A Method for Combining Inference Across Related Nonparametric Bayesian Models

Peter Müller

University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, U.S.A.

Fernando Quintana

Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile.

Gary Rosner

University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, U.S.A.

Summary. We consider the problem of combining inference in related nonparametric Bayes models. Analogous to parametric hierarchical models, the hierarchical extension formalizes borrowing strength across the related sub-models. In the nonparametric context, modelling is complicated by the fact that the random quantities over which we define the hierarchy are infinite dimensional. We discuss a formal definition of such a hierarchical model. The approach includes a regression at the level of the nonparametric model. For the special case of Dirichlet process mixtures, we develop a Markov chain Monte Carlo scheme to allow efficient implementation of full posterior inference in the given model.

1. Introduction

Hierarchical models with nonparametric extensions at various levels of the hierarchy have been defined and used successfully in the recent literature. MacEachern (1994), Escobar (1994), and Escobar and West (1995) discuss computations in Dirichlet process (DP) mixture models where a parametric prior in a hierarchical model is replaced by the nonparametric DP model. Bush and MacEachern (1996) use a DP prior as random effects distribution in an ANOVA setup. Müller and Rosner (1997) use similar DP mixture models to introduce nonparametric population distributions for random effects in longitudinal data models. West et al. (1994) consider normal hierarchical models with DP mixture priors for density estimation. Quintana (1998) uses hierarchical models with DP priors to assess homogeneity in contingency tables. A recent collection of related review papers can be found in Dey et al. (1998).

In this paper we consider extension of such models to produce combined inference over related nonparametric Bayes models, i.e., hierarchical models where each sub-model is of nonparametric type. A byproduct of this extension is the resulting meta-analysis over models, restricted to the case where the full datasets are available. The approach we introduce is valid independently of the specific nonparametric model chosen for the individual submodels. However, the discussion of implementation details and the example are specific to Dirichlet process mixtures of normals.

One solution to achieve combined inference over related nonparametric models is to link separate nonparametric models at the level of the hyperparameters only, i.e., independent sub-models conditional on hyperparameters. For example, the base measure in a Dirichlet process prior for the i-th sub-model could include a regression on covariates specific to the submodel. This construction is introduced in Cifarelli and Regazzini (1978) as mixture of products of Dirichlet process. The model is used, for example, in Muliere and Petrone (1993). They define dependent nonparametric models for a set of random distributions $\{F_x,\ x \in \mathcal{X}\}$ by assuming marginally for each F_x a Dirichlet process prior, and introducing a regression in the base measures of these Dirichlet process priors. Similar models are discussed in Mira and Petrone (1996), Giudici et al. (2002) and Carota and Parmigiani (2002). While straightforward, this strategy is strictly limited to learning about features that can be represented by the hyperparameters. For example, consider mixtures of normal sub-models where the hyperparameters are the number of terms in the mixture and mean and variance of a hyperprior on the cluster locations. If we learn in the first study that observations are clustered in a certain way, the only information that is formally shared with the analysis of the other study is the number of terms and the overall location and variance as represented by the hyperparameters. In other words, learning about specific features of the second study, such as location of given terms in the mixture, is not improved by the information available from the first study. Tomlinson and Escobar (1999) mitigate this constraint by using a hyperparameter which itself is a random measure, i.e., a model with nonparametric hyperprior. MacEachern (1999) discusses an alternative approach for dependent DP models based on introducing correlations across the point masses in Sethuraman's stick-breaking construction (Sethuraman, 1994) of DP models.

Many applications that would naturally lead to nonparametric modelling include covariates at the level of the nonparametric model. For example, consider a longitudinal model for drug concentrations over time with a nonparametric prior for patient-specific random effects. It is important that the model incorporates the dependence of the random effects distribution on known patient-specific covariates, like treatment levels. One approach is discussed in Mallick and Walker (1997) who introduce regression in DP models. They propose a model that includes a finite partition of the covariates space, and for each subset of the partition they consider a different DP. Of course, this approach only works for finite categorical covariates. Alternatively, a straightforward generic strategy for introducing regression in a nonparametric model is to include the covariates in the nonparametric distribution. Consider a nonparametric model for an unknown distribution $p(\theta)$, for example the random effects distribution in a longitudinal data model, as mentioned above. To make the model $p(\theta)$ depend on covariates x, one could consider a joint distribution $p(x, \theta)$. The implied conditional distribution $p(\theta|x)$ formalizes the desired density estimation on θ as a function of x. This approach is used, for example, in Mallet et al. (1988) and Müller and Rosner (1998). However, the approach can be criticized from a modelling perspective for using the wrong likelihood. Including a joint distribution $p(x, \theta)$ in the model implies a marginal distribution $p(x)$. Although x is fixed by design, the model introduces a factor $p(x)$ in the likelihood. In Section 3.3 we discuss a justification of this approach as correct posterior inference under an alternative prior probability model.

Section 2 outlines an approach to combining inference over related nonparametric models. In Section 2.2 we consider the specific case of a hierarchical model with DP mixtures as nonparametric submodels. Section 3 discusses posterior simulation in the proposed model using Markov chain Monte Carlo simulation. Section 4 shows an example of combined inference over related Dirichlet process mixture models. Section 5 concludes with a final discussion.

2. A Hierarchical Model over Related Studies

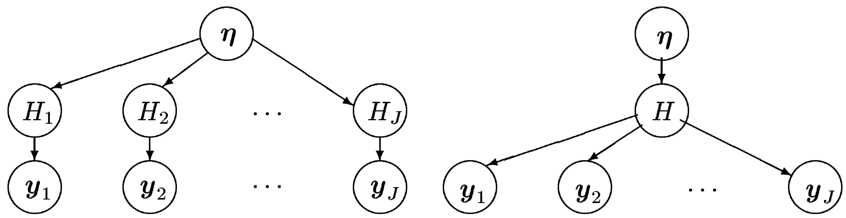
2.1. Combining Nonparametric Models

Consider a generic Bayesian model consisting of likelihood $y_i \sim p(y_i|H)$ and prior probability model $H \sim p(H|\boldsymbol{\eta})$, with possible hyperparameters $\boldsymbol{\eta}$. The model is referred to as nonparametric if H can not be indexed by finitely many parameters, i.e., $p(H|\boldsymbol{\eta})$ is a probability measure on a function space. Although the term "nonparametric" for these models is traditional, a possibly more appropriate terminology would be "massively parametric." In this paper we restrict the discussion to the case where H is a random probability measure. Typical examples are DP's (Ferguson, 1973; Antoniak, 1974), Polya trees (Lavine, 1992, 1994), Gaussian processes (O'Hagan, 1992; Angers and Delampady, 1992), beta-Stacy processes (Walker and Muliere, 1997), beta processes (Hjort, 1990), or extended gamma processes (Dykstra and Laud, 1981). See Walker et al. (1999) for a recent review.

