Nonparametric Bayes Applications to Biostatistics

David B. Dunson

Biostatistics Branch, National Institute of Environmental Health Sciences, U.S. National Institutes of Health
dunson@stat.duke.edu

1. Introduction

Biomedical research has clearly evolved at a dramatic rate in the past decade, with improvements in technology leading to a fundamental shift in the way in which data are collected and analyzed. Before this paradigm shift, studies were most commonly designed to be simple and to focus on relationships among a few variables of primary interest. For example, in a clinical trial, patients may be randomized to receive either the drug or placebo, with the analysis focusing on a comparison of means between the two groups. However, with emerging biotechnology tools, scientists are increasingly interested in studying how patients vary in their response to drug therapies, and what factors predict this variability. Such questions are of fundamental interest in personalizing medicine, so that the physician prescribes the most appropriate therapy given the patient's history, genetics and lifestyle factors. Given this focus, it has become routine to collect large amounts of information for each study subject, with the statistician's challenge then being to perform inferences and develop predictive models based on the massive amount of data available. Clinical trials and personalized medicine are just one example of a growing trend in biomedicine towards embracing emerging technologies for collection, storage and analysis of massive amounts of data.

To address big problems of this type, it is crucial for statisticians to have an appropriate toolbox at their disposal. Certainly, classical statistical methods were developed with simpler data structures and problems in mind. Hence, it has become necessary to consider new statistical paradigms that perform well in characterizing complex data from a broad variety of study designs. In complex settings, it is seldom if ever the case that one has a defensible parametric model at their disposal, and it can be very challenging to check modeling assumptions in high-dimensions. Hence, non- or semiparametric models seem required. However, classical nonparametric methods often perform poorly in complex settings due to the curse of dimensionality and to difficulties in accommodating complicating features of the data, such as censoring and missing data. Nonparametric Bayes methods provide a widely useful paradigm that gains some key advantages of a fully model-based probabilistic framework, while being highly flexible and adaptable. In addition, a key to the success of nonparametric Bayes methods in applications is the incorporation of a sparseness-favoring structure, which combats the curse of dimensionality. This is accomplished automatically through the Bayesian penalty for model complexity (Jeffreys and Berger, 1992) and is aided through centering on a base parametric model.

The goal of this chapter is to provide a brief review and motivation for the use of nonparametric Bayes methods in biostatistical applications. Clearly, the nonparametric Bayes biostatistical literature is increasingly vast, and it is not possible to properly present or even mention most of the approaches that have been proposed. Instead, the focus here is entirely on methods utilizing random probability measures, with the emphasis on a few approaches

that seem particularly useful in addressing the considerable challenges faced in modern biostatistical research. In addition, the emphasis will be entirely on practical applications-motivated considerations, with the goal of moving the reader towards implementing related approaches for their own data. Readers interested in the theoretical motivation, which is certainly a fascinating area in itself, are referred to the cited papers and to the later chapters in the book.

Section 2 describes the use of Dirichlet process (DP) priors in formulating semi-parametric Bayes hierarchical models. Section 3 considers methods for functional data analysis using DP-based methods, while also considering extensions for joint modeling with functional predictors. Section 4 describes approaches for local shrinkage and clustering. Section 5 considers methods for hierarchical borrowing of information across studies, centers or exchangeable groups of data. Section 6 overviews the recent work on flexible modeling of conditional distributions using priors for collections of random probability measures that evolve with predictors, time and spatial location. Section 7 highlights some recent applications in bioinformatics. Section 8 outlines methods for nonparametric Bayes hypothesis testing, and Section 9 contains a brief discussion.

2. Hierarchical Modeling with Dirichlet Process Priors

2.1 *Illustration for simple repeated measurement models*

Hierarchical modeling has become the standard tool for accommodating dependence in longitudinal and nested data structures and for combining information from different studies or data sources. One of the simplest hierarchical models has the form:

$$\begin{aligned} y_{ij} &= \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \\ \mu_i &\sim P, \end{aligned} \tag{1}$$

where y_{ij} is the jth observation within subject (or blocking factor) i, μ_i is a subject-specific mean, ϵ_{ij} is an observation-specific residual, σ^2 is the within-subject variance, and P is the distribution of the subject-specific means, with $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$. The typical parametric specification of (1) lets $\mu_i = \mu + b_i$, with μ an overall mean, b_i a deviation or random effect for subject i, and $b_i \sim N(0, \psi)$ to characterize heterogeneity among subjects. In this case, P is chosen to correspond to the $N(\mu, \psi)$ distribution.

Although (1) is very appealing in allowing random variability among subjects while borrowing information, one may question the appropriateness of the normality assumption on P. It is well known that borrowing of information across subjects is quite sensitive to departures from this assumption. In particular, the normal distribution has light tails and does not allow some subjects to be very different from other subjects or to have groups of subjects that cluster close together. Hence, outlying subjects tend to have their means over-shrunk towards the population mean, and the data from such subjects may be overly-influential in estimation of μ.

Although one could potentially choose a heavier-tailed alternative to the normal random effects distribution, such as a t distribution, it is appealing to instead use a more flexible form that allows for the possibility of skewness and multimodality. Even a heavy-tailed distribution, such as the t, has a very restrictive unimodal and symmetric shape, and in most applications there is no reason *a priori* to believe that such a shape is required. Refer to Lee and Thompson (2007) for a recent Bayesian approach for flexible parametric modeling

of random effects distributions. Parametric models, such as extensions of the t distribution that allow skewness, are still restrictive, and do not allow multi-modality, which may arise due to latent sub-populations.

Bayesian nonparametric models incorporate infinitely-many parameters in order to more flexibly represent uncertainty in P. Hence, the term "nonparametric" is something of a misnomer in that Bayesian nonparametric models are massively parametric. However, by including infinitely-many parameters within a prior that is centered on a base parametric model, one can allow a great deal of flexibility, while regularizing through favoring shrinkage towards a simpler parametric form. For example, in the absence of other knowledge, it may be reasonable to guess that the random effects distribution resembles a Gaussian, while allowing substantial uncertainty in this guess. From a Bayesian perspective, in the absence of parametric knowledge of P, one should choose a prior for P with support on the set of distributions on the real line, with this prior effectively corresponding to a distribution over distributions.

Bush and MacEachern (1996) proposed to address this problem by choosing a Dirichlet process (DP) prior for P (Ferguson, 1973, 1974). Readers are referred to the chapter by Ghosal for detailed background on properties of the DP. In order to allow P to be an unknown distribution on the real line, let $P \sim DP(\alpha P_0)$, with $\alpha > 0$ a concentration parameter characterizing prior precision and clustering and P_0 a base distribution on $\Re$. More formally, P would correspond to a random probability measure, and P_0 to a fixed baseline probability measure. However, to make the presentation more generally accessible, I follow the common convention of using P to refer to both the random probability measure and the corresponding distribution function.

By choosing a DP prior for P, one allows P to be an unknown distribution, with P_0 corresponding to one's best guess for P *a priori* and α expressing confidence in this guess. In particular, P_0 is often chosen to correspond to the normal distribution, $N(\mu_0, \psi_0)$, often with a normal-inverse gamma hyperprior then chosen for (μ_0, ψ_0) to allow the base distribution to have unknown mean and variance. In addition, α is commonly assigned a gamma hyperprior to allow the data to inform more strongly about clustering in the data and the extent to which P is similar to P_0, with gamma(1,1) providing a commonly used choice.

An applied statistician may like to know what choosing a DP(αP_0) prior for P implies about their prior beliefs regarding P. For this reason, the constructive stick-breaking representation of Sethuraman (1994) is extremely helpful. The stick-breaking representation implies that $P \sim DP(\alpha P_0)$ is equivalent to letting

$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \overset{iid}{\sim} P_0, \tag{2}$$

where $\pi_h = V_h \prod_{l<h}(1 - V_l)$ is a probability weight that is formulated from a stick-breaking process, with $V_h \overset{iid}{\sim} \text{beta}(1, \alpha)$, for $h = 1, \ldots, \infty$, and δ_θ is a point mass at θ. Note that the "stick-breaking" terminology arises, because starting with a unit probability stick, V_1 is the proportion of the stick broken off and assigned to θ_1, V_2 is the proportion of the remaining $1 - V_1$ length stick allocated to θ_2, and so on.

Using the stick-breaking formulation, Figure 1 plots realizations from the DP(αP_0) prior for P for a range of different values of α. For values of α close to zero, $V_1 \approx 1$ and essentially

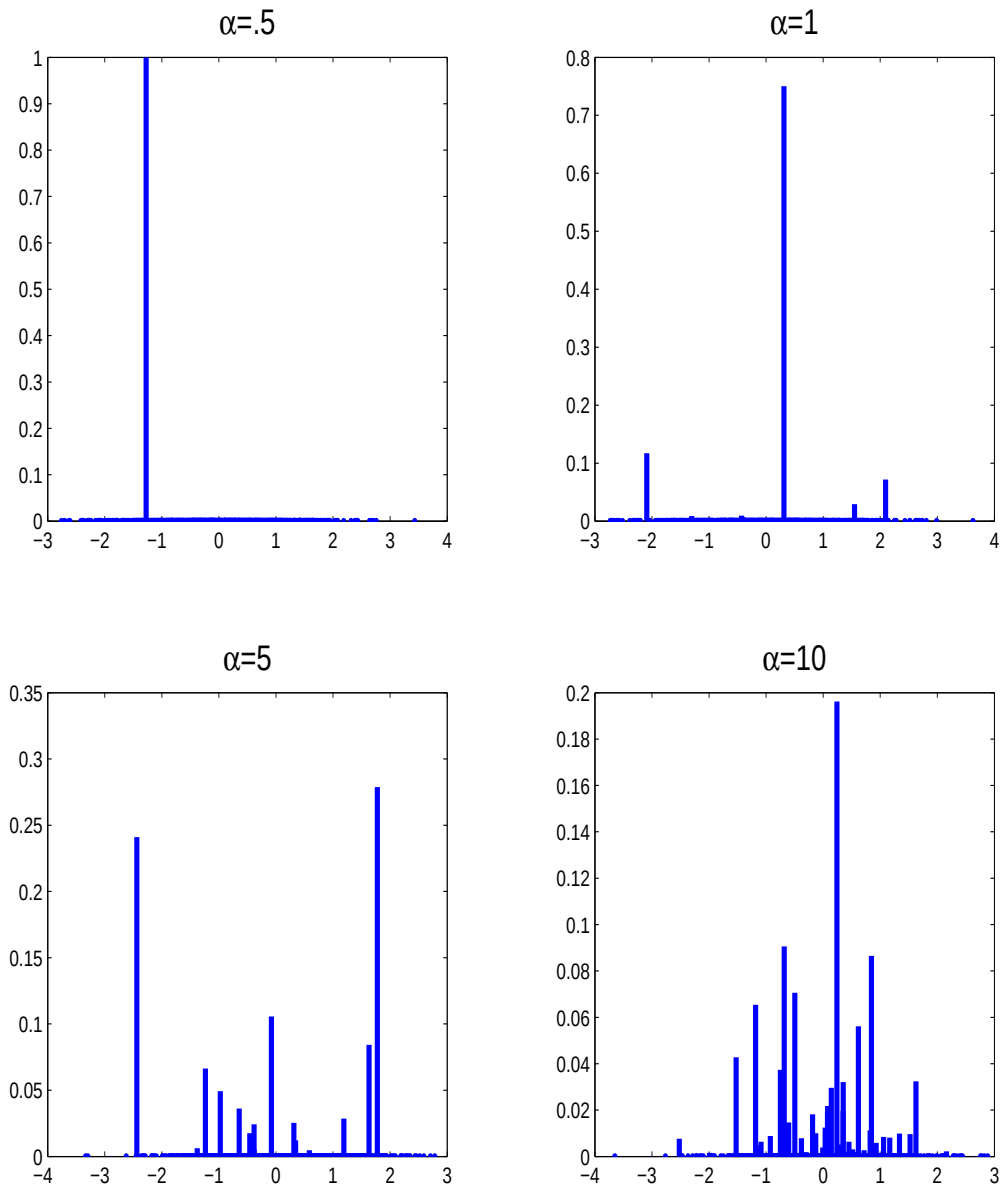

Figure 1. Realizations from the DP(αP_0) prior for P_0 corresponding to the standard normal and for a range of values of α.

all the probability weight will be assigned to a single atom. For small values of α, such as $\alpha = 1$, most of the probability is allocated to the first few atoms, while for large α, each of the atoms is assigned vanishingly-small weight, so that P resembles P_0. Because the probability weights assigned to the atoms decrease stochastically as the index h grows, we were able to accurately represent realizations from P in Figure 1 with only the first 1,000 atoms. In fact, even for $\alpha = 10$, approximately 99% of the probability was allocated to the first 50 atoms,

$\{\theta_1, \ldots, \theta_{50}\}$. As a default choice that is widely-used in applications and favors a sparse representation having a few dominating atoms, it is quite common to let $\alpha = 1$. For more details on the stick-breaking representation of the DP and broader classes of stick-breaking priors, refer to the later chapters in the book.

Returning to the discussion of expression (1), an important implication of (2) is the discrete nature of P. This creates ties among μ_i, $i = 1, \ldots, n$, with the configuration of ties defining clusters, with subjects in a cluster having the same random effect value. Letting $S_i = j$ denote that subject i belongs to cluster j, we have $\mu_{S_i} = \theta^*_{S_i}$, for $i = 1, \ldots, n$, where θ^*_j denotes the value of the random effect for all subjects in cluster j. Note that the $*$ superscript is included to distinguish θ^*_h, the hth of the clusters represented in the sample of n subjects, from θ_h, the hth of the infinitely-many atoms in the stick-breaking representation. Choosing a DP prior for the random effects distribution P has two important consequences: (1) allowing the random effects distribution to be unknown, avoiding a parametric specification; (2) clustering the n individuals in a sample into $k \leq n$ clusters defined by the subjects random effects values. For a comparison of the DP approach with Bayesian models of finite mixture models with an unknown number of components, refer to Richardson and Green (2001).

The clustering property of the DP has been widely exploited in recent years, since it has some appealing practical properties relative to alternative clustering procedures. In particular, it avoids assuming that all individuals can be clustered into a fixed number of groups, k. Instead, as is again clear from the stick-breaking form in (2), the DP assumes that there are infinitely many clusters represented in the overall population, with an unknown number observed in a finite sample of n subjects. When an $(n+1)$st subject is added, there is a positive probability, $\alpha/(\alpha + n)$, that the subject is assigned to a new cluster not yet represented in the sample (Blackwell and MacQueen, 1973).

In clustering there must be some implicit or explicit penalty for model complexity to avoid assigning everyone in the sample to their own cluster to obtain a higher likelihood. Hence, it is important not to view the DP model as a magic clustering approach, which avoids assuming a fixed number of clusters and specification of an arbitrary penalty. Instead, one must carefully consider how the penalty for model complexity or overfitting arises in the DP implementation, while also assessing the role of the hyperparameters, α and parameters characterizing P_0. Under prior (2), the prior expectation for the number of clusters is proportional to $\alpha \log n$, so that the number of clusters tends to increase slowly with the sample size at a rate determined by α. However, in addition to α and n, there are other more subtle factors that play a large role in determining the posterior distribution on the number of clusters and the penality for over-fitting.

To build an intuition, it is useful to consider two extreme clustering cases: a partition with all singletons versus a single cluster. In the first case, each subject is allocated to their own cluster, so that $S_i = i$, for $i = 1, \ldots, n$. This clearly implies that $\mu_i \sim P_0$, for $i = 1, \ldots, n$, which is equivalent to assuming that the random effects distribution is P_0. Interestingly, in this case in which all the subjects are assigned to their own clusters and we may be very concerned about over-fitting, we are effectively fitting a parametric base model (e.g., a normal hierarchical model). Hence, by assuming that each of the cluster-specific parameters is drawn from the common distribution P_0, we have avoided the over-fitting problem characteristic of assigning subjects to many small clusters. At the other extreme,

consider $S_i = 1$, for $i = 1, \ldots, n$, so that all the subjects are assigned to the same cluster and we have $\mu_i = \mu$, for $i = 1, \ldots, n$. In this case, we are effectively fitting a normal model that assumes no heterogeneity among subjects (i.e., the random effects distribution has zero variance). If $P \not\approx P_0$ and $P \neq \delta_\mu$, then to appropriately characterize P, we must allocate individuals to few clusters.

The question is then what determines the posterior distribution of k, the number of clusters. As mentioned previously, the DP induces a prior on k, which is stochastically increasing in n and α, so that α plays a key role. However, it tends to be the case that the data inform strongly about α, so that one can obtain robust results by simply choosing a hyperprior for α. A bigger problem, which is not as widely known, is sensitivity of the posterior distribution of k to the choice of P_0. The importance of this choice was noted in detail by Steve MacEachern in his lecture at the 2007 Nonparametric Bayes Workshop at the Isaac Newton Institute.

Assuming for sake of discussion that P_0 corresponds to a $N(\mu_0, \sigma_0^2)$ distribution, a poor choice of μ_0, σ_0^2 will tend to lead to fewer clusters being chosen. This is clear in examining the conditional posterior probability that subject i is allocated to a new cluster not occupied by subjects $(-i) = \{1, \ldots, n\} \setminus i$,

$$c\left(\frac{\alpha}{\alpha + n - 1}\right) \int \prod_{j=1}^{n_i} N(y_{ij}; \mu_i, \sigma^2) dP_0(\mu_i), \tag{3}$$

where c is a normalizing constant, while the probability of allocating subject i to the hth of the $k^{(-i)}$ existing clusters is

$$c\left(\frac{n_h^{(-i)}}{\alpha + n - 1}\right) \prod_{j=1}^{n_i} N(y_{ij}; \theta_h^{*(-i)}, \sigma^2), \tag{4}$$

where $n_h^{(-i)}$ is the number of subjects $l \in (-i)$ having $\mu_l = \theta_h^{*(-i)}$, for $h = 1, \ldots, k^{(-i)}$, with $k^{(-i)}$ the number of unique values of $\{\mu_l\}_{l \in (-i)}$. Note that (3) is proportional to the marginal likelihood of the data for subject i integrating over P_0.

