Independent and Identically Distributed Monte Carlo Algorithms for Semiparametric Linear Mixed Models

Hemant ISHWARAN and Glen TAKAHARA

Hybrid versions of independent and identically distributed weighted Chinese restaurant (WCR) algorithms are developed for inference in semiparametric linear mixed models under minimal assumptions for the random-effects distributions. The WCR method of working with the posterior partition structure leads to Rao–Blackwell estimates for higher-order moments of random effects, such as skewness and kurtosis, and can be used to estimate densities for random effects. A key feature of our approach is the manner in which we incorporate external estimates into our algorithms. The use of such information leads to simplified computational procedures, reduces the amount of user input required for specifying models, and results in numerical stability and accuracy. The resulting procedures are automated and can be readily used in standard statistical software. Our methods are tested by simulation and illustrated by application to a longitudinal study involving chronic renal disease.

KEY WORDS: Dirichlet process; Moment; Random effect; Rao–Blackwellization, Restricted maximum likelihood; Sequential importance sampling; Weighted Chinese restaurant.

1. INTRODUCTION

This article introduces a new class of independent and identically distributed (iid) Monte Carlo algorithms which can be used for inference in semiparametric linear mixed models under minimal assumptions for the random-effects distribution. An important feature of our algorithms is that they combine frequentist parametric estimates with Bayesian nonparametric techniques and as such seek to exploit the strengths of both approaches in producing an overall more flexible and efficient method for inference. Our iid algorithms are extensions of the weighted Chinese restaurant (WCR) algorithms discussed by Ishwaran, James, and Lo (2001) for fitting semiparametric models. Lo, Brunner, and Chan (1996) have provided general methodology related to iid WCR algorithms, and Ishwaran and James (2000) and Ishwaran, James, and Sun (2001) have given extensions to generalized Chinese restaurant processes and algorithms. Related work on sequential importance sampling (SIS) has been done by MacEachern, Clyde, and Liu (1999), Quintana (1998), and Quintana and Newton (2000).

One of the primary focuses of the article is the widely used linear mixed model for continuous longitudinal data, where models include fixed-effects terms for population parameters and random-effects terms for subject-specific variation (Laird and Ware 1982). However, here we relax the usual parametric assumption of normal random effects to focus on more general inference for the random-effects distribution, such as estimation of higher-order moments and detection of skewness and multimodality. Although estimation for the fixed-effects parameters is relatively robust to misspecification of the random effects, for example, by empirical best linear unbiased estimators (EBLUE) obtained from restricted maximum likelihood (REML) estimation (see Butler and Louis 1992 for some empirical evidence and Jiang 1998 for theoretical results), it is becoming widely recognized that inference for random effects can be misleading when normal distributional assumptions do not hold (Verbeke and Lesaffre 1996). Moreover, even though certain features for random-effects distributions, such as variances, can be estimated accurately even when the assumption of normality is violated (Richardson and Welsh 1994; Jiang 1996), detecting higher-order properties of distributions such as skewness and multimodality necessitates dispensing with normality assumptions. Identifying such deviations from normality can be important, sometimes leading to critical insight into the data. For example, Zhang and Davidian (2002) used a semiparametric approach to identify population differences in cholesterol levels by detecting skewness in random intercept terms, while Greene (2001) was able to identify clinical differences in the progression of renal kidney disease by identifying skewed and thick-tailed random-slope distributions. (We return later in more detail to this last example as one illustration of our approach.) For other non-Bayesian semiparametric approaches to linear mixed models see Verbeke and Lesaffre (1996) and Aitkin (1999) who used a finite-mixture approach with implementation by the EM algorithm.

Although the longitudinal problem is a key application, our methods also apply to single-measurement data when only one observation is recorded per subject; for example, when it is anticipated that subject variation is larger than can be explained by measurement error. If error distributions are unknown, then such data can be viewed more generally as arising from a semiparametric regression model and cannot always be fit properly using standard methods such as REML. The assumption throughout, for both single and repeated measurements, is that random-effects distributions are unknown. Our approach, discussed in detail in Section 2, is to use a Bayesian nonparametric framework in which random effects are modeled using a Ferguson Dirichlet process.

