Biometrics 59, 66–75
March 2003

Bayesian Meta-analysis for Longitudinal Data Models Using Multivariate Mixture Priors

Hedibert Freitas Lopes,[1,*] Peter Müller,[2] and Gary L. Rosner[3]

[1]Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Caixa Postal 68530 - 21945-970, Rio de Janeiro, RJ, Brazil
[2]The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.
[3]The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.
[*]*email:* hedibert@im.ufrj.br

Summary. We propose a class of longitudinal data models with random effects that generalizes currently used models in two important ways. First, the random-effects model is a flexible mixture of multivariate normals, accommodating population heterogeneity, outliers, and nonlinearity in the regression on subject-specific covariates. Second, the model includes a hierarchical extension to allow for meta-analysis over related studies. The random-effects distributions are decomposed into one part that is common across all related studies (common measure), and one part that is specific to each study and that captures the variability intrinsic between patients within the same study. Both the common measure and the study-specific measures are parameterized as mixture-of-normals models. We carry out inference using reversible jump posterior simulation to allow a random number of terms in the mixtures. The sampler takes advantage of the small number of entertained models. The motivating application is the analysis of two studies carried out by the Cancer and Leukemia Group B (CALGB). In both studies, we record for each patient white blood cell counts (WBC) over time to characterize the toxic effects of treatment. The WBCs are modeled through a nonlinear hierarchical model that gathers the information from both studies.

Key words: Markov chain Monte Carlo; Mixture model; Model averaging; Model selection; Pharmacodynamic models; Reversible jump.

1. Introduction

We analyze longitudinal data from population pharmacodynamic studies. The data come from two related studies. Two important features of the data are heterogeneous populations, and relatively small sample sizes in each study. Both are typical features that arise in anticancer drug development. Quite often, multiple studies at different institutions examine the same or similar drugs or drug combinations in small, early-phase clinical studies that enroll cancer patients who differ greatly in disease and personal characteristics. It would strengthen inference if one could combine information across studies in a way that allowed for the high degree of heterogeneity within and between the studies.

We introduce a class of flexible random-effects models to accommodate this heterogeneity and propose a hierarchical extension to allow for borrowing strength across the two studies. From a data analysis point of view, the proposed approach provides two important advantages over separate, independent modeling for each study. First, we can make inference in the more recent study more precise by including data from the earlier study. Second, we can make predictions for a future third study by incorporating between-study and within-study heterogeneity in the overarching model. This is critical for study design and planning.

Models for Bayesian inference in longitudinal data models with random effects are reviewed, for example, in Wakefield, Aarons, and Racine-Poon (1998), with a focus on population pharmacokinetic and pharmacodynamic (PK/PD) studies similar to the application in this paper. Let y_{ij} denote the jth measurement on the ith patient, let $\boldsymbol{\theta}_i$ denote a random-effects vector for patient i, and let $\boldsymbol{x}_i$ denote patient-specific covariates, including treatment dose. The usual structure of population PK/PD models is

$$p(y_{ij} \mid \boldsymbol{\theta}_i), \quad p(\boldsymbol{\theta}_i \mid \boldsymbol{x}_i, \boldsymbol{\phi}), \quad p(\boldsymbol{\phi}). \tag{1}$$

Here $p(y_{ij} \mid \boldsymbol{\theta}_i)$ is typically a parameteric nonlinear regression for expected response over time. For example, $\boldsymbol{\theta}_i$ could be the parameters in a compartmental model for drug concentrations. The second level of the model specifies the prior distribution for the random-effects vectors $\boldsymbol{\theta}_i$, possibly including a regression on patient-specific covariates $\boldsymbol{x}_i$, with $\boldsymbol{\phi}$ denoting the hyperparameters. Bayesian models similar to (1) have been considered in Zeger and Karim (1991) for generalized linear mixed models, in Wakefield et al. (1994) for the general population model assuming a multivariate normal population

distribution, in Dellaportas and Smith (1993), and in Wakefield (1996) with multivariate t prior distributions.

Heterogeneity in the patient population, outliers, and overdispersion make a strict parametric model for the population distribution $p(\boldsymbol{\theta}_i \mid \boldsymbol{x}_i, \boldsymbol{\phi})$ unreasonable in our application. Instead, we consider an essentially nonparametric extension. In maximum likelihood–based inference, popular nonparametric extensions to the population model (1) are the Nonparametric Maximum Likelihood (NPML) method (Mallet, 1986), with no restrictions on the distribution of the random-effects in the model and yielding a discrete estimate of this distribution; and the Semi Nonparametric (SNP) approach (Davidian and Gallant, 1993), a method that assumes a smooth density for the random-effects distribution. Bayesian approaches to nonparametric extensions are described in Rosner and Müller (1997), Müller and Rosner (1997), Kleinman and Ibrahim (1998), and Walker and Wakefield (1998). Walker and Wakefield (1998) use Dirichlet process priors, the other references use Dirichlet process mixtures. In this article, we propose an alternative approach based on finite mixture-of-normals models. A similar idea is proposed by Magder and Zeger (1996), where the mixing distributions in linear mixed-effects models are estimated as mixtures of Gaussians. Unrelated with the application in population models, Green and Richardson (2001) argue for the use of finite mixture-of-normals models in place of Dirichlet process mixtures, citing issues of computational efficiency, flexibility of prior specifications, and interpretability.

Another important extension of the basic model (1) is to allow joint analysis of several related studies. Such meta-analysis is a popular theme in statistical inference, often modeled with a hierarchy as in (1). But there is little work on hierarchical models relating nonparametric models. One approach is discussed by Müller, Quintana, and Rosner (1999), who extend Dirichlet process mixtures to allow for hierarchical extensions suitable for meta-analysis with nonparametric submodels. In contrast, in this article we propose finite mixtures of normals with an emphasis on interpretation and parsimony.

The article is organized as follows. In Section 2, we introduce the motivating application. The random-effects model based on finite mixtures of multivariate normals is presented and discussed in Section 3. We propose a hierarchical model across related studies to allow for meta-analysis. In Section 3.4, we extend the model to include uncertainty about the number of terms in the finite mixtures. Section 4 presents results, and Section 5 concludes with a discussion.

2. Data

We analyze 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 anticancer agent cyclophosphamide one could give cancer patients every two weeks. Patients also received the drug GM-CSF to help reduce the ill effects of cyclophosphamide on the patients' marrow. The other study, CALGB 9160, was built upon the experience gained in 8881, and included an additional treatment, amifostine. The drug amifostine had been shown in some studies to reduce some of the toxic side effects of anticancer therapies, such as cyclophosphamide and radiation therapy (Spencer and Goa, 1995). A common toxicity of cancer therapy is myelosuppression, in which the immune system is suppressed by killing cells involved in immune functions. The objective of CALGB 9160 was to determine if adding amifostine would reduce the myelosuppressive side effects of aggressive chemotherapy with cyclophosphamide and GM-CSF. CALGB 9160 randomized patients to receive or not receive amifostine, along with cyclophosphamide (3 grams per square meter of body surface area) and GM-CSF (5 micrograms per kilogram of body weight). CALGB 9160 studied mainly the effect of amifostine on various measures of hematologic toxicity such as nadir (i.e., minimum) blood cell counts, or days of life-threatening myelosuppression (see Rosner and Müller (1997) for more details about the design of this experiment). 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 particular, the investigators were interested in gaining a better understanding of how clinically relevant characteristics of the time course of the blood counts change in response to the treatment. Rather than relying solely on the observed blood counts that were sampled only two—or sometimes three—times a week for patients in CALGB 9160, we can make inference more precise by incorporating information learned from the previous study. We therefore set about combining the data from CALGB 9160 with data available from CALGB 8881—to model the entire time course of myelosuppression (reduced blood counts). The enrolled patient population in each study was diverse in terms of baseline characteristics, and the analysis of the combined data has to accommodate this heterogeneity—both within and between the two studies.

