Computing Nonparametric Hierarchical Models

Article · August 1998

DOI: 10.1007/978-1-4612-1732-9_1 · Source: CiteSeer

2 authors, including:

Michael David Escobar
University of Toronto
63 PUBLICATIONS **5,440** CITATIONS

Computing Nonparametric Hierarchical Models

Michael D Escobar
Mike West

ABSTRACT Bayesian models involving Dirichlet process mixtures are at the heart of the modern nonparametric Bayesian movement. Much of the rapid development of these models in the last decade has been a direct result of advances in simulation-based computational methods. Some of the very early work in this area, circa 1988-1991, focused on the use of such nonparametric ideas and models in applications of otherwise standard hierarchical models. This chapter provides some historical review and perspective on these developments, with a prime focus on the use and integration of such nonparametric ideas in hierarchical models. We illustrate the ease with which the strict parametric assumptions common to most standard Bayesian hierarchical models can be relaxed to incorporate uncertainties about functional forms using Dirichlet process components, partly enabled by the approach to computation using MCMC methods. The resulting methology is illustrated with two examples taken from an unpublished 1992 report on the topic.

1 Introduction

It is news to no-one that the practice of hierarchical modeling – the mainstay of the applied Bayesian statisticians toolbox – has simply exploded in the last decade due to the development of Markov chain Monte Carlo (MCMC) methods for simulation-based implementation and analysis. At the conceptual heart of hierarchical modeling lies the structurally simplifying nesting/linking of submodels that allows the statistician to break a complex data structure and problem into components that are then naturally assembled in the tree structure of an encompassing hierarchical model. The methods of nested MCMC simulation analysis now standard in the Bayesian world are simply tailor-made for this hierarchical structure (Gilks, Richardson, and Spiegelhalter 1996, for example).

In applications of hierarchical models, as with all parametric models, there are almost inevitable concerns about the sensitivity of resulting inferences to assumed forms of component distributions. Hierarchical models require specification of distributions for parameters and, often, hyperparameters, about which there is usually considerable uncertainty. Hence it is of interest to combine the developments of modern approaches to nonparametric, and semiparametric, modeling of distribution functions with

hierarchical models. The advances in applied nonparametric modeling in recent years have also been driven, in the main, by the development of MCMC methods, and the synthesis of models is eased and indeed naturally facilitated by these approaches to implementation and computation. This chapter describes and illustrates this synthesis using Dirichlet process models for uncertain functional forms of distributions for parameters in hierarchical models.

This chapter first provides some discussion and background on Dirichlet processes and of the implications and interpretations of the priors for the parameters of Dirichlet processes in specific modeling frameworks. We then describe MCMC algorithms for posterior sampling in models with such components, and some strategies for choosing prior distributions for the parameters of Dirichlet processes. Specific extensions to hierarchical models are keyed and illustrated in two data analyses.

We conclude this introduction with a collection of references related to the developments in this area in recent years. Much of the work in nonparametric modeling traces its origins back to the original papers of Freedman (1963), Ferguson (1973, 1974), and Blackwell and MacQueen (1973), though applications were quite limited until the late 1980s. The introduction of MCMC methods in this area began with Escobar (1988), later published in Escobar (1992, 1994). Various applied developments in the early 1990s, and more recently, include works on density estimation and related matters (Escobar and West 1995, based on an original technical report from 1991; West 1990, 1992), mixture deconvolution (Turner and West 1992, West and Cao 1993, West and Turner 1994), applications in hierarchical and prior modeling (Escobar 1995, West 1997), regression and multivariate density estimation (Müller, Erkanli and West 1996, Erkanli, Stangl and Müller 1993, West, Müller and Escobar 1994), design (Bush and MacEachern 1996), and time series (Müller, West and MacEachern 1997), among others. More recent work has extended the range of applications in regression contexts and hierarchical/prior modeling in significant applications (Kuo and Mallick 1998, Mukhopadhyay and Gelfand 1997, West 1997).

Underlying these more applied developments were innovations in computation and the development of new MCMC schemes, including key contributions by Doss (1994), Bush and MacEachern (1996), MacEachern (1994), MacEachern and Müller (1994), West, Müller and Escobar (1994), and alternative, non-MCMC methods such as Liu (1996). Much more detail and perspective on algorithms and computation is provided in MacEachern (1998, this volume).

2 Notation and Perspectives

Consider a collection of random quantities $Y_i, (i = 1, \ldots, I)$, where Y_i is modeled as drawn from a specific distribution $F(Y_i|\theta_i, \zeta_i)$, independently over i given parameters θ_i and ζ_i. In any specific problem, the partition of parameters into the two sets θ_i and ζ_i will be specified. Write $Y = (Y_1, \ldots, Y_I)$, $\theta = (\theta_1, \ldots, \theta_I)$ and $\zeta = (\zeta_1, \ldots, \zeta_I)$. Each Y_i might be thought of as representing observations for an "individual". Let the parameters θ_i be independent draws from some prior distribution $G(\cdot|\lambda)$, characterized by a parameter λ. Denote the prior for ζ by H_1 and the prior for λ by H_2. It is assumed that Y and ζ are conditionally independent of λ given θ. It is also assumed that given Y, ζ and λ, quantities θ and G are conditionally independent of everything else in the hierarchical model. Here the term hierarchical model is meant in the broadest sense: it simply means a model with many levels, structured in terms of sequences of nested conditional distributions. A key point of this paper is to show how to model uncertainties about the functional form of G using Dirichlet processes.

A Dirichlet process prior for G is determined by two parameters: a distribution function $G_0(\cdot)$ that defines the "location" of the Dirichlet process prior, and a positive scalar precision parameter α. The precision parameter determines the concentration of the prior for G around the prior "guess" G_0, and therefore measures the "strength of belief" in G_0. By way of notation we write $G \sim \mathcal{D}(G|G_0, \alpha)$. The Dirichlet process, therefore, adds a further stage to the hierarchical model, and formally allows for the modeling of deviations away from the specific distribution G_0. So G_0 may be viewed as a "baseline" prior such as might be used in a typical parametric analysis, e.g., a normal prior in a normal linear model, and the framework enables the analysis of sensitivity to the assumptions of the baseline parametric model. For large values of α, a sampled G is very likely to be close to G_0. For small values of α, a sampled G is likely to put most of its probability mass on just a few atoms. In the latter case, when the mass of G is concentrated on a few atoms, the Dirichlet process is often used to model a finite mixture model. This is shown in Escobar and West (1995) and West, Müller, and Escobar (1994), and is discussed in more detail in Section 3.2.