Clearly, the probability of allocating subject i to a new cluster is strongly dependent on P_0. For example, suppose that subject i's data provide compelling evidence that subject i is an outlier, with the observations systematically higher than those for the other subjects. However, if P_0 is chosen so that $\mu_0 < \theta_h^{*(-i)}$, for all $h = 1, \ldots, k^{(-i)}$, and $\sigma_0^2 < \sigma^2$, then the probability of allocating subject i to a new cluster may be quite small. Choosing σ_0^2 to be high, so that P_0 is supposedly "non-informative" does not solve this problem, because then the probability of allocating subject i to a new cluster is critically dependent on exactly how large σ_0^2 is. If σ_0^2 is extremely large, then the tendency is to allocate all the subjects to the same cluster, since in that case, we assign very low probability to the introduction of new clusters. This behavior is not surprising, as clustering is a type of model selection, and it is well known that a very high variance prior for coefficients that are not shared across models tends to favor smaller models.

Given these issues, it is clearly important to choose P_0 with careful thought. A partial solution is to specify a hyperprior for parameters, μ_0, σ_0^2, characterizing P_0. However, one then faces the same issues discussed above in choosing the hyperparameters in this hyperprior. For example, if one chooses a normal inverse-gamma hyperprior for (μ_0, σ_0^2), then one

could view the hyperparameters in the normal inverse-gamma as fixed parameters characterizing a heavier-tailed P_0. Hence, the clustering will be still be sensitive to hyperparameter choice. There are two reasonable solutions for addressing this issue. The first is to think carefully about plausible values for μ_0, σ_0^2 in the application being considered and to choose an informative prior. This is an appealing choice, and it tends to be the case that one is locally robust to the chosen informative prior. The second option, which has been widely used, is to standardize the data prior to analysis, and then use a prior with location zero and scale one. For example, one can normalize the y_{ij}'s by subtracting the overall mean and dividing by the overall variance. Although this is unappealing in lacking a fully Bayesian justification, it can be viewed as an empirical Bayes approach that tends to have good practical performance. For an article on nonparametric empirical Bayes estimation of the base measure, P_0, refer to McAuliffe, Blei and Jordan (2006).

2.2 *Posterior Computation*

After specifying prior distributions, it is necessary to consider how to update these prior distributions with information in the data to obtain posterior distributions. In this Section, we consider posterior computation for hierarchical model (1) in the case in which $P \sim DP(\alpha P_0)$, with $P_0 = N(\mu_0, \sigma_0^2)$ and (μ_0, σ_0^2) assigned a normal inverse-gamma hyperprior. Even in this simple hierarchical model, the posterior distribution is not analytically tractable and we must rely on approximations. Markov chain Monte Carlo (MCMC) algorithms are the standard approach, though a wide variety of alternatives have been proposed, including partial predictive recursion (Newton and Zhang, 1999), sequential importance sampling (MacEachern, Clyde and Liu, 1999), weighted Chinese restaurant sampling (Ishwaran and Takahara, 2002), and variational Bayes approximations (Blei and Jordan, 2006; Kurihara, Welling and Vlassis, 2006; Kurihara, Welling and Teh, 2007).

There are three main types of MCMC algorithms that have been proposed for posterior computation in DPMs, including the collapsed Gibbs sampler (MacEachern, 1994), the blocked Gibbs sampler (Ishwaran and James, 2001), and reversible jump-type approaches (Jain and Neal, 2004; Dahl, 2007). The collapsed, or Polya urn, Gibbs sampling algorithm avoids updating the infinitely many parameters characterizing P (refer to expression 2) by marginalizing out P and relying on the Polya urn scheme of Blackwell and MacQueen (1973). In particular, if we let $\mu_i \sim P$, with $P \sim DP(\alpha P_0)$, then the joint distribution of $\mu_1, \ldots, \mu_n$ marginalizing out P can be expressed as

$$p(S_1, \ldots, S_n, \theta_1^*, \ldots, \theta_k^*) = \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha + n)} \prod_{j=1}^{k} p_0(\theta_j^*)\, \Gamma(n_j), \tag{5}$$

where $(\theta_1^*, \ldots, \theta_k^*)'$ are the unique values of $(\mu_1, \ldots, \mu_n)'$, $S_i = j$ denotes that $\mu_i = \theta_j^*$, $n_j = \sum_{i=1}^{n} 1(S_i = j)$ is the number of subjects having value θ_j^*, and P_0 is assumed to have density p_0 with respect to Lesbesgue measure (refer to Petrone and Raftery, 1997, for a derivation of (5)). Then, instead of updating P in the MCMC algorithm, we update $\mathbf{S} = (S_1, \ldots, S_n)'$ and $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)'$ using a simple Gibbs sampling algorithm proposed by Bush and MacEachern (1996). Refer to West et al., 1994 for a simple description of this type of approach.

A useful alternative to the collapsed Gibbs sampler is the blocked Gibbs sampler of

Ishwaran and James (2001). The blocked Gibbs sampler relies on approximating P through truncation of the stick-breaking representation in (2). In particular, noting that the probability weights assigned to the atoms tend to decrease rapidly as the index h increases, it is reasonable to replace the infinite sum with a sum across the first N terms, which can be accomplished by simply letting $V_N = 1$. Truncations of stick-breaking representations to DPs were originally proposed by Muliere and Tardella (1998). Conveniently, if the base measure P_0 is conditionally conjugate, as is true in the case we consider above, the full conditional distributions needed for Gibbs sampling have simple conjugate forms. In particular, the blocked Gibbs sampler cycles through steps for (1) allocating each individual to one of the components by sampling the index S_i from a closed form multinomial conditional posterior, with probabilities:

$$\Pr(S_i = h \,|\, -) = \frac{\left\{V_h \prod_{l<h}(1-V_l)\right\} \prod_{j=1}^{n_i} N(y_{ij}; \theta_h, \sigma^2)}{\sum_{r=1}^{N} \left\{V_r \prod_{s<r}(1-V_s)\right\} \prod_{j=1}^{n_i} N(y_{ij}; \theta_r, \sigma^2)}, \quad h = 1, \ldots, N.$$

(2) updating the stick-breaking weights from conditionally conjugate beta posterior distributions:

$$(V_h \,|\, -) \overset{ind}{\sim} \text{beta}\left(1 + \sum_{i=1}^{n} 1(S_i = h), \alpha + \sum_{i=1}^{n} 1(S_i > h)\right), \quad h = 1, \ldots, N-1,$$

with $V_N = 1$. (3) updating the atoms (random effects specific to each cluster) by independent sampling from normal posteriors:

$$(\theta_h \,|\, -) \overset{ind}{\sim} \text{N}\left(\frac{\sigma_0^{-2}\mu_0 + \sigma^{-2}\sum_{i:S_i=h}\sum_{j=1}^{n_i} y_{ij}}{\sigma_0^{-2} + \sigma^{-2}\sum_{i:S_i=h} n_i}, \frac{1}{\sigma_0^{-2} + \sigma^{-2}\sum_{i:S_i=h} n_i}\right), \quad h = 1, \ldots, N,$$

(4) updating the hyperparameters (μ_0, σ_0^{-2}) by sampling from the conditionally-conjugate normal-gamma posterior:

$$(\mu_0, \sigma_0^{-2} \,|\, -) \sim \text{N}(\mu_0; \widehat{\mu}_0, \widehat{\tau}\sigma_0^2) G(\sigma_0^{-2}; \widehat{a}_0, \widehat{b}_0),$$

with $\text{N}(\mu_0; \mu_{00}, \tau\sigma_0^2) G(\sigma_0^{-2}; a_0, b_0)$ the prior, $\widehat{\tau} = 1/(\tau^{-1} + N)$, $\widehat{\mu}_0 = \widehat{\tau}(\tau^{-1}\mu_{00} + \sum_{h=1}^{N} \theta_h)$, $\widehat{a}_0 = a_0 + N/2$, and $\widehat{b}_0 = b_0 + 1/2(\tau^{-1}\mu_{00}^2 + \sum_{h=1}^{N} \theta_h^2 - \widehat{\tau}^{-1}\widehat{\mu}_0^2)$, and (5) updating the within-subject precision σ^{-2} from its conditionally-conjugate gamma posterior:

$$(\sigma^{-2} \,|\, -) \sim G\left(a_1 + \frac{1}{2}\sum_{i=1}^{n} n_i, b_1 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} (y_{ij} - \theta_{S_i})^2\right),$$

where $G(a_1, b_1)$ is the prior for σ^{-2}, with $G(a, b)$ corresponding to the gamma distribution parameterized to have mean a/b and variance a/b^2.

Each of these steps is quite easy to implement, so MCMC computation in the analysis of variance model with a Dirichlet process prior on the random effects distribution is essentially no more difficult than posterior computation in the parametric case in which a normal distribution is assumed for the random effects. One potential issue with the specification in which we let $P \sim DP(\alpha P_0)$ is that it assumes a discrete distribution for the random effects,

so that different subjects have exactly the same random effects value. This may be a useful approximation, but it may be more realistic to suppose that each subject has their own unique random effect value. Hence, we may instead want to characterize the random effects distribution using an unknown continuous density. This can be easily accomplished using a minor modification to the above specification to let $\mu_i \sim N(\mu_{0i}, \sigma^2_{0i})$, with $(\mu_{0i}, \sigma^2_{0i}) \sim Q$, and $Q \sim DP(\alpha Q_0)$. In this case, the random effect distribution, P, is characterized as a DP mixture (DPM) of normals (Lo, 1984; Escobar and West, 1995).

The DPM of normals for the random effects distribution can be used for clustering of subjects into groups having similar, but not identical, random effects. The blocked Gibbs sampler is easily modified to accommodate this case. One common concern with the blocked Gibbs sampler, and other approaches that rely on truncation of the stick-breaking representation to a finite number of terms, is that in bypassing the infinite-dimensional representation, we are effectively fitting a finite (and hence parametric) mixture model. For example, if we let $N = 25$ as a truncation level, a natural question is how this is better or intrinsically different than fitting a finite mixture model with 25 components. One answer is that N is not the number of components occupied by the subjects in your sample, but is instead an upper bound on the number of subjects. In most cases, taking a conservative upper bound, such as $N = 25$ or $N = 50$, should be sufficient, since mixture models are most useful when there are relatively few components. In addition, because the weights in the infinite stick-breaking representation (2) decrease rapidly for typical choices of α, we also obtain an accurate approximation to the DP for modest N.

However, there have been some recent approaches that avoid the need for truncation. Walker (2007) proposed a slice sampling approach. Papaspiliopoulos and Roberts (2007) proposed an alternative retrospective MCMC algorithm, which is an easy to implement modification to the blocked Gibbs sampler that allows one to adaptively add, but not delete. components as needed as the MCMC algorithm progresses. This can actually result in substantially improved efficiency in some cases. To clarify, note that one would typically choose a conservative truncation level in implementing the blocked Gibbs sampler, which would then require updating of the stick-breaking weights and atoms for many components that are not needed in that they are assigned very low probabilities and are not occupied by any subjects in the sample. The retrospective sampling approach instead allows one to conduct computation for the number of components that are needed, though to take advantage of this efficiency gain it is typically necessary to run a short preliminary chain of 10-100 iterations to choose good starting values. Otherwise, the retrospective MCMC approach may add a large number of components in the first few sampling steps, and then one is unable to delete these components later in the sampling, resulting in a large computational burden.

2.3 *General Random Effects Models*

Until this point, I have focused for illustration on the simple variance component model in (1). However, it is straightforward to extend the ideas to much richer classes of random effects models. For example, Kleinman and Ibrahim (1998a) placed a DP prior on the distribution of the random effects in a linear mixed effects model, while Kleinman and Ibrahim (1998b) extended this approach to the broader class of generalized linear mixed models. Mukhopadhyay and Gelfand (1997) propose a wide class of DP mixtures of GLMs. Müller and Rosner

(1997) used a DPM to obtain a flexible non-linear hierarchical model for blood count data, while in more recent work, Müller, Quintana and Rosner (2007) proposed a semiparametric model for multilevel repeated measurement data. Walker and Mallick (1997) used Polya tree (PT) priors for nonparametric modeling of random effect and frailty distributions. The PT prior is another popular and computationally attractive nonparametric Bayes prior. From a data analysis perspective, it could be used as an alternative to a DP prior on P in (2). For a recent article on mixtures of PT priors, refer to Hanson (2006).

The linear mixed effects model (Laird and Ware, 1982) is used routinely for the analysis of data from longitudinal studies and studies having multilevel designs (e.g., patients nested within study centers). Focusing on the longitudinal data case, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{i,n_i})'$ denote the response data for subject i, with y_{ij} the observation at time t_{ij}, for $j = 1, \ldots, n_i$. Then, the linear mixed effects model has the form:

$$\begin{aligned} y_{ij} &= \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \\ \mathbf{b}_i &\sim P, \end{aligned} \tag{6}$$

where $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ are fixed effect predictors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\mathbf{z}_{ij} = (z_{ij1}, \ldots, z_{ijq})'$ are random effect predictors, and P is a random effect distribution on $\Re^q$.

It is straightforward to allow P to be unknown through the use of a DP prior to induce a discrete random effects distribution, or a DP mixture of Gaussians to induce an unknown continuous random effects density. In fact, such models have been increasingly used in applications. For example, van der Merwe and Pretorius (2003) applied a linear mixed effects model with a DP prior for the random effects distribution in an animal breeding application. Ohlssen, Sharples and Spiegelhalter (2007) provide a recent overview of the literature on Bayesian semiparametric random effects models, and provide a tutorial on routine implementation in WinBUGS. In addition, there is an R package, DPpackage, which provides R functions for efficiently fitting a broad variety of semiparametric Bayes hierarchical models, including not only DPMs but only Polya tree models (Jara, 2007).

However, there are some subtle issues that arise in semiparametric modeling of random effects distributions, which should be carefully considered in fitting such models. In particular, in expression (6), the posterior distribution of the random moments of P may impact inferences on $\boldsymbol{\beta}$. In parametric models, it is standard practice to use a multivariate normal distribution with mean zero for P, so that the coefficients $\boldsymbol{\beta}$ are then interpretable as fixed effects. In particular, we require that $\mathrm{E}(y_{ij} \,|\, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta}$, which is not true if the posterior expectation of the mean of P is non-zero. Constraining the base measure P_0 to have zero mean is not sufficient to ensure that P is centered on zero a posteriori.

One way to get around this problem is to use a centered parameterization. For example, one can let $y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_i + \epsilon_{ij}$, with $\boldsymbol{\beta}_i \sim P$. In this case, the fixed effect regression coefficients correspond to the mean of the distribution P. One limitation of this is that one needs to assume that $\mathbf{x}_{ij} = \mathbf{z}_{ij}$, though the extension to allow $\mathbf{z}_{ij}$ to correspond to a subset of the predictors in $\mathbf{x}_{ij}$ is straightforward. Another limitation of using a centered parameterization is that the fixed effect coefficients $\boldsymbol{\beta}$ corresponding to the mean of P, which may not be directly available if one implements a computation approach, such as the collapsed Gibbs sampler, that marginalizes out P.

Two recent alternatives were proposed by Li, Lin and Müller (2007) and Dunson, Yang and Baird (2007). The Li et al. (2007) approach uses post-processing to adjust for bias in

using a DP prior for the random effects distribution. The Dunson, Yang and Baird (2007) approach instead induces a centered DP or centered DP mixture prior by considering the DP or DPM prior for the random effects distribution as a parameter-expanded version of a centered process with mean and/or variance constraints. The centered DPM is an alternative to previously proposed approaches, which constrain a random effects distribution to have median 0 (e.g., Burr and Doss, 2005).

2.4 *Latent Factor Regression Models*

Linear mixed effects models and generalized linear mixed effects models are appropriate when the same type of response is measured repeatedly over time but not when data on a subject consist of a multivariate vector of different types of responses. In many biomedical studies, one may measure multiple surrogates of a latent predictor or health response of interest. For example, single cell gel electrophoresis measures the frequency of DNA strand breaks on the individual cell level through different surrogates for the amount of DNA in the tail of an image that resembles a comet. As these different surrogates are on different scales, one can consider a latent factor regression model, such as

$$\begin{aligned} y_{ij} &= \mu_j + \lambda_j \eta_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_j^2), \\ \eta_i &= \mathbf{x}_i' \boldsymbol{\beta} + \delta_i, \quad \delta_i \sim P, \end{aligned} \tag{7}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$ are p different measures of the frequency of strand breaks in cell i, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ are intercept parameters for the different surrogates, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ are factor loadings, with $\lambda_j > 0$ for $j = 1, \ldots, p$, η_i is a continuous frequency of DNA strand breaks latent variable, $\boldsymbol{\epsilon} = (\epsilon_{i1}, \ldots, \epsilon_{ip})'$ are idiosyncratic measurement errors, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_p)$ is a diagonal covariance matrix, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ are predictors of frequency of strand breaks (e.g., dose of a possible genotoxic agent), and P is an unknown latent variable residual distribution.

