

PART I
Biomedical and Health Sciences

.1.

Flexible Bayes regression of epidemiologic data

David B. Dunson

1.1 Introduction

Epidemiology is focused on the study of relationships between exposures and the risk of adverse health outcomes or diseases. A better understanding of the factors leading to disease is fundamental in developing better strategies for disease prevention and treatment. Hence, the findings from epidemiology studies have the potential to have a fundamental impact on public health. However, this potential is often not fully realized due to distrust among the general public and physicians about the reliability of conclusions drawn from epidemiology studies. This distrust stems in part from inconsistencies in findings from different studies.

Although statistics cannot solve certain problem in epidemiology, such as the occurrence of unmeasured confounders, many of the problems with lack of reproducibility stem from the overly simplistic statistical analyses that are routinely used. In particular, it is standard practice to reduce a potentially complex health outcome to a simple 0/1 indicator of disease and then apply a logistic regression model. By also categorizing exposures, one can then obtain exposure group-specific odds ratios, which form the primary basis for inference on exposure – disease relationships. Such an approach leads to transparent analyses, with results easily interpretable by clinicians having minimal expertise in statistics.

However, there are important limitations to this paradigm, which can lead to misinterpretations of exposure – disease relationships. The first is the obvious lack of efficiency that results from discarding data. For example, if one has a continuous health response and a continuous exposure, then categorizing both the response and exposure can lead to a reduction in power to detect an association. In addition, the results of the analysis will clearly be very sensitive to the number of categories chosen and the cutpoints for defining these categories (Boucher *et al.*, 1998). For continuous exposures, an obvious alternative to categorization is to use splines to estimate the unknown dose-response curve (Greenland, 1995). However, for continuous responses,

it is not clear how to best assess risk as a function of exposures and other factors.

This chapter focuses on flexible Bayesian methods for addressing this problem motivated by pregnancy outcome data on birth weight and gestational age at delivery. As argued by Wilcox (2001), birth weight is not particularly meaningful in itself as a measure of health of the baby. However, there is often interest in assessing factors predictive of intra-uterine growth restriction (IUGR). IUGR is best studied using longitudinal ultrasound data to assess fetal growth over time as in Slaughter, Herring and Thorp (2009). However, such data are typically not available, so it is most common to study IUGR using small for gestational age (SGA) as a surrogate. SGA is defined as a 0/1 indicator that the baby is below the 10th percentile of the population distribution of birth weight stratified on gestational age at delivery. One is also concerned about large-for-gestational age (LGA) babies, which are more likely to be born by Cesarean delivery or with a low Apgar score (Nohr *et al.*, 2008). The standard approach analyses risk of SGA, LGA and preterm birth, defined as a delivery prior to 37 weeks completed gestation, in separate logistic regression analyses.

A natural question that arises is whether one can obtain a coherent picture of the impact of an exposure on pregnancy outcomes from such analyses. Although SGA is meant to provide a surrogate of growth restriction, which is adjusted for gestational age at delivery, it is not biologically plausible to assume that fetal growth and the biological process of initiating delivery are independent. Hence, it seems much more natural to consider birth weight and gestational age at delivery as a bivariate response, while avoiding the loss of information that accompanies categorization (Gage, 2003). Even if the focus is only on assessing predictors of risk of premature delivery, the cutoff of 37 weeks eliminates the possibility of assessing risk on a finer scale. In particular, babies born near the 37 week cutoff experience limited short and long term morbidity compared with babies born earlier.

Figure 1.1 shows data on gestational age at delivery and birth weight for $n = 2313$ pregnancies from the Longnecker *et al.* (2001) substudy of the US Collaborative Perinatal Project. The vertical dashed line at 37 weeks shows the cutoff used to define preterm births, while the solid lines show the cutoff for SGA depending on gestational age at delivery for male and female babies. Interest focuses on assessing how predictors, such as the level of exposure to chemicals in the environment (DDT, PCBs, etc.), impact the risk of adverse pregnancy outcomes, which correspond to lower values of birth weight and gestational age at delivery. As reducing the data to binary indicators has clear disadvantages, I propose to instead let the response for pregnancy i correspond to $\gamma_i = (\gamma_{i1}, \gamma_{i2})'$, with $\gamma_{i1} =$ gestational age at delivery and $\gamma_{i2} =$ birth weight. Although this seems very natural, it does lead to some complications in that

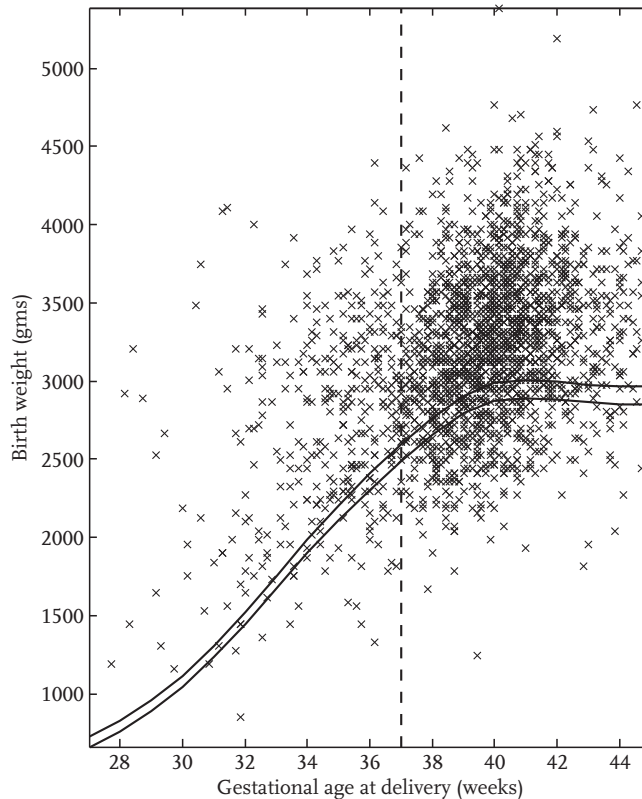


Fig. 1.1 Data on gestational age at delivery and birth weight for the Longnecker *et al.* (2001) substudy of the Collaborative Perinatal Project.

standard parametric models for bivariate continuous data, such as the bivariate normal or t -distributions, provide a poor fit to the data. This is clear in examining Figure 1.2, which shows an estimate of the marginal density of y_{i1} . The density is left skewed and standard transformations fail to produce a Gaussian shape.

As adverse health outcomes correspond to values in the tails of the response distribution, the main interest is in studying how predictors impact the distributional tails. For example, in studying the impact of DDE levels in maternal serum on risk of premature delivery using the Longnecker *et al.* (2001) data, we would like to assess how the left tail of the distribution in Figure 1.2 changes with dose of DDE and other predictors. Potentially, this interest can be addressed using quantile regression. However, this would necessitate choosing a particular percentile of the distribution that is of primary interest, which is in some sense as unappealing as categorization. As an alternative, one can allow the conditional distribution of y_i given predictors $x_i = (x_{i1}, \dots, x_{ip})'$ to be unknown and changing flexibly over the predictor space.