An important aspect of the desired analysis is inferring about patients in the population at large, i.e., those who might be in a future third study. The hierarchical meta-analysis of the related two studies is critical for deriving such an inference. Without the hierarchical model, there could be no learning about which features of the data are study-specific and which features are common to the population at large.

Let $K = 2$ be the number of studies under consideration, and n_k be the number of patients in study k. In study 8881, we have data on $n_1 = 52$ patients. The second study, CALGB 9160, includes data on $n_2 = 46$ patients. In both studies, the main outcome measure was the white blood cell count (WBC) for each patient over time. We will use y_{kij} to denote the jth blood count measurement on the ith patient in study k on day t_{kij}, recorded on a log scale of thousands, i.e., $y_{kij} = \log(\text{WBC}/1000)$. In CALGB 8881 and 9160, we had a total of 674 and 706 observations, respectively. The number of observations for one patient varies between 2 and 19. Some patients' measurements are shown in Figure 1. Rosner and Müller (1997) used a nonlinear regression model to fit the data:

$$y_{kij} \sim N\left\{f(\boldsymbol{\theta}_{ki}, t_{kij}), \sigma^2\right\}. \tag{2}$$

The vector $\boldsymbol{\theta} = (z_1, z_2, z_3, \tau_1, \tau_2, \beta_0, \beta_1)$ parameterizes a mean function $f(\boldsymbol{\theta}, t)$ that is defined piecewise as (i) a horizontal line for $0 \le t < \tau_1$; (ii) a straight line connecting parts (i) and

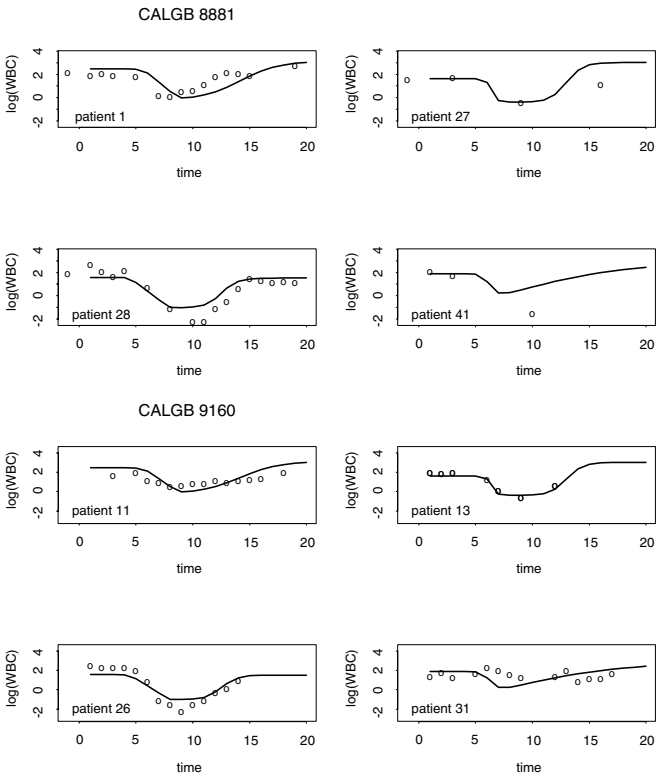

Figure 1. Observed white blood cell count (log(WBC/1000)) for some typical patients from studies CALGB 8881 and CALGB 9160. The solid line represents the posterior means for $f(\boldsymbol{\theta}_{ki}, t)$ for $t = 1, 2, \ldots, 20$ under the "best model" (see Table 1 from Section 4).

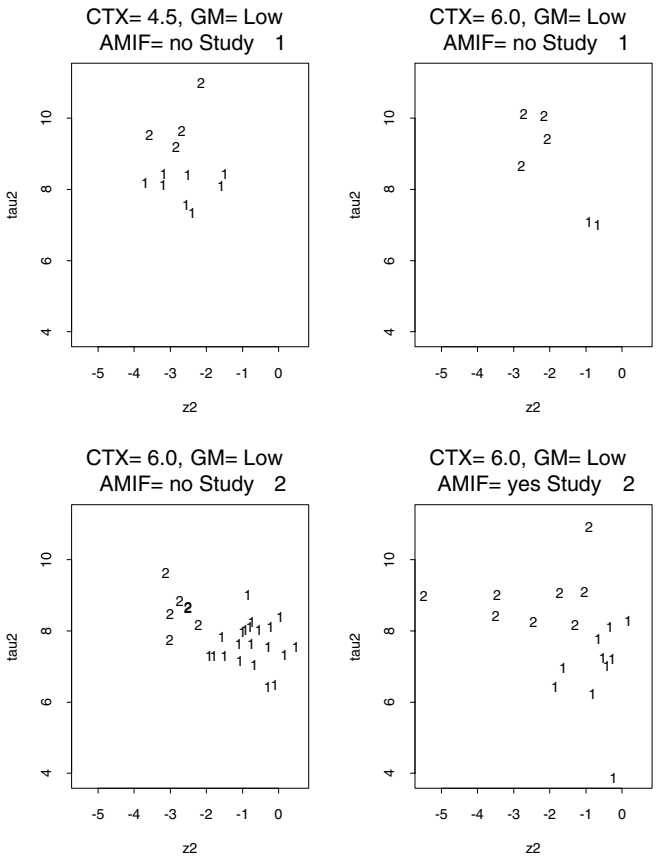

Figure 2. Scatterplot of the maximum likelihood estimators of z_2 and τ_2 for certain combinations of CTX, GM-CSF and amifostine from both studies. See equation (3) for the definition of the parameters. Some pairs are identical, which causes some hidden points.

(iii) ($\tau_1 \le t < \tau_2$); and (iii) a shifted and scaled logistic curve ($t \ge \tau_2$):

$$f(\boldsymbol{\theta}, t) = \begin{cases} z_1 & t < \tau_1 \\ rz_1 + (1-r)g(\boldsymbol{\theta}, \tau_2) & \tau_1 \le t < \tau_2 \\ g(\boldsymbol{\theta}, t) & t \ge \tau_2, \end{cases} \qquad (3)$$

where $r = (\tau_2 - t)/(\tau_2 - \tau_1)$ and $g(\boldsymbol{\theta}, t) = z_2 + z_3/[1 + \exp\{-\beta_0 - \beta_1(t - \tau_2)\}]$. The horizontal line (i) represents the initial baseline count; the steep decline (ii) corresponds to the sudden drop of WBC during chemotherapy; and (iii) models an S-shaped recovery.