The article is organized as follows. Section 2 provides the necessary background material on the iid WCR procedure and

Hemant Ishwaran is Associate Staff, Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation, Cleveland, OH 44195 (E-mail: *ishwaran@bio.ri.ccf.org*). Glen Takahara is Associate Professor, Department of Mathematics and Statistics, Queen's University, Kingston, Ontario K7L 3N6, Canada (E-mail: *takahara@mast.queensu.ca*). The authors thank Tom Greene for providing access to the MDRD data and for helpful discussions surrounding its analysis. They also thank Jiming Jiang for help with technical questions concerning mixed models.

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Theory and Methods
DOI 10.1198/016214502388618951

provides an overview of our approach. Connections with this work to the Bayesian nonparametric literature are discussed. Section 3 characterizes the posterior of Dirichlet process mixture models in terms of the partition structure, and thus provides the blueprint for estimating random effects. In particular, Rao–Blackwell estimates for higher-order moments for random effects, such as the skewness and kurtosis, estimates for standard errors, and methods for density estimation are discussed. Sections 4 and 5 discuss the WCR procedures for the single-measurement case. The methods developed in these sections are then extended to the longitudinal setting of Section 6, which describes how to combine REML estimates with the iid WCR algorithms for inference in Laird–Ware random-effects models. Section 7 discusses the extensive testing of the method by simulation and then illustrates its application to a study involving chronic renal disease. Section 8 summarizes key computational features.

2. SEMIPARAMETRIC HIERARCHICAL MODELS: OVERVIEW OF METHODS

For ease of presentation, it is simpler for us to first consider the single-measurement case, and later extend the methods for longitudinal data in Section 6. With single-measurement data, inference for the fixed-effects parameter $\boldsymbol{\beta}$ (p-dimensional) and random-effects parameters $\boldsymbol{\alpha}_i$ (s-dimensional) are based on data $\mathbf{Y} = (Y_1, \ldots, Y_n)$, where

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{Z}_i'\boldsymbol{\alpha}_i + \epsilon_i, \qquad i = 1, \ldots, n. \tag{1}$$

Thus we observe one random-effects parameter $\boldsymbol{\alpha}_i$ per subject i. In (1), the $\mathbf{X}_i$ are the p-dimensional fixed-effects covariates, $\mathbf{Z}_i$ are the s-dimensional random-effects covariates, and ϵ_i are iid normal random variables (the measurement errors) with mean 0 and variance σ^2. We assume that $\boldsymbol{\alpha}_i$ are iid from an unknown distribution Q_0; thus (1) can be viewed generally as a semiparametric regression model. For example, if $s = 1$ and $Z_i = 1$, then

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \xi_i, \qquad i = 1, \ldots, n,$$

where $\xi_i = \alpha_i + \epsilon_i$ are iid measurement errors with unknown distribution.

An illustrative example in Section 5 we look at is what we call the "two-slope" problem, which can be described as follows. In group 1, the data are assumed to follow a slope–intercept model with a fixed-effects term for the slope and a random-effects term for the intercept, whereas in group 2, both slopes and intercepts are assumed to be random. For convenience, assuming that group 1 corresponds to observations $1, \ldots, m$, and group 2 corresponds to observations $m+1, \ldots, n$, the model is

$$Y_i = \alpha_{i,0} + X_i\beta_1 + \epsilon_i, \qquad i = 1, \ldots, m, \tag{2}$$

and

$$Y_i = \alpha_{i,0} + X_i\alpha_{i,1} + \epsilon_i, \qquad i = m+1, \ldots, n. \tag{3}$$

It is clear this can be written as (1).