The above notation is meant to be very general in order to accommodate almost any hierarchical model. To illustrate this, consider the following typical linear model. Assume one observes $X_i = \beta_{1i}W_{1i} + \beta_{2i}W_{2i} + \epsilon_i$ where the W's are known covariates. For $j = 1, 2$, define $\beta_j = (\beta_{j1}, \ldots, \beta_{jI})$. Assume that, for each i, the β_{1i} and β_{2i} are independently normally distributed, and that β_{ji} has mean and variance μ_j and τ_j^2, respectively, for each $j = 1, 2$. Assume the ϵ_i to be zero-mean, independent normals with variance σ^2. Also, assume that (μ_j, τ_j^2) has the usual normal/inverse gamma prior distribution with parameter vector A_j, and let σ^2 have an inverse gamma prior with parameter vector B. In this example, the X_i's corresponds to the Y_i's; we

might now ask what corresponds to the θ_i's? The answer depends on which distribution one is uncertain about. For example, presently the β_{1i}'s are assumed to come from a normal population. However, maybe the normal assumption seems too restrictive – it might be desirable to have the β_{1i}'s from a distribution which is "near" a normal distribution. So, in this case θ_i would be β_{1i}, λ would be A_1, and ζ would be (β_2, σ^2) with $\zeta_i = (\beta_{2i}, \sigma^2)$. If one wishes to model uncertainty about the joint distribution of β_1 and β_2, then θ_i would be (β_{1i}, β_{2i}), λ would be a vector concatenating the elements of A_1 and A_2, and ζ would simply be the scalar σ^2 with each $\zeta_i = \sigma^2$.

Now suppose that, instead of being concerned about the distribution of the β's, one wishes to model uncertainty about the normality assumption for the ϵ_i's. One way of doing this is to first change the model by assuming that the ϵ_i's are from a normal distribution with mean 0 and variance σ_i^2 and the σ_i^2's are a sample from an uncertain distribution $G(\cdot)$ that does not necessarily put all its weight on a single atom (Brunner 1995). This generates the class of scale mixture of normals (see Andrews and Mallows 1974, West 1987). Here then, θ_i would be σ_i^2, the Y_i's would still be the X_i, and ζ would be the matrix with the j^{th} column being $(\beta_{ji}; i = 1, \ldots, I)$ with $\zeta_i = (\beta_{1i}, \beta_{2i})$. The details of this are contained in Section 3.2, after a more technical description of the Dirichlet process.

The notation is designed so that given ζ, Y, and λ, then θ and G are conditionally independent of all remaining variables. Once θ and G are isolated, then the methods below can be used to model the uncertainty in the functional form of G for many hierarchical models. Note, from the example, that the variables θ, Y, λ, and ζ do not have to be scalars, but are allowed to be multidimensional. For an example where θ_i is uncountably infinite dimensional, see Tomlinson (1998). Also note that there is a symmetry between the prior and the likelihood. Both the prior and the likelihood are basically distribution functions and a Dirichlet process can be used to model uncertainty about either. Finally, note that Y does not need to be observed. Since everything in this paper is eventually conditional upon Y, it is clear that the framework can be embedded in a super-model which is being analysed using MCMC. Examples of this appear in Roeder et al (1998), Mukhopadhyay and Gelfand (1997), and West (1997).

3 Posterior Sampling in Dirichlet Process Mixtures

A specific MCMC algorithm is discussed in three parts. First, the computational scheme for the Dirichlet process itself is discussed, involving simulations of the parameters θ_i. This lies at the heart of more standard nonparametric models in density estimation and other applications (Escobar and West 1995, West, Müller and Escobar 1994). Then we discuss computational issues arising when we include learning about the precision

parameter α of the Dirichlet process. Here we provide some additional commentary on the role of this parameter in a Dirichlet process model more generally. Finally, we describe how the overall hierarchical model analysis is structured by combining these components into the larger model structure.

3.1 Posterior sampling under a Dirichlet process

In our analysis, the random distribution G is integrated out of consideration, the analysis focusing entirely on the θ parameters assumedly generated from G. There are two basic components of the algorithm and approach. The first step uses an algorithm based on Escobar (1988, 1992, 1994) which uses a Polya urn scheme representation; the second step relies on the cluster structure of the θ's and is based on the work of MacEachern (1994). The combination of these two steps is introduced in West, Müller, and Escobar (1994), and Bush and MacEachern (1996).

Consider the distribution of $(Y_i|\theta_i, \zeta, \lambda)$ with ζ and λ held fixed in this subsection. Recall that the θ_i's come from G, and that $G \sim \mathcal{D}(G|G_0, \alpha)$. Assume here that both α and G_0 are known. This structure results in a posterior distribution which is a mixture of Dirichlet processes (Antoniak 1974). Using the Polya urn representation of the Dirichlet process (Blackwell and MacQueen 1973), the joint posterior distribution of $[\theta|Y, \zeta, \lambda]$ has the form

$$[d\theta|Y, \zeta, \lambda] \propto \prod_{i=1}^{n} f(Y_i|\theta_i, \zeta) \frac{\alpha\, G_0(d\theta_i|\lambda) + \sum_{k<i} \delta(d\theta_i|\theta_k)}{\alpha + i - 1},$$

where $f(Y_i|\theta_i, \zeta)$ is the density or the probability function of $F(Y_i|\theta_i, \zeta)$ at θ_i and ζ, and where $\delta(d\theta_i|\theta_k)$ is the distribution which is a point mass on θ_k. At this point, the random distribution G has been integrated out. Integrating out G simplifies the algorithm since we now sample only the the θ's.

The above equation clarifies the effect of the precision parameter α. In the limiting case $\alpha \to \infty$ we have

$$[d\theta|Y, \zeta, \lambda] = \prod_{i=1}^{n} G_b(d\theta_i|Y_i, \zeta, \lambda) \propto \prod_{i=1}^{n} f(Y_i|\theta_i, \zeta_i)\, G_0(d\theta_i|\lambda).$$

Here $G_b(d\theta_i|Y_i, \zeta, \lambda) \propto f(Y_i|\theta_i, \zeta_i) G_0(d\theta_i|\lambda)$ is the "baseline" posterior, that is, the posterior assuming θ_i to come from the baseline prior G_0. As α goes to zero, the estimation of θ_i is a little more complicated to understand. The posterior for θ_i is based largely on other θ_k's which are near Y_i, so that inference for θ_i heavily depends on Y_i and nearest "neighboring" Y_k's.