In the next level of a probability model for the WBC data, we need to assume some population distribution $p(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}_i, \boldsymbol{\phi})$ for patient-specific random effects. The model should include a regression on patient-specific covariates $\boldsymbol{x}_i$. Here, $\boldsymbol{\phi}$ generically denotes the hyperparameters of such a model. To explore suitable modeling approaches for the population distribution, we start with some exploratory data analysis. In Figure 2, we plot maximum likelihood estimates (m.l.e.) $\hat{\boldsymbol{\theta}}_{ki}$ for the random effects $\boldsymbol{\theta}_{ki}$ for all 98 patients, arranged by CTX and GM. The m.l.e.'s were computed for each patient separately using (2) only, without any hierarchical modeling across patients. A simple multivariate normal random-effects distribution $p(\boldsymbol{\theta}_i | \, \boldsymbol{x}_i, \boldsymbol{\phi})$ is inappropriate, since there seem to be subgroups of patients, within each study, with distinct patterns—as Figure 2 suggests. We propose a model with mixture-of-normals random-effects distributions. To informally investigate how such mixture models might accommodate the observed heterogeneity, we applied a cluster analysis to the maximum likelihood estimates $\hat{\boldsymbol{\theta}}_{ki}$. The resulting clusters are indicated by different plotting symbols in Figure 2.

3. The Random-Effects Model

3.1 *A Mixture-of-Normals Model*

We start by describing the random-effects model for just one study, i.e., assuming $K = 1$. To simplify notation, we will drop the k subindex until we discuss extension to $K > 1$. Also, we shall first consider a prior distribution without a regression on covariates $\boldsymbol{x}_i$.

First, we generalize a multivariate normal prior $p(\boldsymbol{\theta}_i \,|\, \boldsymbol{\phi})$ to a mixture-of-normals model $p(\boldsymbol{\theta}_i \,|\, \boldsymbol{\phi}) = \sum_{k=1}^{L} \pi_l N(\boldsymbol{\mu} + \boldsymbol{d}_l, \boldsymbol{S})$. The mixture is parameterized by $\boldsymbol{\phi} = (\pi_l, \boldsymbol{\mu}, \boldsymbol{d}_l \; l = 1, \ldots, L)$, including an overall location parameter $\boldsymbol{\mu}$ and offsets $\boldsymbol{d}_l$ for the individual terms, with the constraint $\boldsymbol{d}_1 = 0$. Under this parameterization, we can assume a noninformative prior for $\boldsymbol{\mu}$. The same parameterization is used, for example, in Mengersen and Robert (1995) and Roeder and Wasserman (1997). As pointed out by Celeux, Hurn, and Robert (2000),

the parameterization of mixture models can affect the convergence of the MCMC algorithm.

With sufficiently large L, the mixture-of-normals model can approximate any desired random-effects distribution (Dalal and Hall, 1983; Diaconis and Ylvisaker, 1985). As in any nonlinear, nonnormal modeling context, the specific choice of L is guided by two competing principles. Low L leads to parsimony and easier estimation. In the extreme case of a single normal distribution, $L = 1$, it reduces to the single multivariate normal prior. On the other hand, by choosing a large L, we can increase flexibility when fitting the population distribution $p(\boldsymbol{\theta}_i \mid \boldsymbol{\phi})$—at the expense of model parsimony and computational simplicity. Later, in Section 3.4, we discuss formal inference on L as a model-selection problem.

We follow an approach used, for example, in Mallet et al. (1988) and Müller and Rosner (1998) to include the desired regression on covariates $\boldsymbol{x}_i$. We augment the prior mixture-of-normals model to a probability model in $(\boldsymbol{\theta}_i,\ \boldsymbol{x}_i)$ jointly, i.e.,

$$p(\boldsymbol{\theta}_i, \boldsymbol{x}_i \mid \boldsymbol{\phi}) = \sum_{l=1}^{L} \pi_l N(\boldsymbol{\mu} + \boldsymbol{d}_l, \boldsymbol{S}). \tag{4}$$

The implied conditional distribution $p(\boldsymbol{\theta}_i \mid \boldsymbol{x}_i,\ \boldsymbol{\phi})$ formalizes the desired regression, and takes the form of a locally weighted mixture of linear regressions.

We now extend the mixture-of-normals random-effects model to a hierarchical model across related studies to allow the desired meta-analysis. The construction of this hierarchical extension is driven by the following considerations. First, the model should include, on the one hand, the extreme cases of one common random-effects distribution across all studies, and, on the other hand, random-effects distributions for each study that are conditionally independent across studies (given the hyperparameters). Second, the hierarchical extension should not unreasonably complicate posterior simulation. Third, the hierarchical model should allow interpretation of the additional parameters required in the hierarchical extension. Based on these considerations, we propose a model

$$\begin{aligned} p(\boldsymbol{\theta}_{ki}, \boldsymbol{x}_{ki} \mid \boldsymbol{\phi}) &= \varepsilon p_c(\boldsymbol{\theta}_{ki}, \boldsymbol{x}_{ki} \mid \boldsymbol{\phi}_0) \\ &\quad + (1-\varepsilon) p_k(\boldsymbol{\theta}_{ki}, \boldsymbol{x}_{ki} \mid \boldsymbol{\phi}_k), \end{aligned} \tag{5}$$

where p_c represents a commom measure shared among all studies and p_k, for $k = 1, \ldots, K$, is a study-specific measure. The vector of hyperparameters $\boldsymbol{\phi}$ is split into subvectors $\boldsymbol{\phi}_0$ and $\boldsymbol{\phi}_k$, $k = 1, \ldots, K$. The additional mixing parameter ε determines the amount of borrowing strength across the related studies. We shall refer to p_c as the common measure, and p_k as the idiosyncratic measure. Figure 3 illustrates this split of the random-effects distributions for the K studies. By assuming a prior distribution $p(\varepsilon)$, with support including point masses at 0 and 1, one can include the special cases of one common random-effects distribution and conditionally independent random-effects distributions, respectively.

For p_k and p_c we assume mixtures of multivariate normal models, as in (6)

$$\begin{aligned} p_c(\boldsymbol{\theta}_{ki}, \boldsymbol{x}_{ki} \mid \boldsymbol{\phi}_0) &= \sum_{l=1}^{L_1} \pi_l N(\boldsymbol{\mu} + \boldsymbol{d}_l, \boldsymbol{S}), \\ p_k(\boldsymbol{\theta}_{ki}, \boldsymbol{x}_{ki} \mid \boldsymbol{\phi}_k) &= \sum_{l=1}^{L_2} \pi_{kl} N(\boldsymbol{\mu} + \boldsymbol{d}_{kl}, \boldsymbol{S}), \end{aligned} \tag{6}$$