If we want to analyze several related studies, $j = 1, \ldots, J$, we require a hierarchical extension of the model. Let $\boldsymbol{y}_j = (y_{ji},\ i = 1, \ldots, n_j)$ denote the data vector in study j, so that

$$\boldsymbol{y}_j \sim p(\boldsymbol{y}_j \mid H_j), \qquad H_j \sim p(H_j|\boldsymbol{\eta}), \tag{1}$$

$j = 1, \ldots, J$ and $i = 1, \ldots, n_j$. In the context of fully parametric inference, the use of hierarchical models to "borrow strength" across different but related sub-models is a common theme in statistical modelling. But in the case of hierarchically linking related studies where each sub-model $p(y_{ji}|H_j)$ is a nonparametric model, the nonparametric nature of H_j complicates modelling. There are two exceptions when the model simplifies, as shown in Figure 1. If the sub-models H_j are independent given the hyperparameters, then the problem reduces to analyzing J separate studies linked only by the finite dimensional hyperparameter vector. At the other extreme, if the observations y_{ji} can be considered exchangeable across studies, then the problem reduces to estimating one random measure H $(= H_1 = \ldots = H_J)$. For



(a) Independent sub-models (b) Exchangeable subjects

Fig. 1. Combining data from related studies assuming independent sub-models (left panel), and exchangeable subjects across studies (right panel). The desired level of borrowing strength across the sub-models is in-between these two extremes. See also Figure 2.

many applications, the first case allows too little borrowing of strength across studies, and the latter enforces too much borrowing by assuming essentially one population.

Instead, we consider a model which allows linking the sub-models at an intermediate level. A graphical representation is given in Figure 2. The model includes a common measure F_0, representing a baseline model which is common to all studies and random probability measures F_j that characterize the idiosyncratic behavior in study j. The split into a common effect and study-specific effects is akin to the setup of ANOVA models which include a similar distinction between overall means and study-specific offsets. We assume

$$H_j = \epsilon\, F_0 + (1-\epsilon)\, F_j, \qquad j = 1, \ldots, J, \tag{2}$$

with random measures

$$F_j \sim p(F_j|\boldsymbol{\eta}), \qquad j = 0, 1, \ldots, J \tag{3}$$

The weight ϵ, $0 \le \epsilon \le 1$, represents the level of borrowing strength across studies. A fraction ϵ of the total mass is shared by all studies, and the rest $(1-\epsilon)$ remains specific to each particular study. Thus, the data collected from each study contributes to the global learning about F_0, but learning on F_j can be accomplished only through $\boldsymbol{y}_j$. By a slight abuse of notation we will use H_j and F_j to generically indicate the probability models, as well as to denote the corresponding probability density functions.

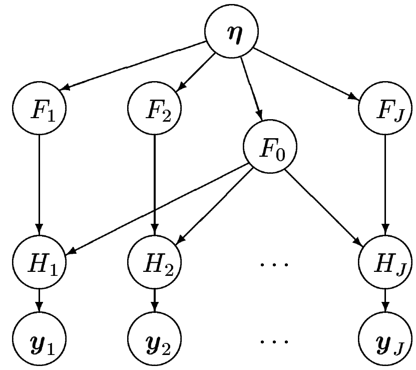


Fig. 2. Full hierarchical model. Equations (2) and (3) define a hierarchical model which assumes the random measure H_j in study j to be a mixture of the common measure F_0, shared by all studies, and an idiosyncratic measure F_j specific to each study.

As in any mixture model, one might wonder about identifiability of the model defined in (2). Since we use proper prior distributions the posterior distributions are guaranteed to be proper. Still, there might be practical concerns related to arbitrary rearrangements of the mixture, throwing into question the interpretation of the terms as idiosyncratic and common measures. Let M^o denote model (1) – (3) and let $\omega^o = (\epsilon, F_0, F_1, \ldots, F_J)$ denote a given parametrization. Could we fit the data equally well with alternative parameterizations defined by moving mass from the idiosyncratic measures F_j into the common measure? Or vice versa, by moving mass from the common measure into each of the idiosyncratic measures?

The first concern is easily addressed. Consider, for example, the following reparametrization which moves a fraction α, $0 < \alpha \leq (1-\epsilon)$ of F_1 into the common measure: $\omega^* = (\epsilon^*, F_0^*, F_1^*, \ldots, F_J^*)$ with $\epsilon^* F_0^* = \epsilon F_0 + \alpha F_1$, $\epsilon^* = \epsilon + \alpha$, and $F_j^* = F_j$. A change from ω^o to ω^* changes the likelihood $p(y_{ji}|\omega)$ for all but the observations in study $j=1$, leaving no concern about identifiability.

The second type of reparameterization needs more discussion. As an (extreme) example of moving mass from F_0 to F_j consider the alternative model M^{**} defined by $\epsilon^{**} = 0$. Consider the specific reparametrization $\omega^{**} = (\epsilon^{**} = 0, F_0^{**}, F_1^{**}, \ldots, F_J^{**})$ with $F_0^{**} = F_0$ and $F_j^{**} = (1-\epsilon) F_j + \epsilon F_0$. The likelihood remains invariant under the change from ω^o to ω^{**}, i.e., model M^{**} can fit the data at least as well as M^o. Still, unless the more complex model M^{**} provides a better fit to the data, the posterior distribution will put higher probability on the simpler model M^o. This is due to a general property of Bayesian posterior inferences. Assuming equal fit to the data, posterior distributions typically favor a more parsimonious model over a more complicated model. Jefferys and Berger (1992) interpret this as an automatic implementation of Ockham's razor. A formal discussion is easiest after marginalizing over the random measures F_j. Since this requires notation introduced in section 3.2 we will revisit the issue at the end of that section. Also, see the discussion there for a formal definition of model complexity, as well as the more general case $0 < \epsilon^{**} < \epsilon$.

2.2. A Hierarchical DP Mixture Model

In many applications nonparametric models are used to generalize traditional models with fully parametric assumptions. For example, in Müller and Rosner (1997) we replace a conventional multivariate normal random effects distribution with a Dirichlet process (DP) mixture of normal distributions. DP mixture models are attractive because of their computational simplicity (MacEachern and Müller, 1998). As we will show in Section 3 this computational simplicity extends to our hierarchical formulation.