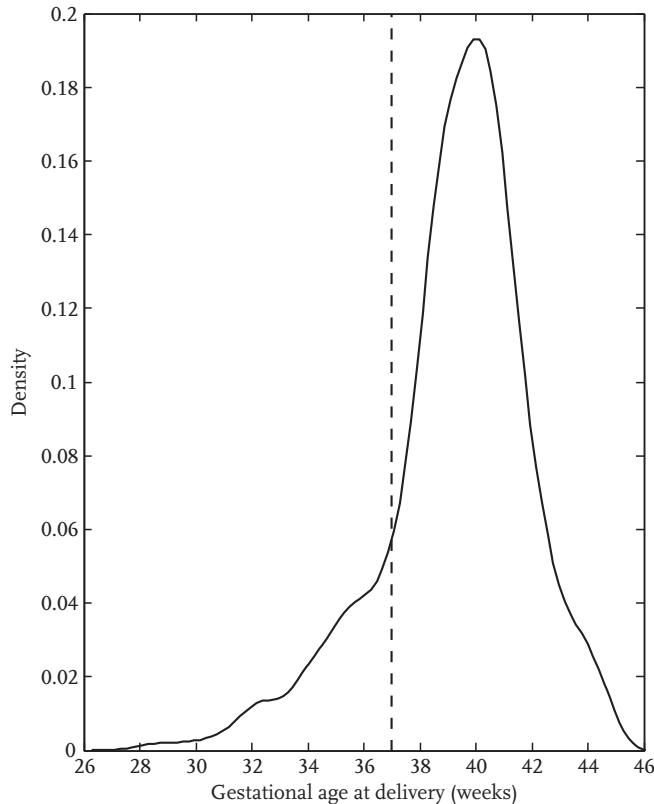


Fig. 1.2 Frequentist kernel estimate of the density of gestational age at delivery for the Longnecker *et al.* (2001) data. The vertical dashed line is the cutoff for defining premature delivery.

As further motivation, Figure 1.3 plots frequentist kernel estimates of the conditional distribution of y_{i1} given x_{i1} = maternal serum level of DDE, with x_{i1} categorized into quintiles and estimates then obtained separately for each category. As categorized DDE increases, there is an increasingly heavy left tail but the right side of the distribution remains essentially unchanged. This type of pattern is not surprising given the biological constraints limiting the possibility of a very long gestational length. I expect that similar constraints occur for many other continuous health outcomes. For such data, standard regression approaches which rely on modelling predictor effects on the mean response are completely inappropriate. Instead, what is needed is flexible methods for density regression, which allow the entire density to change flexibly with predictors, while allowing variable selection and hypothesis testing. This chapter focuses on applying such methods to the pregnancy outcome application.

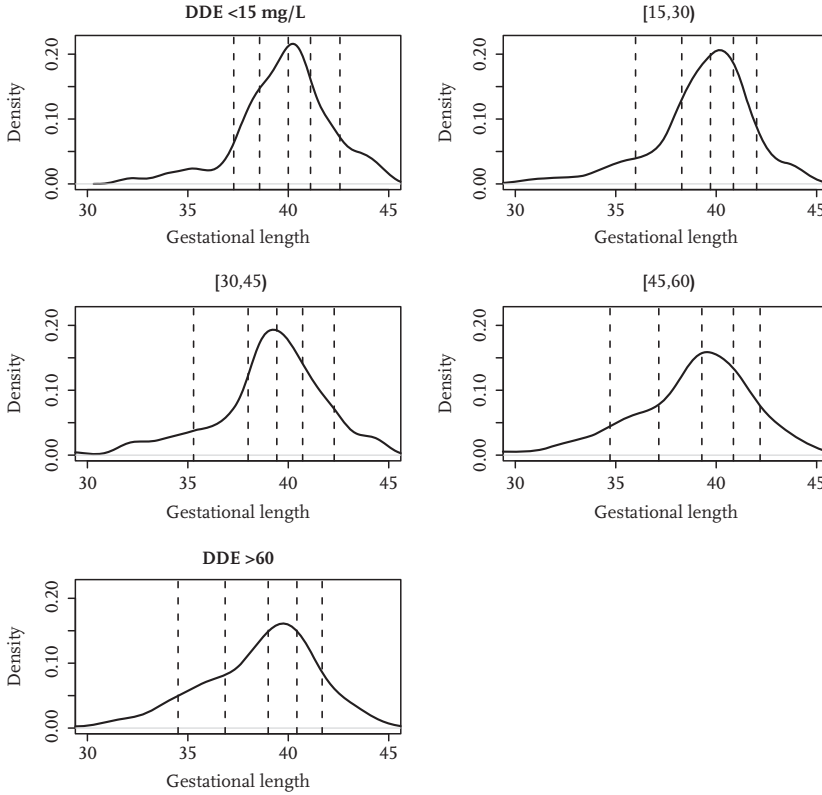


Fig. 1.3 Frequentist kernel estimate of the density of gestational age at delivery conditional on categorized DDE level. The vertical dashed lines correspond to quantiles of the empirical distribution in each group.

1.2 Mixture models

1.2.1 Finite mixture models

As a starting point, it is useful to consider finite mixture models for characterizing the conditional distribution of y_i given x_i . One possibility is to use a latent class model,

$$f(y | x) = \sum_{h=1}^k \pi_h(x) N_2(y; \mu_h, \Sigma_h), \quad (1.1)$$

where $h = 1, \dots, k$ indexes mixture components, $\pi_h(x)$ is the probability of allocation to component h conditionally on the predictors x , and the distribution of $y = (y_1, y_2)'$ among pregnancies in latent class h corresponds to the bivariate normal distribution, $N(\mu_h, \Sigma_h)$. Note that model (1.1) relies on categorizing individuals into groups based on their pregnancy outcomes, with the probability of allocation to each group dependent on predictors. Hence, there are some

clear similarities with standard epidemiologic analysis approaches. However, the fundamental difference is that instead of predefining hard thresholds on the y s, the latent class model adaptively allocates individuals into groups probabilistically. There are no hard thresholds and the allocation to latent classes relies on the data at hand without discarding information.

For example, in order to fit the data in Figure 1.3, one could potentially use a mixture of four normal distributions, with components corresponding to full term births, preterm births near the 37 week cutoff, early preterm births, and late term births. The need for this final component may be reduced in more recent data sets having ultrasound to accurately date the pregnancy, as many of the values in the right tail are likely due to measurement error in pregnancy dating. In order to complete a specification of the model, we can choose a regression model for the probability weights, $\pi_h(x)$. In order to aid interpretation for epidemiologists, the polytomous logistic regression model provides a convenient choice,

$$\pi_h(x) = \frac{\lambda_h \exp(x' \beta_h)}{\sum_{l=1}^k \lambda_l \exp(x' \beta_l)}, \quad (1.2)$$

where $\lambda = (\lambda_1, \dots, \lambda_k)'$ are parameters characterizing the baseline probabilities of allocation to each class, β_h are parameters characterizing the effect of predictors on the probability of allocation to class h , for $h = 1, \dots, k$, and $\lambda_1 = 1$, $\beta_1 = 0$ for identifiability.

As there is nothing in the model distinguishing classes $1, \dots, k$, we face the well-known label ambiguity problem, which is omnipresent in mixture models and clustering (Stephens, 2000; Jasra, Holmes and Stephens, 2005). Frequentist analyses of model (1.1)–(1.2) can be implemented using the EM algorithm for maximum likelihood estimation. The EM algorithm will converge to a local mode, which corresponds to one possible configuration of the label indices on classes $1, \dots, k$. As all other configurations lead to the same maximum likelihood estimate, the EM algorithm is in some sense not sensitive to the label switching problem. Each run will produce the same maximum likelihood estimates after convergence and relabelling. However, the EM algorithm just produces a point estimate, while a Bayesian analysis implemented with MCMC can allow a full probabilistic treatment of uncertainty in estimation of $f(y|x)$. Properly accounting for uncertainty is important in conducting valid inferences about predictor effects, which is the primary goal of epidemiologic analyses.

