University of California, Santa Cruz.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Society* **85**, 398–409.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

Giudici, P., Mezzetti, M. and Muliere, P. (2003). Mixtures of Dirichlet process priors for variable selection in survival analysis. *Journal of Statistical Planning and Inference* **111**, 101–115.

Gopalan, R. and Berry, D. (1998). Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association* **93**, 1130–1139.

Griffin, J. and Steel, M. (2006). Order-based dependent Dirichlet process. *Journal of the American Statistical Association* .

Grubbs, C., Eto, I., Juliana, M. and Whitaker, L. (1991). Effect of canthaxanthin on chemically induced mammary carcinogenesis. *Oncology* **48**, 239–245.

Gustafson, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics* **53**, 230–242.

Gustafson, P., Aeschliman, D. and Levy, A. (2003). A simple approach to fitting Bayesian survival models. *Lifetime Data Analysis* **9**, 5–19.

Hans, C. and Dunson, D. (2005). Bayesian inferences on umbrella orderings. *Biometrics* **61**, 1018–1026.

Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of Pólya trees. *Journal of the American Statistical Society* **97**, 1020–1033.

Härkänen, T., Hausen, H., Virtanen, J. and Arjas, E. (2003). A non-parametric frailty model for temporally clustered multivariate failure times. *Scandanavian Journal of Statistics* **30**, 523–533.

Harlow, S., Lin, X. and Ho, M. (2000). Analysis of menstrual diary data across the reproductive life span: applicability of the bipartite model approach and the importance of within-woman variance. *Journal of Clinical Epidemiology* **53**, 722–733.

Hartigan, J. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., New York.

Henderson, R. and Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika* **90**, 355–366.

Hjort, N. (1990). Nonparametric Bayes estimators based on beta processes in models of life history data. *Annals of Statistics* **18**, 1259–1294.

Hoff, P. (2003). Bayesian methods for partial stochastic orderings. *Biometrika* **90**, 303–317.

Huang, P., Tilley, B., Woolson, R. and Lipsitz, S. (2005). Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens-Fisher problem. *Biometrics* **61**, 532–539.

Ibrahim, J., Chen, M. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer Verlag, New York.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Ishwaran, H. and James, L. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* **11**, 508–532.

Ishwaran, H. and Rao, J. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* **98**, 438–455.

Jain, S. and Neal, R. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics* **13**, 158–182.

Jonckheere, A. (1954). A distribution free k-sample test against ordered alternatives. *Biometrika* **41**, 133–145.

Kadane, J. and Wolfson, L. (1998). Experiences in elicitation. *The Statistician* **47**, 3–19.

Kalbfleisch, J. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* **40**, 214–221.

Kiefer, J. (1959). k-sample analogues of the Kolmogorov-Smirnov and Cramér-v. Mises tests. *Annals of Mathematical Statistics* **30**, 420–447.

Kleinman, K. and Ibrahim, J. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics* **54**, 921–938.

Kokoska, S., Hardin, J., Grubbs, C. and Hsu, C. (1993). The statistical analysis of cancer inhibition/promotion experiments. *Anticancer Research* **13**, 1357–1364.

Kottas, A. and Gelfand, A. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association* **96**, 1458–1468.

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

Lam, K., Lee, Y. and Leung, T. (2002). Modeling multivariate survival data by a semiparametric random effects proportional odds model. *Biometrics* **58**, 316–323.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Annals of Statistics* **20**, 1161–1176.

Lavine, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *Annals of Statistics* **22**, 1222–1235.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Liu, J. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Annals of Statistics* **24**, 911–930.

Liu, L., Wolfe, R. and Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* **60**, 747–756.

Lovell, D., Thomas, G. and Dubow, R. (1999). Issues related to the experimental design and subsequent statistical analysis of in vivo and in vitro comet studies. *Tetratogenesis, Carcinogenesis, and Mutagenesis* **19**, 109–119.

MacAuliffe, J., Blei, D. and Jordan, M. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing* to appear.

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

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA. American Statistical Association.

MacEachern, S. (2000). Dependent Dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University.

MacEachern, S., Clyde, M. and Liu, J. (1999). Sequential importance sampling for nonparametric Bayes models: the next generation. *Canadian Journal of Statistics* **27**, 251–267.

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

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Madigan, D., Raghavan, N., DuMouchel, W., Nason, M., Posse, C. and Ridgeway, G. (2002). Likelihood-based data squashing: a modeling approach to instance construction. *Data Mining and Knowledge Discovery* **6**, 173–190.