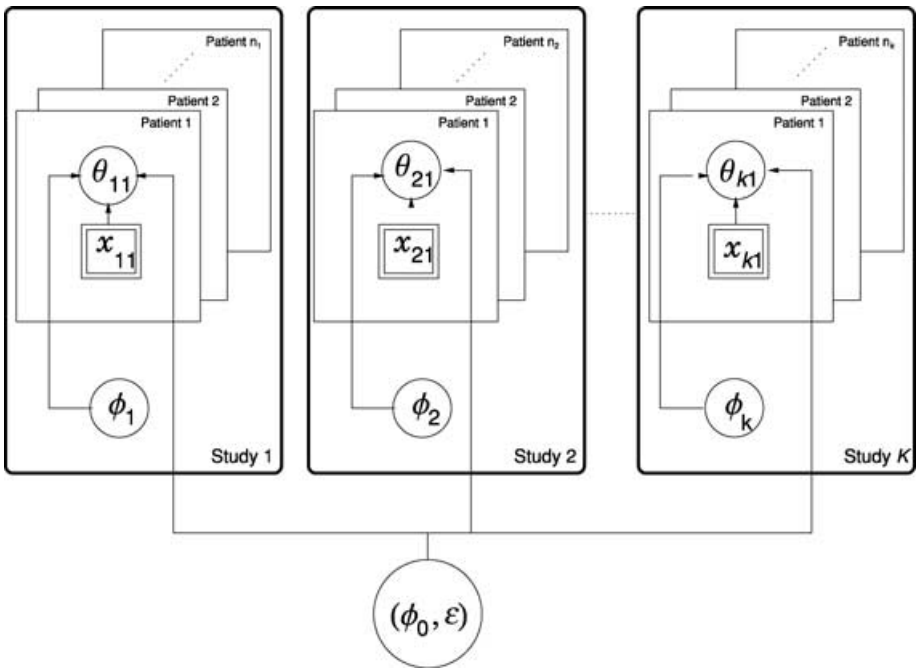

Figure 3. Graphical representation of the hierarchical model across related studies in equation (5). The vector of random-effects, $\boldsymbol{\theta}_{ki}$, parameterizes the mean profile for the ith patient from the kth study; $\boldsymbol{x}_{ki}$ are the patient's covariates; $\boldsymbol{\phi}_k$ and $\boldsymbol{\phi}_c$ are study-specific and common hyperparameters (see equation 2).

with possibly different numbers (L_1 and L_2) of components in the mixtures. For the moment, we assume fixed-size mixtures, i.e., L_1 and L_2 are fixed hyperparameters. Later, in Section 3.4, we shall describe how to extend the model to random-size mixtures. Also, $\boldsymbol{\mu}$ represents the overall mean, while $\boldsymbol{d}_l$ and $\boldsymbol{d}_{kl}$ are deviations from $\boldsymbol{\mu}$. As before, $\boldsymbol{d}_1$ is set to 0, and the variance $\boldsymbol{S}$ is assumed equal across terms of the mixtures. Alternatively, one could assume, without complicating the analysis, covariance matrices $\boldsymbol{S}_l$ and $\boldsymbol{S}_{kl}$ specific to each term in the mixture. As in many mixture models, the tradeoff is between a smaller number of terms in the mixture versus simplicity.

3.2 *Identification Issues*

Using a mixture-of-normals model to represent the unknown measures $p_c(\cdot)$ and $p_k(\cdot)$, we face identification problems common to any mixture model. To avoid lack of likelihood identifiability due to arbitrary permutations of indices, we impose an order constraint on the weights π_l and π_{kl}. But there are at least two more sources of possible identifiability concerns. One could move one term $N(\boldsymbol{\mu} + \boldsymbol{d}_l, \boldsymbol{S})$ from p_c into each of the study-specific measures p_k by defining $\boldsymbol{d}_{k,L_2+1} = \boldsymbol{d}_l$, incrementing L_2 by one, removing the lth term from the mixture for p_c, and adjusting the weights accordingly to leave $p(\cdot \,|\, \boldsymbol{\phi})$ unchanged. The likelihood would remain invariant under such reparameterization. The problem could be mitigated by choosing an informative prior for L_2 that favors smaller-size idiosyncratic mixtures. Also, by replacing one term in p_c by K copies in each of the idiosyncratic measures p_k, we substantially increase the model complexity. Posterior inference intrinsically favors the more parsimonious alternative, i.e., putting the common probability mass into p_c, as desired. This is known as Occam's razor principle (Jefferys and Berger, 1992). Now consider the reverse, i.e., moving a term from p_k into the common measure p_c, and adjusting the weights accordingly to keep $p(\boldsymbol{\theta}_{ki}, \boldsymbol{x}_{ki} \,|\, \boldsymbol{\phi}_k)$ unchanged. However, moving terms in this way, from the idiosyncratic model for study k into the common measure p_c, would change the random-effects distributions $p(\boldsymbol{\theta}_{ri}, x_{ri} \,|\, \boldsymbol{\phi}_r)$ for the other studies $r \neq k$, and thus change the overall likelihood.

To investigate the impact of these identifiability issues on inference under the proposed model, and to illustrate the general discussion above, we have set up a simulation study. We assume two studies, $k = 1, 2$, with 50 patients each. Each patient was monitored daily for two weeks. In a first simulation, $\boldsymbol{\theta}_{ki}$ were chosen such that the common and idiosyncratic components were well separated. In a second simulation, the subpopulations are less clearly separated. In both cases, we fixed $L_1 = L_2 = 2$.

We summarize some of the simulation results. In both simulations, posterior inference in the proposed model identified the true structure. This remained true when we overparameterized with $L_1 = L_2 = 3$. The extra terms in the mixture were fitted with negligible weight. In particular, we did not find in the common measure any duplication of terms of the idiosyncratic measure. We computed conditional predictive ordinates (CPO) (Pettit and Young, 1990) that were favorable to the simpler structure.

Finally, when the focus is on predictive inference, as is the case in this application, the relevance of identifiabilty concerns is limited to practical issues related to efficiency of the MCMC simulation, and general understanding of the model structure. See, for example, Raftery, Madigan, and Hoeting (1997) and Hoeting et al. (1999) for a related discussion.

3.3 *Implementation*

With mixture models such as the proposed random-effects model, a common device in posterior simulation is the introduction of indicator variables to break the mixture (Diebolt and Robert, 1994). Consider indicators $w_{ki} = (j, l)$ to break the mixture in (5) and (6). In words, j indicates whether the random effect is sampled from either the common ($j = 0$) or the idiosyncratic part ($j = 1$) of the random-effects distribution, and l indicates the term of the respective mixtures of normals. Therefore,

$$Pr\{w_{ki} = (0, l)\} = \varepsilon\pi_l \text{ and } Pr\{w_{ki} = (1, l)\} = (1 - \varepsilon)\pi_{kl}.$$

Conditional on the indicators w_{ki},

$$p(\boldsymbol{\theta}_{ki}, \boldsymbol{x}_{ki} \,|\, w_{ki} = (j, l)) = \begin{cases} N(\boldsymbol{\mu} + \boldsymbol{d}_l, \boldsymbol{S}) & \text{if } j = 0, \\ N(\boldsymbol{\mu} + \boldsymbol{d}_{kl}, \boldsymbol{S}) & \text{if } j = 1. \end{cases}$$