The distribution of the frequency of strand breaks tends to be right skewed, often with a secondary mode in the right tail. Hence, in order to limit parametric assumptions, one can use a DPM of normals for the latent variable residual distribution, P. However, latent factor regression models require some restrictions for identifiability. In the parametric case of expression (7), one would typically let P correspond to the standard normal distribution, which is automatically restricted to have mean 0 and variance 1. Then, the coefficient β_j would have a simple interpretation as the number of standard deviations the latent trait is shifted for each unit change in the jth predictor. To induce mean 0 and variance 1 constraints on the latent variable distribution in the semiparametric case, one can use a centered DPM prior for P, as in Dunson, Yang and Baird (2007). Such an approach is very easy to implement, since a blocked Gibbs sampler can be implemented as if a DPM of normals were used for P, followed by a simple post-processing step. One can also apply related approaches in a much broader class of latent variable models that allow mixed categorical and continuous measurements and multiple latent variables.

3. Nonparametric Bayes Functional Data Analysis

3.1 *Background*

In many applications, interest focuses on studying variability in random functions. Some examples of random functions include hormone trajectories over time and brain images

collected using MRI technology. Functional data analysis (FDA) methods are used when data consist of error-prone observations on random functions that may differ for the different subjects under study (Ramsay and Silverman, 1997). In order to study heterogeneity among subjects and to borrow strength across the different subjects in estimating their functions, one may consider hierarchical models of the form:

$$\begin{aligned} y_i(t) &= \eta_i(t) + \epsilon_i(t), \quad \epsilon_i(t) \sim N(0, \sigma^2) \\ \eta_i &\sim P, \end{aligned} \tag{8}$$

where $y_i(t)$ is an error-prone observation of the function η_i for subject i at time t, $i = 1, \ldots, n$, $\epsilon_i(t)$ is a measurement error, and P is a distribution on Ω, the space of $\mathcal{T} \to \Re$ functions. In practice, it is not possible to observe η_i directly at any time, and we only have measurements of $y_i(t)$ for $t \in \mathbf{t}_i = (t_{i1}, \ldots, t_{i,n_i})'$.

In this section, we consider a variety of semiparametric Bayes approaches for functional data analysis. Section 3.2 describes methods based on basis function expansions of η_i. Section 3.3 reviews methods that avoid explicit basis function representations using functional Dirichlet processes. Section 3.4 provides an overview of recent kernel-based approaches, and Section 3.5 considers methods for joint modeling of related functions and of functional predictors with response variables.

3.2 *Basis Functions and Clustering*

In nonlinear regression and functional data analysis, it is common to simplify modeling by assuming that the unknown functions fall in the linear span of some pre-specified set of basis functions. For example, focusing on hierarchical model (8), suppose that

$$\eta_i(t) = \sum_{h=1}^{p} \beta_{ih} b_h(t), \quad \forall t \in \mathcal{T}, \tag{9}$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})'$ are basis coefficients specific to subject i, and $\mathbf{b} = \{b_h\}_{h=1}^t$ is a set of basis functions. For example, if η_i could be assumed to be a smooth function, cubic spline basis functions of the following form may be reasonable:

$$\mathbf{b}(t) = \{1, t, t^2, t^3, (t-\xi_1)_+^3, (t-\xi_2)_+^3, \ldots, (t-\xi_q)_+^3\},$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)'$ are knot locations, and x_+ returns 0 for negative x and x for positive x.

Assuming the basis functions are pre-specified (e.g., by choosing a grid of a modest number of equally-spaced knots), models (8) and (9) imply that

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{10}$$

where $y_{ij} = y_i(t_{ij})$, $\mathbf{x}_{ij} = [b_1(t_{ij}), b_2(t_{ij}), \ldots, b_p(t_{ij})]'$, and $\epsilon_{ij} = \epsilon_i(t_{ij})$, for $i = 1, \ldots, n, j = 1, \ldots, n_i$. Hence, letting $\boldsymbol{\beta}_i \sim Q$, one can simply use a linear mixed effects model for functional data analysis. As a flexible semiparametric approach, one can place a DP prior on Q. As noted by Ray and Mallick (2006) in the setting of a wavelet model, such an approach induces functional clustering.

To clarify, note that the DP prior for Q implies that each subject is allocated into one of $k \leq n$ clusters, with subjects in a cluster having identical values for the basis coefficients. In particular, letting $S_i = h$ denote that subject i is allocated to cluster h, we would have $\boldsymbol{\beta}_i = \boldsymbol{\theta}_h^*$ for all subjects having $S_i = h$. Hence, all the subjects in a cluster would also have identical functional trajectories, with subjects in cluster h having $\eta_i(t) = \mathbf{b}(t)\boldsymbol{\theta}_h^*$, for all $t \in \mathcal{T}$. Note that this provides a semiparametric Bayes alternative to frequentist latent class trajectory models (Muthén and Shedden, 1999) and growth mixture models (Jones, Nagin and Roeder, 2001). Such approaches rely on finite mixture models, with the EM algorithm typically used to obtain maximum likelihood estimates.

Assuming a DP prior for Q implies that individuals in a functional cluster have exactly the same value of the measurement error-corrected function, η_i. This assumption may be overly-restrictive and may result in estimation of a large number of functional clusters in some cases. Hence, it may be more realistic to suppose that every individual has a unique function, η_i, but that the functions for individuals in a cluster are similar to each other. This can be accomplished by using a DP mixture of multivariate Gaussians as the prior for Q.

Model (8) can be easily modified to include fixed and random effect covariates. Posterior computation is straightforward using a very similar approach to that described in Section 2.2 for the simple variance component model (1). However, some complications can arise in interpretation of the MCMC output. In particular, the Bayesian semiparametric approach has the appealing property of allowing uncertainty in the number of clusters and the allocation of subjects to these clusters. This has the side effect that the number of clusters and the meaning of the clusters will change across the MCMC iterations. Hence, it can be quite challenging to obtain meaningful posterior summaries of cluster-specific parameters. This problem is not unique to the functional clustering application, and is typically referred to in the literature as the label switching problem (Stephens, 2000; Jasra, Holmes and Stephens, 2005).

Frequentist analyses of mixture models that are fitted with the EM algorithm do not face this issue, because the EM algorithm converges to a point estimate corresponding to a local mode. This point estimate includes the cluster probabilities and each of the cluster-specific parameters (e.g., basis coefficients). In performing inferences on the cluster-specific parameters, one ignores the numeric labels for each of the clusters. However, EM algorithm-based analyses of mixture models also face problems in locating a global mode even when multiple starting points are used. In addition, such methods rely on pre-specification or selection of a fixed number of clusters, while the Bayesian semiparametric approach automatically allows uncertainty in the number of clusters.

Fortunately label-switching is only a problem if one is interested in performing cluster-specific inferences instead of simply accounting for uncertainty in clustering when conducting predictions or performing inferences on global features. I use the term "global features" to denote any functional of interest that is not cluster-specific. For example, global features may include fixed effect regression coefficients and values of $\eta_i(t)$, for specific subjects or averaged across subjects. Such features can be estimated easily from the MCMC output without worrying about the fact that there are latent cluster indices that are changing in meaning and dimension over the MCMC iterates. For example, one can obtain a posterior mean and 95% credible interval by collecting $\eta_i(t)$ for each of a large number of MCMC iterates, and then averaging the samples and calculating the 2.5th and 97.5th percentiles.

The real problem occurs when one want to estimate the functional trajectories specific to each cluster, and a variety of strategies have been proposed. One technique is to attempt to relabel the clusters at each MCMC iteration using a post-processing algorithm. For examples of post-processing approaches, refer to Stephens (2000) and Jasra, Holmes and Stephens (2005). Such approaches tend to be time-consuming to implement and do not seem to fully address the problem. For example, in running an MCMC algorithm under models (8) and (9) with a DP prior on the distribution of the basis coefficients, the number of clusters may vary substantially over the MCMC iterations. However, in re-labeling to line up the clusters from the different iterations, one needs to assume that there is some coherence in the clusters after re-labeling. Typically, this would at least require fixing of the number of clusters.

Given the very high dimensional set of possible partitions of subjects into clusters, it is not at all unlikely that one may visit dramatically different configurations over the MCMC iterations, particularly if an efficient MCMC is used. Hence, instead of attempting to align clusters that are in some sense unalignable, it is useful to view the partition of subjects as a model selection problem, with the MCMC approaches outlined in Section 2.2 providing an approach for model averaging. As is well known in the literature on Bayesian model selection, model averaging is most useful for prediction, and one may commonly encounter difficulties in interpretation when averaging over models in which the parameters have different meanings. In such settings, it is necessary to perform model selection to maintain interpretability.

Carrying over this idea to the setting of DP mixture models, one could attempt to identify an optimal partition of subjects into clusters, with this optimal clustering then used to obtain some insight into how the clusters differ. Before reviewing some of the approaches available to identify an optimal clustering in this setting, it is important to note that one should be very careful to avoid over-interpretation of the estimated partition. Even if one is able to identify the optimal partition from among the very high dimensional set of possible partitions, this partition may have extremely low posterior probability, and there may be a very large number of partitions having very similar posterior probability to the optimal partition. This is essentially the same problem that is faced in model selection in high-dimensional settings. That said, it is sometimes impossible to bypass the need for selection given the scientific interests in a study. In such settings, the approaches of Medvedovic and Sivaganesan (2002), Dahl (2006) and Lau and Green (2007) are quite useful.

3.3 *Functional Dirichlet Process*

The basis function expansion shown in (9) clearly requires an explicit choice of a set of basis functions. For well behaved, smooth functions of time or a single predictor, it may be sufficient to choose splines, with knots specified at a modest-dimensional grid of equally spaced locations. However, when $\mathcal{T}$ is multi-dimensional (e.g., corresponding to a subset of $\Re^2$ in image or spatial applications or to $\Re^r$ in multivariate regression applications), it can be difficult to pre-specify an appropriate basis. Bigelow and Dunson (2005) proposed a modification, which allows unknown numbers and locations of knots, while placing a DP prior on the distribution of the basis coefficients. An alternative is to avoid using an explicit basis function representation entirely by instead relying on a functional Dirichlet process (FDP).

The FDP provides a direct approach to specify a prior for P in (8) by letting $P \sim$

$DP(\alpha P_0)$, where P_0 corresponds to a Gaussian process (GP). Hence, using the stick-breaking representation, we have

$$\eta_i \sim P = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \sim GP(\mu, \mathcal{C}), \tag{11}$$

where $\{\pi_h\}_{h=1}^{\infty}$ are defined as in (2) and $\boldsymbol{\theta} = \{\theta\}_{h=1}^{\infty}$ are functional atoms sampled independently from a Gaussian process with mean function μ and covariance function $\mathcal{C}$, for $h = 1, \ldots, \infty$. The FDP has the same mathematical structure as the dependent DP proposed by MacEachern (1999), which will be discussed in Section 6.2. Gelfand et al. (2005) used the FDP for modeling of spatial data.

Under the FDP in (11) subjects will be allocated to functional clusters. Letting $S_i = h$ denote that subject i is allocated to the hth of the $k \leq n$ clusters represented in the data set, we have $\eta_i = \theta^*_{S_i}$, for $i = 1, \ldots, n$. In this case, instead of using a finite vector of basis coefficients to characterize the functions specific to each cluster, we take independent draws from a Gaussian process. This avoids the need to explicitly specify a set of basis functions. However, it is necessary to choose mean and covariance functions in the GP, and the number of functional clusters and the cluster-specific estimates can be quite sensitive to the particular choice made. For a recent book on Gaussian processes, including a discussion of the role of the covariance function, refer to Rasmussen and Williams (2006).

In implementing posterior computation, it is clearly not possible to estimate the functions across the infinitely-many locations in $\mathcal{T}$. Hence, in practice one performs computation for finitely-many points, typically corresponding to locations at which data are collected along with a tightly-spaced grid of additional locations. The GP base measure then implies a multivariate normal base measure across this finite grid, and posterior computation can proceed as for DP mixtures of Gaussians. However, instead of updating the finite-dimensional mean and covariance in the base multivariate normal distribution, one estimates parameters characterizing the mean and covariance functions. For the covariance function, this would typically involve Metropolis-Hastings steps, after assuming a particular form, such as exponential, Gaussian or Matérn.

3.4 *Kernel-Based Approaches*

As there is often concern in practice about the impact of basis and covariance function specification on inferences, estimation and prediction, it is appealing to consider alternatives. In the frequentist literature on function estimation, it is common to consider kernel-based approaches. In particular, focusing initially on the mean regression function estimation problem in which $\mathrm{E}(Y \mid X = x) = \eta(x)$, there is a rich literature on estimation of η subject to the constraint that $\eta \in \mathcal{H}_K$, where $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) defined by the uniformly bounded Mercer kernel K.

In their representer theorem, Kimeldorf and Wahba (1971) show that the solution to a least squares minimization problem subject to an RKHS norm penalty lies in a subspace of $\mathcal{H}_K$ represented as follows:

$$\eta(x) = \sum_{i=1}^{n} w_i K(x, x_i), \tag{12}$$

where $\mathbf{w} = (w_1, \ldots, w_n)'$ are unknown coefficients. Tipping (2001), Sollich (2002) and Chakraborty et al. (2005) consider Bayesian kernel-based methods based on choosing a prior for $\mathbf{w}$. Such approaches implicitly assume that the support of the prior lies in a subspace of $\mathcal{H}_K$ represented as in (12), which is somewhat unnatural in that (12) was derived in solving an optimization problem.

Pillai et al. (2007) and Liang et al. (2006) noted that a fully Bayesian solution would instead place a prior for η with large support in $\mathcal{H}_K$. Pillai et al. (2007) accomplished this through the integral representation:

$$\eta(x) = \int_{\mathcal{T}} K(x,u) d\gamma(u), \forall x \in \mathcal{T}, \tag{13}$$

where $\gamma \in \Gamma$ and $\eta \in \mathcal{G}$. When Γ corresponds to the space of all signed Borel measures, then $\mathcal{G} = \mathcal{H}_K$. Pillai considered a variety of specific possibilities for γ, focusing on Lévy process priors, while Liang et al. (2006) instead used a decomposition that expressed γ as a product of a GP and a DP. In the Liang et al. (2006) specification, the DP component essentially places a random probability measure on the locations of the kernels, while the GP places a prior on the coefficients at the resulting countably infinite collection of locations.

MacLehose and Dunson (2008) generalized the Liang et al. (2006) formulation to the functional data analysis setting by letting $\eta_i(x) = \int K(x,u) d\gamma_i(u)$, and then choosing a nonparametric Bayes hierarchical prior for $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_n\}$. In particular, their specification relied on functional DP and hierarchical DP (HDP) (Tomlinson, 1998; Teh et al., 2006) components. The HDP is a prior for modeling of related distributions through incorporating dependence by assuming a common base measure in DPs for each of the distributions, with this base measure allowed to be unknown through use of a DP prior. For further details on the HDP, refer to the chapter by Teh and Jordan.

An appealing feature of the MacLehose and Dunson (2008) approach relative to the approaches described in Sections 3.2 and 3.3 is the allowance for local borrowing of information through local selection of kernels and locally-dependent weights. To illustrate this, consider an application to progesterone trajectory data previously analyzed in Brumback and Rice (1998). Data were available for 51 women who provided samples over 91 cycles, of which 22 were conception cycles. Taking results from MacLehose and Dunson (2008), Figure 2 shows the raw data and estimated posterior mean progesterone curves for 3 women randomly selected from the non-conception group. This figure demonstrates the borrowing of information. In particular, during the baseline phase prior to ovulation (day 0 in the figure), the progesterone values are quite similar, so there is strong borrowing of information and the estimates are essentially equivalent. However, following ovulation the curves smoothly deviate, with the approach favoring similar shapes across the curves. Note that the methods of Section 3.2 - 3.3 instead borrow information only through global clustering and through the parametric base model. Alternative methods for local borrowing of information will be discussed in detail in Section 4.