Mauldin, R., Sudderth, W. and Williams, S. (1992). Pólya trees and random distributions. *Annals of Statistics* **20**, 1203–1221.

Meyer, M. and Booker, J. (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide*. ASA/Society of Industrial and Applied Mathematics, Philadelphia.

Mira, A. and Petrone, S. (1996). Bayesian hierarchical nonparametric inference for change-point problems. In Berger, J. O., Bernardo, J. M., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 5*, pages 609–620. Oxford University Press.

Mukhopadhyay, S. and Gelfand, A. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* **92**, 633–639.

Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet process priors. *Canadian Journal of Statistics* **26**, 283–297.

Muliere, P. and Walker, S. (1997). A Bayesian non-parametric approach to survival analysis using Pólya trees. *Scandinavian Journal of Statistics* **24**, 331–340.

Müller, P. and Quintana, F. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 95–110.

Müller, P., Quintana, F. and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society, Series B* **66**, 735–749.

Müller, P. and Rosner, G. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association* **92**, 1279–1292.

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

Newton, M. and Zhang, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika* **86**, 15–26.

Nickerson, R. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods* **5**, 241–301.

Nieto-Barajas, L. and Walker, S. (2002). Markov beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics* **29**, 413–424.

O'Brien, P. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

Owen, A. (2003). Data squashing by empirical likelihood. *Data Mining and Knowledge Discovery* **7**, 101–113.

Paik, M., Tsai, W. and Ottman, R. (1994). Multivariate survival analysis using piecewise gamma frailty. *Biometrics* **50**, 975–988.

Quintana, F. and Newton, M. (2000). Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences. *Journal of Computational and Grapical Statistics* **9**, 711–737.

Ridgeway, G. and Madigan, D. (2003). A sequential Monte Carlo method for Bayesian analysis of massive datasets. *Data Mining and Knowledge Discovery* **7**, 301–319.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Sahu, S., Dey, D., Aslanidou, H. and Sinha, D. (1997). A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis* **3**, 123–137.

Sargent, D. (1997). A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Analysis* **3**, 13–25.

Sargent, D. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics* **54**, 1486–1497.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Sinha, D. (1998). Posterior likelihood methods for multivariate survival data. *Biometrics* **54**, 1463–1474.

Sinha, D., Ibrahim, J. and Chen, M. (2002). Models for survival data from cancer prevention studies. *Journal of the Royal Statistical Society, Series B* **64**, 467–477.

Sinha, D. and Maiti, T. (2004). A Bayesian approach for the analysis of panel-count data with dependent termination. *Biometrics* **60**, 34–40.

Sneath, P. (1957). The application of computers to taxonomy. *Journal of General Microbiology* **17**, 201–226.

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

Teh, Y., Jordan, M., Beal, M. and Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* to appear.

Terpstra, T. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indigationes Mathematicae* **14**, 327–333.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussions). *Annals of Statistics* **22**, 1701–1762.

Tomlinson, G. and Escobar, M. (1999). Analysis of densities. Technical report, University of Toronto.

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

Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness of fit testing using infinite dimensional exponential families. *Annals of Statistics* **20**, 1215–1241.

Walker, S., Damien, P., Laud, P. and Smith, A. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B* **61**, 485–528.

Walker, S. and Mallick, B. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society, Series B* **59**, 845–860.

West, M. (1990). Bayesian kernel density estimation. ISDS Discussion Paper 90-A02, Duke University.

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

West, M., Müller, P. and Escobar, M. (1994). Hierarchical priors and mixture models with application in regression and density estimation. In Smith, A. and Freeman, P., editors, *Aspects of Uncertainty: A Tribute to D.V. Lindley*, pages 363–386, New York. Wiley.

West, M. and Turner, D. (1994). Deconvolution of mixtures in analysis of neural synaptic transmission. *The Statistician* **43**, 31–43.

Westfall, P., Johnson, W. and Utts, J. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* **84**, 419–427.

Williams, D. (1986). A note on Shirley's nonparametric test for comparing several dose levels with a zero-dose control. *Biometrics* **42**, 183–186.

Wolfinger, R., Tobias, R. and Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for generalized linear mixed models. *SIAM Journal on Scientific Computing* **15**, 1294–131.

Yau, K. and McGilchrist, C. (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine* **17**, 1201–1213.

Yue, H. and Chan, K. (1997). A dynamic frailty model for multivariate survival data. *Biometrics* **53**, 785–793.

Zeger, S. and Karim, M. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86.