We complete the model with prior distributions on σ^2, ε, $\boldsymbol{\pi}$, $\boldsymbol{\pi}_k$, $\boldsymbol{S}$, $\boldsymbol{\mu}$, $\boldsymbol{d}_l$, $\boldsymbol{d}_{kl}$. For σ^2, we assume a conditionally conjugate inverse gamma distribution: $\sigma^2 \sim IG(\alpha_0/2,\, \beta_0/2)$, with fixed hyperparameters α_0 and β_0. The prior on $\boldsymbol{S}^{-1}$ is a Wishart distribution, with hyperparameters ν_0 and $\boldsymbol{S}_0^{-1}$, i.e., $\boldsymbol{S}^{-1} \sim W(\nu_0,\, \nu_0^{-1}\boldsymbol{S}_0^{-1})$. For ε, we use a beta prior, $\varepsilon \sim Beta(a_0,\, b_0)$. Alternatively, we can include prior point mass at 0 and 1, without significantly complicating posterior simulations. For $\boldsymbol{\mu}$ and $\boldsymbol{d}$ we choose conjugate multivariate normal priors, $\boldsymbol{d}_{kl} \sim N(\tilde{\boldsymbol{d}}, \tilde{\boldsymbol{V}})$, $\boldsymbol{d}_l \sim N(\tilde{\boldsymbol{d}}, \tilde{\boldsymbol{V}})$, and $\boldsymbol{\mu} \sim N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mathrm{V}}})$.

Typically, the hypermean $\tilde{\boldsymbol{d}}$ will be zero. Finally, the prior distributions for $\boldsymbol{\pi}$ are Dirichlet: $(\pi_{k1}, \ldots, \pi_{kL_2}) \sim Dir(\boldsymbol{\alpha}_k)$, for $k = 1, \ldots, K$, and $(\pi_1, \ldots, \pi_{L_1}) \sim Dir(\boldsymbol{\alpha})$, with fixed hyperparameters $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kL_2})$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{L_1})$. Lopes (2000) describes in full detail the Markov chain Monte Carlo algorithm designed for simulating from the full posterior and predictive distributions.

3.4 *Model Selection for* L_1 and L_2

We now allow for uncertainty about L_1 and L_2. There are two main aspects to model uncertainty. The first is related to model selection. In many situations, even if no particular model is thought to be the true one, it is convenient to select one for scientific reporting. On the other hand, in many applications there are quantities of interest—such as predictive inference for future patients—that do not depend on a particular model and that might be averaged across models. General issues of model averaging and model selection are discussed in a recent overview and tutorial on Bayesian model averaging by Hoeting et al. (1999).

Let $\mathcal{M} = \{1, 2, \ldots, M\}$ denote the set of indices representing all models under consideration. In our context $\mathcal{M}$ is the set of all combinations of L_1 and L_2. Assume that Δ is an outcome of interest, such as the future profile of a new patient from the population, or the time at which the white blood cell count of a new patient drops below a critical threshold. Let $\boldsymbol{\theta}_m$ denote the parameter vector under model m. The posterior

distribution for Δ is $p(\Delta \mid D) = \sum_{m=1}^{I} p(\Delta \mid m,\, D) Pr(m \mid D)$, where D denotes the data,

$$Pr(m \mid D) = p(D \mid m) Pr(m) \left\{ \sum_{\tilde{m}=1}^{M} p(D \mid \widetilde{m})\ Pr(\widetilde{m}) \right\}^{-1}$$

is the posterior probability of model m, $p(D \mid m) = \int p(D \mid \boldsymbol{\theta}_m,\, m) p(\boldsymbol{\theta}_m \mid m)\ d\boldsymbol{\theta}_m$ is the marginal likelihood conditional on model m, and $Pr(m)$ is the prior probability of model m. For the finite mixture-of-normals models used here, we assume a uniform prior $Pr(m) = 1/M$.

In the proposed population model, neither $p(\boldsymbol{\theta}_m \mid m,\, D)$ nor $p(\Delta \mid m,\, D)$ are analytically tractable. Additionally, $p(D \mid m)$ involves the solution of a possibly high-dimensional integral. Several alternative approaches for computing marginals $p(D \mid m)$ have been proposed in the literature. We refer to Lopes (2000) for references and further discussions.

In this article, we will use reversible-jump Markov chain Monte Carlo (RJMCMC) simulation to compute $Pr(m \mid D)$. RJMCMC is a Markov chain simulation that jumps between models of different dimensions (Green, 1995). Dellaportas, Forster, and Ntzoufras (2002) and Godsill (1998) have shown the relationship between alternative methods of model selection. The frequency with which the algorithm visits each particular model approximates $Pr(m \mid D)$. Green (1995) and Richardson and Green (1997) propose RJMCMC algorithms for general mixture-of-normals models. However, these general algorithms are difficult to apply in the high-dimensional context of our application. Instead, we suggest an RJMCMC scheme that exploits the fact that we have only a few competing models to consider. Before starting the RJMCMC, we run a posterior MCMC simulation separately, for each model $m \in \mathcal{M}$, and save the simulated posterior draws. Denote with Θ_m the set of Monte Carlo simulations from the posterior under model m. At each iteration of our RJMCMC scheme, we then draw from the posterior distributions of the current model m and the proposed model $\tilde{m}$, using the saved Monte Carlo samples Θ_m and $\Theta_{\tilde{m}}$, and calculate the usual Metropolis-Hastings acceptance probability. The imputed parameter vectors for all other models are kept unchanged. See the Appendix for a detailed description of the algorithm.

The proposed RJMCMC avoids computing predictive distributions, which, in our case, are analytically intractable. There are many alternative ways to propose the jumps between the models. The main motivation for using this particular algorithm is the easy access to posterior draws from each competing model, and the difficulty of generating proposals $\tilde{\theta}_{\tilde{m}}$ in high-dimensional mixture models like (5). Further references and details can be found in Lopes (2000).

4. Results

We used the proposed model (2) with the hierarchical mixture-of-normals priors (5) and (6) to analyze the data described in Section 2.

As initial values for the MCMC simulation, we used m.l.e. fits to the nonlinear regression (2) for each patient separately. Hyperparameters were set at $\alpha_0 = 4.25$, $\beta_0 = 1.125$ ($E(\sigma^2) = 0.5$ and $V(\sigma^2) = 2$), $\nu_0 = 12$, $\boldsymbol{\phi}_0 = \mathbf{1}$, $\tilde{\boldsymbol{\mu}} = (2, -1.5, 4.5, 5, 8, -2, 0.5, -0.9, -2, -0.5)'$, $\tilde{\boldsymbol{d}} = (0, -0.5, 0, 0, 0, 0, 0, -1, 0, 0)$, and $\tilde{\boldsymbol{V}} = 5\boldsymbol{I}$. These prior hyperparameters were chosen to reflect fairly uninformative—but still proper—priors centered at reasonable parameter values. The Dirichlet parameters α_l and α_{kl} were set to 1.0. We studied four combinations of L_1 and L_2, and used a discrete uniform prior distribution on the models (see Table 1).

Table 1

Posterior model probabilities, Bayes factors, and pseudo-Bayes factors relative to model M_0 ($L_1 = 0$, $L_2 = 1$). The posterior model probabilities and the Bayes factors ($BF_{i0} = p(D \mid M_i)/p(D \mid M_0)$) were computed based on the proposed RJMCMC algorithm. The pseudo-Bayes factors were computed from conditional predictive ordinates. M_0 is the default model using a multivariate random-effects distribution for each study, with conditional independence across studies. Posterior probabilities are based on $Pr(M_i) = 0.2, i = 0, 1, \ldots, 4$.

