Sankhyā : The Indian Journal of Statistics
Special Issue on Bayesian Analysis
1998, Volume spl, Series , Pt. 1, pp. 196-214

POPULATION MODELS WITH A NONPARAMETRIC RANDOM COEFFICIENT DISTRIBUTION

By STEPHEN WALKER*
Imperial College of Science, Technology and Medicine, London
and
JON WAKEFIELD
Imperial College School of Medicine at St Mary's, London

SUMMARY. Population data fit very naturally into a hierarchical framework. At the first stage of this hierarchy the data of a particular individual are modelled, typically by a nonlinear regression model, with the same regression model assumed for each individual. Inter-individual variability is accommodated at the second stage by assuming that the parameters of each individual are independently and identically distributed from a population distribution F. Often, interest is in learning about F so that predictions can be made for future individuals from the population. Previous Bayesian work has largely concentrated on F being assigned a specific parametric form, typically the normal. In this paper we propose a Bayesian nonparametric approach using the Dirichlet process (Ferguson, 1973; Antoniak, 1974) as a class of prior distributions for F. We consider the important case where covariate relationships are modelled at the second stage, and allow for errors-in-variables in the measured covariates. Relevant posterior distributions are summarised using Markov chain Monte Carlo methods. A challenging population pharmacokinetic dataset, involving a nonlinear concentration/time relationship and individual-specific covariates, is analysed and the results are compared with those of previous non-Bayesian parametric and non-parametric analyses.

1. Introduction

Population models are used in a wide variety of applications, including pharmacokinetic/pharmacodynamic studies (Wakefield, Aarons and Racine-Poon, 1998) and growth curves (Berkey, 1982). In general, interest may focus on individual units (for example, specific patients) or the population in general.

Population data consist of repeated measurements on a number of individuals with individual-specific covariates (e.g. age, weight and gender) often being

---

AMS (1991) *subject classification.* 62C10
Key words and phrases. Dirichlet process; Errors-in-variables; Markov chain Monte Carlo; Nonlinear random coefficient model; Pharmacokinetics; Random distributions.
* Research supported financially by an EPSRC Realizing Our Potential Award.

available. The modelling of such data is carried out using a hierarchical model where, at the first stage, each individual has their own (typically) non-linear regression model, each identically defined up to a set of unknown individual parameters. At the second stage, possibly after conditioning on individual-specific covariates, each set of parameters is assumed to be independently and identically distributed (iid) from some unknown population distribution. The first stage can therefore be thought of as modelling intra-individual variability and the second stage inter-individual variability. A Bayesian approach adds a third stage containing prior distributions on the population parameters. Much of the work on population models to date has focused on parametric distributions for the population distribution, typically the normal is chosen. Davidian and Giltinan (1995) provide a recent review of the various approaches to nonlinear hierarchical modelling whilst Racine-Poon and Wakefield (1998) consider methods specific to population pharmacokinetic modelling.

A parametric approach assumes that all random quantities are from known families of probability distributions and that known functional forms describe the relationship between random quantities and predictor variables. These modelling choices must be made at both the first and the second stages of the hierarchy. The parameters of the distributions and the functional forms are then estimated. Nonparametric approaches do not assume any particular choices and allow the data to select the shape of the unknown distributions/functional forms, whilst semiparametric procedures place some restrictions upon the choices.

Clearly a non- or semiparametric method is preferred when little is known about the underlying process which is generating the data and one does not want to be tied to restrictive assumptions. However, if information is available then this should be included since its exclusion may lead to a less efficient analysis. This paper is concerned with assigning a nonparametric prior (the Dirichlet and mixtures of Dirichlet process) for the unknown population distribution function. We assume a fully parametric model at the first stage since there is typically greater information available at this stage.

The situation in which the first stage is linear and the second stage is modelled using mixtures of Dirichlet process priors has been covered, for example, by Escobar and West (1995). In this paper we extend the approach to consider:

1. a nonlinear first stage model;

2. obtaining posterior distributions for functionals of the population distribution; and

3. the incorporation of individual-specific covariates into the second stage of the model.

The advantage of 2 is that it allows us to directly gain insight into the population distribution, which may be the primary aim of the analysis. A consequence of 3. is that it is no longer possible to assume the individual parameters to be

iid. Instead, interest is now focused on the *joint* distribution of parameters and covariates, which are jointly assumed to be iid. We carry out this joint modelling by specifying a prior (marginal) distribution for the covariates. To extend our basic model we allow for errors-in-variables in the measured covariates.

Muller and Rosner (1997) describe a Bayesian semiparametric population model using mixtures of Dirichlet processes. Their approach, developed independently of our own, is somewhat different.

The paper is structured as follows. In Section 2 we consider the general hierarchical model of interest and non-Bayesian nonparametric approaches. In Section 3 we describe the Bayesian nonparametric approach using a Dirichlet process prior for the population distribution, without covariate information. Section 4 introduces the modifications required to incorporate covariate information into the model. An example concerning the population pharmacokinetics of phenobarbital in new-born babies is presented in Section 5 and Section 6 contains concluding comments. An Appendix contains details of a prior specification for the precision parameter of the Dirichlet process.

2. **Statistical model**

2.1 *General model.* In this section we describe the basic two-stage hierarchical model that is considered in this paper.

Stage 1: model for within-individual variability. Let y_{ij} denote the $j-$th univariate observation on the $i-$th individual, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, and x_{ij} denote an associated explanatory variable which, for ease of explanation, we will refer to as time. Denote by $f(\theta_i, x_{ij})$ the regression model for individual i at time x_{ij}. This model depends on a $p-$dimensional, individual-specific parameter θ_i, the elements of which, possibly after transformation, are assumed to lie on the whole of the real line. We assume independent, additive, homoscedastic normal errors with variance σ^2. Note that this may follow after a transformation of the original observation. For example in pharmacokinetic studies the logarithm of the observed concentration is often taken.

To summarise, we have

$$y_{ij}|\theta_i, x_{ij} \sim N\big(y_{ij}|f(\theta_i, x_{ij}), \sigma^2\big), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i,$$

where N denotes the univariate normal distribution. Additionally, for $i \neq i'$, y_i, given θ_i, is conditionally independent of $\theta_{i'}$ and $y_{i'}$.

Stage 2: model for between-individual variability. We assume that

$$\theta_1, \ldots, \theta_n \sim_{\text{iid}} F,$$

where F denotes the unknown population distribution. A parametric analysis would assume that F was given by a specific parametric family, for example the

normal or Student's t. Wakefield *et al.* (1994) consider the Bayesian analysis of these models.