Gibbs sampling exploits the simple structure of the conditional posteriors for the elements of θ, resulting in the following conditional distribution. For

each $i = 1, \ldots, I$,

$$[d\theta_i | \{\theta_k, k \neq i\}, Y, \zeta, \lambda] \sim q_0 \, G_b(d\theta_i | Y_i, \zeta, \lambda) + \sum_{k \neq i} q_k \, \delta(d\theta_i | \theta_k) \qquad (1.1)$$

with the following ingredients:

- $G_b(\theta_i | Y_i, \zeta, \lambda)$ is the baseline posterior distribution noted above;
- $q_0 \propto \alpha \int f(Y_i | \theta_i, \zeta_i) dG_0(\theta_i | \lambda)$, just α times the density of the marginal distribution of Y_i under the baseline prior $G_0(\cdot | \lambda)$;
- $q_k \propto f(Y_i | \theta_k, \zeta_i)$, the density of the marginal distribution of Y_i;
- the quantities q_i are standardized to unit sum, $1 = q_0 + \sum_{k \neq i} q_k$.

Escobar (1994) and Escobar and West (1995) give the above formulæ in normal models, and they appear from the general case in Escobar (1992).

Iterative posterior simulation generates new values of the θ_i from conditional distributions that are modified forms of the above structure, as discussed further below. The computations are very straightforward if we are assured that G_b is of manageable form, that the integrations required to compute q_0 may be cheaply performed, and that the density f is easily evaluated. There has been some criticism of the above method because of the potential difficulty in calculating the marginal distribution needed to obtain q_0. In regards to that, please note the following:

1. When G_0 is a conjugate prior, then the marginal is known analytically. In many large hierarchical models, conjugate priors are often used in the middle stages of the model. Also, since the interest is in finding distributions "around" the centering distribution G_0, the fact that G_0 is the conjugate prior is often not that strong an assumption because G's sampled from $\mathcal{D}(G_0, \alpha)$ can be fairly flexible.

2. If we must compute the marginal distribution numerically, we can often do so extremely quickly and accurately. Since the integration is over an individual θ_i and not over all of θ, the dimension of the integral is the dimension of the individual θ_i. Also, the parameters ζ and λ are fixed for this integration. Therefore, the marginal is typically a low dimensional integral, and there are many efficient methods to numerically integrate low dimensional integrals. For example, if $f(Y_i | \theta_i, \zeta_i) G_0(d\theta_i | \lambda)$ can be put into the form of a smooth function times a normal or gamma density, then one can use Gaussian quadrature rules. For an overview of these and other integration methods, see Naylor and Smith (1982) or Evans and Swartz (1995).

3. It may be that the marginal distributions cannot be calculated efficiently. In such a case, the methods in this subsection might not

apply directly. There is some work on methods to perform MCMC without calculating the marginal distribution needed for q_0. With such an alternative algorithm, one would use this technique to sample from the posterior conditional distribution for θ, and then the rest of the methods in this paper can be used to include a Dirichlet process enhancement into the hierarchical model. See, for example, MacEachern and Müller (1994) for a proposed alternative method to sample from the conditional distribution of θ.

The required computations are reduced by the fact that the distinct θ_i's typically reduce to fewer than I due to the clustering of the θ_i's inherent in the Dirichlet process (Antoniak 1974). Using the superscript $*$ to denote distinct values, suppose that the conditioning quantities θ_k's concentrate on $I^* \leq I-1$ distinct values θ_k^*, with some n_k taking this common value. Then, the above formula can be rewritten as:

$$[d\theta_i|\{\theta_k, k \neq i\}, Y, \zeta, \lambda] \sim q_0\ G_b(d\theta_i|Y_i, \zeta, \lambda) + \sum_{k=1}^{I^*} n_k q_k^*\ \delta(d\theta_i|\theta_k^*),$$

with $q_k^* \propto f(Y_i|\theta_k^*, \zeta)$, and $1 = q_0 + \sum_k n_k q_k^*$. Besides simplifying notation, the cluster structure of the θ_i can also be used to improve the efficiency of the algorithm.

Simulation of θ_i from the above conditional is now evidently straightforward: with conditional probability proportional to $n_k q_k$, the sampled value is one of those defining the existing "clusters;" otherwise, the sampled θ_i is drawn anew from the conditional posterior $G_b(d\theta_i|Y_i, \zeta, \lambda)$. When using the above conditional distribution in an MCMC algorithm, there may occur problems if the sum of the q_k^*'s becomes very large relative to q_0 on any iteration. This occurs when the Markov chain has "stabilized" on a small number of "clusters," and it is then unlikely to generate a "new" value of θ_k^*. In order to prevent the algorithm from getting stuck on a small set of θ_j^*'s in this way, it is helpful to "remix" the θ_j^*'s after every step. This improvement is used in Bush and MacEachern (1996) and West, Müller, and Escobar (1994), and it is a combination of the above algorithm with the algorithm developed in MacEachern (1994). The latter paper describes a method which considers the cluster membership structure. In that algorithm it is necessary to sample the new cluster membership structure by computing complex expressions which heavily rely on conjugacy, and which therefore relatively limit its applicability in hierarchical models. However, with the original Escobar method described above, each step of that algorithm produces the cluster memberships as a simple by-product. So the computational difficulties with the MacEachern algorithm are solved when combined with the Escobar algorithm. Also, the combined algorithm mixes better than the Escobar algorithm alone because the θ_k^*'s are resampled at each step providing more movement in the MCMC sampler's which in turn improves convergence.

Some notation is now introduced to define the remixing algorithm. Conditioning on I^*, introduce indicators $\mathcal{S}_i = j$ if $\theta_i \equiv \theta_j^*$ so that, given $\mathcal{S}_i = j$ and θ^*, $y_i \sim F_i(.|\theta_j^*)$. The cluster structure, which is called a configuration by West (1990), is defined by the set of indices $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$. The set $\mathcal{S}$ determines a one-way classification of the data $Y = \{y_1, \ldots, y_n\}$ into I^* distinct groups or clusters; the $n_j = \#\{\mathcal{S}_i = j\}$ observations in group j share the common parameter value θ_j^*. Now, define J_j as the set of indices of observations in group j; i.e., $J_j = \{i : \mathcal{S}_i = j\}$. Let $Y_{(j)} = \{Y_i : \mathcal{S}_i = j\}$ be the corresponding group of observations. Once the set $\mathcal{S}$ is known, the posterior analysis of the θ_j^*'s devolves into a collection of I^* independent analyses. Specifically, the θ_j^*'s are conditionally independent with posterior densities

$$p(\theta_j^*|Y, \mathcal{S}, I^*, \zeta, \lambda) \equiv p(\theta_j^*|Y_{(j)}, \mathcal{S}, I^*, \zeta, \lambda) \propto \prod_{i \in J_j} f_i(y_i|\theta_j^*, \zeta_i) dG_0(\theta_j^*, \lambda),$$

for $j = 1, \ldots, I^*$. Note that this is just the posterior of θ_j^* given several Y_i's sampled from the $F(.|\theta_j^*, \zeta_i)$.