Model	L_1	L_2	$Pr(M_i \mid \boldsymbol{y})$	BF_{i0}	PBF_{i0}
M_1	1	1	0.125	1.000	90.1
M_2	1	2	0.141	1.128	292.3
M_3	2	1	0.172	1.367	338.8
M_4	2	2	0.437	3.496	333.8

All reported inferences are based on 10,000 MCMC iterations—beyond a burn-in of 100,000 iterations, and saving only every 10th iteration. The credible intervals for posterior predictive profiles are based on a subsample of size 500, taken from the final 10,000 MCMC samples. Experimenting with different choices, we found similar results over a wide range of MCMC tuning parameters, including, for example, an implementation with a burn-in of 10,000 iterations and not discarding any draws.

We are entertaining four possible models with $1 \leq L_1$, $L_2 \leq 2$ (Table 1). In a preliminary analysis, we let L_1 (and L_2) increase to 4, but the change in the predictive inference was negligible. As starting values for the latent indicator variables, w_{ki}, we applied techniques for model-based cluster analysis (Fraley and Raftery, 1998) to the set of initial values for $\boldsymbol{\theta}_{ki}$, which, in turn, were obtained by maximum likelihood estimation. We found virtually no difference in the final results when we used other, simpler, choices of starting values. Nevertheless, we recommend the described initialization, to avoid the possible danger of the MCMC simulation getting trapped in local modes, and because the involved exploratory data analysis might lead to insights that are of value independent of their use in the initialization.

Table 1 compares the four competing models. The summaries were computed using the proposed RJMCMC and are relative to a default model M_0, defined by a (single) multivariate normal random-effects distribution and conditional independence across studies. Although the RJMCMC points to a particular model, $L_1 = L_2 = 2$ with posterior probability 0.44, it also reports significant posterior probabilities for other models. Those probabilities will later be used for model averaging. A trace plot of model indicators imputed in the RJMCMC simulation (not shown) suggests good mixing over the model space, and indicates no problems with practical convergence. Table 1 also presents pseudo-Bayes factors (PBF) (Gelfand, 1996), based on conditional predictive ordinate (CPO) (Pettit and Young, 1990).

We now consider inference in the a posteriori most likely model, $L_1 = L_2 = 2$. The posterior means for the weights $\{\varepsilon\pi_1, \varepsilon\pi_2, (1-\varepsilon)\pi_{11}, (1-\varepsilon)\pi_{12}, (1-\varepsilon)\pi_{21}, (1-\varepsilon)\pi_{22}\}$ were found to be (0.42, 0.07, 0.43, 0.09, 0.48, 0.04). The random-effects distributions are split into approximately equal parts for the common measure p_c, and the study-specific measures p_k. This is consistent with what might be vaguely expected from inspection of Figure 2. Each measure, p_c, along with the p_k, is dominated by the first term, with a second term adding a minor correction to allow for outliers and lack of normality. The following plots illustrate inference about the high-dimensional random-effects distributions. For each simulated draw from the posterior distribution, we consider the corresponding curves $f(\boldsymbol{\theta} = \boldsymbol{\mu} + \boldsymbol{d}_l,\ t)$ and $f(\boldsymbol{\theta} = \boldsymbol{\mu} + \boldsymbol{d}_{kl},\ t)$ as functions of time t. Figure 4 shows posterior means and percentiles for $f(\boldsymbol{\theta},\ t)$ for each t. That is, we summarize the posterior distributions on $\boldsymbol{\mu}$, $\boldsymbol{d}_l$, and $\boldsymbol{d}_{kl}$, by showing the implied posterior quantiles on the corresponding curves. The lower the weight allocated to a component, the wider are the credible intervals—indicating that less information is present in such components. The predictive profile for the second idiosyncratic component ($l = 2$), from study $k = 1$, is considerably different from the other components—indicating the presence of some patients whose measurements are not in agreement with the rest of the patients in both studies.

Bayesian model averaging is presented in Table 2. CTX and GM doses were fixed at 3.0 mg/m^2 and 5.0 μg/kg, respectively. The entries report posterior predictive white blood cell counts for a new patient taken from study $k = 1$ (CALGB 8881), from study $k = 2$ (CALGB 9160), and from the population, respectively. We obtain posterior predictive inference for a patient from the population by considering a future third study, $k = 3$, with $n_3 = 0$. The right-hand columns of Table 2 show posterior inference after Bayesian model averaging, marginalizing across all four entertained models. Wider credibility intervals were obtained (not shown) by allowing for model uncertainty in the number of mixture components.

Figure 5 compares posterior predictive profiles for a new patient from study CALGB 9160, based on two competing models. The levels of CTX and GM-CSF were fixed at 3.0 mg/m^2 and 5.0 μg/kg, respectively. One model is the proposed mixture-of-normals hierarchical model. The alternative model assumes that the random-effects distributions $p(\boldsymbol{\theta}_{ki}, \boldsymbol{x}_{ki} \mid \boldsymbol{\phi})$ are independent given the hyperparameters $\boldsymbol{\phi}$, i.e., there is no borrowing strength across studies—beyond learning about the common hyperparameters.

Up to five days into the treatment, both models give similar results. The reason for the similarity is that patient responses in both studies are similar over the first few days.

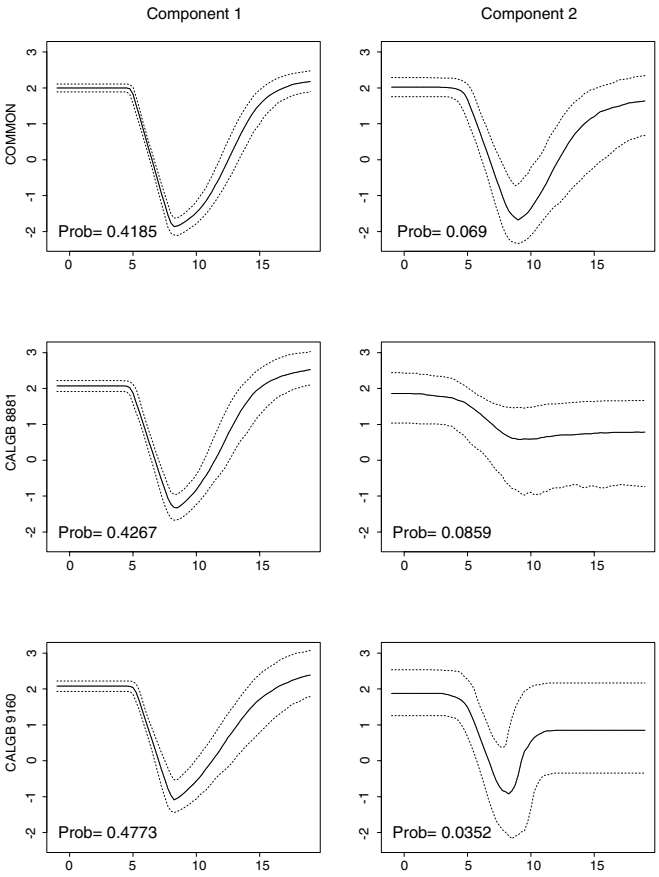

Figure 4. Posterior means, 5th and 95th percentiles for WBC profiles (as log WBC/1000) corresponding to the individual terms in p_c and p_k, $k = 1, 2$. The two panels in the top row correspond to the two terms in p_c. The panels show profiles corresponding to a $N(\boldsymbol{\mu} + \boldsymbol{d}_l,\ \boldsymbol{S})$ random-effects distribution, with $L = 1$ (left) and $L = 2$ (right). The second and third row show profiles corresponding to an $N(\boldsymbol{\mu} + \boldsymbol{d}_{kl},\ \boldsymbol{S})$ random-effects distribution with $k = 1$ (second row) and $k = 2$ (third row), $l = 1$ (left) and $l = 2$ (right), respectively.