Let $\varphi_{\boldsymbol{m},\boldsymbol{S}}(\boldsymbol{x})$ denote a (multivariate) normal p.d.f. with moments $(\boldsymbol{m}, \boldsymbol{S})$, evaluated at $\boldsymbol{x}$, and let $\mathcal{D}(M, G_{\boldsymbol{\eta}})$ denote the DP with centering probability measure $G_{\boldsymbol{\eta}}$ and weight (total mass) parameter M. Typically the centering measure $G_{\boldsymbol{\eta}}$ includes some unknown hyperparameters $\boldsymbol{\eta}$ which are given a hyperprior $p(\boldsymbol{\eta})$, detailed below. The DP mixture of normal model defines a nonparametric model $p(F_j|\boldsymbol{\eta})$ as a mixture of normal distributions with respect to a random mixing measure G_j generated by a DP prior:

$$F_j(\cdot|\boldsymbol{\eta}, M_j) = \int \varphi_{\boldsymbol{\mu},\boldsymbol{S}}(\cdot)\, dG_j(\boldsymbol{\mu}), \quad G_j \sim \mathcal{D}(M_j, G_{\boldsymbol{\eta}}), \quad j = 0, 1, \ldots, J. \tag{4}$$

We build on (4) to define a hierarchical model for random distributions H_j, $j = 1, \ldots, J$ in J related studies. Using the structure introduced in (2) and assuming that the relevant sampling model in each study is i.i.d. sampling from H_j, we have

$$H_j = \epsilon\, F_0 + (1-\epsilon)\, F_j, \qquad j = 1, \ldots, J, \tag{5}$$

$$y_{ji} \sim H_j(y_{ji}). \tag{6}$$

We refer to (4) – (6) as hierarchical DP mixture model. The sampling model could be more general than (6) without changing much in the following discussion. In fact, the example in Section 4 uses a sampling model where the H_j play the role of a random effects distribution in each sub-model.

Model (4) – (6) includes commonly used models as special cases. With $J = 1$ and $\epsilon = 0$ the model reduces to a Dirichlet process mixture model as used, for example, in Kleinman and Ibrahim (1998). If $\epsilon = 0$ and the DP mixture of normals is replaced by a single multivariate normal, $y_{ji} \sim N(\mu_j, S)$ and $\mu_j \sim G_{\boldsymbol{\eta}}$, then the model becomes a one-way ANOVA model with a normal sampling distribution and random effects distribution $G_{\boldsymbol{\eta}}$.

A DP prior with a small total mass parameter M approximates this special case. If (4) is replaced by a finite mixture of normals then we obtain a flexible parametric alternative model. Such models are explored in Lopes et al. (2003).

We choose the following hyperpriors on the various hyperparameters that are present in our model. First, the centering probability measure $G_{\boldsymbol{\eta}}(\cdot)$ is chosen as a normal distribution $N(\boldsymbol{m}, \boldsymbol{B})$ with moments $\boldsymbol{\eta} = (\boldsymbol{m}, \boldsymbol{B})$. Let $W(\cdot|\cdot,\cdot)$ denote the Wishart distribution. We assume a conjugate hyperprior $p(\boldsymbol{\eta}) = \varphi_{\boldsymbol{m}_0,\boldsymbol{A}}(\boldsymbol{m}) \cdot W[\boldsymbol{B}^{-1}|c,(c\boldsymbol{C})^{-1}]$, with fixed hyperparameters $\boldsymbol{m}_0$, c, $\boldsymbol{A}$, and $\boldsymbol{C}$. Next, we choose conjugate-style hyperpriors for $\boldsymbol{S}$ and M_j: $\boldsymbol{S}^{-1} \sim W[\boldsymbol{S}^{-1}|q,(q\boldsymbol{R})^{-1}]$ and $M_j \sim \Gamma(a_0, b_0)$, where $\Gamma(\cdot,\cdot)$ is the Gamma distribution, and $\boldsymbol{R}$, a_0, b_0 and q are fixed hyperparameters. Alternatively, $\boldsymbol{S}$ could be indexed with study j.

Finally, for the weight ϵ we assume a prior distribution which allows for positive prior probability on $\epsilon = 0$ and $\epsilon = 1$:

$$p(\epsilon) = \pi_0\delta_0(\epsilon) + \pi_1\delta_1(\epsilon) + (1 - \pi_0 - \pi_1)\mathrm{Beta}(\epsilon|a_\epsilon, b_\epsilon), \tag{7}$$

where $a_\epsilon, b_\epsilon > 0$, and $0 \leq \pi_0, \pi_1 < 1$ are fixed hyperparameters such that $0 \leq \pi_0 + \pi_1 < 1$, and $\delta_x(\cdot)$ is a point mass distribution at x. The distribution in (7) assigns positive probability to the two extreme models shown in Figure 1, represented by $\delta_0(\epsilon)$ and $\delta_1(\epsilon)$, but it also allows all the intermediate combinations. We note here that π_0 and π_1 are treated as fixed, because little is gained by putting prior distributions on these quantities.

3. Posterior Simulation

3.1. Latent Variables and Indicators

We implement posterior and posterior predictive inference in the proposed model by Markov chain Monte Carlo (MCMC) simulation. Posterior MCMC simulation for DP mixture models is developed, for example, in MacEachern and Müller (1998) for models without the additional hierarchy defined in (5), i.e.:

$$\boldsymbol{y}_i \sim \int \varphi_{\boldsymbol{\mu},\boldsymbol{S}}(\boldsymbol{y}_i)\, dG(\boldsymbol{\mu}), \qquad G \sim \mathcal{D}(G|M, G_{\boldsymbol{\eta}}),$$

or, replacing the mixture by a latent variable $\boldsymbol{\mu}_i$:

$$\boldsymbol{y}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{S}), \qquad \boldsymbol{\mu}_i \sim G, \qquad G \sim \mathcal{D}(G|M, G_{\boldsymbol{\eta}}), \tag{8}$$

$i = 1, \ldots, n$. See Walker and Damien (1998), Neal (2000), and Green and Richardson (2001) for alternative approaches.

Implementing posterior simulation in (8) we can marginalize over the unknown measure G, and consider only the latent variables $\boldsymbol{\mu}_i$. Due to the discrete nature of G, some of the $\boldsymbol{\mu}_i$ can be identical. Denote by $\boldsymbol{\phi} = \{\boldsymbol{\phi}_h,\ h = 1, \ldots, K\}$, $K \leq n$, the set of distinct $\boldsymbol{\mu}_i$'s.

Implementation of the MCMC simulation for (8) proceeds by introducing latent indicator variables which identify clusters of equal $\boldsymbol{\mu}_i$'s, say $s_i = h$ if and only if $\boldsymbol{\mu}_i = \boldsymbol{\phi}_h$. A critical step in the MCMC simulation is the resampling of these indicators. Conditional on the configuration indicators $\boldsymbol{s} = (s_i,\, i = 1, \ldots, n)$, the conditional posterior of $\boldsymbol{\phi}_h$ given $\boldsymbol{s}$ and all other parameters is exactly the same as in a corresponding parametric model

$$\boldsymbol{y}_i \sim p(\boldsymbol{y}_i | \boldsymbol{\phi}_h), \qquad i \in \{i : s_i = h\},$$

with prior $\boldsymbol{\phi}_h \sim G_{\boldsymbol{\eta}}$. Details are discussed in MacEachern and Müller (1998).