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

(i) how to sample from the posterior distribution of θ_i given all the other parameters if G_0 were the prior distribution for θ_i,

(ii) how to evaluate the marginal distribution for Y_i assuming that G_0 is the prior distribution of θ_i,

(iii) how to evaluate the density function f of F, and

(iv) how to sample from the posterior of θ_j^* when there are several observations sampled from $f(Y_i|\theta_j^*, \zeta_i)$.

3.2 Learning about the precision parameter

The precision parameter, α, of the Dirichlet process is extremely important for the model. When α is small, then G tends to concentrate on a few atoms of probability. When α is large, then G is a distribution with many support points and the nonparametric model is "closer" to G_0, the baseline model. These features are to be borne in mind when considering priors for α. In this subsection, we discuss various effects of the parameter α and then issues arising in learning about α within the MCMC analysis.

Consider the following simple example to illustrate these points. Suppose the baseline parametric model is $Y_i = \mu_i + \epsilon_i$ with $\mu_i \sim N(\xi, \tau^2)$, $\epsilon_i \sim N(0, \sigma^2)$, $\sigma^2 \sim H_4$, and with some appropriate hyperpriors for ξ and τ^2. Now, if we were uncertain about the normality assumption for μ_i, we could identify μ_i as the θ parameter and adopt the Dirichlet process framework with $G_0 = N(\xi, \tau^2)$. A high degree of belief in normality would imply a prior

on α favoring large values. Now, if instead of modeling uncertainty about the functional form of the distribution of μ_i, consider modeling uncertainty about the constancy of variance of the ϵ_i. This can be done by relaxing the assumptions so that $\epsilon_i \sim N(0, \sigma_i^2)$, identifying $\theta_i = \sigma_i^2$ and basing the Dirichlet process around $G_0 = H_4$. To assign some prior weight to the original baseline model, the prior for α should have mass near zero. For more discussion about the prior for α and the model interpretations, see Escobar (1995, 1997) and Escobar and West (1995).

Learning about α from the data may be addressed with a view to incorporating α into the Gibbs sampling analysis. Assume a prior density $p(\alpha)$ (which may depend on the sample size I); this implies a prior for I^*, the number of distinct components, namely $P(I^*|I) = E[P(I^*|\alpha, I)]$. Using results in Antoniak (1974), the distribution of number of components may be written as

$$P(I^*|\alpha, I) = c_I(I^*)I!\ \alpha^{I^*}?(\alpha)/?(\alpha + I), \qquad (I^* = 1, 2, \ldots, I),$$

and $c_I(I^*) = P(I^*|\alpha = 1, I)$, not involving α. If required, the factors $c_I(I^*)$ are easily computed using recurrence formulæ for Stirling numbers. This is important, for example, in considering the implications for priors over I^* of specific choices of priors for α (and *vice-versa*) in the initial prior elicitation process. As an aside, note that there is a great deal of flexibility in representing prior opinions about I^* through choices of prior for α. Also, note that posterior distribution of α given I and I^* is conditionally independent of all the other variables in the model.

If the prior for α is a gamma distribution, or a mixture of gamma distributions, we have access to a neat data augmentation device for sampling α (Escobar and West 1995, West 1992). For example, let α have a gamma prior with a shape parameter $a > 0$ and scale parameter $b > 0$ (which we may extend to include a 'reference' prior by letting $a \to 0$ and $b \to 0$), namely $\alpha \sim Ga(a, b)$. A data augmentation technique (Tanner and Wang 1987) can be used to generate conditional samples. At each Gibbs iteration, the currently sampled values of I^* and α allow us to draw a new value of α by first sampling, conditionally on α and I^* fixed at their most recent values, a latent variable η from the simple beta distribution $(\eta|\alpha, I^*) \sim Be(\alpha + 1, I)$, a beta distribution with mean $(\alpha + 1)/(\alpha + I + 1)$. Then, a new α value is sampled from a mixture of two gamma distributions based on the same I^* and the new η; in detail,

$$(\alpha|\eta, I^*) \sim \pi_\eta Ga(a + I^*, b - log(\eta)) + (1 - \pi_\eta)Ga(a + I^* - 1, b - log(\eta))$$

where the weights π_η are defined in odds form by

$$\pi_\eta/(1 - \pi_\eta) = (a + I^* - 1)/(I(b - log(\eta))).$$

Note that these distributions are well defined for all gamma priors, all η in the unit interval and all $I^* > 1$.

The above method of sampling α is used in the remaining sections of this paper. Before ending this section, we note an alternative prior for α used in Escobar (1994, 1997). Here $p(\alpha)$ is discrete, supported on a grid of points, $A_l, l = 1, \ldots, L$, with prior weights defined as $P(\alpha = A_l)$. Using Bayes theorem and the formula for $P(I^*|\alpha, I)$ defined above, the conditional posterior distribution is

$$P(\alpha = A_l|I^*, I) \propto P(\alpha = A_l)I!\ A_l^{I^*}?\,(A_l)?\,(A_l + I).$$

With the grid of points, it is necessary that L be large enough so that the MCMC chain can possible travel between different support points A_l. One choice of grid values is to let $\log(A_l)$ be equally spaced, with equal weight, and ranging from about n^{-1} to about n^2. See Escobar (1997) for an implementation of this algorithm.

As a final note about estimating α, Liu (1996) shows how to obtain the maximum likelihood estimate of α. Although this would not be useful to insert in the middle of a MCMC, the reader might use it in an empirical Bayes' approximation.