In MCMC analyses, label ambiguity will lead to a multimodal posterior and MCMC algorithms can experience a label switching problem. For example, for the first several 100 iterations, classes 1–4 may correspond to full-term, late preterm, early preterm and late term, but then the labels may switch (corresponding to a different mode in the posterior) so that classes 1–4 correspond

to late term, full term, early preterm and late preterm. During an MCMC run, there may be many such relabelings, so that it becomes very difficult to summarize the results from the MCMC output in a coherent manner. Certainly, posterior means and credible intervals for component-specific parameters (e.g. λ_2, β_2) are meaningless. Several solutions have been proposed to this problem: (1) place restrictions on the parameters to avoid label ambiguity; (2) apply a post-processing relabelling algorithm to the MCMC output in order to align draws from the same component, so that one can calculate posterior summaries for component-specific parameters after realignment (Stephens, 2000; Jasra *et al.*, 2005); and (3) ignoring the label ambiguity problem, while avoiding mixture component-specific inferences.

I start by commenting on approach (1), which is by far the most widely used in the literature. In order to implement this strategy for model (1.1), the most common approach would be to place an order restriction on the means. For example, letting $\mu_h = (\mu_{h1}, \mu_{h2})'$, a possible identifying restriction would let $\mu_{11} < \mu_{21} < \dots < \mu_{k1}$. Hence, the components would be restricted in advance to be ordered by length of gestation. Although this seems reasonable on the surface, there are a number of pitfalls making this approach unreliable. The most notable is that it does not solve the label ambiguity problem unless the components are very well separated. For example, suppose that μ_{11} and μ_{21} are close together, with the difference between classes 1 and 2 occurring in Σ_1 and Σ_2 and/or μ_{21} and μ_{22} . Then, the constraint has accomplished nothing. Such problems are well motivated in Jasra *et al.* (2005), who advocate in favour of strategy (2). In my experience, such relabelling approaches tend to work well, though they add substantially to the computational burden.

My own view is that latent class modelling and clustering is most appropriately used as a tool to obtain a flexible model, and one should avoid interpreting the clusters obtained as biologically meaningful. It is well known that clusters are entirely sensitive to parametric assumptions, with such assumptions seldom if ever justifiable by substantive theory in an application area (Bauer, 2007). That said, in some cases, clustering is useful as a hypothesis generating tool, to simplify interpretations of complex data and as a dimensionality reduction device. Even if there is no reason to suppose that biologically meaningful clusters exist in the pregnancy outcome data, one can still use model (1.1)–(1.2) in conducting inferences about changes in the distribution of the pregnancy outcomes with predictors. Such inferences are unaffected by label switching. For example, suppose that one wants to obtain an estimate of $f(y|x)$. A Gibbs sampling algorithm can be implemented for model (1.1)–(1.2), relying on data augmentation to obtain conjugate full conditional posterior distributions without any order constraints on the parameters. From this Gibbs sampler, one can obtain samples from the posterior of $f(y|x)$. Although the labels on

clusters underlying each sample will vary, this has no effect on the value of $f(y|x)$, and one can calculate posterior means and credible intervals without complication.

Model (1.1) is a type of finite mixture model, and a fundamental problem that arises in applications of such models is choice of the number of mixture components, k . The most common strategy in the literature relies on fitting the mixture model for different choices of k , using the EM algorithm to obtain maximum likelihood estimates. The BIC is then used to choose the ‘best’ k . There are two problems with this approach. First, the BIC was originally developed by Schwarz (1978) based on a Laplace approximation to the marginal likelihood. His justification does not hold for finite mixture models, so that BIC lacks theoretical grounding, though it does seem to work well in practice in many cases (Fraley and Raftery, 1998). The second problem is that the approach of fitting models for many different choices of k and then basing inferences on the final selected model ignores uncertainty in selection of k . In my experience, such uncertainty is often substantial, with the data not providing compelling evidence in favour of any single k . Hence, it seems more appropriate to use Bayesian model averaging to allow for uncertainty in k . This can potentially be accomplished using reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). However, RJMCMC tends to be quite difficult to implement efficiently, so I instead advocate bypassing the model selection problem through the use of a nonparametric Bayes approach.

1.2.2 Nonparametric Bayes

Nonparametric Bayes methods allow for uncertainty in distributional assumptions within Bayesian hierarchical models. For example, initially excluding predictors from consideration, suppose that interest focuses on estimating the joint density of gestational age at delivery and birth weight. Then, one possibility is to use a Dirichlet process mixture (DPM) of normals (Lo, 1984; Escobar and West, 1995), which can be expressed in hierarchical form as

$$\begin{aligned} y_i &\sim N_2(\mu_i^*, \Sigma_i^*) \\ (\mu_i^*, \Sigma_i^*) &\sim P, \quad P \sim DP(\alpha P_0), \end{aligned} \tag{1.3}$$

where P is an unknown mixture distribution, which is assigned a Dirichlet process prior (Ferguson, 1973; 1974), parametrized in terms of a precision α and base distribution P_0 . This model can accurately characterize any smooth bivariate density.

To obtain additional insight and relate the DPM model to the latent class model in (1.1), one can use the stick-breaking representation of the DP

(Sethuraman, 1994), which implies that

$$f(y) = \sum_{h=1}^{\infty} \pi_h N_2(y; \mu_h, \Sigma_h), \quad (\mu_h, \Sigma_h) \sim P_0,$$

$$\pi_h = V_h \prod_{l < h} (1 - V_l), \quad V_h \sim \text{Beta}(1, \alpha). \quad (1.4)$$

Hence, the DPM of normals model is equivalent to (1.1) except with predictors excluded, the number of mixture components (classes) equal to $k = \infty$, the component-specific parameters sampled iid from P_0 , and a stick-breaking prior placed on the component weights. This stick-breaking prior, which is described in the second line of (1.4) favors allocating non-negligible weight only to the first few components for small α . This leads to a sparse representation in which only a few components are occupied by the subjects in a given sample, though infinitely many components are available, so that the model can be made more complex as subjects are added.

There is a wide variety of easy-to-implement MCMC and fast approximation algorithms available for fitting of DPMs. Refer to Dunson (2008) for a recent review of applications of nonparametric Bayes and Dirichlet process mixtures to biostatistical applications. In order for such approaches to be useful in epidemiologic applications, it is necessary to incorporate predictors. For example, a natural extension of (1.4) would let

$$f(y | x) = \sum_{h=1}^{\infty} \pi_h(x) N_2(y; \mu_h x, \Sigma_h), \quad (\mu_h, \Sigma_h) \sim P_0, \quad (1.5)$$

where μ_h is a $2 \times p$ matrix of coefficients specific to mixture component h , for $h = 1, \dots, \infty$. This model is a type of bivariate infinite mixtures of experts model, which generalizes parametric mixtures of experts models that assume a finite k and focus on univariate responses (Jordan and Jacobs, 1994; Peng, Jacobs and Tanner, 1996; and Jacobs, Peng and Tanner, 1997). By mixing linear regression models with predictor-dependent weights, we allow for a sparser characterizations of complex data than possible using the framework of (1.1) with $k = \infty$. In particular, we find that it is possible to use many fewer mixture components by mixing normal linear regressions instead of normals without a regression component.