3.5 *Joint Modeling*

In biomedical studies, there is very commonly interest in studying the relationship between functional predictors and response variables. For example, the functional predictor may correspond to the longitudinal trajectory in the level of an environmental exposure, such as

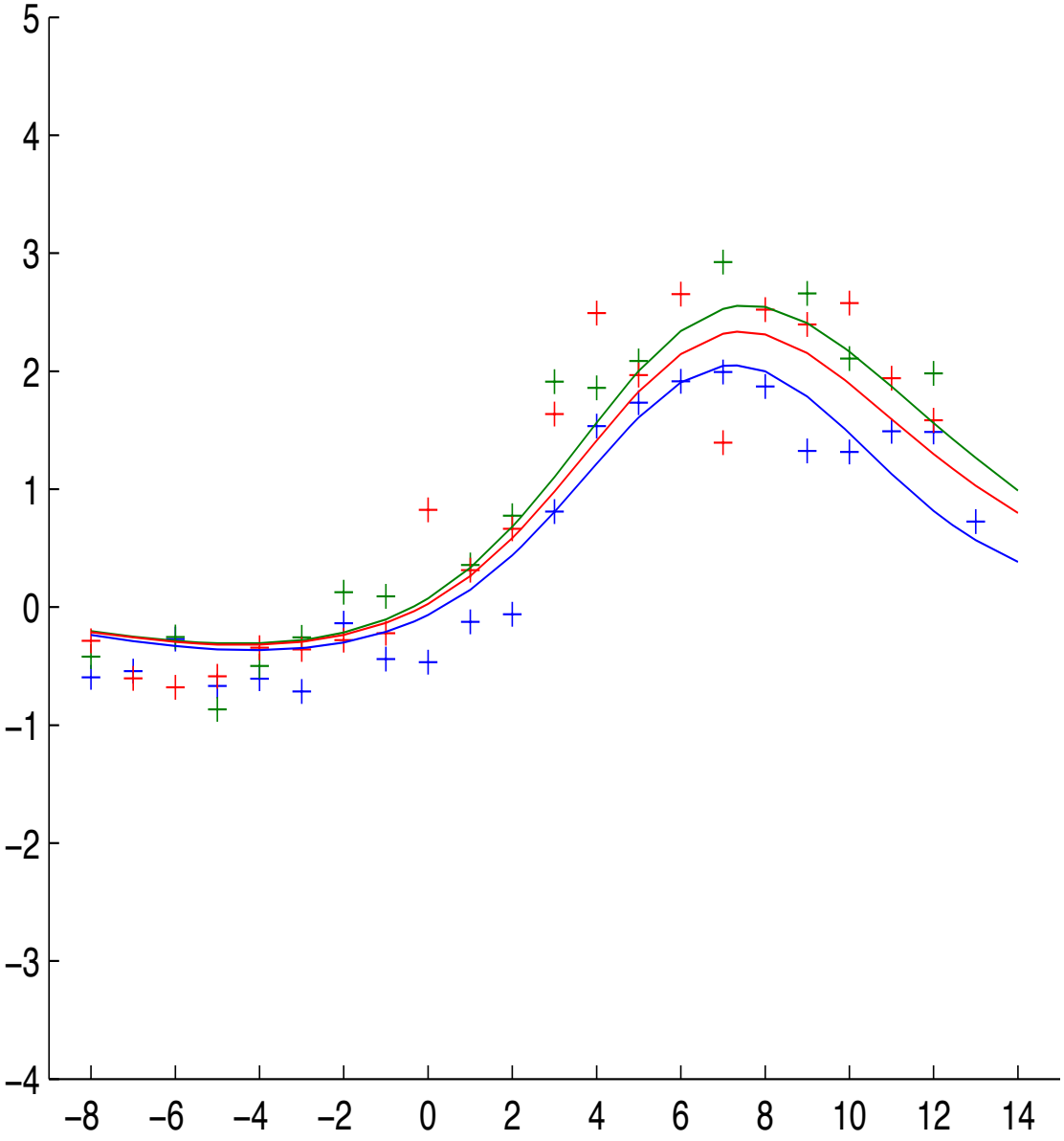

Figure 2. Posterior mean progesterone curves (solid lines) and observed progesterone levels for 3 selected non-conceptive cycles. Estimated progesterone curves and observed data are linked by color.

air pollution, or to a diagnostic image, while the response corresponds to an indicator of an adverse health condition. In such cases, there is substantial interest in building flexible joint models for relating a subject's functional predictor to their health status, adjusting for possible confounding factors, such as age and demographic variables.

To focus on a motivating application, we consider an epidemiologic study of early preg-

nancy loss (EPL) in which daily urine samples were collected in order to measure levels of hormone metabolites over time prior to conception and through early pregnancy (Wilcox et al., 1988). Our interest is in studying how progesterone trajectories following ovulation predict EPL. EPLs are identified when hCG rises soon after implantation but then declines back to baseline levels instead of continuing to rise. Progesterone plays a critical role in maintaining the pregnancy, so some clinicians have even suggested treatment with exogenous progesterone as a possible intervention to reduce risk of EPL. There are affordable home devices available for measuring progesterone metabolites in urine, so an algorithm could potentially be programmed into such a device to alert the woman when she is at risk of impending loss.

Motivated by this application, Bigelow and Dunson (2007) develop a Bayesian nonparametric approach for joint modeling of functional predictors with a response variable. Their proposed approach relies on a simple extension of the method proposed in Section 3.2. In order to facilitate applications to other settings, I will initially present the approach in more generality than considered in Bigelow and Dunson (2007). In particular, suppose for subject i, we have data $\mathbf{y}_i = (\mathbf{y}_{i1}', \ldots, \mathbf{y}_{ip}')'$, where $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ij,n_{ij}})'$ is a vector of observations of type j, for $j = 1, \ldots, p$. For example, $\mathbf{y}_{i1}$ may consist of error-prone measurements of a functional predictor, while y_{i2} is 0/1 indicator of a health response. More generally, the $\mathbf{y}_{ij}$'s may correspond to several different types of information collected on a subject.

We define separate models for each of the components of the data vector as follows:

$$\begin{aligned} \mathbf{y}_{ij} &\sim f_j(\boldsymbol{\beta}_{ij}; \boldsymbol{\phi}_j), \quad j = 1, \ldots, p, \\ \boldsymbol{\beta}_i = (\boldsymbol{\beta}_{i1}', \ldots, \boldsymbol{\beta}_{ip}')' &\sim P, \end{aligned} \tag{14}$$

where $f_j(\boldsymbol{\beta}_{ij}; \boldsymbol{\phi}_j)$ is the likelihood for component j, defined in terms of the subject-specific parameters $\boldsymbol{\beta}_{ij} = (\beta_{ij1}, \ldots, \beta_{ij,p_j})$ and population parameters $\boldsymbol{\phi}_j$, and the different component models are linked through P, the joint distribution for $\boldsymbol{\beta}_i$. In parametric joint modeling of multivariate data having a variety of measurement scales, it is common to use latent factor and structural equation models that incorporate shared latent variables in the different component models. By allowing P to be unknown through a nonparametric Bayes approach, we obtain a more flexible class of joint models.

Bigelow and Dunson (2007) propose to use a DP prior for P with the following structure:

$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\Theta_h}, \quad \Theta_h \sim P_0 = \otimes_{j=1}^{p} P_{0j}, \tag{15}$$

where $\boldsymbol{\pi} = \{\pi_h\}_{h=1}^{\infty}$ are as defined in (2), $\Theta_h = \{\Theta_{hj}\}_{j=1}^{p}$ is a collection of atoms corresponding to the parameters in each of the p different components, and the base measure P_0 used in generating these atoms is characterized as a product measure of probability measures for each component. For example, in joint modeling of a longitudinal trajectory with a 0/1 response, P_{01} may correspond to a Gaussian process and P_{02} to a beta distribution, so that each DP cluster then contains a random function along with a corresponding probability of an adverse response. In this manner, a semiparametric joint model is defined for data having different scales, with dependence in the different types of data collected for a subject induced through allocating individuals to clusters having different parameters for each component model.

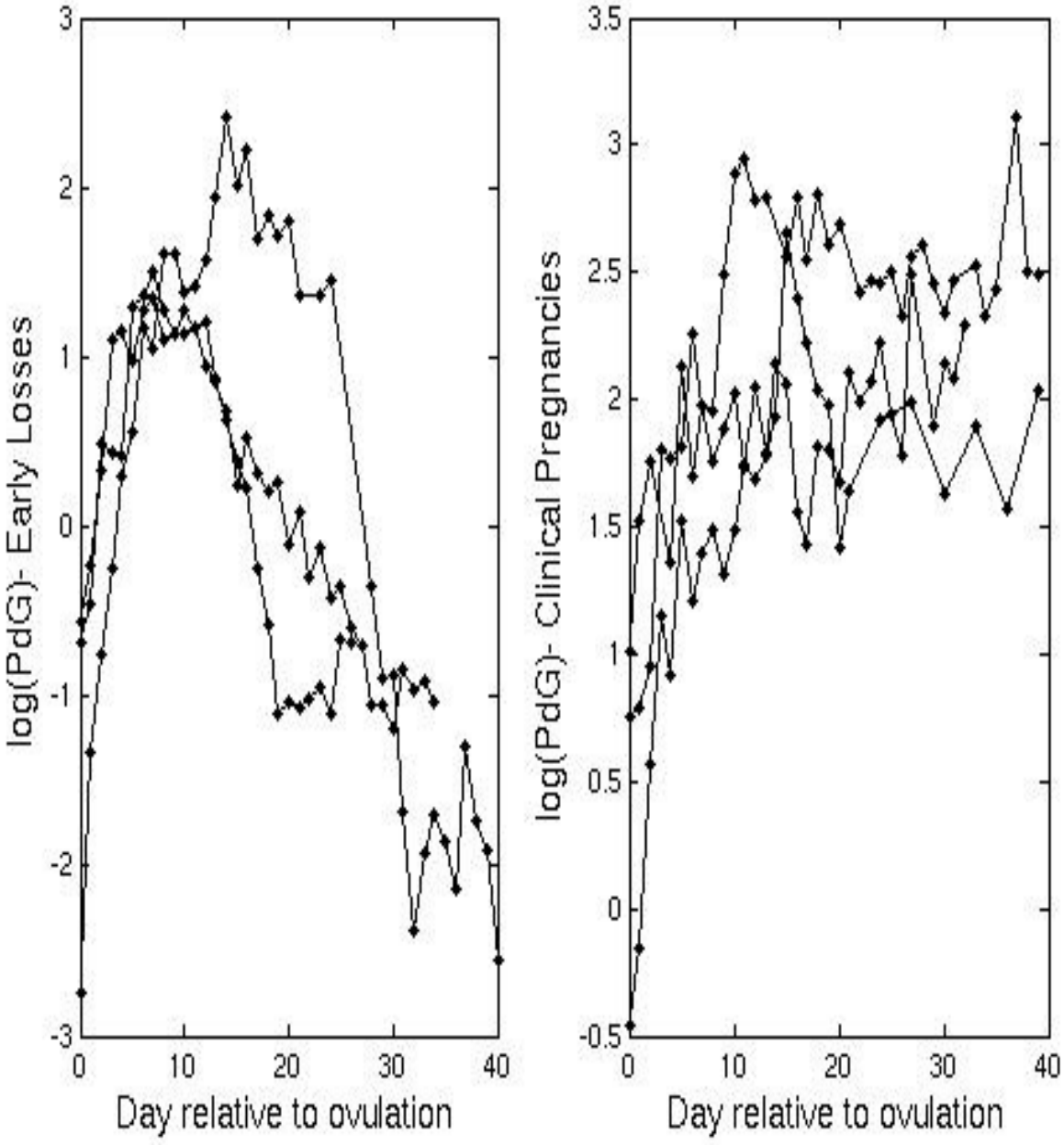

Figure 3. Progesterone data beginning at the estimated day of ovulation for three early losses and three clinical pregnancies.

Instead of using a GP for the functional predictor component, Bigelow and Dunson (2007) use multivariate adaptive splines with unknown numbers and locations of knots. They applied this approach to progesterone metabolite and EPL data from Wilcox et al. (1988). Figure 3 shows the progesterone data for three randomly selected early losses and three clinical pregnancies. Figure 4 shows the progesterone data for each of the 16 identified clusters containing more than one subject, along with the number of pregnancies in the

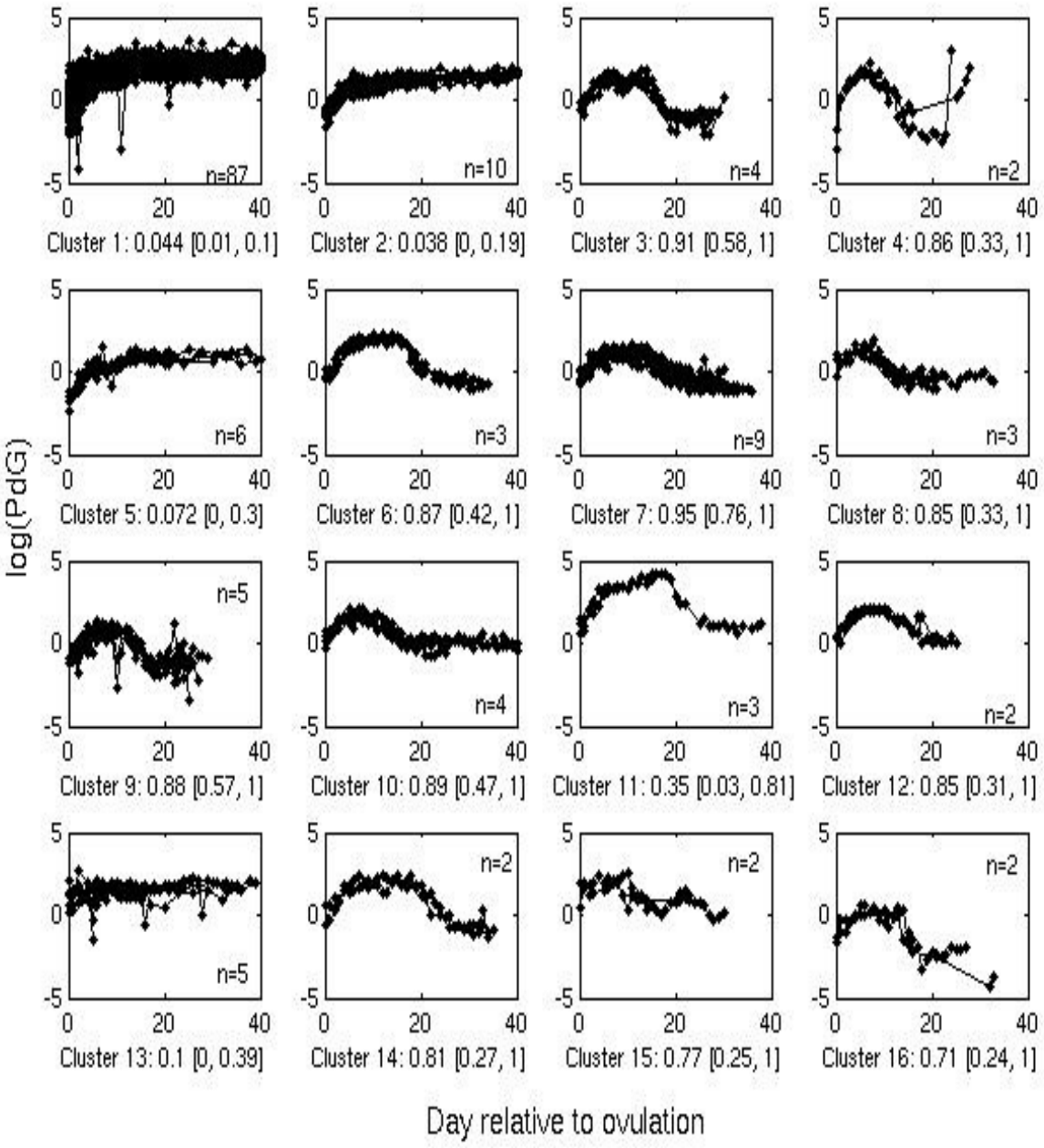

Figure 4. Progesterone data for pregnancies in each of the 16 clusters identified in the Wilcox et al. (1988) data. Only those clusters containing more than one pregnancy are shown. Below each plot is the estimated probability of early pregnancy loss within the cluster, along with a 95% credible interval.

cluster and the estimated probability of early pregnancy loss. Note that it is very clear that the identified clusters match the data from these plots. In addition, the early loss probabilities varied dramatically across the clusters. This resulted in accurate out of sample prediction of impending losses soon after implantation (results not shown).

4. Local Borrowing of Information and Clustering

Until this point, I have focused primarily on methods that rely on different variants of the DP for modeling of an unknown random effects distribution in a hierarchical model. As discussed above, such methods have the side effect of inducing clustering of subjects into groups, with groups defined in terms of unique random effects values or unique parameters in a parametric model. One characteristic of this type of specification is that clustering occurs globally, in that two individuals are clustered together for all their random effects or none.

For example, focus on the joint model specification in (15) and suppose $\boldsymbol{\beta}_i \sim P$, for $i = 1, \ldots, n$. Then, it is clear that either $\boldsymbol{\beta}_i = \boldsymbol{\beta}_{i'}$, with prior probability $1/(1+\alpha)$, or $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_{i'}$ correspond to two independent draws from P_0, so that none of the elements of $\boldsymbol{\beta}_i$ can cluster together with the corresponding elements of $\boldsymbol{\beta}_{i'}$ (assuming non-atomic P_0). Such global clustering is quite restrictive in that two different subjects may be quite similar or even identical for most of their random effects, while having important deviations in certain components. An alternative, which allows local clustering, is to let $\boldsymbol{\beta}_{ij} \sim P_j$, with $P_j \sim DP(\alpha_j P_{0j})$, independently for $j = 1, \ldots, p$. However, this approach does not allow for accrual of information about similarities between subjects. In particular, if $\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_{i'j}$ then the probability should intuitively be increased that $\boldsymbol{\beta}_{ij'} = \boldsymbol{\beta}_{i'j'}$ compared with the case in which $\boldsymbol{\beta}_{ij} \neq \boldsymbol{\beta}_{i'j}$.

Motivated by this problem, Dunson, Xue and Carin (2007) proposed a matrix stick-breaking process (MSBP), which generalizes the DP stick-breaking structure to allow row and column stick-breaking random variables to induce dependent local clustering. Posterior computation for the MSBP can proceed using a simple modification to the blocked Gibbs sampler of Ishwaran and James (2001). The MSBP has particular practical advantages over joint DP and independent DP priors in high-dimensional settings in which subjects can be very similar for most of their coefficients, while having distinct local deviations. This occurs, for example, when modeling of multiple, related functions or images using a basis representation. Often, most of the function or image is quite similar across subjects, suggesting that most of the basis coefficients are effectively identical. However, it is those local regions of heterogeneity that are most scientific interest.