5. Conclusion

We proposed an approach to meta-analysis over random-effects models that allows borrowing of strength across related studies. The mixture-of-normals random-effects models allow considerable flexibility, along with exploiting the computational simplicity of conjugate models. The proposed implementation includes Bayesian model averaging over mixtures of different size. We discussed a variation of RJMCMC that enables us to implement model averaging for the high-dimensional parameter vectors of the mixture model (5). Being able to combine information from related studies involving longitudinal monitoring of complex biomedical phenomena strengthens inference and allows improved study planning.

The proposed approach has some practical and methodological limitations. If the number L_1 and L_2 of terms in the mixture is allowed to vary over more than a moderately small number of possible models, the computational effort of upfront MCMC simulation for each model becomes prohibitive. This could be overcome by appropriate modifications, such as only initiating the separate MCMC simulation when a model is actually proposed, and dynamically increasing the Monte Carlo samples as and when needed. From a modeling perspective, hierarchical structure—beyond exchangeable studies—is ruled out. For example, the model does not accommodate a scenario where some subset of K related studies share more similarities than do the remaining studies. This could possibly be achieved by considering variations with study-specific

Table 2

Predictive median for a new patient's white blood cell count ($log(WBC/1000)$). For each study/model, τ_1 and τ_2 represent the time at which the patient's WBC starts to decline and recover, respectively. Also, $\tau^* = 15$ days is two weeks after start of chemotherapy. The left columns report the results for the best model, $M_4 = (L_1 = 2, L_2 = 2)$. The right columns report results for Bayesian model averaging.

Best model		Model averaging	
Days	Medians		Days
	CALGB 8881		
$\tau_1 = 3.00$	1.98	1.93	$\tau_1 = 3.75$
$\tau_2 = 9.25$	−0.59	−0.65	$\tau_2 = 8.75$
$\tau^* = 15.0$	1.14	0.46	$\tau^* = 15.0$
	CALGB 9160		
$\tau_1 = 2.75$	1.98	1.97	$\tau_1 = 3.25$
$\tau_2 = 9.25$	−0.59	−0.87	$\tau_2 = 9.00$
$\tau^* = 15.0$	1.14	0.53	$\tau^* = 15.0$
	Population		
$\tau_1 = 3.75$	1.96	2.04	$\tau_1 = 3.25$
$\tau_2 = 8.75$	−0.60	−0.88	$\tau_2 = 8.75$
$\tau^* = 15.0$	1.15	0.97	$\tau^* = 15.0$

mixture weights ε_k. Another variation occurs when related studies only share some features. For example, assume that in our application a third trial is conducted with similar but less toxic treatments. It would still be reasonable to link the random-effects distributions for some parameters representing the shape of the profiles, but other parameters—for example, parameters related to the level of toxicity—might not be shared across studies.

The discussion was in the context of meta-analysis of longitudinal data models, with the hierarchy over the random-effects distributions. The proposed methods are more generally applicable, though. The approach is relevant whenever hierarchical modeling over related random probability distributions is required.

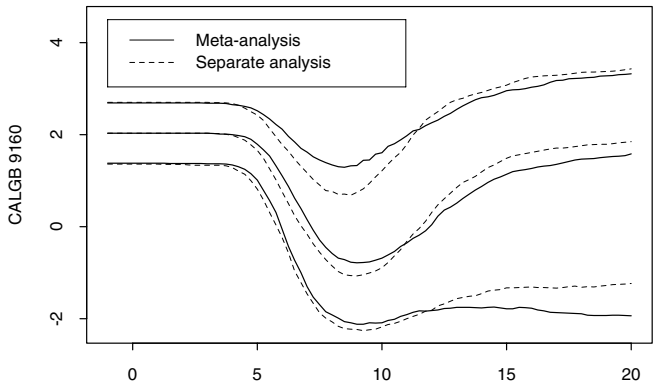

Figure 5. Predictive profile for new patients from study CALGB 9160. The solid lines represent posterior means, 5th and 95th percentiles based on our meta-analysis model ($L_1 = L_2 = 2$). The dashed lines represent the same percentiles for the alternative model where $Pr(\varepsilon = 0) = 1$ (see the last paragraph of Section 4 for further details). CTX = 3.0 mg/m^2 and GM-CSF = 5.0 μg/kg.

Acknowledgements

This research was supported, in part, by grants CA75981 and CA33601 from the National Cancer Institute, and by CAPES, Brazil, who supported the first author during his Ph.D. studies at the Institute of Statistics and Decision Sciences, Duke University, U.S.A.

Résumé

Nous proposons une classe de modèles pour données longitudinales avec effets aléatoires qui étendent de deux façons essentielles les modèles actuellement utilisés. D'abord le modèle à effets aléatoires est un mélange adaptable de modèles multinormaux, qui permet de traiter l'hétérogénéité des populations, les données atypiques et la non linéarité des régressions sur les variables spécifiques aux sujets. Ensuite le modèle inclut une extension hiérarchique qui permet la méta-analyse de plusieurs études connexes. Les distributions des effets aléatoires se décomposent en une partie commune aux différentes études (mesure commune) et une partie spécifique à chaque étude décrivant la variabilité inter sujets à l'intérieur d'une étude donnée. On paramétrise et la mesure commune et la part spécifique à l'étude à l'aide d'un mélange de modèles normaux. Les inférences sont conduites à l'aide de simulation a posteriori à saut réversible. L'échantillonneur tire parti du petit nombre de modèles mis en æuvre.

L'application à l'origine de ce travail est l'analyse de deux études conduites par le Cancer and Leukemia Group B (CALGB). Dans les deux études, on a enregistré pour chaque patient les comptages de globules blancs (WBC) au cours du temps pour caractériser les effets toxiques du traitement. Les WBC sont modélisés par un modèle non linéaire hiérarchique qui rassemble les informations des deux études.

References

Celeux, G., Hurn, M., and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* **95,** 957–970.

Dalal, S. and Hall, W. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society, Series B* **45,** 278–286.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80,** 475–488.

Dellaportas, P., Forster, J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing* **12,** 27–36.

Dellaportas, P. and Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics* **42,** 443–459.

Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion. In *Bayesian Statistics,* Volume 2, J. Bernardo, M. DeGroot, D. Lindley, and A. Smith (eds), 163–175. New York: John Wiley.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B* **56,** 163–175.