In order to complete a specification of model (1.5), it is necessary to choose a model for the unknown weight function $\pi_h(x)$. Ideally, this would be done in a highly flexible manner, while favouring a sparse representation with a few dominate components having relatively high weights. There are several possibilities that have been proposed in the literature, including the order-based dependent Dirichlet process (Griffin and Steel, 2006), the kernel stick-breaking process (KSBP) (Dunson and Park, 2008) and the local Dirichlet process (Chung

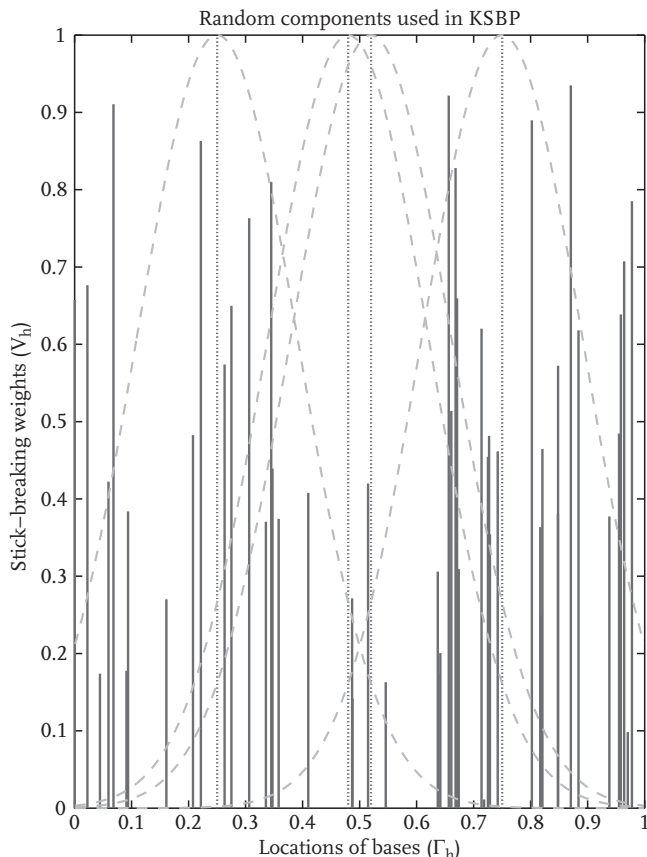


Fig. 1.4 Random components used in the KSBP. The red dotted lines represent predictor locations of interest, the blue lines correspond to weights V_h at locations Γ_h , and the green dashed line shows the kernel, K .

and Dunson, 2009a). Here, I focus on the KSBP, since it provides a relatively simple generalization of the DP stick-breaking process in (1.4), while leading to straightforward posterior computation and inferences.

1.2.3 Kernel stick-breaking process

Under the KSBP, the probability of allocation to component h conditionally on predictors x is

$$\pi_h(x) = V_h K(x, \Gamma_h) \prod_{l < h} \{1 - V_l K(x, \Gamma_l)\}, \quad h = 1, \dots, \infty, \quad (1.6)$$

where $\Gamma_l \sim H$ is a random basis location and $V_l \sim \text{beta}(1, \alpha)$ is a random weight on that location, for $h = 1, \dots, \infty$. Figure 1.4 plots the first 50 random components for a single draw from the KSBP in a simple case in which there is a single predictor $x \in [0, 1]$. The red vertical dotted lines correspond to predictor

values of interest, which are set to $x = 0.25, 0.48, 0.52, 0.75$ for illustration. For example, consider calculating $\pi_h(0.25)$ using expression (1.6) applied to the components shown in Figure 1.4. Clearly, the locations Γ_h receiving highest weight will correspond to those that receive higher weights V_h , particularly if they are close to $x = 0.25$, so are not downweighted too much by the kernel (shown with a green dashed line), and if the index h is small, so that too much of the probability stick is not already used up by other locations.

The resulting values of $\pi_h(x)$ for $x = 0.25, 0.48, 0.52, 0.75$ are shown in Figure 1.5. Note that the probability weight, $\pi_h(x)$, at location Γ_h is the weight on $N_2(\mu_h(1, x)', \Sigma_h)$ in the mixture of bivariate normals shown in (1.5). Hence, it is appealing to have similar weights for x values that are close together, while allowing different weights for x s that are far apart. Figure 1.5 illustrates this behaviour, as the weights for $x = 0.48$ and $x = 0.52$ are close together, with both predictor values having the same two dominant mixture components. The tendency for the mixture distributions to change smoothly with predictors induces smoothness in modelling of the unknown conditional response distributions, with small changes in the predictors implying small changes in the response density.

Theoretical properties of the KSBP and an efficient MCMC algorithm for posterior computation are presented in Dunson and Park (2008). Here, I focus on the pregnancy outcome application. Note that the MCMC algorithm produces draws from the posterior distribution for $f(y|x)$ for any x value of interest. Such draws can be summarized in a broad variety of ways. For example, one can obtain nonlinear dose response curves characterizing the change in mean gestational age at delivery and birth weight with increasing dose of DDE adjusting for the other predictors. One can also obtain quantile response curves characterizing the change in a percentile of the response distribution with DDE. The next section illustrates some of the possibilities focusing on gestational age at delivery data from the Longnecker *et al.* (2001) study. The same type of approach can be applied directly in the bivariate case to jointly model gestational age at delivery and birth weight according to DDE exposure and other predictors.

1.3 Density regression for pregnancy outcomes

I focus on gestational age at delivery, DDE and age data from the Longnecker *et al.* (2001) study. DDE is the most persistent metabolite of the pesticide DDT. Although DDT is no longer in use in much of the developed world, including the United States, DDT is still used broadly in some parts of the world due to its effectiveness against malaria transmitting mosquitoes. Decisions of whether or not to continue use of DDT must necessarily weigh this benefit against the

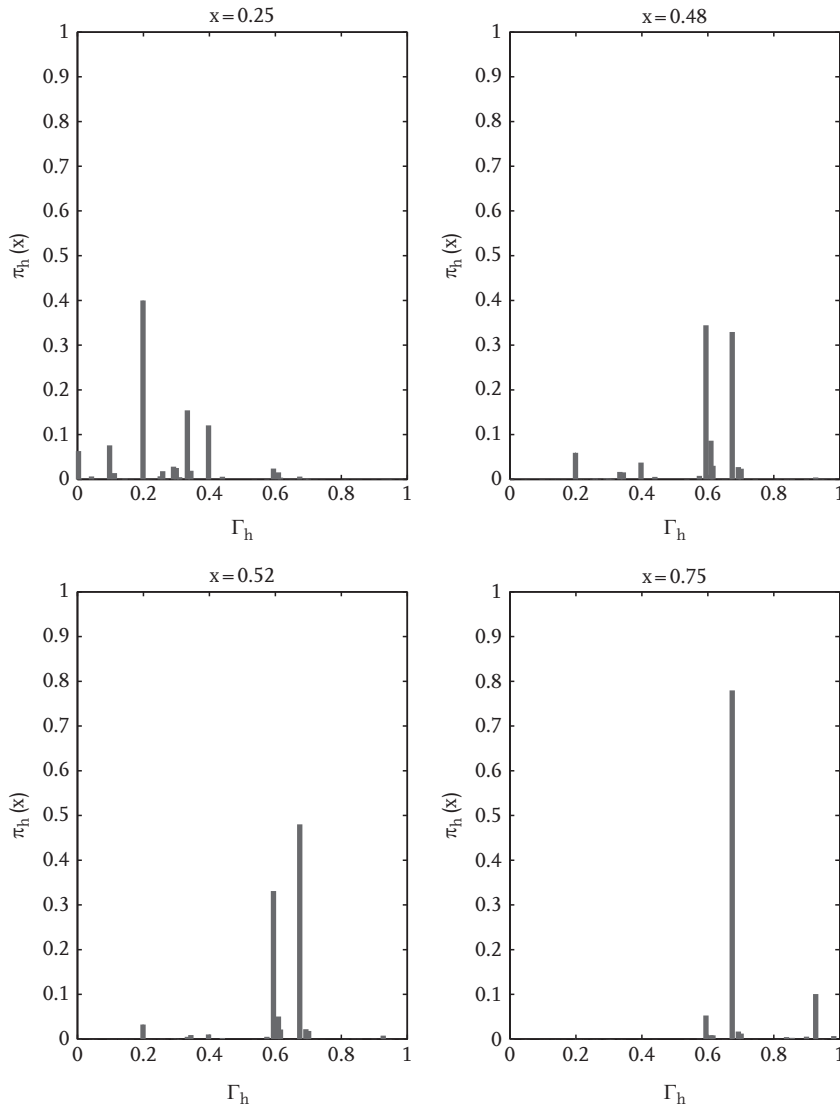


Fig. 1.5 Probability weights, $\pi_h(x)$, on each of the mixture components for different choices of x including $x = 0.25, 0.48, 0.52, 0.75$. These realizations correspond to one draw from the kernel stick-breaking process.