2.2 *Nonparametric approaches.* At the first stage of a population model there is typically more information available on functional and distributional forms, either based on subject-matter considerations or on empirical observation. For example in pharmacokinetic applications compartmental models (Gibaldi and Perrier, 1982) have been empirically found to provide an adequate description of the concentration/time profile. Similarly the assumption of normal errors has been found to be adequate, often based on assay-validation data. Consequently, there is no need, in general, to consider nonparametric approaches at the first stage.

The area of nonlinear population pharmacokinetic models has provided the motivation for a number of approaches to the nonparametric estimation of the second stage distribution, both non-Bayesian (Mallet, 1986; Davidian and Gallant, 1993) and Bayesian (Muller and Rosner, 1997; Wakefield and Walker, 1997). In general there is less certainty in the second stage distribution in part because the modelling concerns unobservable quantities. In the pharmacokinetic context there are particular reasons to believe that multimodalities, for example, may be present in the population distribution. These may occur due to the existence of an important unmeasured explanatory variable. Consequently, flexible modelling is required.

Non-Bayesian nonparametric approaches are concerned with obtaining maximum likelihood estimates for F over specific classes of distributions.

2.3 *NPML.* Mallet (1986) proposed a nonparametric maximum likelihood (NPML) method for estimating the whole population distribution when there is no *a priori* information about the shape. Mallet (1986) shows that the maximum likelihood estimate of F is obtained by maximising

$$l(F) = \prod_{i=1}^{n} \int_{\Theta} p(y_i|\theta_i, \sigma^2) dF(\theta_i)$$

with respect to F. The estimator for F turns out to be discrete with the number of atoms bounded by the sample size, that is, n. The algorithm for the determination of the atoms of this discrete distribution arises from D-optimal design theory. Predictive inferences from such an estimate are straightforward. A disadvantage of the original approach was that the first stage likelihood function had to be specified completely, with, in particular, the intra-individual variance term being known. The algorithm becomes much more computationally expensive when this restriction is relaxed. Another disadvantage is that no measure of the *uncertainty* in the estimate of F is obtained. This means that it is difficult to determine whether features of the estimate are real or merely artifacts. Also, since no measure of uncertainty is available, all inference (for example, moment estimation and prediction) is made "as if" the true estimate

of F has been found. This means that the estimation uncertainty in F is not acknowledged and interval estimates will, in general, be too narrow. Finally the maximisation is difficult and the value which is converged to may be sensitive to the initial parameter ranges that are specified (Wakefield and Walker, 1997). The method has been applied to many population pharmacokinetic datasets (an early example is Mallet *et al.*, 1988).

Mentré and Mallet (1994) extend the method to carry out joint estimation of pharmacokinetic parameters *and* individual-specific covariates, making the approach fully nonparametric.

2.4 *SNP.* The smooth nonparametric (SNP) maximum likelihood method of Davidian and Gallant (1993) imposes smoothness on the population distribution. With such an assumption the density function can be represented by an infinite series expansion. Davidian and Gallant (1993) approximate this by a finite expansion so that for some $k = 0, 1, 2, \ldots$,

$$F^{(k)}(\delta) = \frac{\{q^{(k)}(R^{-1}\delta)\}^2 N_p(\delta|0, RR^t)}{\int \{q^{(k)}(u)\}^2 N_p(u|0, I)du},$$

where N_p denotes the $p-$dimensional normal distribution, $\delta = \theta - \mu$, with $\mu = E[\theta]$, and $q^{(k)}$ denotes a polynomial of degree k. The estimation problem, for a given k, becomes one of estimating μ, the coefficients of the polynomial $q^{(k)}$, and the upper triangular matrix R. These are obtained by maximum likelihood where the likelihood function is obtained by integrating over the random effects, δ, so that the population likelihood can be written as

$$l(\mu, F^{(k)}) = \prod_{i=1}^{n} \int_{\Theta} p(y_i|\delta, \mu) dF^{(k)}(\delta).$$

The integration is performed approximately using either Gauss-Hermite numerical integration or a more computationally expensive importance sampling Monte Carlo method.

Covariates z are incorporated into the SNP method parametrically. A typical form is given by

$$\theta_i = \mu_0 + \mu_1 z + \delta_i,$$

with the random effects δ_i assumed to be iid from a smooth density. The method therefore allows no nonparametric element to this relationship.

The parameter k is of crucial importance since it determines the number of normal mixtures used to estimate the population distribution. The strategy suggested by Davidian and Gallant is to start with $k = 0$, that is a single normal distribution, and then to increase k until no further changes in the estimate of μ, or visual appearance in the distribution, are seen. Selection of k is then decided upon via informal visual examination of univariate and bivariate marginal estimates or formally using a number of proposed criteria, including

the Akaike, Schwarz and Hannan-Quinn measures. Covariate relationships are modelled parametrically, for example via loglinear forms. Confidence intervals for μ and functionals for $\hat{F}$ can also be obtained. The approach has been implemented on population pharmacokinetic data by Davidian and Gallant (1992, 1993). The maximisation is in general difficult and it is recommended that a "wave of starting values", be used. The crucial choice of the number of terms to take in the expansion is also fixed in the sense that the uncertainty in this parameter is not accounted for in the inference.

Wakefield and Walker (1997) provide a comparison of NPML, SNP and a Bayesian nonparametric method.

3. Bayesian nonparametric model

3.1 *The Dirichlet process prior.* In this paper we describe a Bayesian nonparametric method in which *a priori* the distribution of interest is assumed to be from a rich class of distributions. Such a class is provided by the Dirichlet process (Ferguson, 1973) or, more generally, by mixtures of Dirichlet processes (Antoniak, 1974).