Consider the simulation presented in Figure 1 based on $n = 1{,}000$ observations (with data evenly distributed between the two groups). In group 2, random slopes $\alpha_{i,1}$ were simulated from a skewed distribution with mean chosen to coincide with the fixed-effects slope β_1 of group 1, whereas random intercepts $\alpha_{i,0}$ were drawn from a discrete two-point mixture distribution. Analysis of this data will be challenging to methods that do not relax assumptions of normality. For example, if intercept distributions are assumed normal, then it becomes theoretically impossible to separate the variance of intercepts from the variance σ^2 for the measurement error. Thus a method such as REML will be inconsistent here (see Sec. 5 for simulation results). Conventional methods will also have difficulty estimating the nonnormal random-effects distributions. For example, random effects typically estimated by empirical best linear unbiased prediction (EBLUP) will have poor performance here. Consider Figure 2, which plots the EBLUP estimates obtained from REML. It is clear that EBLUP cannot pick up on the discreteness of the intercept distribution, whereas the slope distribution, although skewed, still looks somewhat normal. Compare this to the WCR posterior estimates $E[\alpha_{i,0}|\mathbf{Y}]$ and $E[\alpha_{i,1}|\mathbf{Y}]$ plotted in Figure 3, which clearly identify the bimodal and skewed shape of the random effects. We return to this example in more detail in Section 5. Later, in Section 6, we consider more complex models for longitudinal data.

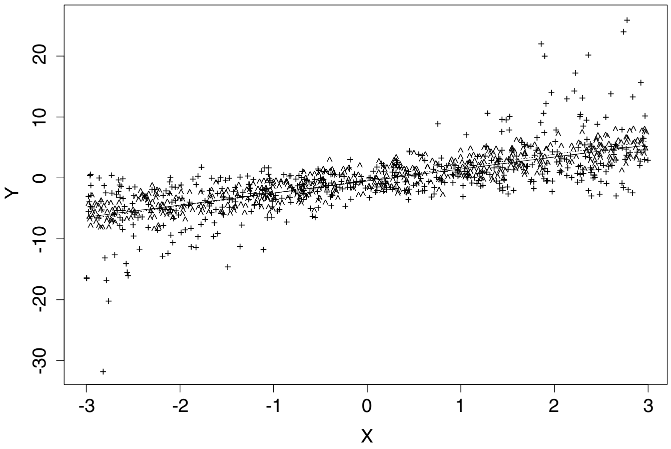

Figure 1. Simulated Data for Two-Slope Problems. Lines correspond to predicted values using least squares. (—— group 1, ······· group 2). Values for group 2 indicated by +.

2.1 Hierarchical Models

Our approach to the linear mixed model uses a Bayesian nonparametric technique of modeling an unknown distribution by a nonparametric prior. In the mixed model, this works by modeling Q_0 using a random measure P with some prior $\mathcal{P}$. The method is best conceptualized by reexpressing (1) in a semiparametric hierarchical framework. Under the prior $\mathcal{P}$, the Bayesian semiparametric model for (1) is (conditioning on fixed effects and the measurement error variance)

$$\begin{aligned} (Y_i|\boldsymbol{\alpha}_i, \boldsymbol{\beta}, \sigma^2) &\overset{\text{ind}}{\sim} \mathrm{N}(\eta_i, \sigma^2), \qquad i = 1, \ldots, n, \\ (\boldsymbol{\alpha}_i|P) &\overset{\text{iid}}{\sim} P, \\ P &\sim \mathcal{P}, \end{aligned} \tag{4}$$