increasing evidence of adverse health effects. Here, the interest is in studying how risk of premature delivery is impacted by maternal exposure to DDT. DDT exposure potentially has a long term impact, as metabolites of DDT are lipophilic and hence bioaccumulate in maternal fat, with body burden often increasing over time. Premature delivery is a major public health problem, being a common condition (approximately 10–13% of pregnancies are delivered prior to 37 weeks completed gestation), which can result in substantial

morbidity. Risks of Infant mortality and morbidity increase substantially for early preterm births, so it is of major public health and clinical interest to separate risk of close to full term births (close to the 37 week cutoff) from risk of early preterm births.

As shown in Figure 1.1, data are available for $n = 2313$ pregnancies after excluding pregnancies having gestational lengths longer than 45 weeks. These values were considered unreliable and to likely arise due to measurement error in pregnancy dating. Letting $y_i =$ gestational age at delivery for pregnancy i and $\mathbf{x}_i = (\text{DDE}_i, \text{age}_i)'$, with DDE_i dose level in mg/L measured in maternal serum during pregnancy, I focus on the following flexible mixture model:

$$f(y_i | \mathbf{x}_i) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}_i) N(y_i; \beta_{h1} + \beta_{h2} \text{DDE}_i + \beta_{h3} \text{age}_i, \tau^{-1}), \quad (1.7)$$

where the predictor-dependent mixture distributions are assumed unknown through the use of a kernel stick-breaking process prior, with a Gaussian kernel, $K(\mathbf{x}, \Gamma_h) = \exp\{-\psi(\mathbf{x} - \Gamma_h)^2\}$. I choose gamma hyperpriors for the parameters α and ψ to allow the data to inform about their choice. The results were found to be robust to the choice of kernel as long as hyperpriors were chosen for the kernel precision. For example, I also tried an exponential kernel, and obtained essentially indistinguishable results.

The MCMC algorithm was run for 30,000 iterations with an 8000 iteration burn-in. The convergence and mixing was good based on examination of trace plots of the conditional densities at different points and of the hyperparameter values. Figure 1.6 shows the data on DDE versus gestational age at delivery. The vertical dashed lines are quintiles of the empirical distribution of DDE. Following standard practice in epidemiologic analyses, Longnecker *et al.* (2001) used these quintiles as thresholds in categorizing DDE prior to conducting a logistic regression analysis, which relied on a 0/1 indicator of premature delivery using the standard 37 week cutoff of gestational age at delivery. Here, I instead avoid categorization of either the predictors or the response. The solid curve is the estimated expectation of y_i conditionally on DDE_i holding age_i fixed at the mean value. The curve is close to linear across much of the data range, which illustrates an appealing property of the analysis, which is a tendency to collapse on a simple submodel when that model holds approximately. In this case, the simple submodel is a linear regression model, with deviations allowed through local mixing.

A simple frequentist alternative to our approach, which also avoids categorization, lets

$$E(y_i | \mathbf{x}_i) = \mu + g_1(x_{i1}) + g_2(x_{i2}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad (1.8)$$

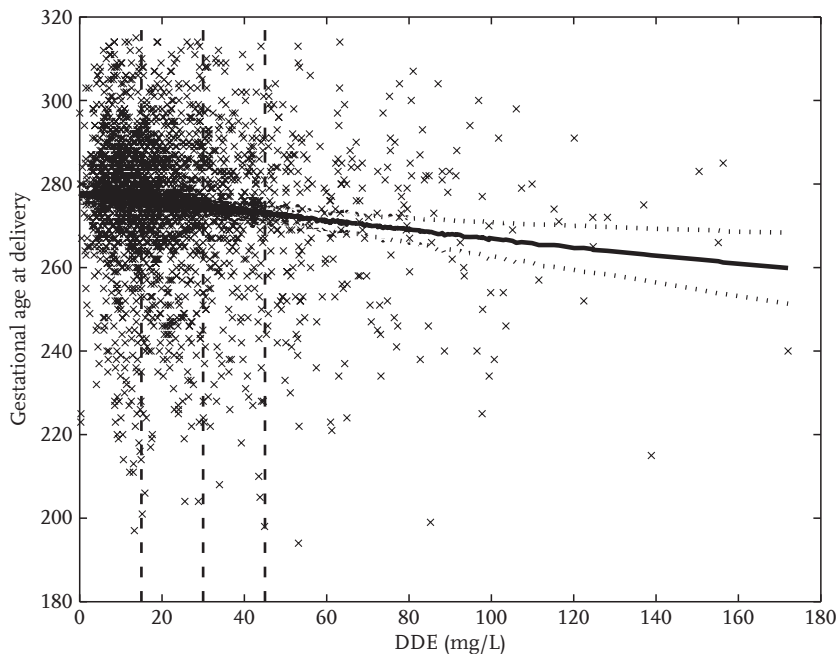


Fig. 1.6 Estimates of the expected gestational age at delivery conditionally on DDE. Gestational age at delivery and DDE data points are shown as \times . Solid lines are posterior means and dashed lines are 99% pointwise credible intervals.

with $g_1(\cdot)$ and $g_2(\cdot)$ unknown functions. This additive model can be fitted easily in standard software packages, such as R, and produces a similar curve to that shown in Figure 1.6, though the frequentist curve estimate is somewhat bumpier, particularly in sparse data regions. Even though the frequentist analysis makes an invalid assumption that the residuals are normally distributed with constant variance, mean estimates tend to be robust to violations of this assumption due to the central limit theorem, as the sample size is not small in this application.

A natural question is then what is gained practically by the flexible Bayesian analysis in this application? The answer is that the interest really does not focus on the mean response, but instead focuses on the tails of the distribution corresponding to adverse health responses. Applying a flexible mean regression does not tell us how the risk of premature delivery changes with level of exposure to DDE. Figure 1.7 shows the estimated conditional densities of gestational age at delivery for a range of values of DDE. It seems clear from this plot that there is an increasingly heavy left tail as DDE dose increases. Note that the estimates tell a similar story to the estimates in Figure 1.3, but categorization is avoided, so a unique density estimate is obtained for every DDE level. In