Our Bayesian approach puts a Dirichlet process prior on F. An advantage over NPML is that even though the estimator is discrete, the number of atoms available is unbounded. In addition the prior distribution can be centred on a particular parametric model which allows information to be incorporated, if required. We use a Dirichlet process prior with parameter $\alpha(\cdot)$, a finite non-null measure defined on the same space as F and let α_0 represent the total mass of α. Following Escobar and West (1995), we assign a hyper-prior to $\alpha(\cdot)$. Consequently, the Bayesian nonparametric population model can be written as

$$y_{ij}|\theta_i, x_{ij} \sim N\big(y_{ij}|f(\theta_i, x_{ij}), \sigma^2\big), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i,$$

$$\theta_1, \ldots, \theta_n | F \sim_{\text{iid}} F,$$

$$F|\alpha \sim \mathcal{D}(\alpha)$$

and

$$\alpha \sim \pi$$

where $\mathcal{D}(\alpha)$ represents the Dirichlet process prior with parameter α. Typically F is integrated out, leaving the marginalised model given by

$$y_{ij}|\theta_i, x_{ij} \sim N\big(y_{ij}|f(\theta_i, x_{ij}), \sigma^2\big), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i,$$

$$\theta_1, \ldots, \theta_n | \alpha \sim \mathcal{P}(\alpha)$$

and

$$\alpha \sim \pi,$$

where $\mathcal{P}(\alpha)$ represents the general Polya-urn scheme (Blackwell and MacQueen, 1973) with parameter α. The joint distribution $p(\theta_1, \ldots, \theta_n)$ is given by

$$p(\theta_1, \ldots, \theta_n) \propto \prod_{i=1}^{n} \left\{ \alpha(\theta_i) + \sum_{j<i} \delta_{\theta_j}(\theta_i) \right\},$$

where $\delta_\theta(\cdot)$ is the measure putting mass 1 at θ.

We now specify the underlying measure α. Let $G(\cdot) = \alpha(\cdot)/\alpha_0$, so that G is a probability distribution and $E[F] = G$. We take G to be the p-variate normal distribution $N_p(\theta|\mu, \Sigma)$, where μ is a $p \times 1$ location vector and Σ is a $p \times p$ positive-definite covariance matrix. Independent priors for μ and Σ^{-1} are then assigned as normal $N_p(\mu|\phi, \Phi)$ and Wishart $W_p(\rho, (\rho R)^{-1})$, respectively. The vector ϕ and the matrix R are then taken to be *a priori* estimates of μ and Σ, respectively. The diagonal elements of the matrix Φ represent the strength of belief in ϕ and the scalar ρ the strength of belief in R. Note, that for a nonlinear first stage model it is necessary, in general, to specify a proper prior for μ in order to ensure that the posterior is also proper. In the absence of prior information the diagonal elements of Φ^{-1} may be taken to be small with the off-diagonal elements being taken as zero. Taking $\rho = p$ makes the prior for Σ^{-1} as flat as possible whilst retaining propriety of the prior and hence the posterior. A gamma prior, $\mathrm{Ga}(\chi_1, \chi_2)$, is specified for σ^{-2}. The motivation is to have regions of high prior probability corresponding to the regions of high probability of the true distribution.

A prior for α_0 is more troublesome. Escobar and West (1995) describe how the gamma prior $\alpha_0 \sim \mathrm{Ga}(\omega, \varrho)$ is suitable. West, Muller and Escobar (1994) take $\omega = \varrho = 1$ in their example to represent vague *a priori* information. See the Appendix for the justification for using an alternative prior for α_0.

It is well known that, if $F \sim \mathcal{D}(\alpha)$ then, with probability 1, F is discrete. Muller and Rosner (1997) therefore proposed the following model which is quite different to that described above and take the distribution of the θ_i's to be continuous:

$$\theta_i|\nu_i, V \sim N_p(\nu_i, V), \quad i = 1, \ldots, n$$

and

$$\nu_1, \ldots, \nu_n \sim \mathcal{P}(\alpha).$$

This implies that

$$\theta_1, \ldots, \theta_n|V \sim_{\mathrm{iid}} \sum_{k=1}^{\infty} \omega_k N_p\big(\nu_k', V\big),$$

where $\{\omega_k, \nu_k'\}$ are obtained from α via

$$\omega_k = v_k \prod_{l=1}^{k-1} (1 - v_l), \qquad v_1, v_2, \ldots \sim_{\mathrm{iid}} \mathrm{Be}\big(1, \alpha_0\big)$$

where Be$(\cdot,\cdot)$ denotes the Beta distribution and

$$\nu_1^{'}, \nu_2^{'}, \ldots \sim_{\text{iid}} \alpha(\cdot)/\alpha_0.$$

Therefore, F is taken to be an infinite mixture of normals.

3.2 *Implementation via MCMC.* With the marginalised model and a first stage linear model, Escobar and West (1995) show that the conditional distribution for θ_i, required for Gibbs sampling (Smith and Roberts, 1993), is given by

$$p(\theta_i|\theta_{(i)}, \sigma^2, \mu, \Sigma, \alpha_0, y) \propto q_0 N_{n_i}\left(y_i|f(\theta_i; x_i), \sigma^2 I_{n_i}\right) N_p(\theta_i|\mu, \Sigma) + \sum_{j\neq i} q_j \delta_{\theta_j}, \quad \ldots (1)$$

where $q_0 \propto \alpha_0 \int p(y_i|\theta)p(\theta|\mu,\Sigma)d\theta$, $q_j \propto p(y_i|\theta_j)$, $\theta_{(i)} = \{\theta_j : j = 1, \ldots, n, j \neq i\}$ and

$$q_0 + \sum_{j=1, j\neq i}^{n} q_j = 1.$$

With a nonlinear first stage model it is not possible to evaluate q_0. Consequently, we introduce the following step into the Markov chain, using a Metropolis-Hastings algorithm (Tierney, 1994). The full conditional for θ_i is given by

$$p(\theta_i|\theta_{(i)}, \sigma^2, \mu, \Sigma, \alpha_0, y) \propto N_{n_i}\left(y_i|f(\theta_i; x_i), \sigma^2 I_{n_i}\right) \left\{ \alpha_0 G(\theta_i) + \sum_{j\neq i} \delta_{\theta_j}(\theta_i) \right\}.$$