3.3 Complete Gibbs sampling schemes

We first revisit the model definition and structure. The primary parameters θ_i follow a prior distribution G which itself follows a Dirichlet process prior. The baseline prior G_0 depends on additional hyperparameters λ, and the priors H_1, H_2, and H_3 are adopted for the parameters ζ, λ and α, respectively. Inserting the Dirichlet process prior extends the standard hierarchical model as displayed in the following diagram:

Standard Model				Dirichlet Enhanced		
$[Y_i\|\theta_i, \zeta]$	$\sim$	$F(Y_i\|\theta_i, \zeta)$		$[Y_i\|\theta_i, \zeta]$	$\sim$	$F(Y_i\|\theta_i, \zeta)$
$[\theta_i\|G, \lambda]$	$\sim$	$G(\theta_i\|\lambda)$		$[\theta_i\|G]$	$\sim$	$G(\theta_i)$
$[\zeta]$	$\sim$	$H_1(\zeta)$	$\stackrel{\mathcal{D}}{\Longrightarrow}$	$[G(\cdot)\|\lambda, \alpha]$	$\sim$	$\mathcal{D}(G\|G_0(\cdot\|\lambda), \alpha)$
$[\lambda]$	$\sim$	$H_2(\lambda)$		$[\zeta]$	$\sim$	$H_1(\zeta)$
				$[\lambda]$	$\sim$	$H_2(\lambda)$
				$[\alpha]$	$\sim$	$H_3(\alpha)$

Since the conditional posterior distributions for θ and α are presented in the earlier sections of this paper and since the random distribution function G is integrated out in the calculations, what remains to be specified is the conditional posterior distribution for ζ and λ. Since ζ is conditionally independent of G given Y and θ, the Dirichlet process enhancement does not change the conditional posterior distribution for ζ. With regard to the random variable λ, Blackwell and MacQueen (1973) show that the distinct values of θ_j^* are independent and identically distributed samples from $G_0(\cdot|\lambda)$. Hence, under the Dirichlet process enhancement, the conditional posterior distribution for λ is simply that from the standard model

with the vector θ replaced by the vector of distinct elements θ^*. The Gibbs sampling analysis may now be implemented using the following conditional distributions of the full joint posterior for $[\theta, \zeta, \lambda, \alpha|Y]$:

(a) $[\theta_i|Y, \{\theta_k, k \neq i\}, \zeta, \lambda, \alpha]$ for each $i = 1, \ldots, n$.

(b) $[\theta_j^*|Y, \mathcal{S}, I^*, \zeta, \lambda]$ for each $j = 1, \ldots, I^*$.

(c) $[\zeta|Y, \theta, \lambda, \alpha]$, which, on observing the conditional independence structure of the posterior, reduces to $[\zeta|Y, \theta]$.

(d) $[\lambda|Y, \theta, \zeta, \alpha]$ which, again using the conditional independence structure, becomes just $[\lambda|\theta^*]$ where θ^* is the set of distinct values of θ.

(e) $[\alpha|Y, \theta, \zeta, \lambda]$, reducing to $[\alpha|\theta]$ as described in the previous section. This in turn is structured by first sampling $[\eta|\alpha, I^*]$ and then $[\alpha|\eta, I^*]$, where I^* is just the number of distinct elements in θ.

Gibbs sampling proceeds by simply iterating through (a)–(e) in order, sampling at each stage based on current values of all the conditioning variates. The following are some important, general features of these distributions:

- The distributions for generating the θ_i's in (a) are just those outlined in Section 3.1 for the simple Dirichlet process prior model. Thus, algorithms written to sample posteriors in such models may be directly implemented in the more complex hierarchical modeling framework without modification.

- The conditional posterior distribution (b) of θ_j^* is obtained by combining the baseline prior $G_0(\cdot|\lambda)$ with the likelihood $\prod_{i \in J(j)} f(Y_i|\theta_j^*, \zeta_i)$. That is, the posterior based on a random sample of the Y_i's which are drawn from the same θ_j^*'s. This is discussed in Section 3.1.

- The posterior (c) for sampling the parameter ζ of the sampling model is simply that found in the traditional, baseline hierarchical model used in the usual Gibbs sampling for such models. The Dirichlet process enhancement changes nothing here. This is true because ζ is conditionally independent of G given Y and θ.

- For λ, the hyperparameter of the baseline prior G_0, the required posterior (d) is exactly that arising from the prior $H_2(\lambda)$ combined with the likelihood $\prod_{i=1}^{I^*} G_0(d\theta_i^*|\lambda)$. That is, the posterior based on a random sample of (the distinct) θ_j^*'s drawn from the baseline prior G_0. This is true because the distinct values, θ_j^*, are independent and identically distributed from $G_0(\cdot|\lambda)$.

- The distributions for sampling α in (e) are precisely as discussed in Section 3.2.

Note that the above features simplify the inclusion of a Dirichlet process into a hierarchical model. The fact that there are no changes to the posterior samplers for ζ and very little change to the sampler for λ means that the Dirichlet process prior can be included in the middle of most Bayesian hierarchical models when computation uses MCMC methods – or essentially all applied models. Choice of priors for the Dirichlet process is simplified since G_0 is already modeled by the baseline parametric model. Hence prior specifications reduce to priors for α. Some examples and illustrations are now discussed.

4 An Example with Poisson-Gamma Structure

The numbers of a certain type of eye tracking anomaly for 101 subjects with schizophrenia are reported in Table 1, and a histogram of the data appears in Figure 1. The number of anomalies is believed to be associated with dysfunctions in the prefrontal cortex of the brain. A fuller description of the data along with a frequentist's mixture analysis of this data is contained in Sweeney et al (1993) and Pauler et al (1996).

Table 1

Outcome:	0	1	2	3	4	5	6	7	8	9
Count:	46	14	9	4	2	3	3	3	1	2

Outcome:	10	11	12	14	15	17	22	24	34
Count:	2	2	2	1	2	2	1	1	1

Consider modeling the number of anomalies, Y_i, as Poisson with individual means θ_i. This determines $F(Y_i|\theta_i)$, there being no additional ζ parameter. Further, take a baseline gamma prior G_0, under which the θ_i are independent gamma with specified shape λ_1 and scale λ_2, and set $\lambda = \{\lambda_1, \lambda_2\}$. Then the posterior distribution for known λ has the θ_i independent with $[\theta_i|Y_i, \lambda] \sim Ga(\lambda_1 + Y_i, \lambda_2 + 1)$. In the special case of a degenerate prior $[\lambda]$ and $\alpha \to \infty$, this model is complete.