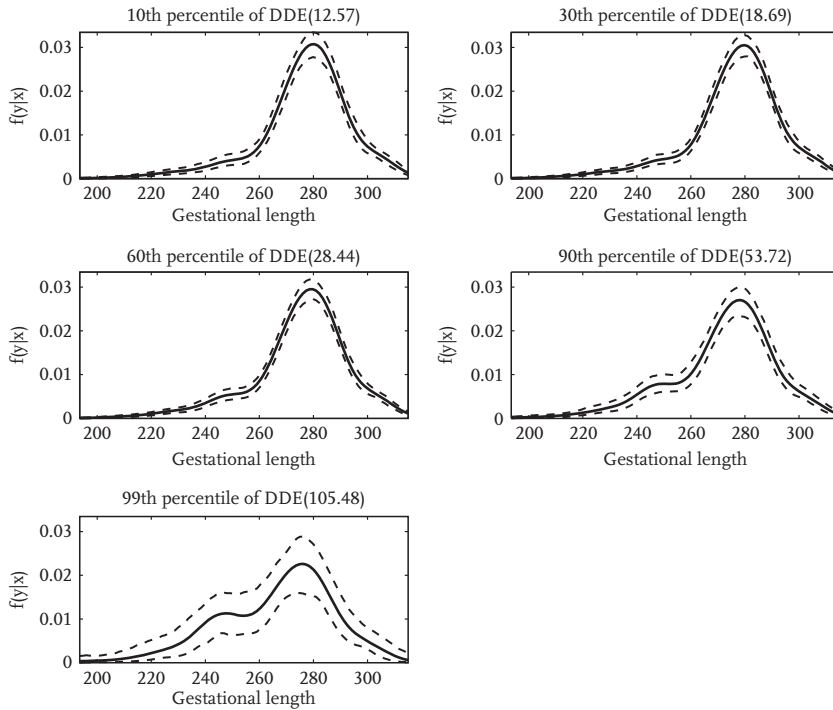


Fig. 1.7 Estimated conditional density of gestational age at delivery given DDE for a range of DDE values. Solid lines are posterior means and dashed lines are 99% pointwise credible intervals.

addition, by borrowing information across different DDE values in a flexible manner, smoother and more realistic density estimates are produced. At the higher DDE values, there is limited data available, so the width of the credible intervals increase substantially.

As it is hard to gauge subtle changes in risk of prematurity from examination of side-by-side conditional density plots, I also provide plots of the risk of falling below various thresholds for defining preterm birth in Figure 1.8. Note that unlike typical procedures which collapse the data into binary indicators and conduct separate analyses for each choice of threshold, these estimates are all based on one coherent underlying model for the conditional response distribution. The use of such a coherent model can lead to substantial efficiency gains relative to separate binary response analyses. Such efficiency gains are very important practically in attempting to estimate risk of falling within the extreme tails of a distribution, since even if the sample size is large overall, one may have limited data in the tails.

Figure 1.8 tells an interesting and disturbing story. First consider the plot in the lower left panel, which shows the risk of having a delivery prior to 37 weeks of completed gestation as a function of DDE dose. Even at low doses of

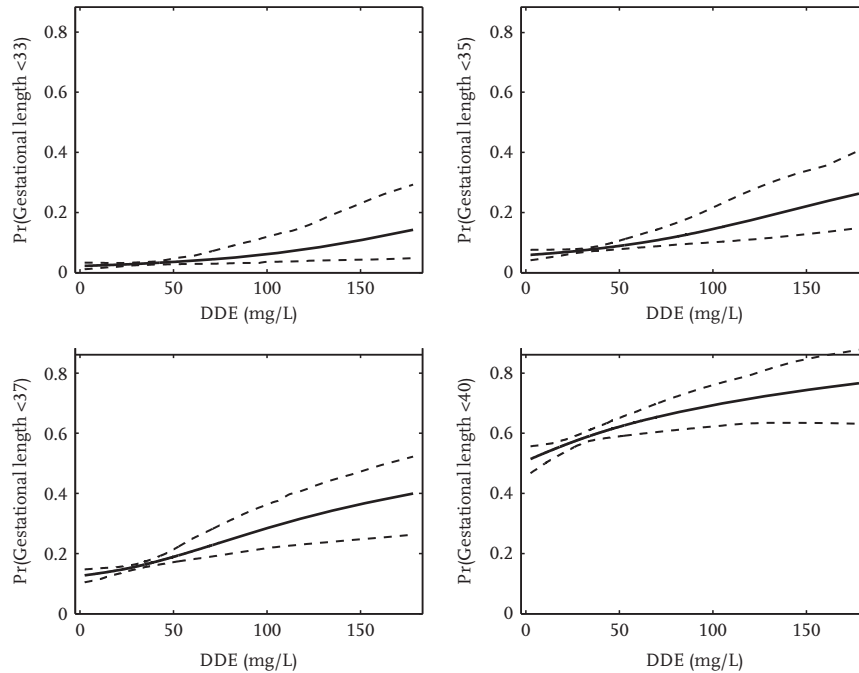


Fig. 1.8 Dose response curves for the probability of observing a gestational age at delivery below 33, 35, 37 or 40 weeks. Solid lines are posterior means and dashed lines are 99% pointwise credible intervals.

DDE, this curve has a positive slope, suggesting an increasing trend in risk of premature delivery with increasing level of exposure. The result is consistent with the Longnecker *et al.* (2001) frequentist logistic regression analysis, but their analysis followed standard practice and only considered the 37 week cutoff for defining preterm birth. From Figure 1.8, there is also an increasing dose response trend for the 33 and 35 week cutoffs. These results are of great clinical and public health importance, since neonatal mortality and morbidity rates for babies born at 34 or 35 weeks are substantially higher than those for a baby born at 36 weeks (McIntire and Leveno, 2008). Hence, variability in gestational length even among late preterm births is an important determinant of risk of subsequent adverse outcomes, which should not be ignored in the analysis by focusing entirely on the 37 week cutoff. Although data become increasingly sparse as we move further into the left tail of the distribution and attempt to assess risk of early preterm births, the Bayesian approach can partly address problems with data sparsity through borrowing of information. That said, credible interval widths are wide for the dose response curve for cutoffs much below 33 weeks, since the data are so sparse in this region, as is apparent from Figure 1.6.

1.4 Discussion

This chapter has described a Bayesian approach for flexible analysis of epidemiologic data, motivated by the clear limitations of current standard approaches. In particular, it is quite common to be interested in relationships between a continuous exposure and multiple continuous health responses. As logistic regression is the default analysis strategy in epidemiology, it is very often the case that continuous responses and exposures are categorized prior to analysis. Although this approach has the advantage of resulting in simple summaries of the relationships between exposures and risk of an adverse health response, there are a number of important disadvantages, which have been highlighted in this chapter through application to studies of pregnancy outcomes. In this and many other biomedical applications, individuals with adverse responses tend to fall in the tails of the distribution. Hence, to simplify the analysis, it is tempting to choose a threshold and categorize the response as extreme/not extreme. For example, with very few exceptions, the thousands of articles in the epidemiologic literature on premature delivery use less than 37 weeks of completed gestation as the threshold, and then discard finer scale information on gestational age at delivery in conducting inferences. The main concern with premature delivery is the increased risk of subsequent adverse health outcomes in the developing child, and risk and severity of these outcomes can increase substantially with each week change within the ≤ 37 week interval. Hence, it is clearly most appropriate to assess how the entire tail of the gestational age at delivery changes with exposures.

Another important issue is how to deal with multivariate responses, such as gestational age at delivery and birth weight in the pregnancy application. Again, in epidemiology almost all analyses rely on simplifying the data, so that a univariate logistic regression can be applied. In particular, risk of premature delivery is analysed separately from risk of small-for-gestational age (SGA), which is defined as a birth that falls within the lower 10th percentile of the birth weight distribution stratified on gestational age at delivery. SGA is arguably not a meaningful summary of growth restriction, and it is difficult to interpret results from analyses assessing relationships between exposures and SGA. The conceptual problem is that SGA represents an attempt to adjust for gestational age at delivery in analyses of birth weight. However, gestational age at delivery and birth weight are so tied together that they are much more appropriately viewed as a bivariate outcome. By using the flexible mixture models proposed in this article, one can jointly model, while avoiding parametric assumptions, as normality is typically violated. If one wants to adjust for gestational length in analysing birth weight, this can be done coherently by estimating the birth weight density profile. Such approaches should be broadly useful in epidemiology.