At the $t-$th iteration, let the current sample for θ_i be $\tilde{\theta}_i$. We approximate the first stage log-likelihood using a Taylor series about the maximum likelihood estimate (see Racine-Poon, 1985 for the use of this in another context). If the calculation of the maximum likelihood estimates is troublesome, due to sparse or noisy data, then alternative Bayesian estimates (for example the posterior mean) can be found using trial runs of the Markov chain. We therefore take the proposed sample from

$$\theta_i^* \sim N_p(\hat{\theta}_i|\theta_i^*, C_i) \left\{ \alpha_0 G_0(\theta_i^*) + \sum_{k\neq i} \delta_{\theta_k}(\theta_i^*) \right\},$$

where $\hat{\theta}_i$ is the maximum likelihood estimate of θ_i and C_i the corresponding asymptotic covariance matrix. Thus,

$$\theta_i^* \begin{cases} = \theta_k \ (k \neq i) & \text{with probability } cN_p(\hat{\theta}_i|\theta_k, C_i) \\ \sim G_i & \text{with probability } c\gamma_i, \end{cases}$$

where G_i is the normal distribution $N_p(\cdot|\nu_i, \Phi_i)$, $\Phi_i^{-1} = C_i^{-1} + \Sigma^{-1}$, $\nu_i = \Phi_i(\Phi_i^{-1}\hat{\theta}_i + \Sigma^{-1}\mu)$, $\gamma_i = \alpha_0 \int N_p(\hat{\theta}_i|\theta, C_i)N_p(\theta|\mu, \Sigma)d\theta$ and c is the normalising constant. Accordingly, the acceptance probability is given by

$$\min\left\{1, \frac{N_{n_i}\left(y_i|f(\theta_i^*, x_i), \sigma^2 I_{n_i}\right) N_p(\hat{\theta}_i|\tilde{\theta}_i, C_i)}{N_{n_i}\left(y_i|f(\tilde{\theta}_i, x_i), \sigma^2 I_{n_i}\right) N_p(\hat{\theta}_i|\theta_i^*, C_i)}\right\}.$$

The sampling of the full conditionals for σ^2, μ, Σ and α_0 are described, for example, in Escobar and West (1995).

The above describes how we may carry out a population analysis with a non-parametric prior for the second stage distribution. Suppose we are particularly interested in gaining some insight into the unknown distribution F, specifically a functional $\phi(F)$. It is possible to obtain estimates of the moments of $\phi(F)$ but in order to obtain samples from the posterior $p(\phi(F)|y)$, it is necessary to obtain samples from $p(F|y)$. Now

$$p(F|y) = \int p(F|\alpha_0, \mu, \Sigma)dp(\alpha_0, \mu, \Sigma|y),$$

and, since samples $\{\alpha_0^{(t)}, \mu^{(t)}, \Sigma^{(t)}\}$ are available from $p(\alpha_0, \mu, \Sigma|y)$, we can take F from $p(F|\alpha_0^{(t)}, \mu^{(t)}, \Sigma^{(t)})$. Therefore, we need to generate random distributions from the Dirichlet process which can only be achieved approximately.

Here we consider generating a random distribution from the Dirichlet process with parameter $\alpha(\cdot) = \alpha_0 G(\cdot)$. Take a sample $\theta_1', \ldots, \theta_L'$ iid from G with L large and let G_L denote the empirical distribution of this sample. Generate an exact Dirichlet process random distribution from $\mathcal{D}(\alpha_L)$, where $\alpha_L(\cdot) = \alpha_0 G_L(\cdot)$, by taking

$$(Z_1, \ldots, Z_L) \sim \text{Dirichlet}(\alpha_0/L, \ldots, \alpha_0/L)$$

and defining

$$F_L = \sum_{l=1}^{L} Z_l \delta_{\theta_l'}.$$

It is straightforward to show the almost sure weak convergence $\mathcal{D}(\alpha_L) \to \mathcal{D}(\alpha)$ as $L \to \infty$ (Sethuraman and Tiwari, 1982). To illustrate the approximation, for any measurable partition $(B_1, \ldots, B_R)$ of the parameter space, a representation of the joint distribution of $(F_L(B_1), \ldots, F_L(B_R))$ is given by

$$\big(F_L(B_1), \ldots, F_L(B_R)|T_1, \ldots, T_R\big) \sim \text{Dirichlet}\big(\alpha_0 T_1/L, \ldots, \alpha_0 T_R/L\big)$$

with

$$(T_1, \ldots, T_R) \sim M\big(L; G(B_1), \ldots, G(B_R)\big),$$

where M denotes the multinomial distribution. This approach should be compared with the correct distribution which is given by

$$\big(F(B_1),\ldots,F(B_R)\big) \sim \text{Dirichlet}\big(\alpha_0 G(B_1),\ldots,\alpha_0 G(B_R)\big).$$

Generating an approximate random distribution from $\mathcal{D}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$ can be achieved in a similar way: take $\theta_{n+1},\ldots,\theta_L$ iid from G and approximate $\alpha + \sum_{i=1}^n \delta_{\theta_i}$ by

$$\alpha_L = \frac{\alpha_0}{(L-n)} \sum_{l=n+1}^{L} \delta_{\theta_l} + \sum_{i=1}^{n} \delta_{\theta_i}.$$

Generating an exact random distribution from $\mathcal{D}(\alpha_L)$ follows by defining

$$F_L = \sum_{i=1}^{n} Z_i \delta_{\theta_i} + \sum_{l=n+1}^{L} Z_l \delta_{\theta_l}, \qquad \ldots (2)$$

where

$$(Z_1,\ldots,Z_L) \sim \text{Dirichlet}(\alpha_1^{'},\ldots,\alpha_L^{'})$$

and

$$\alpha_l^{'} = \begin{cases} 1 & \text{if } 1 \le l \le n \\ \alpha_0/(L-n) & \text{if } n+1 \le l \le L. \end{cases}$$

The samples $\{F^{(k)}\}$, obtained via (2), can be used to obtain samples from $p(\phi(F)|y)$, from which inference about $\phi(F)$ can be made. An alternative approach to sampling a random F is to use the constructive definition of Sethuraman and Tiwari (1982).

4. Bayesian nonparametric model with covariates

With the inclusion of covariates at the second stage, the Bayesian nonparametric model is not so straightforward to implement, since it is no longer possible to assume that $\theta_1,\ldots,\theta_n$ are exchangeable. The method outlined in Section 3 therefore needs to be modified. We let z represent the vector of continuous covariates (e.g. height, weight), assumed to be measured with error, and c represent the vector of discrete covariates (e.g. gender, smoking status), which are assumed to be observed without error.

The aim is to obtain information about the joint distribution of $\lambda = (\theta, z, c)$. We take $(\lambda_1,\ldots,\lambda_n)$ to be exchangeable and, following Section 3, assume

$$\lambda_1,\ldots,\lambda_n | F \sim_{\text{iid}} F,$$

with

$$F|\alpha \sim \mathcal{D}(\alpha).$$

The baseline measure α is allowed to depend on the linear model that is typically used in parametric approaches via:

$$\alpha(\lambda) = \alpha_0 N_p\big(\theta|\mu_0 + \mu_1 z + \mu_2 c, \Sigma\big)\pi(z, c).$$

Here $\pi(z, c)$ is the joint distribution of the covariates in the patient population, for which there will often be sufficient external information to construct informative priors. If not, noninformative priors may be adopted. Priors for μ, Σ and α_0 can be taken as described in Section 3.