Extending to a Dirichlet process analysis as outlined in Section 3 results in

$$\begin{aligned} [Y_i|\theta_i] &\sim F(Y_i|\theta_i), \\ [\theta_i] &\sim G(\theta_i), \\ [G|\alpha] &\sim \mathcal{D}(G_0, \alpha), \\ \alpha &\sim Ga(a, b). \end{aligned}$$

The required conditional posteriors (a)-(e) of Section 3.3 are easily derived. Note the special circumstances of no ζ parameter and assuming the prior for λ degenerate at the chosen values of λ_1 and λ_2; therefore steps (c) and (d) are vacuous. Step (e) is covered in Section 3.2, and the remaining steps (a) and (b) follow Section 3.1 to give

$$\begin{aligned}
[\theta_i|Y, \{\theta_k, k \neq i\}, \alpha, \lambda] &\sim q_0\, Ga(\lambda_1 + Y_i, \lambda_2 + 1) + \sum_{k \neq i} q_k \delta(d\theta_i|\theta_k), \\
[\theta_j^*|Y, \mathcal{S}, \lambda] &\sim Ga(\lambda_1 + \sum_{i \in J_j} Y_i, \lambda_2 + n_j),
\end{aligned}$$

with

$$q_0 \propto \alpha \binom{Y_i + \lambda_1 - 1}{Y_i} \left(\frac{\lambda_2}{1+\lambda_2}\right)^{\lambda_1} \left(\frac{1}{1+\lambda_2}\right)^{Y_i}$$

and, for each $k > 0$, $q_k \propto \theta_k^{Y_i} \exp(-\theta_k)/Y_i!$.

Two separate analyses are illustrated, based on baseline priors with λ set at either (0.4,0.1) or (1,1). The values (0.4,0.1) correspond to a method of moment estimate of (λ_1, λ_2) obtained from the complete data set. For the precision parameter, α, we consider two priors: $Ga(.01, .01)$ and $Ga(1, 1)$. The former prior is fairly noninformative, giving reasonable mass to both high and low values of α. The $Ga(1, 1)$ prior favors relatively low values of α. The posteriors under these two priors are quite similar, it being the case that the choice of G_0 is more important for this (model:data) combination. For more on the effect of the prior values for the Dirichlet process on the model, see Escobar (1997).

In the analysis with $\lambda = (0.4, 0.1)$, Figures 2a, 2b, and 2c display the estimated marginal posteriors for θ_i when Y_i is equal to 4, 8, and 12. Under this prior, the inclusion of the Dirichlet process does not change the posterior distribution very much. The baseline model produces posterior distributions which are very close to the posteriors obtained under the two Dirichlet process models. The use of the Dirichlet process sometimes does lead to marginally increased spread in posteriors, reflecting additional uncertainty implied by the use of the Dirichlet structure. In part the concordance arises as the $Ga(0.4, 0.1)$ prior for the Poisson rates is not in conflict with the observed data.

Under the alternative $Ga(1, 1)$ prior, the data observed in the right tail of Figure 1 are "unusual." In this case the marginal distribution of Y_i under the standard model is negative binomial with parameters $p = .5$ and $r = 1$. That is, the marginal distribution of Y_i is equivalent to tossing a fair coin and counting the number of tails before the first appearance of a head. There is only a 0.1% chance of values greater than 9, yet 14% of the observations are greater than 9. When the Dirichlet process prior is used, the data downweights the $Ga(1, 1)$ component and results in posteriors very

different than obtained under the standard/baseline model. Figures 2d, 2e, and 2f display the posterior margins for those θ_i for which Y_i is equal to 4, 8, and 12. Note the posterior distributions for $(\theta_i|Y = 12)$ in particular. Since the baseline prior puts very little probability on data values higher than 10, the baseline posterior distribution puts most of its mass below 10. However, the posteriors under each of the Dirichlet enhanced models are bimodal. The smaller modes appears as "reduced" versions of the baseline posterior, while the major modes favour values in the 10 to 20 range in each case. Hence the nonparametric analysis naturally adapts to and reflects the inherent conflict in the (model:data) match. For related discussion, see Escobar (1997).

Note finally that extensions to include learning about λ are straightforward. For an example of such a baseline parametric model, see George, Makov, and Smith (1994), who show how to construct conditional samplers to perform the analysis. Once this is done, the Dirichlet process enhancement may use the same algorithm to sample λ as the baseline model, but the vector of θ_j^*'s would be substituted in for the vector of θ_i's. We do something along these lines in the following example.

5 An Example with Normal Structure

Gelfand et al (1990) present a data set of weights of rats measured at the same five time points. There are two groups of rats, a control group and a treatment group, with 30 rats in each group. Let Y_{ij} be the weight of the ith rat at the jth time point, with $i = 1, \ldots, I$ and $j = 1, \ldots, J_i$. First, each rat is modeled by a linear growth curve with normal errors,

$$[Y_{ij}|\theta_i, \sigma^2] \sim N(\theta_{i0} + \theta_{i1}X_j, \sigma^2),$$

where X_j is the age at the jth point. The line is defined by the parameter vector $\theta_i = (\theta_{i0}, \theta_{i1})$ and all rats have the same within-group variance σ^2. For G_0 we take the normal distribution

$$[\theta_i|\mu, \Sigma] \sim N(\mu, \Sigma).$$

Note here that $\lambda = \{\mu, \Sigma\}$ and $\zeta = \sigma^2$. Finally, uncertainty about the parameters σ^2, μ, and Σ is modeled with independent priors,

$$\begin{aligned} [\sigma^2] &\sim IG(0.5\nu_1, 0.5\nu_1\nu_2), \\ [\mu] &\sim N(\xi_1, \Xi_2), \\ [\Sigma^{-1}] &\sim W((\xi_3\Xi_4)^{-1}, \xi_3), \end{aligned}$$

where IG is the inverse gamma distribution and W is the Wishart distribution, and the ξ's, Ξ's, and ν's are prior parameters that are specified

in any analysis. As shown in Gelfand et al (1990), the standard/baseline hierarchical model has conditional posterior distributions (where we drop the conditioning variates in each case, for clarity of display):

$$\begin{aligned}
\sigma^2 &\sim IG(0.5(\nu_1+n), 0.5S),\\
\theta_i &\sim N(D_i[\sigma^{-2}X_i^TY_i+\Sigma^{-1}\mu], D_i),\\
\mu &\sim N(V[I\Sigma^{-1}\bar{\theta}+\Xi_2^{-1}\xi_1], V),\\
\Sigma^{-1} &\sim W(U^{-1}, I+\xi_3),
\end{aligned}$$

where $S=\sum_{ij}(Y_{ij}-\hat{Y}_{ij})^T(Y_{ij}-\hat{Y}_{ij})+\nu_1\nu_2$, $\hat{Y}_{ij}=\theta_{i0}+\theta_{i1}X_j$, $n=\sum_{i=1}^I J_i$, $\bar{\theta}=I^{-1}\sum_{i=1}^I\theta_i$, $D_i=(\sigma^{-2}X_i^TX_i+\Sigma^{-1})^{-1}$, $V=(I\Sigma^{-1}+\Xi_2^{-1})^{-1}$, and $U=\sum_{i=1}^I(\theta_i-\mu)(\theta_i-\mu)^T+\xi_3\Xi_4$.