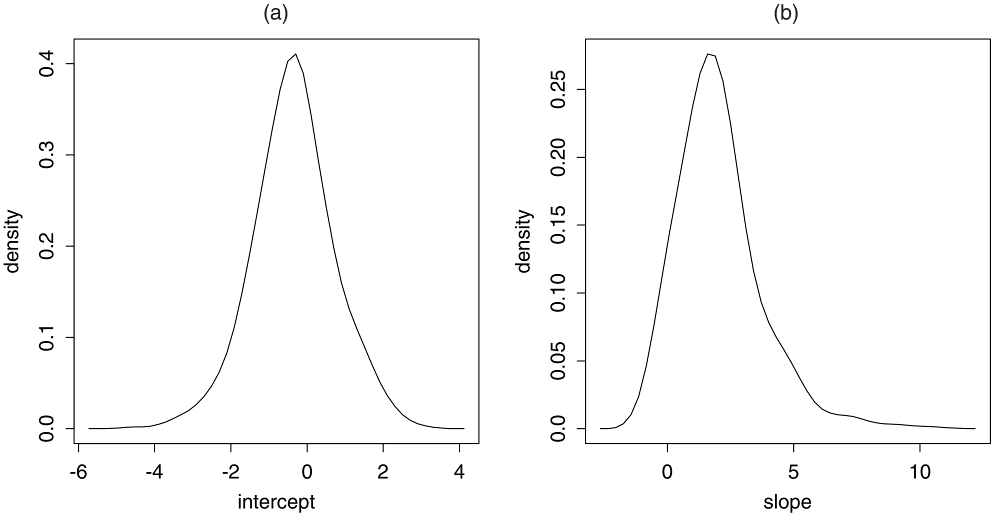

Figure 2. Estimated Random Effects Using EBLUP. (a) Intercept; (b) slope.

where $\eta_i = \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{Z}_i'\boldsymbol{\alpha}_i$. Inference for Q_0 is thus based on the posterior for P from (4). The WCR methods discussed later can be applied to general discrete random measures $\mathcal{P}$ (Ishwaran and James 2000), although here we illustrate its use applied to the Ferguson (1973, 1974) Dirichlet process. Thus throughout, we write $\mathcal{P}$ to refer to $\mathrm{DP}(a_0H)$, the Ferguson Dirichlet process with finite measure $a_0H(\cdot)$, where $a_0 > 0$ is some constant and $H(\cdot)$ is a probability measure over $\Re^s$. The measure $H(\cdot)$ can be thought of as the guess of the distribution of the random effects, and a_0 as a measure of the strength of this belief. In our applications, it is convenient to take $H(\cdot)$ to be a normal distribution. These details are spelled out later in the article.

Hierarchical models subject to the Dirichlet process, similar in nature to (4), are now increasingly used for inference in nonparametric and semiparametric problems (see West, Müller, and Escobar 1994; Escobar and West 1998 for background and examples). Applications of such models specifically to linear mixed models have been considered in various contexts; for example, Bush and MacEachern (1996) considered a semiparametric randomized block design, and Kleinman and Ibrahim (1998) looked at mixed models for longitudinal data. Such approaches typically rely on the use of Markov chain Monte Carlo for model fitting, typically using a Pólya urn Gibbs sampler, a general Gibbs sampling technique for fitting Dirichlet process mixture models (Escobar 1988, 1994). (See, however, Tao, Palta, Yandell, and Newton 1999 for a predictive recursive algorithm for fitting models.) But little work has been done along the lines of fitting linear mixed models based on iid sampling, or on methods that rely on the partition structure. Here we consider such an approach, the iid WCR algorithm, an SIS technique based on the partition structure for the Dirichlet process (Lo et al. 1996; Brunner, Chan, James, and Lo 2001). As we discuss later (Sec. 3), such a technique can be used to produce stable Rao–Blackwell estimates for functionals of the random effects, such as estimates for moments and density estimates.