Motivated by the problem of local clustering in functional data analysis, Petrone, Guindani and Gelfand (2007) proposed a hybrid functional Dirichlet process prior. The hybrid DP is based on the clever idea of introducing a collection of global species, with each individual function formulated from a patchwork of these global species. In particular, a latent Gaussian process is introduced, with the level of this latent process controlling local allocation to the global species. This results in local clustering of the functions for different individuals, with discontinuities occurring in the functions at changepoints in which the latent process crosses thresholds so that allocation switches to a different species. Petrone et al. (2007) applied this approach to an interesting brain image application. Rodriguez, Dunson and Gelfand (2007) proposed an alternative latent stick-breaking process (LaSBP). The LaSBP is defined by introducing a fixed unknown marginal distribution, which is assigned a stick-breaking prior, and then using a latent Gaussian process copula model to control local allocation to the atoms. The LaSBP avoids problems faced by the hybrid DP in performing spatial interpolations by introducing an order constraint on the atoms. This constraint also induces skewness, which is appealing in many applications.

Friedman and Meulman (2004) defined a concept of clustering on subsets of attributes

(COSA) in which two subjects can be partially clustered by having identical values for a subset of a vector of parameters. Hoff (2006) proposed a simple but clever Bayesian nonparametric version of COSA relying on a DP. In particular, if $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})'$ represents a vector of subject-specific parameters in a hierarchical model, then Hoff (2006) first lets $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{r}_i \times \boldsymbol{\delta}_i$, with $\mathbf{r}_i \in \{0,1\}^m$, $\boldsymbol{\delta}_i \in \Re^p$ and $\times$ denoting the elementwise product. Then, letting $(\mathbf{r}_i, \boldsymbol{\delta}_i) \sim P$, with $P \sim DP(\alpha P_0)$, results in clustering of the n subjects into $k \leq n$ groups. Note that subjects in a group will deviate from the baseline in only those attributes having $r_{ij} = 1$. In this manner, subjects in different DP clusters can be identical for a subset of their random effects, effectively accommodating a type of local clustering.

5. Borrowing Information Across Studies and Centers

In biomedical studies, there is very commonly interest in combining information across data from different sources. Classical examples include multi-center studies, in which data are collected for individuals in different study centers, and meta-analysis, in which interest focuses on combining results for studies with similar study designs, endpoints and predictors. In recent years, there has also been increasing interest in attempting to borrow strength across data from disparate sources. For example, one may want to combine results from cell assay, animal and epidemiologic studies in drawing conclusions about the health effects of an environmental exposure.

In such settings, hierarchical models provide a standard tool for borrowing of information, but flexible approaches are needed to avoid inappropriate borrowing. In this section, I provide an overview of several hierarchical extensions of the DP that have been proposed for borrowing of information across data from different sources, including mixtures of DPs, dependent DPs, hierarchical DPs and nested DPs.

To provide motivation, I will focus on a multi-center study application. In particular, the National Collaborative Perinatal Project (NCPP) was a large prospective epidemiologic study conducted from 1959-1974. Pregnant women were enrolled through different study centers and were followed over time, with the children given examinations at birth, 4, 8 and 12 months, and 3, 4, 7 and 8 years. Although a variety of data were collected, I will focus here on gestational age at delivery (gad) and birth weight (bw), with $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2})'$ for the jth woman in study center i, where y_{ij1} =gad and y_{ij2} =bw.

The joint distribution of gad and bw is distinctly non-Gaussian and is not well characterized by a parametric model. Hence, one can consider the follow Gaussian mixture model:

$$\begin{aligned} \mathbf{y}_{ij} &\sim N_2(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}) \\ (\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}) &\sim P_i, \end{aligned} \tag{16}$$

where P_i is an unknown mixture distribution. This specification results in a joint distribution for gestational age at delivery and birth weight, which is specific to study center. Certainly it is well known in the reproductive epidemiology literature that pregnancy outcomes can vary substantially for women of different ethnicities and from different socio-economic groups. Hence, it is appealing to allow differences between study centers, since different centers may serve different types of women. However, it is also likely the case that the joint distribution of gad and bw is similar for different centers, so the question is how to borrow information in specifying a prior for $P = \{P_1, \ldots, P_n\}$.

Müller, Quintana and Rosner (2004) proposed an approach for inducing dependence in the P_i's through a mixture of independent DPs. In particular, under their specification,

$$P_i = \pi P_0^* + (1-\pi)P_i^*, \quad P_h^* \overset{iid}{\sim} DP(\alpha P_0), h = 0, 1, \ldots, n, \tag{17}$$

where P_0^* is a global component that is shared across the different centers, $0 \leq \pi \leq 1$ is a probability weight on the global component, and P_j^* is a local component allowing deviations for the jth study. If the joint distribution of gad and bw is very similar for the different study centers, then π will be close to one. The Müller, Quintana and Rosner (2004) approach is appealing in working well in practice and being simple to implement. In particular, a standard MCMC algorithm for a DPM can be used after modification to update latent indicators, $Z_{ij} \in \{0, 1, \ldots, n\}$, with $Z_{ij} = h$ denoting that $(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})$ are drawn from P_h^*. Dunson (2006) modified (17) to define a dynamic mixture of DPs (DMDP) appropriate for time series settings.

An alternative to (17), referred to as the hierarchical DP (HDP), was proposed by Teh et al. (2006). The HDP is specified as follows:

$$P_i \sim DP(\alpha P_0), \quad P_0 \sim DP(\gamma P_{00}), \tag{18}$$

so that the mixture distributions for the different study centers are assigned separate DP priors, with sharing induced through the common base measure P_0, which is also given a DP prior. Note that this specification has the property that all the P_i's will have the same atoms, while having distinct but dependent weights on these atoms. The shared atoms property differs from the specification in (17). However, because the subset of atoms in P_0 that are occupied by the subjects in the sample will differ across study centers, the HDP does effectively allow local clusters. The HDP is covered in detail in the chapter by Teh and Jordan.

Note that the specifications (17) and (18) both allow borrowing of information across study centers through incorporation of common atoms and dependent weights on these atoms. However, both formulations assume that $\Pr(P_i = P_{i'}) = 0$, except in limiting cases in which hyperparameters are chosen so that the distributions are identical for all study centers. In many applications, it is of interest to allow clustering of study centers into groups, with each group having a different nonparametric response distribution. For example, one could then identify medical centers with the same distribution of patient outcomes, while also identifying outlying centers.

With this goal in mind, Rodriguez, Dunson and Gelfand (2007) proposed the nested DP (nDP), which lets

$$P_i \sim \sum_{h=1}^{\infty} \pi_h \delta_{P_h^*}, \quad P_h^* \overset{iid}{\sim} DP(\gamma P_0), \tag{19}$$

where the weights $\boldsymbol{\pi} = \{\pi_h\}_{h=1}^{\infty}$ are as defined in (2). Note that this specification allows P_i to be exactly equal to $P_{i'}$ with prior probability $1/(1+\alpha)$. When $P_i = P_{i'}$, the joint distribution of gad and bw in center i is identical to that for center i', and these centers are clustered together. Hence, unlike the HDP, which clusters patients within and across study centers while assuming study centers are distinct, the nDP allows clustering of both study centers

and patients within centers. Such clustering may be of direct interest or may be simply a tool for flexible borrowing of information in estimating the center-specific distributions.

6. Flexible Modeling of Conditional Distributions

6.1 *Motivation*

In many applications, it is of interest to study changes in the distribution of a response variable Y over time, for different spatial locations, or with variations in a vector of predictors, $\mathbf{x} = (x_1, \ldots, x_p)'$. Depending on the study design, the primary interest may be

1. Prediction of a health response Y for a new subject given demographic and clinical predictors for that subject.

2. Inference on the impact of time, space or predictors on the conditional response distribution.

3. Inverse regression problems involving identification of predictor values associated with an adverse health response.

4. Clustering of subjects based on their health response while utilizing predictor information.

The methods reviewed in Section 5 can be used to address these interests when there is a single unordered categorical predictor, such as the study center. However, alternative methods are needed in the general case. This section reviews some approaches for modeling of predictor-dependent collections of distributions through the use of mixture models incorporating priors for collections of dependent random probability measures indexed by predictors. In particular, let

$$P_{\mathcal{X}} = \{P_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\} \sim \mathcal{P}, \tag{20}$$

where $P_{\mathbf{x}}$ denotes the random probability measure at location (predictor value) $x \in \mathcal{X}$, $\mathcal{X}$ is the sample space for the predictors, and $\mathcal{P}$ is the prior for the collection, $P_{\mathcal{X}}$. Here, I use the term predictor broadly to refer also to time and spatial location.

There are a wide variety of applications in which it is useful to incorporate priors for dependent collections of distributions. For example, one may be interested in modeling of conditional densities, $f(y \,|\, \mathbf{x})$. Revisiting the reproductive epidemiology application from Section 5, suppose that y_i is the gestational age at delivery for woman i and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ is a vector of predictors, including age of the woman and blood level of DDE, a persistent metabolite of the pesticide DDT. Then, it is of interest to assess how the risk of premature delivery, corresponding to the left tail of the distribution of gestational age at delivery, changes with increasing DDE exposure adjusting for age. In making such assessments, it is appealing to limit parametric assumptions, and avoid the common epidemiologic practice of reducing information on gestational age at delivery (gad) to a 0/1 indicator of gad ≤ 37 weeks.

To solve this problem, one can consider a mixture model of the form:

$$f(y \,|\, \mathbf{x}) = \int \int g(y; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) dP_{\mathbf{x}}(\boldsymbol{\theta}) d\pi(\boldsymbol{\phi}), \tag{21}$$

where $g(y; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$ is a parametric model for the conditional density of y given $\mathbf{x}$, and (21) allows deviations from this parametric model through nonparametric mixing. Expression (21) contains both parametric and nonparametric components, with the prior distribution for the parameters $\boldsymbol{\phi}$ treated as known, while the mixture distribution for $\boldsymbol{\theta}$ is nonparametric and predictor-dependent. Some possibilities for $g(y; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$ include $N(y; \mu, \sigma^2)$, with $\boldsymbol{\theta} = (\mu, \sigma^2)$, and $N(y; \mathbf{x}'\boldsymbol{\beta}, \sigma^2)$, with $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)$. The advantage of incorporating a regression component in $g(y; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$ is that a sparse structure is then favored through centering on a base parametric regression model. In this context, a sparser specification allows subjects to be allocated to few mixture components, while maintaining flexibility. In addition, we allow for interpolation across sparse data regions through the base parametric model, addressing the curse of dimensionality.

6.2 *Dependent Dirichlet Processes*

One possibility for $\mathcal{P}$ in (20) is the dependent DP (DDP) originally proposed by MacEachern (1999, 2001) (see also, De Iorio et al., 2004). In full generality, the DDP is specified as follows:

$$P_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\Theta_h(\mathbf{x})}, \quad \Theta_h \overset{iid}{\sim} P_0, \quad \forall \mathbf{x} \in \mathcal{X}, \tag{22}$$

where $\pi_h(\mathbf{x}) = V_h(\mathbf{x}) \prod_{l<h} \{1 - V_l(\mathbf{x})\}$, for $h = 1, \ldots, \infty$, with the stick-breaking weights $\{V_h(\mathbf{x})\}_{h=1}^{\infty}$, at any fixed $\mathbf{x}$, consisting of independent draws from a beta$(1, \alpha)$ distribution. In addition, Θ_h is a stochastic process over $\mathcal{X}$ generated from P_0. For example, P_0 may correspond to a Gaussian process.

Due to complications involved in allowing the weights to depend on predictors, most applications of the DDP have assumed fixed weights, resulting in the specification:

$$P_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h \delta_{\Theta_h(\mathbf{x})}, \quad \Theta_h \overset{iid}{\sim} P_0, \quad \forall \mathbf{x} \in \mathcal{X}, \tag{23}$$

where $\boldsymbol{\pi} = \{\pi_h\}_{h=1}^{\infty}$ are as defined in (2). De Iorio et al. (2004) applied this specification to develop ANOVA-type models for collection of dependent distributions. Gelfand, Kottas and MacEachern (2005) applied the DDP in spatial data analysis applications.

One can also use the fixed π DDP in (23) to develop a method for conditional density modeling as in (21) by letting

$$f(y \,|\, \mathbf{x}) = \sum_{h=1}^{\infty} \pi_h N(y; \mu_h(\mathbf{x}), \sigma_h^2), \quad (\mu_h, \sigma_h^2) \sim P_0 = P_{01} \otimes P_{02} \tag{24}$$

where P_{01} is a Gaussian process over $\mathcal{X}$ and P_{02} is a probability measure on $\Re^+$ (e.g., corresponding to an inverse-gamma distribution). Note that (24) characterizes the conditional density using an infinite mixture of normals, with the component means varying differentially and non-linearly with predictors. This specification is a generalization of typical Gaussian process regression models, which would correspond to letting $\pi_1 = 1$, so that the mean varies flexibly while the residual density is assumed to be constant and Gaussian. In contrast, the DDP mixture of normals in (24) allows multi-modal residual densities that vary with $\mathbf{x}$.

It is useful to consider the gestational age at delivery application. In that case, there are likely a few dominate mixture components that are assigned most of the probability weight, with these components corresponding to early preterm birth, preterm birth and full term birth. Specification (24) assumes that the probability allocated to these components does not vary with age or dde, but the locations of the component can vary. For example, the mean of the preterm birth component can shift to earlier weeks as dde level increases to allow an increasing proportion of babies born too soon at higher exposures.

From this example, it is clear that the fixed π DDP may be overly-restrictive in that biologic constraints on the timing of gestation make it more realistic and interpretable to consider models with fixed locations but weights that depend on predictors. For example, if we have three dominate components corresponding to early preterm, preterm and full term, then a varying weights model would allow the probability of early preterm birth to increase as dde increases without necessarily changing the timing and hence the meaning of the early preterm component. In addition, a varying weights model is necessary to allow the probability that two subjects are assigned to the same cluster to depend on predictors.

Motivated by such considerations, Griffin and Steel (2006) proposed an order-based dependent DP, referred to as the π-DDP. The π-DDP allows for predictor-dependent weights in the DDP in a clever manner by allowing the ordering in the stick-breaking weights to depend on predictors. In more recent work, Griffin and Steel (2007) proposed a simplification of the π-DDP, which they used for modeling of the residual component in a flexible regression model. The resulting approach is referred to as the Dirichlet process regression smoother (DPRS).

An alternative approach that was developed for spatial applications in which it is appealing to allow the weights to vary spatially, was proposed by Duan, Guindani and Gelfand (2007). This approach relies on an approach which places a stochastic process on the weights, which is carefully specified so that the marginal distributions at any fixed spatial location maintain the DP stick-breaking form. De la Cruz-Mesia, Quintana and Müller (2007) recently proposed an alternative extension of the DDP for classification based on longitudinal markers. Their approach incorporated dependence in the random effects distribution across groups.

6.3 *Kernel-Based Approaches*

As an alternative to the DDP, Dunson et al. (2007) proposed an approach based on kernel-weighted mixtures of independent DP basis components. This approach is conceptually related to the kernel regression approach described in Section 3.4 and expression (12). However, instead of specifying a prior for a single function, the goal is to specify a prior for an uncountable collection of predictor-dependent random probability measures. To motivate the general approach, first consider the case in which there is an single continuous predictor with support on [0,1] and interest focuses on modeling the conditional density $f(y|x)$, for all $x \in [0,1]$. In this case, a simple model would let

$$f(y|x) = \frac{1}{K(x,0)+K(x,1)}\big\{K(x,0)f_0^*(y) + K(x,1)f_1^*(y)\big\}, \tag{25}$$

where K is a kernel, with $K(x,x)=1$ and $K(x,x')$ decreasing as the distance between x and x' increases, $f_0^*(y)$ is an unknown basis density located at $x=0$, and $f_1^*(y)$ is an unknown

basis density located at $x = 1$. For example, K may correspond to a Gaussian kernel, and f_0^*, f_1^* to DP mixtures of normals.

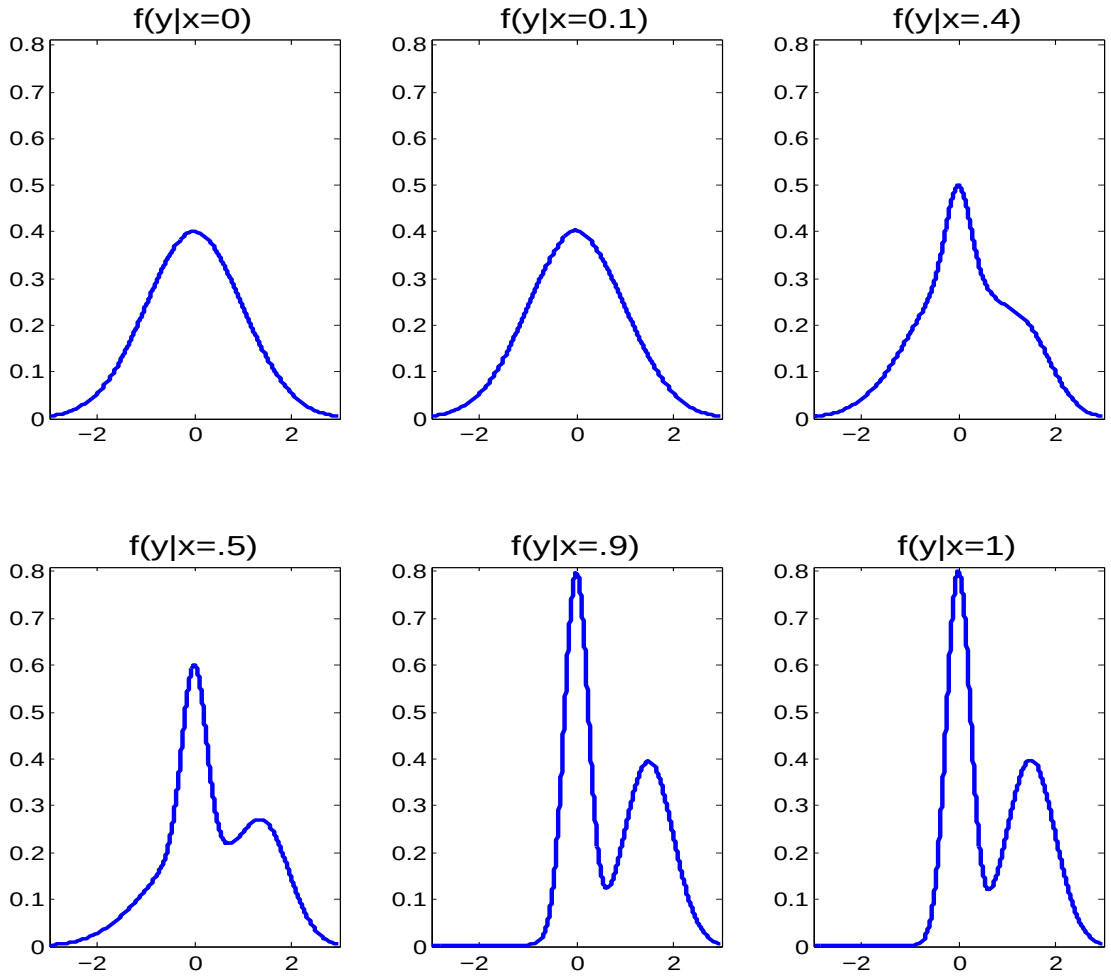

Figure 5. Plot illustrating the kernel mixtures approach in the simple case in which $x_i \in \{0, 1\}$ for all subjects in the sample, and a normal kernel with standard deviation 0.3 is chosen.