Considering MCMC posterior simulation in model (4) – (6) we run into some good luck. Although the hierarchical model (4) – (6) generalizes the basic DP mixture model (8) by allowing for the additional hierarchy corresponding to the studies $j = 1, \ldots, J$, the technicalities of the posterior MCMC simulation change little. The only changes are additional indicators, say r_{ji}, corresponding to the mixture (2) into common and idiosyncratic measure, and an additional constraint in resampling the configuration indicators $\boldsymbol{s}$. Essentially the constraint on $\boldsymbol{s}$ amounts to allowing only indicators corresponding to observations from the same study to share the same cluster. For reference, we restate the complete model, (4) – (5), with indicators (r_{ji}) and latent variables $(\boldsymbol{\mu}_{ji})$ replacing mixtures at all levels

$$y_{ji} \sim N(\boldsymbol{\mu}_{ji}, \boldsymbol{S}) \qquad \text{with} \qquad \boldsymbol{\mu}_{ji} \sim \begin{cases} F_0(\boldsymbol{\mu}_{ji}) & \text{if } r_{ji} = 0 \\ F_j(\boldsymbol{\mu}_{ji}) & \text{if } r_{ji} = 1, \end{cases}$$

$Pr(r_{ji} = 0) = \epsilon$, and $F_j \overset{ind}{\sim} \mathcal{D}(M_j, G_{\boldsymbol{\eta}}),\ j = 0, \ldots, J$.

Implementing the MCMC simulation we proceed by marginalizing over the random measures F_j. Paralleling the discussion of posterior inference in MacEachern and Müller (1998), as summarized above, some of the $\boldsymbol{\mu}_{ji}$'s are identical. Let $\boldsymbol{\phi}_j = \{\boldsymbol{\phi}_{jh},\ h = 1, \ldots, K_j\}$ denote the set of distinct values among the components of $\boldsymbol{\mu}_j = \{\boldsymbol{\mu}_{ji} :\ i = 1, \ldots, n_j \text{ and } r_{ji} = 1\}$. Similarly, let $\boldsymbol{\phi}_0 = \{\boldsymbol{\phi}_{0h},\ h = 1, \ldots, K_0\}$ denote the distinct values in $\boldsymbol{\mu}_0 = \{\boldsymbol{\mu}_{ji} :\ j = 1, \ldots, J,\ i = 1, \ldots, n_j \text{ and } r_{ji} = 0\}$. Here K_j and K_0 are the number of distinct values in $\boldsymbol{\phi}_j$ and $\boldsymbol{\phi}_0$, respectively. We introduce indicators s_{ji} with $s_{ji} = h$ if and only if ($\boldsymbol{\mu}_{ji} = \boldsymbol{\phi}_{jh}$ and $r_{ji} = 1$) or ($\boldsymbol{\mu}_{ji} = \boldsymbol{\phi}_{0h}$ and $r_{ji} = 0$). We will use (ji) and (jh) to refer to patients and clusters with the given indices, respectively. Let $n_{jh} = |\{i :\ \boldsymbol{\mu}_{ji} = \boldsymbol{\phi}_{jh}\}|$ and $n_{0h} = |\{(ji) :\ \boldsymbol{\mu}_{ji} = \boldsymbol{\phi}_{0h}\}|$ denote the number of observations allocated to cluster (jh) and $(0h)$, respectively. Let $n_j = \sum_h n_{jh},\ j = 1, \ldots, J$. Let n_{kh}^- denote the number of observations allocated to cluster (kh), excluding a given observation (ji), and similarly for n_{0h}^- and n_j^-. It will be clear from the context which (ji) is excluded. We will use $\boldsymbol{\mu}$ to denote the vector of all $\boldsymbol{\mu}_{ji}$, and $\boldsymbol{\mu}^-$ for the same vector without a specific $\boldsymbol{\mu}_{ji}$ component. Also, $\boldsymbol{\nu} = (M_0, \ldots, M_J, \boldsymbol{S}, \boldsymbol{\eta}, \epsilon)$ will denote the vector of hyperparameters.

3.2. Markov Chain Monte Carlo Simulation

We only describe here the steps of updating s_{ji}, r_{ji} and ϵ. All other steps in the MCMC remain unchanged as described in MacEachern and Müller (1998) for model (8), above. MCMC simulation proceeds as a Gibbs sampling scheme scanning over the complete conditional distributions for s_i $(i = 1, \ldots, n)$, ϕ_h $(h = 1, \ldots, K)$, M, and η. The only non-standard distribution is the conditional posterior for s_i, which is modified as follows for model (4) – (6) . Recall that we use F_j to denote probability densities (rather than distribution functions).

(i) *Resampling* $(\boldsymbol{\mu}_{ji}, s_{ji}, r_{ji})$. Let $G^*(\boldsymbol{\phi}) \propto \varphi_{\boldsymbol{\phi},\boldsymbol{S}}(y_{ji}) F_0(\boldsymbol{\phi})$ and let g^* be the normalization constant $g^* = \int \varphi_{\boldsymbol{\phi},\boldsymbol{S}}(y_{ji}) dF_0(\boldsymbol{\phi})$ in G^*. Define the probabilities $\pi_{jh} = c\, \varphi_{\boldsymbol{\phi}_{jh},\boldsymbol{S}}(y_{ji})\; (1-\epsilon)\, n^-_{jh}/(M_j + n^-_j)$, $\pi^*_j = c\, g^*\; (1-\epsilon)\; M_j/(M_j + n^-_j)$, $\pi_{0h} = c\, \varphi_{\boldsymbol{\phi}_{0h},\boldsymbol{S}}(y_{ji})\; \epsilon\, n^-_{0h}/(M_0 + n^-_0)$, and $\pi^*_0 = c\, g^*\; \epsilon\, M_0/(M_0 + n^-_0)$, where c is the appropriate constant to standardize the sum of all weights $\pi_{jk}, \pi_{0k}, \pi^*_j, \pi^*_o$ to add up to 1.0. Let $\boldsymbol{\phi}^* \sim G^*(\boldsymbol{\phi})$. To generate a draw $(\boldsymbol{\mu}_{ji}, s_{ji}, r_{ji})$ from the complete conditional $p(\boldsymbol{\mu}_{ji}, s_{ji}, r_{ji} | \theta, \boldsymbol{\nu}, \boldsymbol{\mu}^-, \boldsymbol{y})$ set

$$(\boldsymbol{\mu}_{ji}, s_{ji}, r_{ji}) = \begin{cases} (\boldsymbol{\phi}_{jh}, h, 1),\; h = 1, \ldots, K_j & \text{with probability } \pi_{jh} \\ (\boldsymbol{\phi}_{0h}, h, 0),\; h = 1, \ldots, K_0 & \text{with probability } \pi_{0h} \\ (\boldsymbol{\phi}^*, K_j + 1, 1) & \text{with probability } \pi^*_j \\ (\boldsymbol{\phi}^*, K_0 + 1, 0) & \text{with probability } \pi^*_0. \end{cases} \tag{9}$$

(ii) *Resampling* ϵ. We update ϵ by generating from the complete conditional posterior given the indicators $\boldsymbol{r} = (r_{ji},\, j = 1, \ldots, J,\; i = 1, \ldots, n_j)$. Given $\boldsymbol{r}$ the weight ϵ is conditionally independent of all other parameters. Let $B(a, b)$ denote the beta function evaluated at (a, b). Let $N_1 = \sum r_{ji}$ and $N_0 = n - N_1$, and use $I(A)$ to denote the indicator function of event A. Then

$$p(\epsilon | \boldsymbol{r}) \propto (1 - \pi_0 - \pi_1)\, \frac{B(a^*_\epsilon, b^*_\epsilon)}{B(a_\epsilon, b_\epsilon)}\, Be(a^*_\epsilon, b^*_\epsilon) + \\ + \pi_0\, I(N_0 = n)\, \delta_0(\epsilon) + \pi_1\, I(N_1 = n)\, \delta_1(\epsilon). \tag{10}$$

with $a^*_\epsilon = a_\epsilon + N_0$ and $b^*_\epsilon = b_\epsilon + N_1$.