Now the above model is extended to incorporate a Dirichlet process prior for G with the baseline prior G_0 as the initial guess. This results in sampling from the following conditional distributions (again with conditioning variates supressed, for clarity):

$$\begin{aligned}
\sigma^2 &\sim IG(0.5(\nu_1+n), 0.5S)\\
\theta_i &\sim q_0N(D_i[\sigma^{-2}X_i^TY_i+\Sigma^{-1}\mu], D_i)+\sum_{k\neq i}q_k\delta(\theta_i|\theta_k),\\
\theta_j^* &\sim N(D_j^*[\Sigma^{-1}\mu+\sigma^{-2}\sum_{i\in J(j)}X_i^TY_i], D_j^*),\\
\mu &\sim N(V^*[I^*\Sigma^{-1}\bar{\theta}^*+\Xi_2^{-1}\xi_1], V^*),\\
\Sigma^{-1} &\sim W(U^{-1}, I^*+\xi_3),
\end{aligned}$$

where $\bar{\theta}^*=(I^*)^{-1}\sum_{i=1}^{I^*}\theta_i^*$, $D_j^*=(\Sigma^{-1}+\sigma^{-2}\sum_{i\in J(j)}X_i^TX_i)^{-1}$, $V^*=(I^*\Sigma^{-1}+\Xi_2^{-1})^{-1}$, $U=\sum_{j=1}^{I^*}(\theta_j^*-\mu)(\theta_j^*-\mu)^T+\xi_3\Xi_4$, $q_0\propto\alpha f_i$ where f_i is the density at Y_i of the multivariate normal distribution $N(X_i\mu, X_i\Sigma X_i^T+\sigma^2\mathcal{I})$ (and where $\mathcal{I}$ is the identity matrix), and each q_k is proportional to the multinormal density of $Y_i\sim N(X_i\theta_k, \sigma^2\mathcal{I})$.

In extending from the traditional baseline analysis to the Dirichlet process analysis, note again that there is no change to the conditional distribution of σ^2. Note also that the conditional distributions of μ and Σ just have θ replaced by $\{\theta^*\}$. Also, a sample from the conditional of θ_i is either a sample from the conditional distribution of θ_i under the baseline analysis, or θ_i is set equal to a previously sampled value θ_k. The remixing step requires that we know the posterior distribution of θ_j^* when several subjects are sampled from that θ_j^*. Therefore, once a computer program is written to perform Gibbs sampling in the baseline analysis it is easily extended to perform analysis in the Dirichlet enhanced model.

In analysing this data, the same vague priors for σ, μ, and Σ are used as in Gelfand et al (1990). The parameter α is modeled with either a $Ga(1,1)$ prior or a $Ga(0.01, 0.01)$ prior. Analyses in Gelfand et al (1990) and again

in Carlin and Polson (1991), report posterior inferences on μ. Here, by comparison, we report predictive inferences in terms of the predictive distribution for a future θ vector (which, in the baseline analysis, is simply the posterior for μ convolved with a t-distribution.) Results of analysis of just the rats in the control group are reported in Figure 3. Figure 3a displays the plot of the maximum likelihood estimates of the intercept (θ_{0i}) and the slope (θ_{1i}). The other three plots in display contours of the joint predictive distributions under different models. In the earlier studies with the standard hierarchical model, component, the posterior distribution for μ has appeared as unimodal, as in Figure 3b here. Carlin and Polson (1991) report that one observation has a large influence and maybe should be disregarded from any inferences from this data. These inferences contrast with the following observations in Dirichlet enhanced analyses.

Figures 3c and 3d display contour plots of the predictive distribution of θ for the two Dirichlet process analyses. Figure 3c is for the model with $\alpha \sim Ga(.01, .01)$ and Figure 3d is for that with $\alpha \sim Ga(1, 1)$. The predictive distributions under both priors seem to contain ridges of probability, and the potential outlier discovered by Carlin and Polson appears to have been isolated with it's own "mini-mode" of probability. The $Ga(1, 1)$ prior seems to induce more severe clustering of the θ_i's, while the $Ga(.01, .01)$ induces more of a blend between the baseline predictive distribution and the predictive distribution obtained from a standard model with a straight $Ga(.01, .01)$ prior. It appears that the Dirichlet enhanced analysis leads to a natural and appropriate downweighting of the effect of outliers and influential points. Also, the contour plots of predictives under the enhanced analysis naturally control for population heterogeneity.

Figure 4 displays marginal predictive distributions for the slopes of models for rats from both the control group and the treatment group. The lines are fit with the baseline prior model and with the Dirichlet process model with knowledge of the group membership for each rat. Basically, the control group and the treatment group are fit separately and the predictive distribution is simply the average of the predictive distributions for the two groups. Then, to see if the Dirichlet process model might be useful to control for unknown heterogeneity, the predictive distribution is calculated under the Dirichlet process model without knowledge of group membership. In Figure 4a, when $\alpha \sim Ga(.01, .01)$, the two predictive distributions are bimodal, and the predictive distribution fit without knowledge of group membership is somewhat flat between the modes of the other two predictive distributions. However, in Figure 4b, when $\alpha \sim Ga(1, 1)$, the predictive distribution obtained without knowledge of the group membership is very close to that in the analysis where group membership is known.

Acknowledgements

Michael D. Escobar was partially financed by the National Cancer Institute #RO1-CA54852-01, a National Research Service Award from NIMH Grant #MH15758, and by the Natural Sciences and Engineering Research Council of Canada. Mike West was partially financed by the National Science Foundation under grant DMS-9024793.

6 References

[1] Andrews, D.F. and Mallows, C.L. (1974) "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society, Series B*, 36, 99-102.

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

[3] Blackwell, D., and MacQueen, J.B. (1973), "Ferguson Distribution via Polya Urn Schemes," *The Annals of Statistics*, 1, 353-355.

[4] Brunner, L.J. (1995) Bayesian linear regression with error terms that have symmetric unimodal densities, *Journal of Nonparametric Statistics*, 4, 335-348.

[5] Bush, C.A. and MacEachern, S.N. (1996) "A semi-parametric Bayesian model for randomized block designs," *Biometrika*, 83, 275-285.

[6] Carlin, B.P., and Polson, N.G. (1991) "An expected utility approach to influence diagnostics," *Journal Of the American Statistical Association,* 86, 1013-1021.