As has been mentioned the continuous covariates are assumed to be measured with error. If z is the "true" covariate value then we consider that we observe $w \sim N(z, \tau)$, where τ, the measurement error of the observation, is assumed known. Note that in general we may assume more complex measurement error models and also estimate the parameters of these models (see Fuller, 1987; Carroll, Ruppert and Stefanski, 1995). The only difference to the algorithm outlined in Section 3.2 is that the full conditional for $\phi_i = (\theta_i, z_i)$ replaces the full conditional for θ_i. Now, if there are k continuous covariates,

$$p(\phi_i|\phi_{(i)}, \sigma^2, \mu, \sigma, \alpha_0, y, w, c) \propto \left\{\prod_{l=1}^{k} N(w_{il}|z_{il}, \tau_l)\right\} N_{n_i}\left(y_i|f(\theta_i, x_i), \sigma^2 I_{n_i}\right)$$

$$\times \left\{\alpha_0 N\left(\theta_i|\mu_0 + \mu_1 z_i + \mu_2 c_i, \Sigma\right)\pi(z_i, c_i) + \sum_{j\neq i, c_i = c_j} \delta_{\phi_j}(\phi_i)\right\}$$

where $c = (c_1, \ldots, c_n)$ and $w = (w_1, \ldots, w_n)$. This can be sampled in essentially the same way as we sampled from $p(\theta_i|\theta_{(i)}, \sigma^2, \mu, \Sigma, \alpha_0, y)$ in Section 3.2.

We need to have the continuous covariates measured with error, otherwise we are lead to a parametric analysis, as we now demonstrate. If $\tau = 0$ then

$$p(\phi_i|c_i, z_i, \ldots) \propto N_{n_i}\left(y_i|f(\theta_i, x_i), \sigma^2 I_{n_i}\right)$$

$$\times \left\{\alpha_0 N\left(\theta_i|\mu_0 + \mu_1 z_i + \mu_2 c_i, \Sigma\right)\pi(z_i)\pi(c_i) + \sum_{j\neq i, c_i = c_j, z_i = z_j} \delta_{\theta_j}(\theta_i)\right\}.$$

However, $z_i \neq z_j$ with probability 1 and so essentially we end up with the parametric model since there are no coincident covariates. In practice, however, *all* continuous covariates are measured with error and the difficulty will arise only when the measurement error is so small that there is effectively zero probability of coincident covariates. In the areas of covariate space in which there are few observations the underlying parametric function will be dominant which is intuitively reasonable.

5. An example

The data set we work with was originally presented by Grasela and Donn (1985) and was analysed using a variety of non-Bayesian parametric and non-parametric approaches by Davidian and Giltinan (1995). Data were collected on 59 infants who were administered phenobarbital (to prevent seizures) for the first 16 days after birth. Each infant received an initial single dose, followed by further intravenous doses to sustain the required concentration. Each provided between 1 and 6 concentration measurements and there were 155 observations in total.

The first stage model, with log-normal error structure, is given by

$$\log y_{ij} = \log f(\theta_i, x_{ij}) + \varepsilon_{ij},$$

where x_{ij} represents time and $\varepsilon_{ij} \sim_{\text{iid}} N(0, \sigma^2)$. According to Grasela and Donn (1985), the pharmacokinetics of phenobarbital may be described by a one-compartment model with intravenous administration and first-order elimination (Gibaldi and Perrier, 1982). The pharmacokinetic model f therefore consists of the contributions from all doses prior to x and is given by

$$f(\theta, x) = \sum_{d:x_d<x} D_d V^{-1} \exp\left(-CV^{-1}(x - x_d)\right),$$

where x_d is the administration time of dose D_d and $\theta_i = (\log C_i, \log V_i)$. The parameters C_i and V_i represent the clearance and volume of distribution, respectively, of the $i-$th infant.

Two covariates were recorded for each infant: the Apgar score, which is a general measure of a baby's health, and birth weight (in kgs). We take z_i to be the $i-$th infant's birth weight and $c_i = 1$ if Apgar < 5 for the $i-$th infant and $c_i = 0$ otherwise (see Davidian and Giltinan, 1995). The birth weights are measured with error and the observed birth weights w_i are modelled as

$$w_i \sim N(w_i | z_i, \tau)$$

where the z_i are the "true" birth weights and τ denotes the variance of the measurement error. The birth weights are measured to one decimal place and so we take $6\tau^{1/2} = 0.1$, giving a variance of 0.01.

The prior specifications were as follows: $\Sigma^{-1} \sim W_2(\rho, (\rho R)^{-1})$ with $\rho = 2$ and R diagonal with diagonal elements 2; the elements of μ were taken to have independent normal distributions with zero means and large variances; $\alpha_0 \sim \text{Ga}(1, 1)$ and $\sigma^{-2} \sim \text{Ga}(0, 0)$.