(iii) All other parameters are resampled as described in MacEachern and Müller (1998).

General conditions to ensure convergence of the proposed MCMC scheme are described in Tierney (1994). In the context of the proposed algorithm, the only practically critical condition is irreducibility of the chain. See MacEachern and Müller (1998) for a detailed verification that the proposed algorithm meets the conditions of the results in Tierney (1994).

The latent variables $\boldsymbol{\mu}_{ji}$ and indicators s_{ji} allow now to formalize the argument from the end of Section 2.1 about how posterior inference in the proposed model implements Ockham's razor and favors the more parsimonious model. Paralleling the discussion of models M^o and M^{**} at the end of Section 2.1 we define M^o_M and M^{**}_M to denote two models parametrized by latent variables $(\boldsymbol{\phi}_{jk}, K_0, K_j, s_{ji})$ and $(\boldsymbol{\phi}^{**}_{jk}, K^{**}_0 = 0, K^{**}_j, s^{**}_{ji})$, respectively. Model M^o_M is model M^o, marginalized with respect to the random measures F_j, and model M^{**}_M is a special case corresponding to no common measure, i.e., $\epsilon = 0$. Considering $K^{**}_0 = 0$, $K^{**}_j = K_j + K_0$ and $\boldsymbol{\phi}^{**}_{jk} = \boldsymbol{\phi}_{0h}$ for $k = K_j + h, h = 1, \ldots, K_0$ we find that M^{**}_M provides at least as good a fit to the data as model M^o_M. In fact, under the described reparametrization $p(\boldsymbol{y}|\boldsymbol{\phi}_{jk}, s_{ji}, K_j) = p(\boldsymbol{y}|\boldsymbol{\phi}^{**}_{jk}, s^{**}_{ji}, K^{**}_j)$. But model M^{**}_M is more complex than M^o_M, in the sense that the total number of terms in the mixtures, summed across all random distributions, is $\sum K_j + J\,K_0$, as opposed to $\sum K_j + K_0$ for M^o_M. Jefferys and Berger (1992) show how posterior inference favors the simpler model with fewer parameters unless the more complicated model provides a significantly better fit to the data. They interpret this as an automatic implementation of Ockham's razor in posterior inference. This mechanism is due to the fact that under the more complicated model prior probability mass has to be distributed over a wider range of the additional parameters, implying a reduced marginal distribution.

Model M^{**} represents the extreme case of moving all probability mass from the common measure into the idiosyncratic measures by setting $\epsilon^{**} = 0$. But the same argument holds for $0 < \epsilon^{**} < \epsilon$. To add the remaining probability mass $\alpha = \epsilon - \epsilon^{**}$ to the idiosyncratic measures we need to include additional terms to each of the study specific mixtures. In the context of identifiability considerations it is important to keep in mind that model (4) – (6) includes a representation (and probability model) for F_0. In particular, this does not constrain $\min_j\{F_j(x)\}$ to vanish, as would be the case in an alternative approach based on deterministically defining $F_0(x) \propto \min_j\{F_j(x)\}$. In any case we caution against over-interpreting inference on the individual parameters in the model. The practically relevant inference are the posterior predictive distributions for the observable outcomes.

3.3. Regression in the Nonparametric Model

We now extend the model to nonparametric regression, i.e., inclusion of covariates in (2) and (3). To be specific, consider a density estimation problem, i.e., H_j is an unknown distribution and

$$y_{ji} \sim H_j,\ i = 1, \ldots, n_j, \tag{11}$$

with the prior model (2) and (3) for H_j. Assume now we have covariates $\boldsymbol{x}_{ji}$ available and want to allow the random distribution to depend on $\boldsymbol{x}_{ji}$. A straightforward approach to include a regression on covariates is to extend the random measures $F_j(\boldsymbol{y})$ and $H_j(\boldsymbol{y})$ to probability measures $F_j(\boldsymbol{x}, \boldsymbol{y})$ and $H_j(\boldsymbol{x}, \boldsymbol{y})$ on the joint space of responses y_{ji} and covariates $\boldsymbol{x}_{ji}$. Let $H_j(\boldsymbol{x}_{ji}) = \int H_j(\boldsymbol{x}_{ji}, \boldsymbol{y}) d\boldsymbol{y}$. The extended model implies a conditional probability model

$$H_j(y_{ji}|\boldsymbol{x}_{ji}) = H_j(\boldsymbol{x}_{ji}, y_{ji})/H_j(\boldsymbol{x}_{ji}) \tag{12}$$

which formalizes the desired regression. Although we define H_j to include $\boldsymbol{x}$, the likelihood (12) is strictly limited to a probability measure $H_j(\boldsymbol{y}|\boldsymbol{x})$ in $\boldsymbol{y}$ only. We use the joint distribution $H(\boldsymbol{x}, \boldsymbol{y})$ solely to define a family of conditional distributions indexed by $\boldsymbol{x}$, as desired. Without any further changes in the probability model, posterior inference would be significantly complicated by the need to evaluate the integrals in the denominator of (12). We avoid this with the following modification to the prior. We replace the original prior $p(F_j|\boldsymbol{\eta})$, $j = 0, \ldots, J$, by what would be the posterior if $\boldsymbol{x}_{ji}$ were sampled $\boldsymbol{x}_{ji} \sim H_j(\boldsymbol{x}_{ji})$, independently. Denote with $p(F_0, \ldots, F_J \mid \boldsymbol{x}, \epsilon, \boldsymbol{\eta})$ the posterior conditional on $\boldsymbol{x}_{ji} \sim H_j(\boldsymbol{x}_{ij})$, under the original prior $F_j \sim p(F_j|\boldsymbol{\eta})$. We define a new prior probability model

$$p^*(F_0, \ldots, F_J| \, \epsilon, \boldsymbol{\eta}) \equiv p(F_0, \ldots, F_J \mid \boldsymbol{x}, \epsilon, \boldsymbol{\eta}).$$

Together with the likelihood (12) this leads to a posterior distribution which is identical to the posterior as if the pairs $(\boldsymbol{x}_{ji}, y_{ji}) \sim H_j$ were sampled independently, allowing easy and efficient posterior simulation. Implementing posterior simulation we can proceed as if we had independent samples $(\boldsymbol{x}_{ji}, y_{ji}) \sim H_j$.