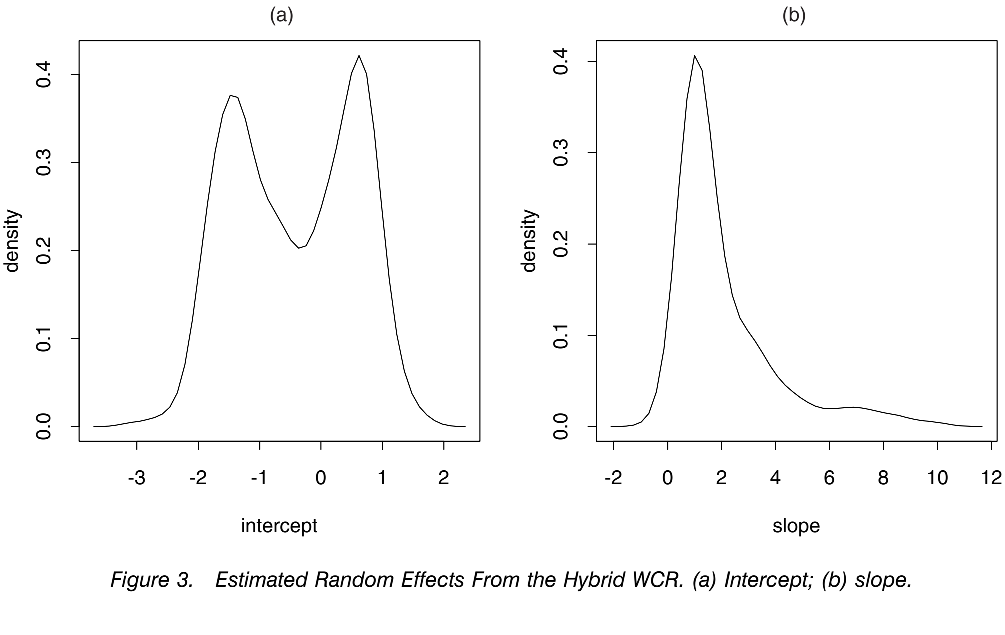

Figure 3. Estimated Random Effects From the Hybrid WCR. (a) Intercept; (b) slope.

2.2 Random Partitions and Clustering

A key to the numerical stability of the WCR algorithm comes from its use of the partition distribution of the Dirichlet process. Let $\mathbf{p} = \{C_1, \ldots, C_{n(\mathbf{p})}\}$ be a partition of size $n(\mathbf{p})$ for the set $\{1, \ldots, n\}$, where each set C_j contains e_j elements. Due to the discrete nature of the Dirichlet process (Blackwell and MacQueen 1973), each realization from the prior induces a clustering of the random effects $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n)$ by some partition $\mathbf{p}$ such that the observations $\{Y_i : i \in C_j\}$ share a common $\boldsymbol{\alpha}_i$ value $\mathbf{u}_j$ for $j = 1, \ldots, n(\mathbf{p})$. Note that $\boldsymbol{\alpha}$ can be equivalently represented as $(\mathbf{u}, \mathbf{p})$, where $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_{n(\mathbf{p})})$. The partition distribution induced by the prior in turn induces a posterior that can be characterized in terms of $\mathbf{p}$ (see Sec. 3), and it is this structure that is sampled by the WCR algorithm. Although one can apply SIS techniques based on working with a posterior characterized by $\boldsymbol{\alpha}$ by, for example, extending the methods discussed by Kong, Liu, and Wong (1994) and Liu (1996), working with a posterior characterized by the minimal information of the partition leads to improved numerical stability due to Rao–Blackwellization. Lo et al. (1996) and Ishwaran and James (2000) provided general discussions on this point. MacEachern et al. (1999) presented a partition-based SIS technique for beta-binomial Dirichlet process mixture models (what they call their “S2 algorithm”). Extensions to the S2 algorithm were given by Quintana (1998), who considered multinomial data models, and by Quintana and Newton (2000), who extended the beta-binomial models of MacEachern et al. (1999) to allow for nonexchangeability.

Observe that $\mathbf{p}$ contains the minimal amount of information for reducing our nonparametric problem parametrically. That is, for a given $\mathbf{p}$, the likelihood for (1) equals

$$\mathrm{Lik}(\mathbf{u}, \boldsymbol{\beta}, \sigma^2|\mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} \prod_{i \in C_j} \phi\big(Y_{i,\boldsymbol{\beta}}|\mathbf{Z}_i'\mathbf{u}_j, \sigma^2\big),$$

where $Y_{i,\boldsymbol{\beta}} = Y_i - \mathbf{X}_i'\boldsymbol{\beta}$ and $\phi(\cdot|\mu, \tau^2)$ denotes a normal density with mean μ and variance τ^2. Thus $\mathbf{p}$ tells us how the data are clustered, reducing the problem to a collection of conditionally independent normal parametric models, from which inference should now be straightforward.