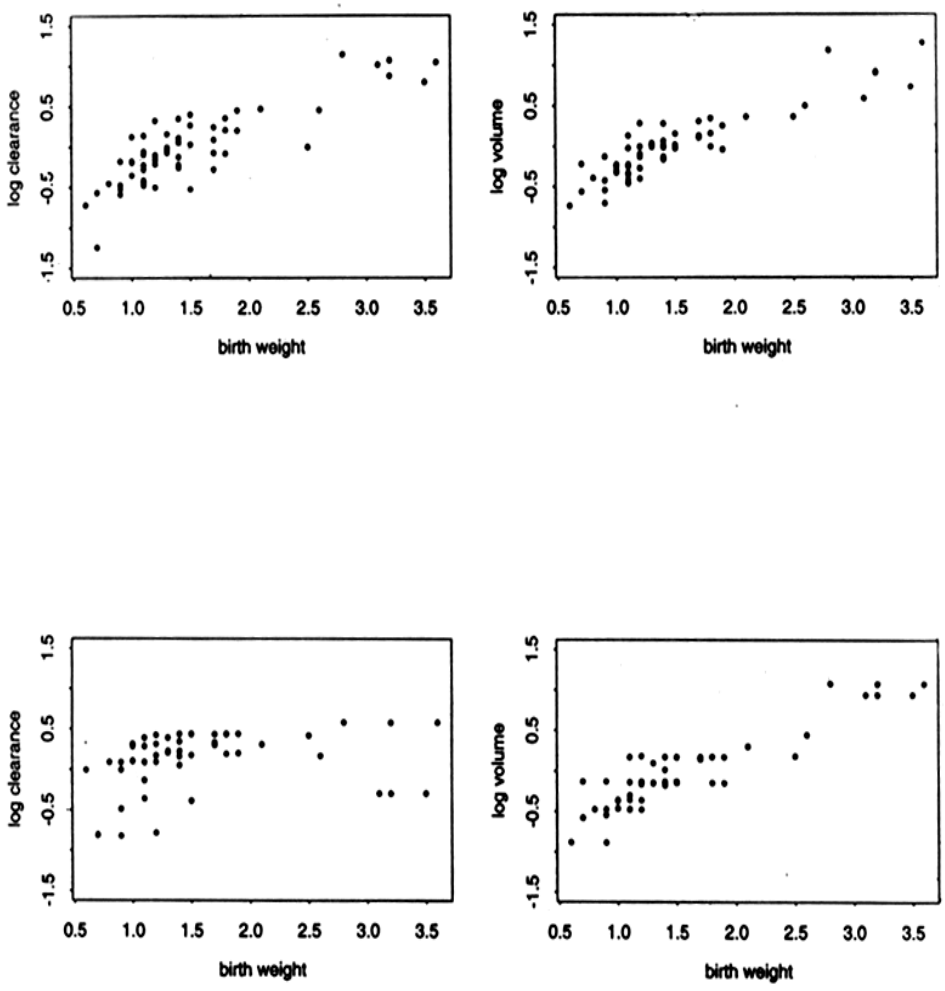

Figure 1: Estimated random effects for log clearance and log volume against birth weight. The top row is from a parametric analysis and the bottom row from a nonparametric analysis.

First we ran the algorithm excluding covariates and obtained estimates for each of the random effects. This first analysis was carried out as an exploratory exercise in order to investigate the relationship between the pharmacokinetic parameters and the covariates, and to examine the shape of the second stage distribution. In Figure 1 we plot the posterior estimates of the random effects against the covariates in order to identify possible relationships. These figures may be directly compared with Figures 6.2 and 7.1 of Davidian and Giltinan (1995). In Figure 6.2 a no covariate analysis using the fully parametric linearisation approach of Lindstrom and Bates (1990) was carried out whilst Figure 7.1 resulted from the semiparametric approach of SNP; it is interesting that the random effect estimates from these two models are almost identical. The top row of Figure 1 is from a fully parametric analysis ($\alpha_0 \to \infty$) and the bottom row is from a nonparametric analysis, with a random α_0. The plots in the top row are virtually identical to those in Davidian and Giltinan (1995), whilst the plot for log clearance in the bottom row is quite different with no strong indication of a relationship with birth weight.

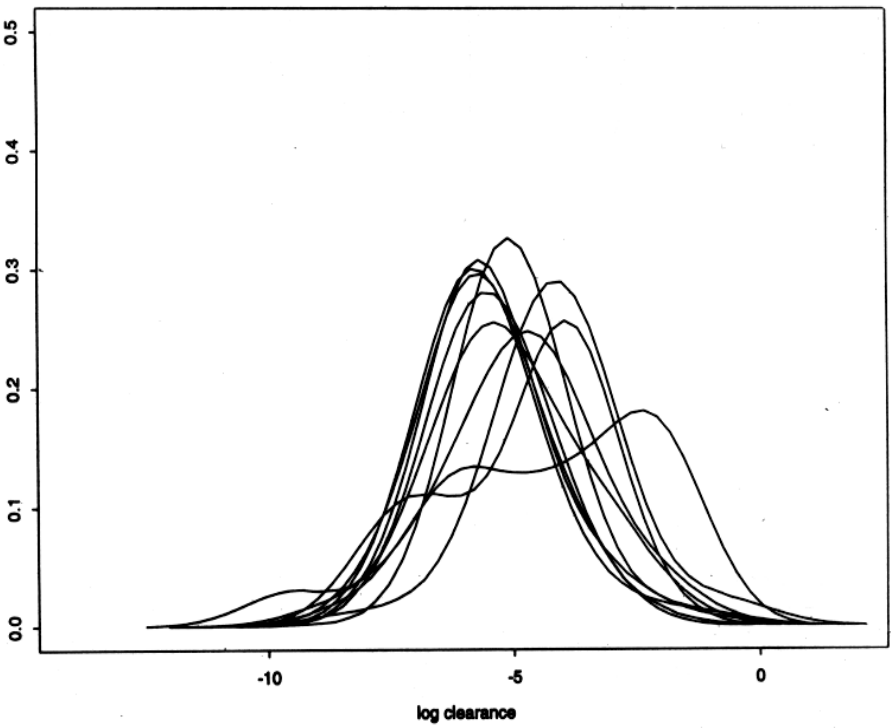

Figure 2: Ten densities for log clearance corresponding to random draws from the posterior distribution of $p(F_{\log C}|y)$.

Figure 2 shows densities for log clearance corresponding to random draws from the distribution function $F_{\log C}$, with the no covariate analysis. We see that there are a variety of shapes including both bimodal, unimodal and skewed densities.

Next we ran the full nonparametric analysis including covariates. In the context considered here the major interest is in determining loading doses for newborn babies with observed birth weight w and Apgar score c. Such dose determination depends on the prior distribution of the pharmacokinetic parameters given w, c, Wakefield (1996). Of primary interest, therefore, is the predictive distribution $p(\theta, z|w, c, y)$ from which we may determine $p(\theta|z, w, c, y)$. It is straightforward to sample from this distribution via the Markov chain, since at each iteration we have $\phi = \{(\theta_i, z_i)\}_{i=1}^n$ and

$$p(\theta, z|w, c, \phi) \propto N(w|z, \tau) \left\{ \alpha(\theta, z) + \sum_{c_i = c} \delta_{(\theta_i, z_i)}(\theta, z) \right\}.$$