Appendix

A. Broader context and background

In this section, we provide additional background and discussion on nonparametric Bayes mixture models. Chapters 6 and 27 present a review of some properties and applications of Dirichlet process priors. Limiting overlap, here I focus specifically on the case in which the focus is on modelling of a collection of unknown distributions indexed by $x \in \mathcal{X}$, with x corresponding to predictors, time or spatial location. A key development was the dependent Dirichlet process (DDP) prior (MacEachern, 1999), which lets

$$P_x = \sum_{h=1}^{\infty} \pi_h(x) \delta_{\theta_h(x)}, \quad \theta_h \sim P_0, \quad (1.9)$$

where $\pi_h(x) = V_h(x) \prod_{l < h} \{1 - V_l(x)\}$, for $h = 1, \dots, \infty$, $\theta_h(x) \sim P_{0x}$ and $V_h(x) \sim \text{Beta}(1, \alpha)$ independently for $h = 1, \dots, \infty$, with P_{0x} the marginal of P_0 at location x . Here, the stick-breaking weights and atoms at a given location x are mutually independent, but dependence is incorporated across locations.

The DDP defines a prior for the collection $\{P_x, x \in \mathcal{X}\}$, with $P_x \sim DP(\alpha P_{0x})$ marginally at each x , while allowing dependence in P_x and $P_{x'}$. Because realizations from a Dirichlet process are almost surely discrete, the DDP is not directly useful for modeling changes in a response distribution with x . However, one can define a DDP mixture model by letting $f(y|x) = \int K(y; \theta) dP_x(\theta)$, with $K(\cdot)$ a kernel (e.g. Gaussian). In practice, it is not immediately obvious how to model dependence of the stick-breaking random variables $\{V_h(x)\}$ while maintaining the necessary marginal property $V_h(x) \sim \text{Beta}(1, \alpha)$ for all x . Hence, most of the applications of DDP mixture models have relied on the fixed- π specification that lets $\pi_h(x) = \pi_h$, while allowing the atoms to vary flexibly (De Iorio *et al.*, 2004; Gelfand *et al.*, 2005; among others).

For continuous \mathcal{X} , a natural approach is to let $\theta_h \sim \text{GP}(\mu, \mathcal{C})$, with $\text{GP}(\mu, \mathcal{C})$ denoting a Gaussian process with mean function μ and covariance function \mathcal{C} . For example, suppose that $x \in [0, 1]$ is a continuous predictor, with $f(y|x) = \int N(y; \mu, \sigma^2) dP_x(\mu)$ and $\{P_x, x \in \mathcal{X}\}$ assigned a fixed- π DDP prior with base measure $\text{GP}(\mu, \mathcal{C})$. Hence, the prior distribution for $f(y|x)$ is marginally a Dirichlet process location mixture of normals, which is highly flexible. As x varies, the probability weights on the different mixture components remain constant, but the locations of the components vary according to a Gaussian process. If one chooses a standard covariance function, such as squared exponential, one then induces smooth changes in the conditional densities with x . This favours similarity in $f(y|x)$ and $f(y|x')$ when x and x' are close together. Instead of a Gaussian process, one can potentially use a spline-based model for

the atoms, which was implemented by Wang and Dunson (2007) with a focus on modeling of stochastically ordered densities and hypothesis testing.

Even in the fixed- π case, DDP mixtures are extremely flexible. However, in many settings one could more parsimoniously characterize changes in the conditional distribution $f(y|x)$ with x by allowing the weights in the mixture to vary with x . The kernel stick-breaking process described in Section 1.2.3 provides one approach to this problem. Another possibility is to induce a model for the conditional distribution $f(y|x)$ through a Dirichlet process mixture model for the joint distribution of y and x (Müller *et al.* (1996). This approach is both simple and highly flexible, leading to a predictor-dependent mixture of linear regressions when y and x are continuous and modeled using a DPM of Gaussians. However, there are two major problems. Firstly, the approach is only appropriate if x can reasonably be considered as a random variable, though one can potentially consider the joint model for y and x as an auxiliary model that is only defined to induce a coherent model on the conditional distribution. Second, there is a pitfall that often arises due to the structure of the Dirichlet process mixture model, which allocates subjects to clusters based on both y and x , while favouring few clusters. When there are more than a few predictors, it tends to be the case that the x component of the likelihood dominates, so that clusters are introduced primarily to improve fit of the predictor component. There is relatively little gain in introducing clusters that improve prediction of y given x . Hence, we have observed poor predictive performance for this approach in a variety of settings, with the performance particularly bad when y is categorical and predictors are continuous and non-Gaussian (unpublished work with Richard Hahn and Deepak Agarwal).

One possibility for addressing the second problem is to use a more flexible prior than the Dirichlet process in order to allow cluster allocation to differ between the y and x components. This can be accomplished using the local partition process (Dunson, 2008), but this approach has not yet been applied in the setting of conditional distribution modeling in regression. Another possibility is to consider x as fixed, and modify the Dirichlet process to allow the weights to vary with x . The order-based DDP (Griffin and Steel, 2006) accomplishes this by using a common set of stick-breaking variables in the Sethuraman (1994) specification of the Dirichlet process, but with the ordering in these variables dependent on x . The local Dirichlet process (Chung and Dunson, 2009a) is a simpler alternative. In this approach, one places an atom $\theta_h \sim P_0$ and stick-breaking weight $V_h \sim \text{Beta}(1, \alpha)$ at location Γ_h for $h = 1, \dots, \infty$. Then, we let $P_x = \sum_{h \in \mathcal{L}_x} V_h \prod_{l < h} (1 - V_l) \delta_{\theta_h}$, with \mathcal{L}_x denoting the subset of locations that fall within a neighborhood of location x . One obtains $P_x \sim DP(\alpha P_0)$ marginally at each location, while inducing dependence in P_x and $P_{x'}$ through including shared stick-breaking weights and atoms within the region of overlap in the neighborhoods around x and x' .

In complex and high-dimensional settings, there can be some difficulties implementing posterior computation in models allowing the mixture weights to vary with x due to lack of conjugacy. One trick that is often used to induce conjugacy in categorical data analysis is data augmentation with underlying normal variables (Albert and Chib, 1993; 2001). Suppose we have a finite mixture model with k levels and with predictor dependent weights. Then, letting $z_i = h$ denote that subject i is allocated to component h , one could use a kernel continuation ratio probit model with

$$\Pr(z_i = h \mid z_i \geq h, x_i) = \Phi\left(\alpha_h + \sum_{l=1}^p \beta_{hl} \eta_l(x_i)\right) = \Phi(w_i' \theta_h), \quad h = 1, \dots, k-1,$$

where η_1, \dots, η_p are prespecified basis functions (e.g. kernels located on a fixed grid). This model can be equivalently expressed in augmented form as:

$$z_i = \arg \min_h \{z_{ih} > 0\}, \quad z_{ih} = 1(z_{ih}^* > 0), \\ z_{ih}^* \sim N(w_i' \theta_h, 1), \quad h = 1, \dots, k-1, \quad z_{ik} = 1.$$

In particular, the allocation of individual i to a mixture component is characterized through a discrete Gaussian process in which we repeatedly sample latent normal variables until obtaining a positive observation. Using this latent variable formulation and specifying Gaussian or mixture of Gaussian priors for the coefficients θ_h , a straightforward Gibbs sampler can be implemented for posterior computation. It is also possible in such a specification to conduct variable or basis selection by using a mixture prior with a mass at zero for the coefficients θ_h . By allowing $k \rightarrow \infty$, we obtain a nonparametric Bayesian specification which avoids the need to choose a finite number of mixture components. This process was proposed by Chung and Dunson (2009b) and is referred to as the probit stick-breaking process (PSBP).