Figure 5 provides an example of the behavior of (25), letting f_0^* correspond to the standard normal density, f_1^* to a mixture of two normals, and K to the Gaussian kernel with standard deviation 0.3. As should be clear, using the mixture model (25) results in a

smoothly morphing profile of conditional densities, with the densities very close for similar predictor values. In order to generalize the kernel mixture approach of expression (25) to allow multiple predictors and more flexibility, one can potentially allow unknown numbers and locations of basis densities, similarly to allowing unknown numbers and locations of knots in a spline model.

Dunson et al. (2007) instead proposed a default approach in which DP bases were automatically placed at the sample predictor values, with each of these bases assigned a weight controlling it's importance. In particular, their proposed weighted mixture of DPs (WMDP) prior had the form:

$$P_{\mathbf{x}} = \sum_{i=1}^{n} \left(\frac{\gamma_i K(\mathbf{x}, \mathbf{x}_i)}{\sum_{l=1}^{n} \gamma_l K(\mathbf{x}, \mathbf{x}_l)} \right) P_i^*, \quad P_i^* \overset{iid}{\sim} DP(\alpha P_0), \quad \forall \mathbf{x} \in \mathcal{X}, \tag{26}$$

where $i = 1, \ldots, n$ indexes subjects in the sample, $K : \mathcal{X} \times \mathcal{X} \to 1$ is a bounded kernel (e.g., Gaussian), $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$ is a vector of weights, and $\{G_i^*\}_{i=1}^n$ are independent DP basis measures.

Dunson et al. (2007) applied the WMDP approach for density regression through use of the model,

$$f(y \,|\, \mathbf{x}) = \int \int N(y; \mathbf{x}'\boldsymbol{\beta}, \sigma^2) dP_{\mathbf{x}}(\boldsymbol{\beta}) d\pi(\sigma^2), \tag{27}$$

which is a mixture of linear regression models with predictor-dependent weights. This model is related to hierarchical mixtures-of-experts models (Jordan and Jacobs, 1994), which are widely used in the machine learning literature, but instead of assuming a finite number of experts (i.e., mixture components), the number of experts is infinite. In addition, instead of a probabilistic decision tree for the weights, a kernel model is used.

From a Bayesian perspective, (26) has an unappealing sample-dependence property, so that the approach is not fully Bayesian and lacks coherent updating and marginalization properties. For this reason, it is useful to consider alternative priors that borrow some of the positive characteristics of the kernel weighted specification but without the sample-dependence. One such prior is the kernel stick-breaking process (KSBP) (Dunson and Park, 2007), which modifies the DP stick-breaking specification shown in (2) as follows:

$$P_{\mathbf{x}} = \sum_{h=1}^{\infty} V_h K_{\psi_h}(\mathbf{x}, \Gamma_h) \prod_{l<h} \left\{ V_l K_{\psi_l}(\mathbf{x}, \Gamma_l) \right\} P_h^*, \quad P_h^* \overset{iid}{\sim} DP(\alpha P_0), \tag{28}$$

where $V_h \overset{iid}{\sim} \text{beta}(1, \lambda)$, K_ψ denotes a kernel having bandwidth ψ, $\{\psi_h\}_{h=1}^{\infty}$ is an infinite sequence of kernel bandwidths sampled from G, and $\{\Gamma_h\}_{h=1}^{\infty}$ is an infinite sequence of kernel locations sampled from H.

The KSBP places random basis probability measures at an infinite sequence of random locations, with the weights assigned to these bases in the formulation for $P_{\mathbf{x}}$ decreasing stochastically with the index h and the distance from $\mathbf{x}$. In using the KSBP as a prior for the mixture distributions in (27), one obtains a flexible, sparseness-favoring structure for conditional distribution modeling. In particular, if the linear regression model provides a

good approximation, then the tendency is to assign most of the weight to few components and effectively *collapse* back to the base parametric normal linear model. This is seen quite dramatically in simulations under the normal linear model in which results obtained in the KSBP mixture analysis and the parametric model analysis are very similar. However, the KSBP mixture model does an excellent job at adaptively capturing dramatic deviations from the normal linear model even when all the subjects are assigned to a modest number of basis locations (e.g., fewer than 10).

The sparseness-favoring property is crucial in obtaining good performance, since the curse of dimensionality makes it very difficult to reliably estimate conditional distributions even with few predictors. One does not obtain nearly as good performance using a KSBP mixture of Gaussians without the regression structure in the Gaussian kernel as in (27). This is partly due to the fact that the base normal linear model allows one to interpolate across sparse data regions much more reliably than a normal with mean that does not vary with predictors.

From a practical perspective, the KSBP can be implemented easily in a wide variety of settings using a simple MCMC algorithm, and produces flexible but sparse models of conditional distributions. In addition, the KSBP results in predictor-dependent clustering of subjects. In the conditional density estimation setting, one can also obtain estimates of quantile regression functions directly from the MCMC output. The KSBP has been used for classification, multitask learning and modeling of multivariate count data in unpublished work.

6.4 *Conditional Distribution Modeling Through DPMs*

Prior to the work on DDPs and kernel-based approaches for modeling of dependent collections of random probability measures, Müller, Erkanli and West (1996) proposed a simple and clever approach for inducing a prior on $\mathrm{E}(y \,|\, \mathbf{x})$ through a joint DP mixture of Gaussians model for $\mathbf{z} = (y, \mathbf{x}')'$. Although they did not consider conditional density estimation, this approach also induces a prior on $f(y \,|\, \mathbf{x})$, for all $y \in \Re, \mathbf{x} \in \Re^p$. In recent work, Rodriguez, Dunson and Gelfand (2007b) showed that the Müller, Erkanli and West (1996) approach results in pointwise consistent estimates of $\mathrm{E}(y \,|\, \mathbf{x})$ under some mild conditions, and the approach can be adapted for functional data analysis.

Given that the Müller, Erkanli and West (MEW) (1996) approach can be implemented routinely through direct use of software for Bayesian multivariate density estimation using DP mixtures of Gaussians, a natural question is what is gained by using the approaches described in Sections 6.2 and 6.3 for conditional density estimation. To address this question, I start by contrasting the two types of approaches. The methods described in Sections 6.2-6.3 provide priors for collections of unknown distributions indexed by spatial location, time and/or predictors. The resulting specification is defined conditionally on the predictors (or time/space). In contrast, the MEW approach relies on joint modeling of the predictors and response to induce a prior on the conditional distributions. Clearly, this only makes sense if the predictors can be considered as random variables, which rules out time, space and predictors that correspond to design points. However, for many observational studies, all the predictors can be considered as random variables.

In such settings, the MEW approach is still conceptually quite different from the conditional approaches of Section 6.2-6.3. All the approaches induce clustering of subjects, and the

MEW approach is similar to the π-DDP, WMDP and KSBP in allowing predictor-dependent clustering. The predictor-dependent clustering of the MEW approach arises through joint clustering of the response and predictors. This implies that the MEW approach will introduce new clusters to better fit the distribution of the predictors even if such clusters are not necessary from the perspective of producing a good fit to the response data. For example, in simulations from a normal linear regression model in which the predictors have non-Gaussian continuous densities, we have observed a tendency of the MEW approach to allocate individuals to several dominate clusters to better fit the predictor distribution. In contrast, the KSBP will tend to collapse on a single dominate cluster containing all but a few subjects in such cases, as a single cluster is sufficient to fit the response distribution, and the KSBP does not model the predictor distribution.

Although the sensitivity of clustering to the predictor distribution has some unappealing characteristics, there are some potential benefits. Firstly, the approach is useful for inverse regression and calibration problems. Griffin and Holmes (2007) recently proposed a MEW-type approach for Bayesian nonparametric calibration in spatial epidemiology. Also, in providing a joint nonparametric model for predictors and response variables, the approach can automatically accommodate predictors that are missing at random through a simple step for imputing these predictors from their full conditional posterior distributions during the MCMC algorithm. Finally, instead of relying entirely on information in the response distribution, the approach tends to automatically change the slope of the regression in regions of the predictor space at which there is a change in the predictor distribution. This may be particularly appealing in semi-supervised learning settings in which the response (labels) are only available for a small subset of the subjects. In order to allow categorical responses and/or predictors, the MEW approach can be easily adapted to use an underlying normal DPM of Gaussians.

6.5 *Reproductive Epidemiology Application*

As an illustration, consider an application to modeling of gestational age at delivery (gad) data from a study of Longnecker et al. (2001) using the KSBP. In epidemiologic studies of premature delivery, it is standard practice to dichotomize the data on gad using 37 weeks as the cutoff for defining preterm delivery. Then, the risk of preterm birth is modeled as a function of exposures of interest and potential confounders using a logistic regression model. Similar analyses are common in studies that collect a continuous health response when the interest focuses on the effect of predictors on the risk of an adverse response. For example, in assessing factors predictive of obesity, one typically dichotomizes body mass index (bmi) to obtain a 0/1 indicator of obesity instead of analyzing the right tail of the bmi distribution. In such settings, adverse responses correspond to values in one or both of the tails of the response density, and it is certainly appealing to avoid sensitivity to arbitrary cutoffs.

Typical regression models that focus on predictor effects on the mean or median or the distribution are clearly not appropriate for characterizing changes in the tails unless the residual distribution is truly homoscedastic. In epidemiology it is often not plausible biologically for the residual distribution to satisfy such an assumption. In particular, most health conditions are multi-factorial, having both genetic and environmental risk factors, with many or most of these factors unmeasured. Hence, in the presence of interactions between measured and unmeasured predictors, we would really expect the response distribution to change

in shape as the values of the predictors vary. Biological constraints on values of the response also play a role.

In the gestational age at delivery setting, such biological constraints make it very unlikely that gestation continues much beyond 45 weeks, because the baby is getting so large at that point. Hence, the risk factors for premature delivery are very unlikely to simply result in a shift in the mean gad, but are more likely to impact the relative proportions of women having gads in the early preterm, preterm or full term intervals. Hence, a mixture of three normals, with the predictor-dependent weights, may be natural for modeling of the gad distribution. However, as we are not certain *a priori* that three components are sufficient, it is appealing to consider a nonparametric Bayes approach that allows infinitely-many components in the general population, while allowing predictor-dependent weights. The KSBP mixture models proposed by Dunson and Park (2007) provide such a framework.

Note that when the focus is on predictor effects on the tails of a distribution, one can potentially apply quantile regression methods. However, most quantile regression methods allow for modeling of a single arbitrarily-chosen quantile (e.g., the 10th percentile) instead or providing a general framework for modeling of all quantiles coherently. By using a nonparametric Bayes conditional distribution modeling approach, one can do inferences on shifts in the entire tail of the distribution instead of focusing narrowly on a selected quantile.

Returning to the Longnecker et al. (2007) application, there were 2313 women in the study, and we implemented the KSBP for model (27), with y_i=gad and $\mathbf{x}_i = (1, \mathrm{dde}_i, \mathrm{age}_i)'$. For the reasons discussed in detail at the end of Section 2.1, we normalized $y_i, \mathrm{dde}_i, \mathrm{age}_i$ prior to analysis, and then chose the base measure in the KSBP to correspond to the unit information-type prior, $N_2(\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1}/n)$. In addition, we fixed $\alpha = 1$ and $\lambda = 1$ to favor a few clusters, and used a Gaussian kernel, with a single kernel precision, $\psi_h = \psi$, assigned a log-normal prior. The retrospective MCMC algorithm described in Dunson and Park (2007) was implemented, with 22,000 iterations collected after a burn-in of 8,000 iterations.

It is worth commenting on convergence assessments for nonparametric Bayes mixture models, with the KSBP providing one example. In particular, due to the label switching problem, the index on the different mixture components can change meaning across the MCMC iterates. This leads to apparently poor mixing in many cases when one examines traces plots of cluster-specific parameters. However, in the analysis of the Longnecker et al. (2001) pregnancy outcome data and in general applications in which the focus is on density estimation, conditional density estimation or inferences on unknowns that are not cluster-specific, one could compellingly argue that the label-switching problem is in fact not a problem at all. It is very commonly the case that trace plots of unknowns that are not cluster-specific exhibit good rates of convergence and mixing properties, while trace plots for cluster-specific parameters show high autocorrelation and fail standard convergence tests. As long as one is not interested in cluster-specific estimates and inferences, this is no cause for concern, as it just represents the cluster index identifiability problem. In KSBP applications, we monitored the value of the conditional density at different locations, and observed good mixing and rapid apparent convergence.

FIgure 6 shows the raw data on dde and gad for the women in the Longnecker et al. (2001) study. The solid line is the posterior mean of the expected value of gad conditionally on dde. This illustrates that the KSBP mixture model induces a flexible non-linear mean regression model, while also allowing the residual distribution to vary with predictors. The

dotted lines correspond to 99% pointwise credible intervals. There is an approximately linear decrease in the mean gad, with the credible intervals becoming considerably wider for high dde values where data are sparse.

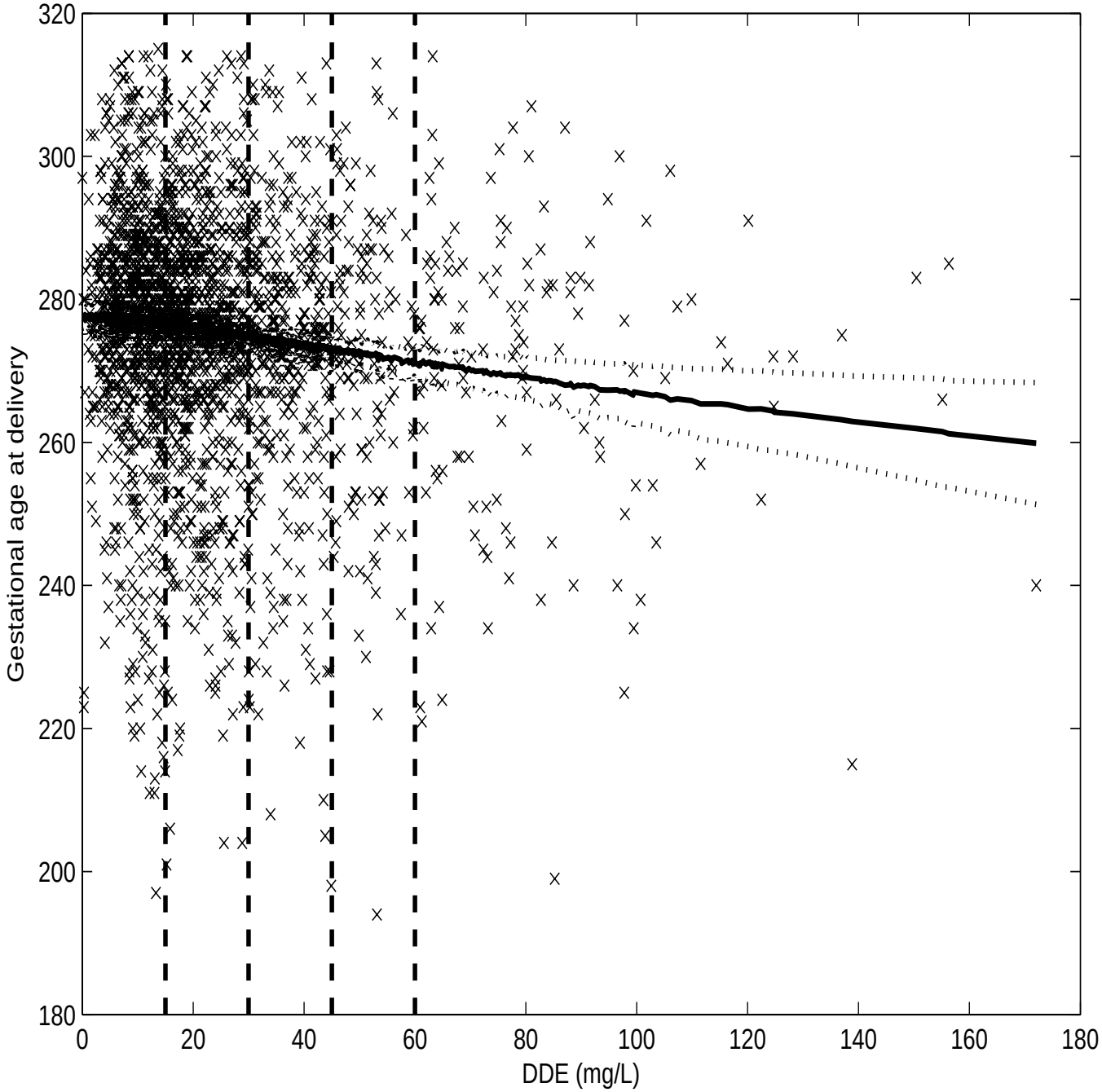

Figure 6. Raw data on dde and gestational age at delivery for 2313 women in the Longnecker et al. (2001) sub-study of the NCPP. Vertical dashed lines are quintiles of the empirical distribution of DDE, the solid line is the posterior mean regression curve, and the dotted lines are 99% pointwise credible intervals.