4. Example: Combined Inference From Related Pharmocological Studies

4.1. Data

The methodology developed in this article was motivated by the analysis of data from two studies carried out by the Cancer and Leukemia Group B (CALGB) (Lichtman et al., 1993). CALGB 8881 was a phase I study that sought the highest dose of the anti-cancer agent cyclophosphamide (CTX) one could give cancer patients every two weeks. Patients also received the drug GM-CSF to help reduce the ill effects of CTX on the patients marrow. The other study, CALGB 9160, built upon the experience gained in 8881 using the resulting doses of CTX and GM-CSF, and investigated the effect of an additional drug, amifostine (AMF). AMF had been shown in some studies to reduce some of the toxic side effects of anticancer agents, such as CTX and radiation therapy (Spencer and Goa, 1995). The objective of CALGB 9160 was to determine if adding AMF would reduce the hematologic side effects of aggressive chemotherapy with CTX and GM-CSF. CALGB 9160 randomized patients to receive AMF or not, along with CTX (3 grams per square meter of body surface area) and GM-CSF (5 micrograms per kilogram of body weight). The main study question in CALGB 9160 concerned the effect of AMF on various measures of hematologic toxicity, such as nadir (i.e., minimum) white blood cell (WBC) counts or days of granulocytopenia. Since only 46 patients entered the randomized trial, we wished to use data already gathered in the earlier study to help make inference in CALGB 9160 more precise.

In both studies, the main response was white blood cell count (WBC) for each patient over time. In study 8881, we have data on $I_1 = 52$ patients. The other study includes data on $I_2 = 46$ patients. We will use y_{jik} to denote the k-th blood count measurement on the i-th patient in study j on day t_{jik}, recorded on a log scale of thousands, i.e., $y_{jik} = \log(\mathrm{WBC}/1000)$. In CALGB 8881 and 9160, we had a total of 674 and 706 observations, respectively, with the number of observations for one patient varying between 2 and 19. Figure 3 shows a few typical patients. In Müller and Rosner (1998), we used a non-linear

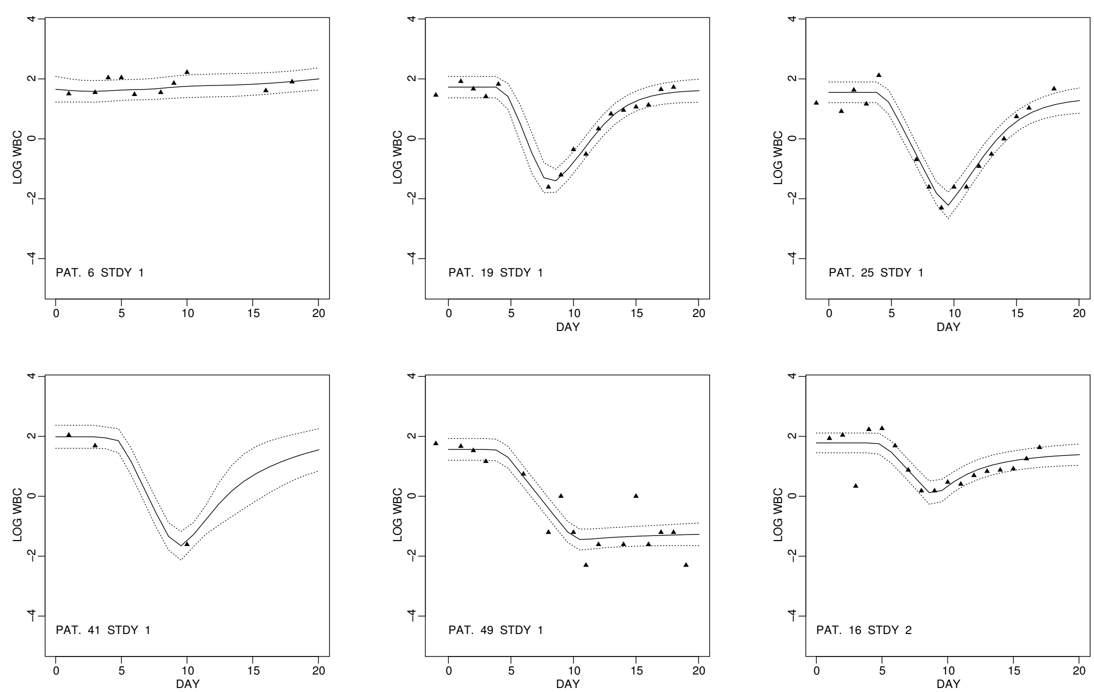


Fig. 3. Some typical patients. The triangles are the observed WBC. The solid line is the posterior fitted curve $E[f_{ji}(t)|\boldsymbol{y}]$ as a function of t. The dotted lines indicate one posterior standard deviation margins.

regression model,

$$y_{jik} = f_{ji}(t_{jik}) + \epsilon_{jik}, \quad \epsilon_{jik} \sim N(0, \sigma^2), \tag{14}$$

to fit these profiles. Let $\boldsymbol{\theta}_{ji} = (z_{1ji}, z_{2ji}, z_{3ji}, \tau_{1ji}, \tau_{2ji}, \beta_{0ji}, \beta_{1ji})$ denote patient-specific regression parameters, and let $\rho_{jik} = (\tau_{2ji} - t_{jik})/(\tau_{2ji} - \tau_{1ji})$, and $g_{ji}(t) = z_{2ji} + z_{3ji}/(1 + \exp\{-\beta_{0ji} - \beta_{1ji}(t - \tau_{2ji})\})$. We define

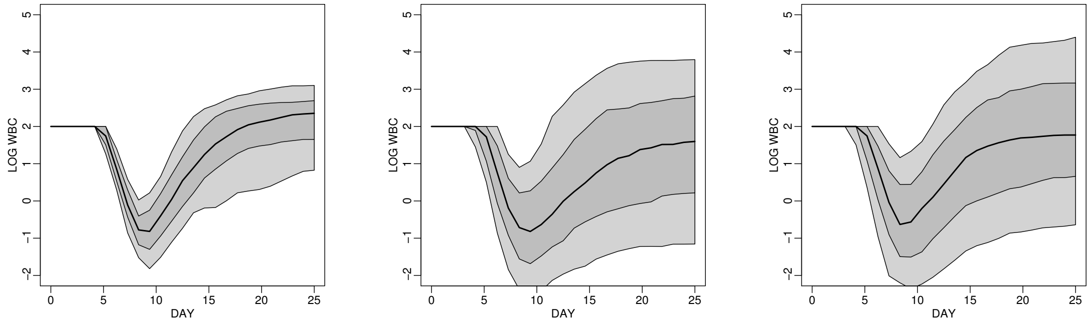
$$f_{ji}(t_{jik}) = \begin{cases} z_{1ji} & \text{if } t_{jik} < \tau_{1ji} \\ \rho_{jik} z_{1ji} + (1 - \rho_{jik}) g_{ji}(\tau_{2ji}) & \text{if } \tau_{1ji} \le t_{jik} < \tau_{2ji} \\ g_{ji}(t_{jik}) & \text{if } t_{jik} \ge \tau_{2ji}, \end{cases} \tag{15}$$

for $k = 1, \ldots, n_{ji}$. The curve defined by (15) consists of a horizontal line up to $t = \tau_{1ji}$, a logistic regression curve starting at $t = \tau_{2ji}$, and a straight line connecting these.