Fraley, C. and Raftery, A. (1998). *How many clusters? Which clustering method? Answers via model-based cluster*

analysis. Technical Report 329, Department of Statistics, University of Washington.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice,* W. Gilks, S. Richardson, and D. Spiegelhalter (eds). London: Chapman and Hall.

Godsill, S. (1998). *On the relationship between MCMC model uncertainty methods*. Technical Report 305, Signal Processing Laboratory, Department of Engineering, University of Cambridge.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82,** 711–732.

Green, P. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28,** 355–376.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14,** 382–417.

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

Kleinman, K. and Ibrahim, J. (1998). A semi-parametric Bayesian approach to the random effects model. *Biometrics* **54,** 921–938.

Lichtman, S., Ratain, M., Van Echo, D., Rosner, G., Egorin, M., Budman, D., Vogelzang, N., Norton, L., and Schilsky, R. (1993). Phase I trial of 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,** 1319–1325.

Lopes, H. (2000). Bayesian analysis in latent factor and longitudinal models. Ph.D. Thesis, Institute of Statistics and Decision Sciences, Duke University.

Magder, L. and Zeger, S. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association* **91,** 1141–1151.

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

Mallet, A., Mentré, F., Steimer, J.-L., and Lokiec, F. (1988). Handling covariates in population pharmacokinetics with an application to gentamicin. *Biomedical Measurement Informatics and Control* **2,** 138–146.

Mengersen, K. and Robert, C. (1995). Testing for mixtures: A Bayesian entropic approach. In *Bayesian Statistics,* Volume 5, J. Bernardo, J. Berger, A. Dawid, and A. Smith (eds). New York: John Wiley.

Müller, P., Quintana, F., and Rosner, G. (1999). *Hierarchical meta-analysis over related non-parametric Bayesian models*. Discussion Paper 99-22, Institute of Statistics and Decision Sciences, Duke University.

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

Müller, P. and Rosner, G. (1998). Semiparametric PK/PD models. In *Practical Nonparametric and Semiparametric Bayesian Statistics,* D. Dey, P. Müller, and D. Sinha (eds). New York: Springer-Verlag.

Pettit, L. and Young, K. (1990). Measuring the effect of observations on Bayes factors. *Biometrika* **77,** 455–466.

Raftery, A., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92,** 179–191.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B* **59,** 731–792.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92,** 894–902.

Rosner, G. and Müller, P. (1997). Bayesian population pharmacokinetics and pharmacodynamic analyses using mixture models. *Journal of Pharmacokinetics and Biopharmaceutics* **25,** 209–233.

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.

Wakefield, J. C. (1996). The Bayesian analysis of population pharmacokinetic models. *Journal of the American Statistical Association* **91,** 62–75.

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

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

Walker, S. and Wakefield, J. C. (1998). Population models with a nonparametric random coefficient distribution. *Sankhya, Series B* **60,** 196–212.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association* **86,** 79–86.

Received July 2001. *Revised May* 2002.
Accepted July 2002.

Appendix

We describe details of the RJMCMC used for posterior simulation across models $m \in \{1, \ldots, M\}$ in Section 3.3. For $L_1 \leq 2$ and $L_2 \leq 2$, we have $M = 4$ competing models to consider. Assume we have sets of Monte Carlo simulations $\Theta_1, \ldots, \Theta_M$ available, with $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}_m \,|\, \boldsymbol{y}, m)$ for $\boldsymbol{\theta} \in \Theta_m$. Since we only consider a small number M of models, it is practically feasible to generate such Monte Carlo samples once up front, before starting the RJMCMC scheme. In step 2b, below, we will use these Monte Carlo samples Θ_m to obtain posterior draws from $p(\boldsymbol{\theta}_m \,|\, \boldsymbol{y}, m)$. The proposed algorithm proceeds as follows:

Step 1 Suppose that the current state of the Markov chain is $\{m, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, i.e., the current model is m and $\boldsymbol{\theta}_i$ is a draw from the posterior distribution, under model $i = 1, 2, \ldots, M$.

Step 2 Propose a move:

Step 2a Propose a new model, $\widetilde{m}$, by generating from $q(\widetilde{m}\,|\,m)$. Here, $q(\cdot\ |\ \cdot)$ is an $M \times M$ transition matrix whose rows add up to one. For example, $q(\widetilde{m}\,|\,m)$ could propose $\widetilde{m} = m + 1$ and $\widetilde{m} = m - 1$, with probabilities 1/2 each when $m = 2, \ldots, M - 1$, and $\widetilde{m} = m - 1$ and $\widetilde{m} = m + 1$ with probability 1.0, for $m = M$ and $m = 1$, respectively.

Step 2b Generate proposals $\tilde{\boldsymbol{\theta}}_m \sim \hat{p}(\boldsymbol{\theta}_m\,|\,\boldsymbol{y}, m)$ and $\tilde{\boldsymbol{\theta}}_{\tilde{m}} \sim \hat{p}(\boldsymbol{\theta}_{\tilde{m}}\,|\,\boldsymbol{y}, \widetilde{m})$, where $\hat{p}(\boldsymbol{\theta}_m\,|\,\boldsymbol{y}, m)$ is based on Θ_m (Lopes, 2000). For all other models i, $i \neq m, \widetilde{m}$, leave $\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i$ unchanged. The proposed new state of the Markov chain is $\{\widetilde{m}, \tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_M\}$.

Step 2c The appropriate Hastings-Metropolis acceptance probability for the proposed new state is

$$\min\left\{1, \frac{Pr(\widetilde{m})}{Pr(m)}\,\frac{q(m\,|\,\widetilde{m})}{q(\widetilde{m}\,|\,m)}\,\frac{p(\boldsymbol{y}\,|\,\boldsymbol{\theta}_{\tilde{m}}, \widetilde{m})}{p(\boldsymbol{y}\,|\,\tilde{\boldsymbol{\theta}}_m, m)}\,\frac{\hat{p}(\boldsymbol{\theta}_{\tilde{m}}\,|\,\boldsymbol{y}, \widetilde{m})\,\hat{p}(\boldsymbol{\theta}_m\,|\,\boldsymbol{y}, m)}{\hat{p}(\tilde{\boldsymbol{\theta}}_{\tilde{m}}\,|\,\boldsymbol{y}, \widetilde{m})\,\hat{p}(\tilde{\boldsymbol{\theta}}\,|\,\boldsymbol{y}, m)}\right\}$$

In the particular case of a uniform prior $Pr(\widetilde{m}) = Pr(m)$ and a symmetric proposal $q(m\,|\,\widetilde{m}) = q(\widetilde{m}\,|\,m)$, the new model $\widetilde{m}$ is automatically accepted every time the current draw from the proposed model has higher likelihood than does the new draw from the current model.

Step 3 Cycle through steps 1 and 2 until convergence has been achieved.

In Step 2, we proposed a new model $\widetilde{m}$, and new parameters for both the current model m and the proposed model $\widetilde{m}$. Of course, we do not need to always keep parameter values for all competing models in memory. Only when a value $\boldsymbol{\theta}_i$ is required in step 2 must we read a value from the appropriate file of posterior Monte Carlo simulations. This procedure assumes that the samples Θ_i are independent. This can be achieved, for example, by using every nth generation of the MCMC simulations for larges values of n.