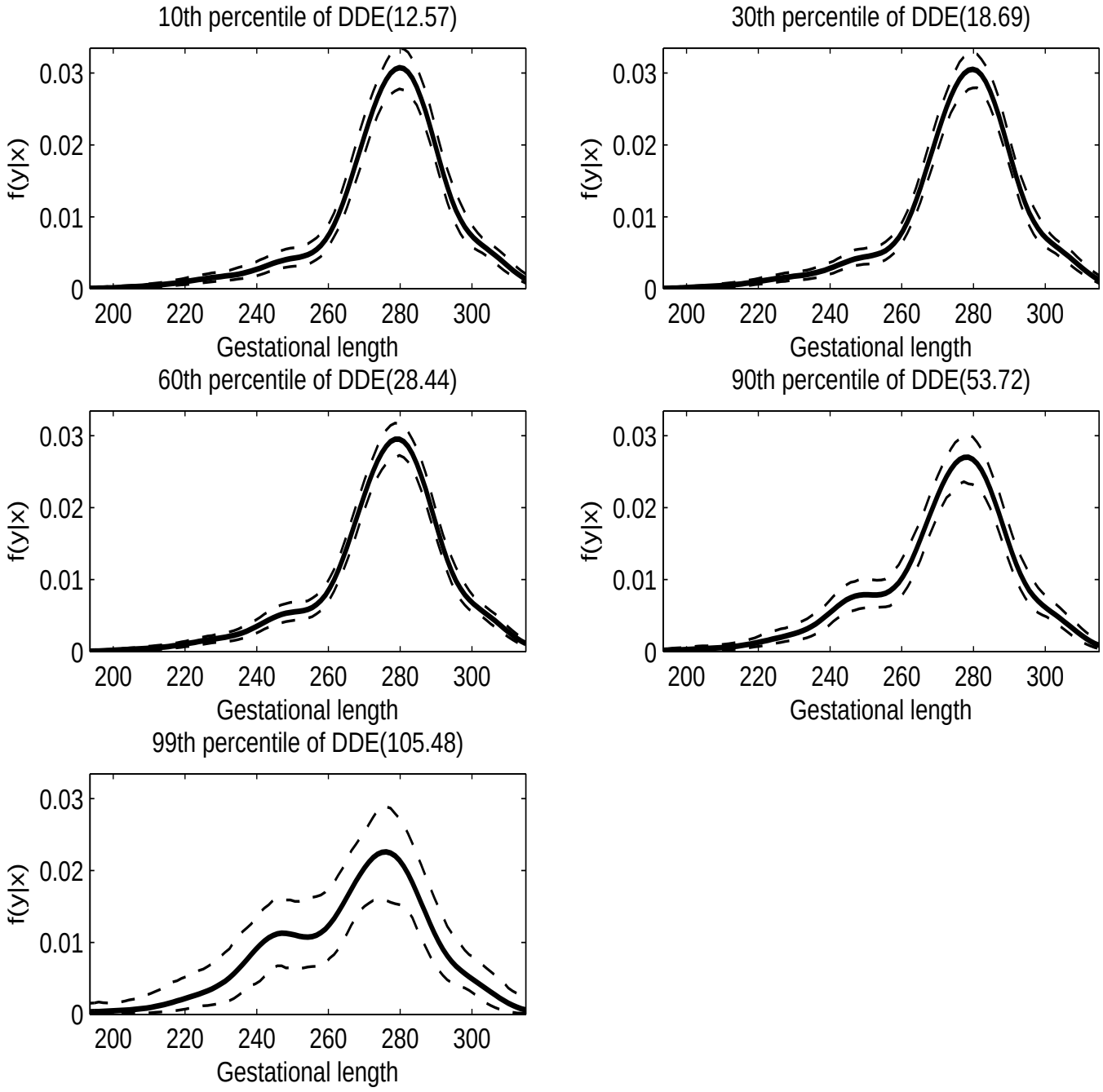

Figure 7. Estimated conditional densities of gestational age at delivery in days for different dde values. Solid lines are posterior means and dashed lines are 99% credible intervals.

Figure 7 shows the estimated conditional densities of gad in days for different dde values, with posterior means shown in solid lines and 99% pointwise credible intervals shown with dashed lines. These figures suggest that the left tail corresponding to premature deliveries is increasingly fat as dde dose increases. However, it is a bit hard to gauge significance of these results based on observing a sequence of conditional density estimates. Figure 8 provides dose response curves for the probability gad falls below different cutoffs, including (a) 33,

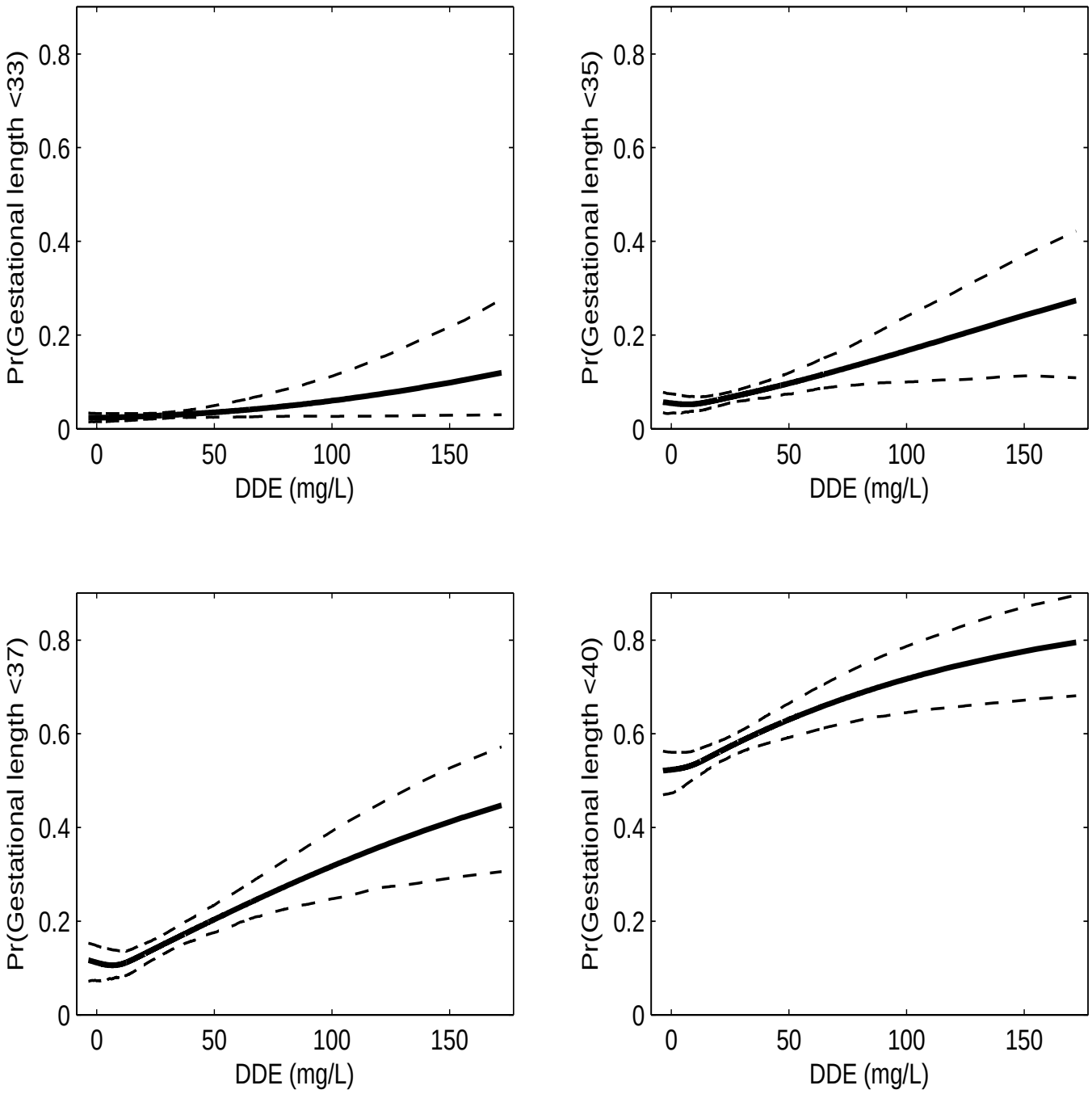

Figure 8. Estimated dose response curves for the probability gestational age at delivery is less than (a) 33, (b) 35, (c) 37 or (d) 40 weeks. Solid lines are posterior means and dashed lines are 99% credible intervals

(b) 35, (c) 37 or (d) 40 weeks. Again, solid lines are posterior means and dashed lines are 99% credible intervals. From these plots, it is clear that risk of premature delivery increases with level of dde. It appears that dde increases risk of early preterm birth prior to 33 weeks, which is an interesting finding given that such births represent a much more adverse response in terms of short and long term morbidity compared with 37 week births.

7. Bioinformatics

In recent years there has been a paradigm shift in biostatistics, and it is now standard to be faced with very high-dimensional data. Hence, there is clearly a need for automated approaches for flexible dimensionality reduction and discovery of sparse latent structure underlying very high-dimensional data. Mixture models have proven to be an extremely useful tool in such settings. For example, there is a vast literature on methods for high-dimensional variable selection using mixture priors, with one component concentrated at zero and another component being more diffuse. Nonparametric Bayes methods have proven extremely useful in this setting in providing a highly flexible framework for limiting sensitivity to arbitrary assumptions made in parametric modeling, such as knowledge of the number of mixture components. In this section, we review some of this work to provide a flavor for the possible applications.

7.1 *Modeling of Differential Gene Expression*

Much of the increased interested in large p, small n problems, which started approximately a decade ago, was initiated by the development and rapidly increasing use of microarray technology for measuring expression levels of large numbers of genes. There has been a particular focus on methods for clustering of gene expression profiles and for identifying genes that are differentially expressed between two groups, with these groups often representing normal and diseased or tumor tissue.

Gene expression analyses are particularly suited for nonparametric Bayes analyses due to the large sample size in terms of the number of genes and to lack of knowledge of good parametric models for approximating the joint distribution of the gene expression values. Nonparametric Bayes methods provide a flexible machinery for characterizing the complex and high-dimensional expression values, inducing a sparse latent structure through partitioning genes into clusters. However, there are some limitations relative to simpler methods in terms of ease of interpretation and computational expense.

Medvedovic and Sivaganesan (2002) proposed a Dirichlet process mixture (DPM) model for clustering genes with similar expression patterns, which they accomplished by calculating the pairwise posterior probabilities that two genes are in the same cluster from the Gibbs sampler output. Qin (2006) proposed a modified approach that relied on an iterative weighted Chinese restaurant seating scheme, designed so that the optimal number of clusters can be estimated simultaneously with assignment to clusters in an efficient manner. Medvedovic, Yeung and Bumgarner (2004) generalize the Medvedovic and Sivaganesan (2002) approach to data containing experimental replicates. Kim, Tadesse and Vannucci (2006) proposed an approach that allowed for selection of the variables to cluster upon in the DPM model. Xiang, Qin and He (2007) developed CRCView, which is a web server providing an easy to use approach for analysis and visualization of microarray gene expression data based on a DPM approach.

Do, Müller and Tang (2005) proposed a DPM of normals for the distribution of gene intensities under different conditions. Their focus was on modeling of differential gene expression between two groups, a problem which has been addressed using a variety of parametric and nonparametric empirical Bayes approaches (Newton et al., 2004). However, Do et al. (2005) demonstrated advantages of the fully Bayes nonparametric approach, including allowance for estimation of posterior expected false discovery rates.

7.2 *Analyzing Polymorphisms & Haplotypes*

In addition to gene expression data, there has been substantial interest in identifying genotypes that are predictive of an increased risk of disease. Data are now routinely collected containing the genotypes at a large number of locations (or loci) along the genome at which there is variability among individuals in the population. Such data are commonly referred to as single nucleotide polymorphisms (SNPs), with a single SNP consisting of a combination of amino acids, with one on the chromosome inherited from the mother and one from the father. Because new SNP chips allow investigators to routinely collect data from hundreds of thousands of loci, such data present quite a challenge to the statistician.

In many cases, instead of casting the net very broadly in searching for disease genes and genotypes, investigators narrow down their search to genes in a specific pathway hypothesized to play a key role in disease risk. However, even in this case, there may be many SNPs under consideration. To provide an example, Mulherin-Engel et al. (2005) related SNPs in cytokine gene regulatory regions to risk for spontaneous preterm birth. The number of loci per cytokine at which SNP data were collected ranged from 1-3, with 22 total across the 12 cytokines. At each loci there are 3 possible genotypes. Hence, in seeking to identify genotypes predictive of an increased risk of preterm birth, there is a very high-dimensional set of models under consideration.

Motivated by this application, Dunson, Herring and Engel (2007) proposed a multi-level DP prior that allowed for borrowing of information across functionally-related genes, while also incorporating a variable selection component. This approach automatically grouped genotypes into null and non-null clusters according the genotypes impact on the risk of disease. In unpublished work, we have found that this type of approach scales nicely to problems involving thousands of genotypes, providing clearly improved performance relative to parametric variable selection mixture priors. One of the reasons for this success is that most variable selection mixture priors shrink the non-null coefficients towards zero, while the DP-based approach allows these coefficients to be shrunk towards other coefficients having similar values.

In order to address the dimensionality problem faced in searching for disease genotypes relying on high-dimensional SNP data, many articles have proposed haplotype-based analyses. Haplotypes represent a sequence of amino acids along a chromosome inherited from one parent. Due to linkage, it tends to be the case that the number of haplotypes observed in the population is substantially less than the number possible. Hence, by using haplotypes as predictors instead of the genotypes formed by a set of SNPs, one clearly obtains a reduction in dimensionality. However, the problem is that current genotyping technology does not allow SNP data to be converted into haplotypes, because the data are un-phased, meaning that the amino acid pairs cannot be allocated to the chromosome for their parent of origin.

This is fundamentally a missing data problem, and a number of approaches have been proposed for imputing the missing haplotypes given the SNP data. Xing, Jordan and Sharan (2007) propose a DPM model to address this missing data problem, with the mixture components corresponding to the pool of haplotypes in the population. Xing et al. (2006) generalize this approach to the multiple population setting through the use of a hierarchical DP (Teh et al., 2006). Xing and Sohn (2007) proposed an alternative approach that used a hidden Markov DP to jointly model genetic recombinations among the founders of a population and subsequent coalescence and mutation events. The goal of this approach is

to identify recombination hotspots and to infer ancestral genetic patterns. An alternative approach for haplotype inference using a Bayesian hidden Markov model was independently developed by Sun et al. (2007).

7.3 *New Species Discovery*

There has clearly been an explosion in the types of high-dimensional data generated, and nonparametric Bayes methods have seen greatly increasing use as a tool for bioinformatics. In addition to the applications presented in Section 7.1-7.2, one very interesting application is to expressed sequence tag (EST) analysis. ESTs provide a useful tool for gene identification in an organism. In searching for genes using this technology, a number of interesting design questions arise. In particular, after obtaining a preliminary EST sample, scientists would like to estimate the expected number of new genes that would be detected from a further sample of a given size. Such information is critical in making decisions about the number of additional samples to sequence.

Lijoi et al. (2007a) addressed this problem using a nonparametric Bayes methodology for estimating the probability of discovery of a new species (Lijoi et al., 2007b). In particular, Lijoi et al. (2007b) derive a closed-form expression for a nonparametric Bayes estimator for the probability of discovery. Their approach is based on a class of priors for species sampling problems that induce Gibbs-type random partitions.

8. Nonparametric Hypothesis Testing

Most of the nonparametric Bayes literature has focused on approaches for estimation under a particular model, and there has been relatively limited consideration of hypothesis testing problems. In biomedical studies, hypothesis testing is often of primary interest. For example, in clinical trials, basic science and toxicology studies, the primary focus is typically on testing the null hypothesis of equalities in the response distribution between treatment groups against the alternative that there are differences. Estimation is often a secondary interest. In this section we review some of the work on nonparametric Bayes hypothesis testing.

Motivated by the problem of testing of equalities between groups in a study with multiple treatment groups, Gopalan and Berry (1998) proposed an approach to adjust for multiple comparisons through use of a DP prior. In particular, letting y_{hi} denote a continuous health response for the ith individual in group h, for $h = 1, \ldots, p$, let $y_{hi} \sim N(\mu_h, \sigma^2)$. Then, the interest is in studying local hypotheses:

$$\mathrm{H}_{0,hl} : \mu_h = \mu_l \quad \text{versus} \quad \mathrm{H}_{1,hl} : \mu_h \neq \mu_l.$$

Clearly, the number of such hypotheses increases rapidly with p, so one faces a multiple comparisons problem. Gopalan and Berry (1998) proposed letting $\mu_h \overset{iid}{\sim} P$, with $P \sim DP(\alpha P_0)$. Then, from basic properties of the DP, $\Pr(\mathrm{H}_{0,hl}) = 1/(1+\alpha)$. Using standard algorithms for posterior computation in DPM models, one can obtain estimates of the posterior hypothesis probabilities.