[7] Cao, G., and West, M. (1996) "Bayesian analysis of mixtures of mixtures," *Biometrics*, 52, 221-227.

[8] Doss, H. (1994) "Bayesian nonparametric estimation for incomplete data via successive substitution sampling," *The Annals of Statistics*, 22, 1763-1786.

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

[10] Escobar, M.D. (1992) Invited comment of "Bayesian analysis of mixtures: some results on exact estimability and identification," by Florens, Mouchart, and Rolin. *Bayesian Statistics 4* (J.M. Bernardo, J.O.Berger, A.P.Dawid and A.F.M.Smith, eds.). Oxford: University press, 142-144.

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

[12] Escobar, M.D., (1995) "Nonparametrics Bayesian Methods in Hierarchical Models," *The Journal of Statistical Inference and Planning*, 43, 97-106.

[13] Escobar, M.D. (1997) "The effect of the prior on nonparametric Bayesian methods," (in preparation).

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

[15] Evans, M. and Swartz, T. (1995) "Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems - with discussion," *Statistical Science*, 10, 254-272.

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

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

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

[19] Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990), "Illustration of Bayesian inference in normal data models using Gibbs sampling.", *Journal Of the American Statistical Association,* 85, 972-985.

[20] George, E.I., Makov, U.E., and Smith, A.F.M. (1994) "Fully Bayesian hierarchical analysis for exponential families via Monte Carlo computation," *Aspects of Uncertainty: A Tribute to D. V. Lindley* (eds: AFM Smith and PR Freeman), London: John Wiley and Sons, 181-199.

[21] Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., eds. (1996). *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

[22] Kuo, L. and Mallick, B. (1998) "Bayesian semiparametric inference for the accelerated failure," *Canadian Journal of Statistics*, to appear.

[23] Liu, J.S. (1996) "Nonparametric hierarchical Bayes via sequential imputations," *The Annals of Statistics*, 24, 911-930.

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

[25] MacEachern, S.N. and Müller, P. (1994) "Estimating mixture of Dirichlet process models," ISDS discussion Paper #94-11, Duke University.

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

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

[28] Müller, P., West, M., and MacEachern, S.N. (1997) "Bayesian models for non-linear auto-regressions," *Journal of Time Series Analysis,* (in press).

[29] Naylor, J.C. and Smith, A.F.M. (1982) "Applications of a method for the efficient computation of posterior distributions," *Applied Statistics*, 31, 214-235.

[30] Pauler, D.K., Escobar, M.D., Sweeney, J.A., and Greenhouse, J. (1996) "Mixture Models for Eye-Tracking Data: A Case Study" *Statistics in Medicine*, 15, 1365-1376.

[31] Roeder, K., Escobar, M., Kadane, J., and Balazs, I., (1998) "Measuring Heterogeneity in Forensic Databases", *Biometrika* to appear.

[32] Erkanli, A., Stangl, D., and Müller, P. (1993) "Analysis of ordinal data by the mixture of probit links," *Discussion Paper #93-A01,* ISDS, Duke University.

[33] Sweeney, J.A., Clementz, B.A., Escobar, M.D., Li, S., Pauler, D.K., and Haas, G.L. (1993) "Mixture analysis of pursuit eye tracking dysfunction in schizophrenia," *Biological Psychiatry*, 34, 331-340.

[34] Tanner, M.A. and Wong, W.H. (1987) "The calculation of posterior distributions by data augmentation (with discussion)," *Journal of the American Statistical Association,* 82, 528-550.

[35] Tomlinson, G. (1998) Analysis of Densities. Unpublished dissertation. University of Toronto.

[36] Turner, D.A., and West, M. (1993) "Statistical analysis of mixtures applied to postsynpotential fluctuations," *Journal of Neuroscience Methods,* 47, 1-23.

[37] West, M. (1987) "On scale mixtures of normal distributions," *Biometrika*, 74, 646-648.

[38] West, M. (1990) "Bayesian kernel density estimation," *ISDS Discussion Paper #90-A02*, Duke University.

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

[40] West, M. (1997) "Hierarchical mixture models in neurological transmission analysis," *Journal Of the American Statistical Association*, 92, 587-608.

[41] West, M., and Cao, G. (1993) "Assessing mechanisms of neural synaptic activity," In *Bayesian Statistics in Science and Technology: Case Studies*, (eds: C.A. Gatsonis, J.S. Hodges, R.E. Kass, and N.D. Singpurwalla), New York: Springer-Verlag.

[42] West, M., Müller, P., and Escobar, M.D. (1994) "Hierarchial Priors and Mixture Models, with Applications in Regression and Density Estimation," *Aspects of Uncertainty: A Tribute to D. V. Lindley* (eds: AFM Smith and PR Freeman), London: John Wiley and Sons, 363-386.

[43] West, M., and Turner, D.A. (1994) "Deconvolution of mixtures in analysis of neural synaptic transmission," *The Statistician*, 43, 31-43.

Figure Legends

Figure 1: Histogram of the number of eye movement anomalies recorded for each of 101 subjects who are diagnosed with schizophrenia.

Figure 2: Fitted posterior distributions of θ under different conditions; (a, b, and c): the models where the baseline prior for θ is assumed to be $Ga(0.4, 0.1)$; (d, e, and f): the models where the baseline prior for θ is assumed to be $Ga(1, 1)$; (a and d): $Y_i = 4$; (b and e): $Y_i = 8$; (c and f): $Y_i = 12$. The solid lines "—": the baseline model; the dotted line "$\cdots$": Dirichlet process with $\alpha \sim Ga(.01, .01)$; the dashed line "$- - -$": with $\alpha \sim Ga(1, 1)$.

Figure 3: The dots are the MLE estimate of the intercept and slope of the linear growth curve of each subject consider alone. Figure (a): only the fitted MLE values are plotted. In the other three plots, the contour plots of the predicted distribution of the (intercept, slope) are shown for three different priors; (b) baseline model; (c) Dirichlet process with $\alpha \sim Ga(.01, .01)$; (d) Dirichlet process with $\alpha \sim Ga(1, 1)$.

Figure 4: The predictive distribution for the slope parameter for the data which contain both treatment and control group. The dashed lines "$- - -$" are for the baseline model where the parameters are fit with knowledge of the grouping variable. The dotted lines "$\cdots$" are for the Dirichlet process model with knowledge of the grouping variable. The solid lines "—" are for the Dirichlet process model when there is no knowledge of the group membership of the subjects. Figure (a): Dirichlet process with $\alpha \sim Ga(.01, .01)$; (b) Dirichlet process with $\alpha \sim Ga(1, 1)$.

Figure 4