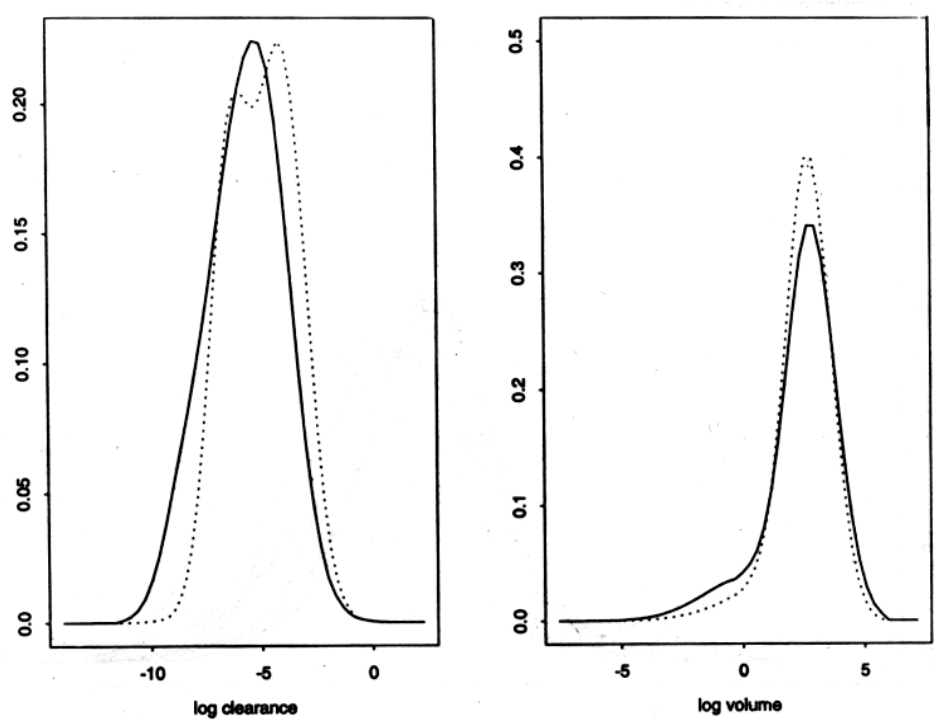

Figure 3: Estimated predictive densities for log clearance and log volume when the observed birth weight is equal to 1.0. The solid line corresponds to Apgar score ≥ 5 and the dotted line Apgar score < 5.

Figures 3 and 4 compare predictive distributions for log clearance (θ_1) and log volume (θ_2) with observed birth weights, w taken, for illustration, to be 1.0 and 2.0. The solid and broken lines represent Apgar ≥ 5 ($c = 0$) and Apgar < 5 ($c = 1$), respectively. A variety of distributional forms are observed. It does not seem that there is a simple monotonic relationship between either log volume or log clearance with birthweight and the spread similarly seems non-constant.

The analysis of Davidian and Giltinan (1995) found evidence of a weak relationship between Apgar score and log volume. In their exploratory plots it appeared that a low Apgar score indicated low volume but the parameter estimate in the covariate analysis suggested a slight relationship in the reverse direction.

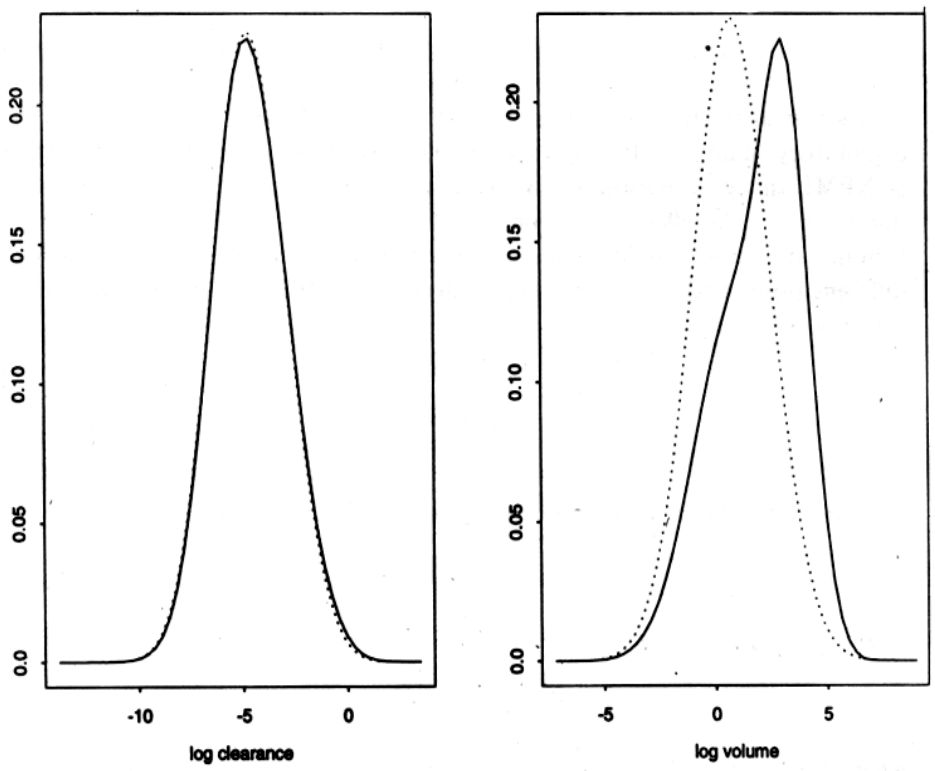

Figure 4: Estimated predictive densities for log clearance and log volume when the observed birth weight is equal to 2.0. The solid line corresponds to Apgar score ≥ 5 and the dotted line Apgar score < 5.

In Figures 3 and 4 we see that Apgar score has little effect on the conditional distributions of log clearance with a suggestion of an effect on log volume in

Figure 4. A higher score produces a slight shift in the distribution of log volume. We note that there is little information about the Apgar score in the dataset since only ten infants had a low score.

We have seen therefore that a full nonparametric approach has highlighted possibly important features, including bimodalities. In general we would suggest that a number of such analyses be carried out with different prior distributions, particularly for α_0, and with different levels of measurement error. With this example we found that as the measurement error variance τ was decreased the posterior for α_0 was shifted upwards, indicating a more parametric model.

6. Concluding comments

In this paper we have presented a fully Bayesian nonparametric analysis of a nonlinear hierarchical model incorporating covariate information. Our model makes few distributional assumptions at the second stage and so is ideal for exploratory analyses. It may in some ways be viewed as a Bayesian version of NPML though the estimates which are obtained fully reflect the estimation uncertainty. The effect of the prior distribution is in general difficult to determine and so we would recommend carrying out a number of analyses with different priors and levels of measurement error. Care must also be taken when convergence is assessed.