Motivated by epidemiologic studies having many predictors that may be highly correlated, MacLehose et al. (2006) proposed an alternative approach which assumed that the regression parameters for the different predictors in a generalized linear model, $\beta_1, \ldots, \beta_p$, were sampled from the prior:

$$\beta_j \overset{iid}{\sim} P = \pi_0 \delta_0 + (1 - \pi_0) P^*, \quad P^* \sim DP(\alpha P_0), \tag{29}$$

where π_0 is the prior probability that the jth predictor has a zero coefficient so can be excluded from the model. This approach allows one to calculate posterior probabilities of $H_{0j} : \beta_j = 0$ versus $H_{0j} : \beta_j \neq 0$, while clustering the β_j's for the non-null predictors. MacLehose et al. (2006) demonstrated improved performance relative to parametric variable selection priors that replace P^* with a normal distribution.

Note that, whenever using Bayes factors and posterior probabilities as a basis for hypothesis testing, it is important to keep in mind the well-known sensitivity to the prior. This sensitivity occurs regardless of whether one is considering a parametric or nonparametric model. Using expression (29) as an example, note that commonly-used parametric variable selection mixture priors would let $P^* \equiv P_0$, with P_0 chosen to correspond to a normal or heavier-tailed density centered at zero. In this parametric special case of (29), choosing a very high variance P_0 will tend to lead to placing high posterior probability on small models having the predictors excluded. In the nonparametric case, this same behavior will occur for very high variance P_0. Hence, far from being non-informative, a high variance P_0 instead corresponds to a very informative prior that overly-favors small models. Motivated by applications to logistic regression selection in genetics and epidemiology, MacLehose et al. (2006) proposed to address the problem of specification of P_0 by using an informative choice, motivated by prior knowledge of plausible ranges for predictor effects. In the absence of such knowledge, one can potentially use default choices of P_0 that are used in parametric variable selection settings, though the theoretical implications of such choices remain to be fully evaluated.

The MacLehose et al. (2006) approach utilizes a DP prior for dimensionality reduction and clustering of the parameters in a parametric model, while assuming that the response distribution belongs to a parametric family. In order to compare two different nonparametric models, one can potentially calculate the marginal likelihoods for each model and then obtain a Bayes factor. For example, suppose that an experiment is run having two groups, and the focus is on testing equalities in the distributions between the two groups without assuming a parametric model for the response distribution or for the types of changes between the two groups. Then, one could fit a null model, which combines the two groups and uses a DPM of Gaussians to characterize the unknown response distribution, and an alternative model, which uses separate DPMs of Gaussians for the two groups. Using the method of Basu and Chib (2003) for estimating marginal likelihoods for DPMs, a Bayes factor can then be obtained.

When there are multiple treatment groups, such an approach faces practical difficulties, since it involves running separate MCMC algorithms for each comparison of interest. In addition, if the treatment groups correspond to increasingly doses, then it is appealing to apply an approach that borrows information across the groups in characterizing the unknown dose group-specific distributions. Pennell and Dunson (2007) propose such an approach based on a dynamic mixture of DPs, which allows adjacent dose groups to be effectively identical. Dunson and Peddada (2007) proposed an alternative approach for testing of partial stochastic ordering among multiple groups using a restricted dependent DP.

In parametric models, Bayesian hypothesis testing and model selection makes the assumption that one of the models under consideration is true. Such an assumption is often viewed as unrealistic, since it seems unlikely that any parametric model is more than a rough approximation of the truth. Walker and Gutiérrez-Pena (2007) proposed an approach that

allows for coherent comparisons of parametric models, while allowing for the fact that none of the models under consideration are true by modeling the truth using a nonparametric Bayes approach. There has also been some focus in the literature on using a nonparametric Bayes alternative for goodness-of-fit testing of a parametric model. To derive Bayes factors, Carota and Parmigiani (1996) embedded the parametric model in a nonparametric alternative characterized as a mixture of Dirichlet processes. They obtained disturbing results showing that results are entirely driven by the occurrence of ties. DP mixtures or Polya trees can be used to bypass this problem, with Berger and Guglielmi (2001) providing a Polya tree-based approach.

8. Discussion

This chapter has provided a brief overview of some recent nonparametric Bayes work that is motivated by or relevant to biostatistical applications. The literature on the intersection between nonparametric Bayes and biostatistics is increasingly vast, so I have focused on only a narrow slice of this literature. In focusing on approaches that explicitly incorporate random probability measures, such as the Dirichlet process and extensions, I have ignored a rich literature on nonparametric Bayes methods for survival analysis. In addition, there are certainly many extremely interesting papers that I have overlooked, have not had space to cover, or appeared after completing this chapter.

Nonparametric Bayes is clearly a new and exciting field, with many important problems remaining to be worked out. One clear problem is computational time and complexity. For example, noting the great potential of the Bayesian approach in pattern recognition for bioinformatics, Corander (2006) also notes the limitation of current Bayesian computational strategies in high-dimensional settings, calling for the evolution of Bayesian computational strategies to meet this need. A number of fast approximations to the posterior have been proposed for DPMs and other nonparametric Bayes models, relying on variational approximations and other strategies. However, such approximations often perform poorly, and there is no general theory or methodology available for assessing approximation accuracy. Some of these issues will be covered in detail in the chapter by Teh and Jordan.

Another issue is hyperparameter choice. In having infinitely-many parameters, nonparametric Bayes methods tend to be heavily parameterized and to require specification of a number of hyperparameters. This subjective component is somewhat counter to the nonparametric spirit of avoiding assumptions, and the hyperparameters can have a subtle role, making them difficult to elicit. For this reason, it is appealing to have new approaches for subjective prior elicitation that allow for the incorporation of scientific and background knowledge. Typically, such knowledge does not take the form of guesses at parameter values, but instead may be represented as functional constraints or knowledge of plausible values for functions of the parameters. It is also appealing to have default priors approaches available for routine use.

There is also a pressing need for new methods and ways of thinking about model selection in nonparametric Bayes. For example, for a given problem (e.g., conditional distribution estimation), one can write down many different nonparametric Bayes models, so how does the applied statistician go about choosing between the different possibilities? Often such a choice may be pragmatic, motivated by the available software for implementation and the plausibility of the results that are produced. However, it would be appealing to have formal

methods for routinely comparing competing approaches in terms of goodness-of-fit versus parsimony, asymptotic efficiency and other criteria. It may be that different formulations are primarily distinguished by sparseness-favoring properties, with "spareness-favoring" in this context meaning that the approach favors using relatively few parameters in characterizing the distributions and functions of interest.

Acknowledgments
The author would like to thank Peter Müller for helpful comments on a draft of this chapter.

References

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.O. 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.L. and Dunson, D.B. (2005). *Discussion Paper*, 2005-, Department of Statistical Science, Duke University.

Bigelow, J.L. and Dunson, D.B. (2007). Bayesian semiparametric joint models for functional predictors. *Journal of the American Statistical Association*, revision invited.

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

Blei, D.M. and Jordan, M.I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1, 121-144.

Brumback, B.A. and Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93, 961-976.

Burr, D. and Doss, H. (2005). A Bayesian semiparametric model for random-effects meta analysis. *Journal of the American Statistical Association*, 100, 242-251.

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

Carota, C. and Parmigiani, G. (1996). On Bayes factors for nonparametric alternatives. *Bayesian Statistics* 5, 507-511, editors J.M. Bernardo et al., Oxford University Press.

Carota, C. (2006). Some faults of the Bayes factor in nonparametric model selection. *Statistical Methods and Application*, 15, 37-42.

Chakraborty, S., Ghosh, M. and Mallick, B. (2005). Bayesian non-linear regression for large p, small n problems. *Journal of the American Statistical Association*, under revision.

Corander, J. (2006). Is there a real Bayesian revolution in pattern recognition for bioinformatics. *Current Bioinformatics*, 1, 161-165.

Dahl, D.B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics*, Kim-Anh Do, Peter Mueller, Marina Vannucci (Eds.), Cambridge University Press.

Dahl, D.B. (2007). Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, under revision.

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

De la Cruz-Mesia, R., Quintana, F.A. and Müller, P. (2007). Semiparametric Bayesian classification with longitudinal markers. *Applied Statistics*, 56, 119-137.

Do, K.-A., Müller, P. and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Applied Statistics*, 54, 627-644.

Duan, J., Guindani, M. and Gelfand, A.E. (2007). Generalized spatial Dirichlet process models. *Biometrika*, to appear.

Dunson, D.B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, 7, 551-568.

Dunson, D.B., Herring, A.H. and Mulherin-Engel, S.M. (2007), "Bayesian Selection and Clustering of Polymorphisms in Functionally Related Genes," *Journal of the American Statistical Association*, to appear.

Dunson, D.B. and Park, J-H. (2007). Kernel stick-breaking processes. *Biometrika*, to appear.

Dunson, D.B. and Peddada, S.D. (2007). Bayesian nonparametric inference on stochastic ordering. *Biometrika*, under revision.

Dunson, D.B., Pillai, N. and Park, J-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society B*, 69, 163-183.

Dunson, D.B., Xue, Y. and Carin, L. (2007). The matrix stick-breaking process: Flexible Bayes meta analysis. *Journal of the American Statistical Association*, to appear.

Dunson, D.B., Yang, M. and Baird, D.D. (2007). Semiparametric Bayes hierarchical models with mean and variance constraints. *Discussion Paper*, 2007-08, Department of Statistical Science, Duke University.

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

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

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

Friedman, J.H. and Meulman, J.J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society*, 66, 815-849.

Gelfand, A.E., Kottas, A. and MacEachern, S.N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100, 1021-1035.

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

Green, P.J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28, 355-375.

Griffin, J.E. and Holmes, C.C. (2007). Bayesian nonparametric calibration with applications in spatial epidemiology. *Technical Report*, Institute of Mathematics, Statistics and Actuarial Science, University of Kent.

Griffin, J.E. and Steel, M.F.J. (2006). Order-based dependent Dirichlet process. *Journal of the American Statistical Association*, 101, 179-194.

Griffin, J.E. and Steel, M.F.J. (2007). Bayesian nonparametric modelling with the Dirichlet process regression smoother. *Technical Report*, Institute of Mathematics, Statistics and Actuarial Science, University of Kent.

Hanson, T. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101, 1548-1565.

Hoff, P.D. (2006). Model-based subspace clustering. *Bayesian Analysis*, 1, 321-344.

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 101, 179-194.

Ishwaran, H. and Takahara, G. (2002). Independent and identically distributed Monte Carlo algorithms for semiparametric linear mixed models. *Journal of the American Statistical Association*, 97, 1154-1166.

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

Jara, A. (2007). Applied Bayesian non- and semi-parametric inference using DPpackage. *Technical Report*, Biostatistical Center, Catholic University of Leuven.

Jasra, A., Holmes, C.C. and Stephens, D.A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20, 50-67.

Jeffreys, W. and Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Statistician*, 80, 64-72.

Jones, B.L., Nagin, D.S. and Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research*, 29, 374-393.

Jordan, M.I. and Jacobs, R.A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6, 181-214.

Kim, S., Tadesse, M.G. and Vannucci, M. (2006), "Variable Selection in Clustering via Dirichlet Process Mixture Models," *Biometrika*, 93, 877-893.

Kimeldorf, G.S. and Wahba, G. (1971). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41, 495-502.

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

Kleinman, K.P. and Ibrahim, J.G. (1998b). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine*, 17, 2579-2596.

Kurihara, K., Welling, M. and Vlassis, N. (2006). Accelerated variational Dirichlet mixture models. *Advances in Neural Information Processing Systems 19*.

Kurihara, K., Welling, M. and Teh, Y.W. (2007). Collapsed variational Dirichlet process mixture models. *Twentieth International Joint Conference on Artificial Intelligence*.

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

Lau, J.W. and Green, P.J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16, 526-558.

Lee, K.J. and Thompson, S.G. (2007). Flexible parametric models for random-effects distributions. *Statistics in Medicine*, advance access.

Li, Y., Lin, X. and Müller, P. (2007). Bayesian inference in semiparametric mixed models for longitudinal data. *Department of Biostatistics Working Paper Series*, UT MD Anderson Cancer Center.

Liang, F., Liao, M., Mao, K., Mukherjee, S. and West, M. (2007). Non-parametric Bayesian kernel models. *Discussion Paper*, 2007-10, Department of Statistical Science, Duke University, Durham, NC.

Lijoi, A., Mena, R.H. and Prünster, I. (2007a). A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinformatics*, to appear.

Lijoi, A., Mena, R.H. and Prünster, I. (2007b). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, advance access.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. 1. Density estimates. *Annals of Statistics*, 12, 351-357.

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

MacEachern, S.N. (1999). Dependent nonparametric process. *ASA Proceeding of the Section on Bayesian Statistical Science*, American Statistical Association, Alexandria, VA.

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

MacLehose, R.F. and Dunson, D.B. (2008). Nonparametric Bayes kernel-based priors for functional data analysis. *Statistica Sinica*, to appear.

MacLehose, R.F., Dunson, D.B., Herring, A.H. and Hoppin, J.A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology*, 18, 199-207.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18, 1194-1206.

Medvedovic, M., Yeung, K.Y. and Bumgarner, R.E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20, 1222-1232.

McAuliffe, J.D., Blei, D.M. and Jordan, M.I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing*, 16, 5-14.

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

Mulherin Engel, S.A., Eriksen, H.C., Savitz, D.A., Thorp, J., Chanock, S.J. and Olshan, A.F. (2005). Risk of spontaneous preterm birth is associated with common proinflammatory cytokine polymorphisms. *Epidemiology*, 16, 469-477.

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

Müller, P., Erkanli, A., West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83, 67-79.

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

Müller, P., Quintana, F. and Rosner, G.L. (2007). Semiparametric Bayesian inference for multilevel repeated measurement data. *Biometrics*, 63, 280-289.

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

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55, 463-469.

Newton, M.A., Noueiry, A., Sarkar, A. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5, 155-176.

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

Ohlssen, D.I., Sharples, L.D. and Spiegelhalter, D.J. (2007). Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Statistics in Medicine*, 26, 2088-2112.

Pennell, M.L. and Dunson, D.B. (2007). Nonparametric Bayes testing of changes in a response distribution with an ordinal predictor. *Biometrics*, OnlineEarly.

Petrone, S., Guindani, M. and Gelfand, A.E. (2007). Hybrid Dirichlet processes for functional data. *Technical Report*, Bocconi University, Milan, Italy.

Petrone, S. and Raftery, A.E. (1997). A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statistics and Probability Letters*, 36, 69-83.

Pillai, N.S, Wu, Q., Liang, F., Mukherjee, S. and Wolpert, R.L. (2007). Characterizing the function space for Bayesian kernel models. *Journal of Machine Learning Research*, 8, 1769-1797.

Papaspiliopoulos, O. and Roberts, G. (2007). Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika*, to appear.

Qin, Z.S. (2006). Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, 22, 1988-1997.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*, Springer.

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*, MIT Press.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society B*, 68, 305-332.

Rodriguez, A., Dunson, D.B. and Gelfand, A.E. (2007a). Latent stick-breaking processes. to be submitted.

Rodriguez, A., Dunson, D.B. and Gelfand, A.E. (2007b). Nonparametric functional data analysis through Bayesian density estimation. *Biometrika*, revision submitted.

Rodriguez, A., Dunson, D.B. and Gelfand, A.E. (2007c). The nested Dirichlet process (with discussion). *Journal of the American Statistical Association*, to appear.

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

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning*, 46, 21-52.

Sun, S., Greenwood, C.M.T. and Neal, R.M. (2007). Haplotype inference using a Bayesian hidden Markov model. *Genetic Epidemiology*, to appear.

Stephens, (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society B*, 62, 795-809.

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

Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 211-244.

Tomlinson, G. (1998). Analysis of densities. Unpublished dissertation. University of Toronto.

van der Merwe, A.J. and Pretorius, A.L. (2003). Bayesian estimation in animal breeding using the Dirichlet process prior for correlated random effects. *Genetics Selection Evolution*, 35, 137-158.

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

Walker, S.G. and Gutiérrez-Pena, E. (2007). Bayesian parametric inference in a nonparametric framework. *TEST*, 16, 188-197.

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

West, M., Müller, P. and Escobar, M.D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. *Aspects of Uncertainty: A Tribute to D.V. Lindley* 363-386.

Wilcox, A.J., Weinberg, C.R., O'Connor, J.F., Baird, D.D., Schlatterer, J.P., Canfield, R.E., Armstrong, E.G. and Nisula, B.C. (1988). Incidence of early loss of pregnancy. *New England Journal of Medicine*, 319, 189-194.

Xiang, Z.S., Qin, Z.H.S. and He, Y.Q. (2007). CRCView: a web server for analyzing and visualizing microarray gene expression data using model-based clustering. *Bioinformatics*, 23, 1843-1845.

Xing, E.P., Jordan, M.I. and Sharan, R. (2007). Bayesian haplotype inference via the Dirichlet process. *Journal of Computational Biology*, 14, 267-284.

Xing, E.P. and Sohn, K. (2007). Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space. *Bayesian Analysis*, 2.

Xing, E.P., Sohn, K., Jordan, M.I. and Teh, Y.W. (2006). Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. *Proceedings of the 23rd International Conference on Machine Learning* (ICML 2006).