We complete the model by assuming a DP mixture model (4) and (5) for the random effects $\boldsymbol{\theta}_{ji}$, including a hierarchical extension over the two studies $j = 1, 2$. Let $\boldsymbol{x}_{ij} =$ (CTX, GM-CSF, AMF) denote the dose levels used for patient i in study j. Proceeding as in Section 3.3. we include a regression on $\boldsymbol{x}_{ij}$ in the random effects model. The non-linear regression (14) adds an additional level to the model, i.e., the random effects $\boldsymbol{\theta}_{ji}$ replace $\boldsymbol{y}_{ji}$ in (6). Conditional on $\boldsymbol{\theta}_{ji}$, the non-linear regression (14) defines the sampling distribution for the observed data $\boldsymbol{y}_{ji}$. The implementation requires an additional step in the MCMC simulation to update the random effects vectors $\boldsymbol{\theta}_{ji}$. See Müller and Rosner (1997) for a description of appropriate MCMC steps.

4.2. Results

Figure 4 shows posterior estimates of F_0, F_1 and F_2. The initial base line z_1 (the first element of the random effects vector $\boldsymbol{\theta}$) was conditioned upon as $z_1 = 2$ to make posterior predictive profiles comparable. The figures visualize the high dimensional distributions by showing the corresponding log WBC profiles for a patient with covariates $\boldsymbol{x}^* =$ (CTX $= 3g/m^2$, GM-CSF $= 5\mu g/kg$, AMF $= 0$). Let $f(t; \boldsymbol{\theta})$ denote the profile parameterized by the random effects vector $\boldsymbol{\theta}$, evaluated at day t. Figure 4a shows the quantiles for $f(t; \boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim F_0(\boldsymbol{\theta} | \boldsymbol{x} = \boldsymbol{x}^*)$, i.e., the quantiles for the mean log WBC for a patient with covariates $\boldsymbol{x}$. Figures 4b and 4c show the same for the random effects distribution F_1 and F_2. Notice how both idiosyncratic measures F_1 and F_2 are more dispersed than the common measure F_0. This can be attributed to the idiosyncratic measure F_j accommodating outliers in study j which do not occur in other studies. Posterior inference on ϵ informs about the proportion of the common measure in the mixture (2). The prior included positive point masses $\pi_0 = \pi_1 = 0.1$. Yet, *a posteriori* we find practically zero probability at the two endpoints. We find marginal posterior summaries $E(\epsilon \mid y) = 0.59$, SD$(\epsilon \mid y) = 0.05$, and



(a) F_0 (b) F_1 (c) F_2

Fig. 4. The common and idiosyncratic measures F_0, F_1 and F_2. Consider a patient with covariates $\boldsymbol{x}^* = (\text{CTX} = 3g/m^2, \text{GM-CSF} = 5\mu g/kg, \text{AMF} = 0)$ and random effects vector generated from F_j ($j = 0, 1, 2$, respectively). That is $\boldsymbol{\theta} \sim F_j(\boldsymbol{\theta}|\boldsymbol{x} = \boldsymbol{x}^*)$. The three panels plot quantiles of $f(t, \boldsymbol{\theta})$ against days t. The dark grey shade indicates the 25% and the 75% quantiles, the light grey shade indicates the 10% and 90% quantiles. The central solid curve plots the median of $f(t, \boldsymbol{\theta})$ against day t.

$Pr(0.45 \leq \epsilon \leq .75 \mid y) = 1.00$, indicating that neither a joint analysis of all data in one population ($\epsilon = 1$), nor an analysis with all studies independent given the hyperparameters ($\epsilon = 0$) is appropriate.

Of particular interest is the posterior predictive distribution for a patient from the population, i.e., for a new patient from a new study. Since the hierarchical model allows us to learn about variation between studies, such inference is meaningfully possible. Figure 5 shows some aspects of such posterior predictive inference. The center and right panels of Figure 5 allow us to infer that the addition of AMF does not appear to add further protection from the effect of CTX on a patient's blood counts. The center panel shows that the probability the patient's WBCs will recover by day 14 to be at least 1,000 $/\mu$L is around 0.65 and is lower than the predictive probability of the same event for the same patient without AMF (around 0.82 from the Figure). This difference in predictive probabilities of a meaningful clinical event is greater than the posterior standard deviation, which is around 0.03. The right-hand panel of Figure 5 shows that the addition of AMF does not appear to make any difference in the predicted number of days a patient's WBCs are below 1,000 $/\mu$L. Thus, the conclusion is that including AMF to CTX and GM-CSF does not reduce the toxic effects of these drugs on the WBCs of these or similar cancer patients receiving this chemotherapy.

Using data from a single study only, inference as in Figure 5 is restricted to the subpopulations from each of the respective studies. For comparison we implemented inference

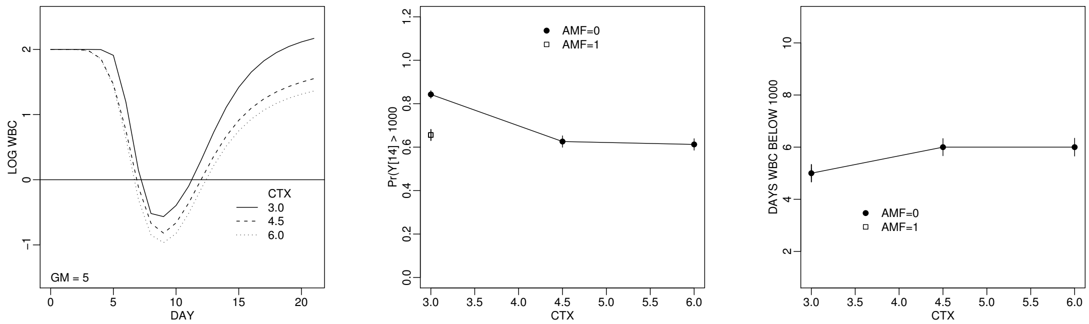


Fig. 5. Some features of the posterior predictive distribution for a patient from the population at large, i.e., $p(\boldsymbol{y}_{J+1,1}|\boldsymbol{y})$. The left panel shows the estimated WBC profiles for different levels of CTX as a function of days. The center panel shows the probability of recovery beyond $WBC = 1000$ by day 14. The right panel plots the expected number of days below the critical level $WBC = 1000$ as a function of the covariates CTX and AMF, keeping GM-CSF at $5\mu g/kg$. For AMF=1, only CTX=3.0 is shown (to avoid extrapolation beyond the range of the data). The point for $(CTX = 3, AMF = 1)$ is overlaid with AMF=0. The vertical bars indicate one posterior standard deviation.

for study 9160 alone, using the same model, but without the additional mixture in (5), or, equivalently, with $\epsilon = 0$. Posterior predictive inference (not shown) for a future patient from the 9160 population looks similar as in Figure 5a (curve for CTX=3), except for a slightly faster recovery, resulting in a reduced posterior predictive mean for the number of days with WBC below 1,000/μL. For AMF=0 we find a posterior predictive mean of 4 days with WBC below 1,000/μL, and slightly below 4 days for AMF=1 (with the other treatments fixed at the only dose used in 9160, CTX=3 g/m^2 and GM-CSF=5 $\mu g/kg$).