B. Posterior computation

One of the hurdles to overcome in implementing nonparametric Bayes analyses is the need to conduct posterior computation for a model with infinitely-many parameters. This initially seems to be an impossibility that cannot be overcome. However, there is a rich literature proposing a variety of simple and efficient computational algorithms that bypass the infinite dimensionality issue in a variety of ways. The most commonly used approach relies on marginalizing out the infinitely many parameters in the unknown distribution P to obtain a prior for the finitely many realizations from P represented in the sample. This strategy is referred to as collapsed, marginal or Polya urn Gibbs sampling, with MacEachern and Müller (1998) providing a good reference in the Dirichlet process case. Although this approach is quite convenient to implement, there are two disadvantages. Firstly, one loses the ability to do inferences

on functionals of P , which may be of interest in certain cases. Second, such approaches typically cannot be applied in more complex settings involving generalizations of DP priors or alternatives, because in such settings it is often not possible to obtain simple expressions for conditional distributions of the realizations marginalizing out P .

Ishwaran and James (2001) proposed an alternative approach, which is referred to as conditional or blocked Gibbs sampling. Their approach, which was designed for stick-breaking priors including the Dirichlet process and many other cases, is based on the approximation:

$$P = \sum_{h=1}^{\infty} V_h \prod_{l < h} (1 - V_l) \delta_{\theta_h} \approx \sum_{h=1}^{N-1} V_h \prod_{l < h} (1 - V_l) \delta_{\theta_h} + \prod_{l < N} (1 - V_l) \delta_{\theta_N}.$$

Truncation approximations to stick-breaking representations of the Dirichlet process were previously proposed by Muliere and Tardella (1998). Using the truncated version, one can apply a standard Gibbs sampler for posterior computation, which alternates between (1) allocating each individual to one of the N clusters by sampling from multinomial conditional distributions; (2) updating the stick-breaking weights V_1, \dots, V_{N-1} from conditionally conjugate Beta distributions; and (3) updating the atom θ_h by sampling from the conditional distribution obtained by updating the prior P_0 with the likelihood for those subjects allocated to cluster h , for $h = 1, \dots, N$. These steps are no more difficult to implement than Gibbs samplers for finite mixtures. However, unlike in the finite mixture case, N can be viewed as an upper bound on the number of clusters, with the number of occupied components varying across the MCMC samples.

Both the marginal and conditional Gibbs samplers can experience mixing problems in which the samples tend to remain for long periods in local regions of the space of configurations of subjects to clusters. This is not a problem unique to nonparametric Bayes models, but also occurs in finite mixture models due to multimodality in the posterior distribution. There has been a broad variety of approaches proposed in the literature for improve mixing in the presence of multimodality. In the setting of DPMS, split-merge algorithms were proposed by Jain and Neal (2004). Evolutionary Monte Carlo is another possibility (Goswami *et al.*, 2007). Algorithms have also been proposed for avoiding the need for truncation in conducting conditional Gibbs sampling, with Walker (2007) proposing a slice sampler and Papaspiliopoulos and Roberts (2008) developing a retrospective sampler, which also includes label switching moves to accelerate mixing. I have observed a tendency of the slice sampler to be slow mixing, while retrospective sampling is somewhat complicated to implement. Papaspiliopoulos (2008) recently proposed an exact block Gibbs

sampler, which combines the two approaches, leading to a simple and efficient algorithm.

Acknowledgements

This work was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

References

- Albert, J.H. and Chib, S. (1993). Bayesian-analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.
- Albert, J.H. and Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics*, **57**, 829–836.
- Bauer, D.J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, **42**, 757–786.
- Boucher, K.M., Slattery, M.L., Berry, T.D., Quesenberry, C. and Anderson, K. (1998). Statistical methods in epidemiology: A comparison of statistical methods to analyze dose-response and trend analysis in epidemiologic studies. *Journal of Clinical Epidemiology*, **51**, 1223–1233.
- Chung, Y. and Dunson, D.B. (2009a). The local Dirichlet process. *Annals of the Institute of Statistical Mathematics*.
- Chung, Y. and Dunson, D.B. (2009b). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, **104**, 1646–1660.
- 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.
- Dunson, D.B. (2008). Nonparametric Bayes applications to biostatistics. *Bayesian Nonparametrics*, (ed. N. Hjort, C. Holmes, P. Müller and S. Walker), Chapter 7. Cambridge University Press, Cambridge.
- Dunson, D.B. (2009). Nonparametric Bayes local partition models for random effects. *Biometrika*, **96**, 249–262.
- Dunson, D.B. and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika*, **95**, 307–323.
- 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.
- Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, **41**, 578–588.
- Gage, T.B. (2003). Classification of births by birth weight and gestational age: An application of multivariate mixture models. *Annals of Human Biology*, **30**, 589–604.
- 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.

- Goswami, G., Liu, J.S. and Wong, W.H. (2007). Evolutionary Monte Carlo methods for clustering. *Journal of Computational and Graphical Statistics*, **16**, 855–876.
- Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
- Griffin, J.E. and Steel, M.F.J. (2006). Order-based dependent Dirichlet process. *Journal of the American Statistical Association*, **101**, 179–194.
- Greenland, S. (1995). Dose-response and trend analysis in epidemiology – alternatives to categorical analysis. *Epidemiology*, **6**, 356–365.
- Jacobs, R.A., Peng, F.C. and Tanner, M.A. (1997). A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks*, **10**, 231–341.
- 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.
- 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.
- Jordan, M.I. and Jacobs, R.A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**, 181–214.
- Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. 1. Density estimates. *Annals of Statistics*, **12**, 351–357.
- Longnecker, M.P., Klebanoff, M.A., Zhou, H.B. and Brock, J.W. (2001). Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *Lancet*, **358**, 110–114.
- MacEachern, S.N. (1999). Dependent nonparametric process. *ASA Proceeding of the Section on Bayesian Statistical Science*, American Statistical Association, Alexandria, VA.
- McIntire, D.D. and Leveno, K.J. (2008). Neonatal mortality and morbidity rates in late preterm births compares with births at term. *Obstetrics and Gynecology*, **111**, 35–41.
- Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of Ferguson–Dirichlet priors. *Canadian Journal of Statistics*, **2**, 283–297.
- Müller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**, 67–79.
- Nohr, E.A., Vaeth, M., Baker, J.L., Sorensen, T.I.A., Olsen, J. and Rasmussen, K.M. (2008). Combined associations of prepregnancy body mass index and gestational weight gain with the outcome of pregnancy. *American Journal of Clinical Nutrition*, **87**, 1750–1759.
- Papaspiliopoulos, O. (2008). A note on posterior sampling from Dirichlet process mixture models. *Working Paper*, 08–20, Centre for Research in Statistical Methodology, University Warwick, Coventry, UK.
- Papaspiliopoulos, O. and Roberts, G.O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95**, 169–186.
- Peng, F.C., Jacobs, R.A. and Tanner, M.A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, **91**, 953–960.
- Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.
- Slaughter, J.C., Herring, A.H. and Thorp, J.M. (2009). A Bayesian latent variable mixture model for longitudinal fetal growth. *Biometrics*, **65**, 1233–1242.
- Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society B*, **62**, 795–809.

- Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics – Simulation and Computation*, **36**, 45–54.
- Wang, L. and Dunson, D.B. (2007). Bayesian isotonic density regression. Discussion Paper, Department of Statistical Science, Duke University.
- Wilcox, A.J. (2001). On the importance – and the unimportance – of birthweight. *International Journal of Epidemiology*, **30**, 1233–1241.