Appendix

Theory for the Dirichlet process says that if $X_0 \sim G$ then

$$\alpha_0 + 1 = \mathrm{var}(X_0)/\mathrm{var}(\phi),$$

where $\phi = \int x dF$. Since

$$\mathrm{var}(X_0) = E[\lambda(F)] + \mathrm{var}(\phi),$$

where $\lambda(F) = \int x^2 dF - \phi^2$, we obtain $\alpha_0 = \sigma_0^2/\sigma_1^2$, where σ_0^2 is the prior expected variance for F and σ_1^2 is the prior variance for the mean of F. Note also that $\sigma^2 = \sigma_0^2 + \sigma_1^2$ is the variance for G. Suppose that we take independent hyper-priors for σ_0^2 and σ_1^2 and desire that the induced hyper-priors for α_0 and σ^2 are also independent (in fact this is quite important for the Dirichlet process). Lukacs' Theorem (Lukacs, 1955) then requires that σ_0^2 and σ_1^2 have independent gamma priors with the same scale parameter. Let us therefore take

$$\sigma_0^2 \sim \mathrm{Ga}(\alpha, \gamma) \quad \text{and} \quad \sigma_1^2 \sim \mathrm{Ga}(\beta, \gamma).$$

The induced prior for σ^2 is $\sigma^2 \sim \text{Ga}(\alpha+\beta,\gamma)$ and the induced prior for α_0 is given up to proportionality by

$$\frac{\alpha_0^{\alpha-1}}{(1+\alpha_0)^{\alpha+\beta}}.$$

This is equivalent to using a $\text{Be}(\beta,\alpha)$ prior for $1/(\alpha_0+1)$. Noninformative specifications would follow by taking $\alpha=\beta=0$ giving α_0^{-1} as the prior for α_0.

References

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **2**, 1152-1174.

Berkey, C.S. (1982). Bayesian approach for a nonlinear growth model. *Biometrics*, **38**, 953-961.

Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Polya-urn schemes. *Ann. Statist.*, **1**, 353-355.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

Davidian, M. and Gallant, A.R. (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. *J. Pharmacokin. Biopharm.*, **20**, 529-556.

– – – – (1993). The non-linear mixed effects model with a smooth random effects density. *Biometrika*, **80**, 475-488.

Davidian, M. and Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measures Data*. Chapman and Hall, London.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.*, **90**, 577-588.

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

Fuller, W.A. (1987). *Measurement Error Models*. John Wiley and Sons, New York.

Gibaldi, M. and Perrier, D. (1982). *Drugs and the Pharmaceutical Sciences, Volume 15 Pharmacokinetics, Second Edition*. Marcel Dekker, New York and Basel.

Grasela, T.H. and Donn, D.M. (1985). Neonatal population pharmacokinetics of phenobarbital derived from routine clinical data. *Dev. Pharmacol. Ther.*, **8**, 374–383.

Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673–687.

Lukacs, E. (1955). A characterisation of the gamma distribution. *Ann. Math. Statist.*, **26**, 319-324.

Mallet, A. (1986). A maximum likelihood estimation method for coefficient regression models. *Biometrika*, **73**, 645–656.

Mallet, A., Mentré, F., Steimer, J-L. and Lokiec, F. (1988). Nonparametric maximum likelihood estimation for population pharmacokinetics, with application to cyclosporine. *J. Pharmacokin. Biopharm.*, **16**, 311-327.

Mentré, F. and Mallet, A. (1994). Handling covariates in population pharmacokinetics, *International Journal of Bio-Medical Computing*, **36**, 25–33.

Muller, P. and Rosner, G.L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. Am. Statist. Assoc.*, **92**, 1279-1292.

Racine-Poon, A. (1985). A Bayesian approach to nonlinear random effects models. *Biometrics*, **41**, 1015–1024.

Racine-Poon, A. and Wakefield, J.C. (1998). Statistical methods for population pharmacokinetic modelling. *Statistical Methods in Medical Research*, **7**, 63–84.

Sethuraman, J. and Tiwari, R.C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Decision Theory and Related Topics III, Vol. 2.* Academic Press Inc, New York.

Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. B* **55**, 3-23.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701-1762.

Wakefield, J.C (1996). Bayesian individualization via sampling based methods. *J. Pharmacokin. Biopharm.*, **24**, 103–31.

Wakefield, J.C., Aarons, L. and Racine-Poon, A. (1998). The Bayesian approach to population pharmacokinetic/pharmacodynamic modeling. In *Case Studies in Bayesian Statistics*, Carlin, B.P., Carriquiry, A.L., Gatsonis, C, Gelman, A., Kass, R.E., Verdinelli, I. and West, M. (editors). Springer, New York.

Wakefield, J.C., Smith, A.F.M., Racine-Poon, A. and Gelfand, A.E. (1994). Bayesian analysis of linear and non-linear population models using the Gibbs sampler. *Appl. Statist.*, **43**, 201-221.

Wakefield, J.C. and Walker, S.G. (1997). Bayesian nonparametric population model: formulation and comparison with likelihood approaches. *J. Pharmacokin. Biopharm.*, **25**, 235–253.

West, M., Muller, P. and Escobar, M.D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty: A Tribute to D.V.Lindley*, P.R. Freeman and A.F.M.Smith (editors), John Wiley and Sons, New York.

Stephen Walker
Department of Mathematics
Imperial College of Science,
Technology and Medicine
180 Queensgate
London, SW7 2BZ
e-mail :s.walker@ic.ac.uk

Jon Wakefield
Department of Epidemiology and Public Health
Imperial College School of Medicine
St Mary's Hospital
Norfolk Place
London, W2 1PG
e-mail : j.c.wakefield@ic.ac.uk

Figure Legends

Figure 1: Estimated random effects for log clearance and log volume against birth weight. The top row is from a parametric analysis and the bottom row from a nonparametric analysis.

Figure 2: Ten densities for log clearance corresponding to random draws from the posterior distribution of $p(F_{\log C}|y)$.

Figure 3: Estimated predictive densities for log clearance and log volume when the observed birth weight is equal to 1.0. The solid line corresponds to Apgar score ≥ 5 and the dotted line Apgar score < 5.

Figure 4: Estimated predictive densities for log clearance and log volume when the observed birth weight is equal to 2.0. The solid line corresponds to Apgar score ≥ 5 and the dotted line Apgar score < 5.