5. Discussion

We defined a framework for hierarchical meta-analysis over related nonparametric models. This general scheme incorporates the ability to represent random measures as functions of certain covariates of arbitrary type. Although the nature of the hierarchical extension is independent of the specific nonparametric model, the discussion of implementation details is necessarily constrained to a specific model. We chose the DP model. We showed how posterior MCMC simulation in the hierarchical model adds only little additional computational difficulty compared to a non-hierarchical model. Essentially the only change is an additional constraint when resampling the cluster indicators s_{ji}.

Generalization of the proposed hierarchical extension to other non-parametric models beyond the DP is possible. For any non-parametric prior based on similar stick-breaking representations as the DP we expect that the same construction and computation efficient posterior simulation remains possible. Such models are proposed, for example, in Muliere and Tardella (1998) and Ishwaran and James (2001). The general structure (2) and (3) remains meaningful also for other, arbitrary non-parametric prior models for the unknown distributions F_j. Conditional on imputed indicators r_{ji} that break the mixture (2) posterior simulation always reduces to the case of the non-hierarchical model. But simulations would typically require separate inference for each of the random distributions F_j, conditional on the indicators r_{ji}. For example, the Polya tree model might be suitable for the described generalization. However, we have not investigated details. A prominent feature of the DP is the particularly simple form of the Polya urn description for the marginal distribution of the observable data, marginalized with respect to the unknown measure F_j. Availability of such simplifications is not a necessary condition for the use of the hierarchical extensions described here, beyond the fact that it simplies inference in the hierarchical model to the same extent as it simplifies inference in the non-hierarchical context.

Another interesting generalization is related to inference on ϵ. As is easily seen from (10), posterior MCMC simulation only includes positive transition probability for a move to $\epsilon = 0$ or $\epsilon = 1$ if all data are allocated to common clusters ($N_0 = n$) or all data are allocated to study specific clusters ($N_1 = n$), respectively. This suggests to consider additional moves in the MCMC algorithm that make a common proposal for all r_{ij}. We did not pursue such extensions since in the motivating application the marginal posterior for ϵ was clearly bounded away from $\epsilon = 0$ and $\epsilon = 1$.

Acknowledgment

Research was partially supported by NIH under grant 1R01CA75981-01, by FONDECYT under grant 1990430 and by DIPUC under a visiting professorship grant. Most of the research was done while the first author was visiting at Pontificia Universidad Católica de Chile.

References

Angers, J.-F. and Delampady, M. (1992) Hiearchical Bayesian curve fitting and smoothing. *The Canadian Journal of Statistics*, **20**, 35–49.

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *Annals of Statistics*, **2**, 1152–1174.

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

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

Cifarelli, D. and Regazzini, E. (1978) Problemi statistici non parametrici in condizioni di scambialbilita parziale e impiego di medie associative. *Tech. rep.*, Quaderni Istituto Matematica Finanziaria, Torino.

Dey, D., Müller, P. and Sinha, D. (eds.) (1998) *Practical Nonparametric and Semiparametric Bayesian Statistics.* New York, NY, USA: Springer-Verlag.

Dykstra, R. L. and Laud, P. W. (1981) A Bayesian nonparametric approach to reliability. *Annals of Statistics*, **9**, 356–367.

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

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

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

Giudici, P., Mezzetti, M. and Muliere, P. (2002) Mixtures of dirichlet process priors for variable selection in survival analysis. *Journal of statistical planning and inference*, to appear.

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

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

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

Jefferys, W. and Berger, J. (1992) Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64–72.

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

Lavine, M. L. (1992) Some aspects of polya tree distributions for statistical modelling. *Annals of Statistics*, **20**, 1222–1235.

— (1994) More aspects of polya tree distributions for statistical modelling. *Annals of Statistics*, **22**, 1161–1176.

Lichtman, S. M., Ratain, M. J., Echo, D. A., Rosner, G., Egorin, M. J., Budman, D. R., Vogelzang, N. J., Norton, L. and Schilsky, R. L. (1993) Phase I trial and granulocyte-macrophage colony-stimulating factor plus high-dose cyclophosphamide given every 2 weeks: a Cancer and Leukemia Group B study. *Journal of the National Cancer Institute*, **85**, 1319–1326.

Lopes, H. F., Muller, P. and Rosner, G. L. (2003) Bayesian Meta-Analysis for Longitudinal Data Models using Multivariate Mixture Priors. *Biometrics*, **59**, 66–75.

MacEachern, S. and Müller, P. (1998) Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 223–239.

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

MacEachern, S. N. (1999) Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, 50–55.

Mallet, A., Mentré, F., Gilles, J., Kelman, A., Thomson, A., S.M., Bryson and Whiting, B. (1988) Handling covariates in population pharmacokinetics with an application to gentamicin. *Biomedical Measurement Informatics and Control*, **2**, 138–146.

Mallick, B. K. and Walker, S. G. (1997) Combining Information from Several Experiments with Nonparametric Priors. *Biometrika*, **84**, 697–706.

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

Muliere, P. and Petrone, S. (1993) A Bayesian predictive approach to sequential search for an optimal dose: Parametric and nonparametric models. *Journal of the Italian Statistical Society*, **2**, 349–364.

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

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

— (1998) Semiparametric pk/pd models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds. D. Dey, P. Müller and D. Sinha), 323–337. New York, NY, USA.

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

O'Hagan, A. (1992) Some bayesian numerical analysis (with discussion). In *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 345–363. Oxford: Oxford University Press.

Quintana, F. A. (1998) Nonparametric bayesian analysis for assessing homogeneity in $k \times l$ contingency tables with fixed right margin totals. *Journal of the American Statistical Association*, **93**, 1140–1149.

Sethuraman, J. (1994) A constructive defnition of the dirichlet process prior. *Statistica Sinica*, **2**, 639–650.

Spencer, C. M. and Goa, K. L. (1995) Amifostine. A review of its pharmacodynamic and pharmacokinetic properties, and therapeutic potential as a radioprotector and cytotoxic chemoprotector. *Drugs*, **91**, 1001–1031.

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

Tomlinson, G. and Escobar, M. (1999) Analysis of densities. *Tech. rep.*, University of Toronto.

Walker, S. and Damien, P. (1998) Sampling methods for bayesian nonparametric inference involving stochastic processes. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds. D. Dey, P. Müller and D. Sinha), 243–254. New York, NY, USA.

Walker, S., Damien, P., Laud, P. and Smith, A. (1999) Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society, Series B (Methodological)*, **61**, 331–344.

Walker, S. and Muliere, P. (1997) Beta-stacy processes and a generalisation of the polya-urn scheme. *Annals of Statistics*, **25**, 1762–1780.

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