2.3 WCR Draws

The key is to be able to draw $\mathbf{p}$ from its posterior, which for a fixed value for $\boldsymbol{\beta}$ and σ^2 equals

$$\pi(\mathbf{p}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{p})\pi(\mathbf{p})}{\sum_{\mathbf{p}} f(\mathbf{Y}|\mathbf{p})\pi(\mathbf{p})},$$

where $\pi(\mathbf{p})$ is the prior probability for $\mathbf{p}$, the previous sum is over all partitions $\mathbf{p}$ and

$$f(\mathbf{Y}|\mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} \int_{\Re^s} \prod_{i \in C_j} \phi\big(Y_{i,\boldsymbol{\beta}}|\mathbf{Z}_i'\mathbf{u}_j, \sigma^2\big) H(d\mathbf{u}_j).$$

However, a direct draw from $\pi(\mathbf{p}|\mathbf{Y})$ is not feasible. Instead, we draw $\mathbf{p}$ from the WCR density $q(\mathbf{p})$, where

$$\Lambda(\mathbf{p})q(\mathbf{p}) = f(\mathbf{Y}|\mathbf{p})\pi(\mathbf{p}).$$

In SIS parlance, $q(\mathbf{p})$ is typically referred to as the importance density (see Lo et al. 1996 for discussion of why this is an appropriate choice), whereas $\Lambda(\mathbf{p})$ are its importance weights. Now observe that

$$\pi(\mathbf{p}|\mathbf{Y}) = \frac{\Lambda(\mathbf{p})q(\mathbf{p})}{\sum_{\mathbf{p}} \Lambda(\mathbf{p})q(\mathbf{p})},$$

from which one can now easily devise a Monte Carlo method for estimating functionals. Thus, for example, to estimate $E[g(\mathbf{p})|\mathbf{Y}]$ for some function g, draw B iid values from $q(\mathbf{p})$ and use the approximation

$$E[g(\mathbf{p})|\mathbf{Y}] \approx \frac{\sum_{b=1}^{B} g(\mathbf{p}^{(b)})\Lambda(\mathbf{p}^{(b)})}{\sum_{b=1}^{B} \Lambda(\mathbf{p}^{(b)})}.$$

This is the essential idea of the WCR procedure.

2.4 External Estimates

Of course, in practice we do not know the value for $\boldsymbol{\beta}$ or σ^2. One method for handling these values is to extend the hierarchy (4) to include priors for these parameters and then estimate their values from the posterior. This technique can be accommodated within the WCR method; however, it will make drawing $\mathbf{p}$ from the WCR density more difficult, because we now need to further integrate over the $(p+1)$-dimensional space for $(\boldsymbol{\beta}, \sigma^2)$ when devising $q(\mathbf{p})$; see Section 4 for details. Moreover, it adds another layer of hyperparameters that must be supplied by the user. On the other hand, reliable and easily computed estimates for $\boldsymbol{\beta}$ are available that are $\sqrt{n}$ consistent under minimal assumptions for the random-effects distributions. For example, with single measurements we can use ordinary least squares (OLS) to estimate $\boldsymbol{\beta}$ (see Sec. 5 for more details), whereas in the longitudinal data setting (see Sec. 6), $\boldsymbol{\beta}$, as well as σ^2, can be consistently estimated using REML. (Unlike in single measurement problems such as the two-slope example, REML is typically consistent with repeated measurements; see Jiang 1996 for details.)

Thus, instead, we follow a general approach discussed by Ishwaran et al. (2001) by incorporating parametric estimates within the WCR algorithm. For single-measurement data, we rely on external estimates for $\boldsymbol{\beta}$, and in the longitudinal setting of Section 6 we also rely on estimates for σ^2 as well as other parametric components, such as hyperparameters used in the specification of the nonparametric prior $\mathcal{P}$. Not all parameters can be estimated using external techniques, however. For example, with single measurements it is not clear in general how to obtain reliable estimates for σ^2. Our approach is to update this value sequentially *within* the WCR algorithm using an idea discussed by Ishwaran et al. (2001), building up the value for σ^2 as the partition structure evolves. This idea can be applied generally to other parameters; see Section 5 for details. Of course, the technique of using external estimates means that our algorithms are really “hybrid” or approximate WCR procedures, and by specifying priors based on such estimates, we are applying a form of empirical Bayes. Even though this is nonstandard, there is growing evidence that the use of data-dependent priors can lead to accurate posterior inference and in some cases higher-order accuracy than is possible with non–data-driven priors (Wasserman 2000). We provide some empirical evidence of the stability and accuracy of the approach by way of simulations in Sections 5 and 6.

3. POSTERIOR CHARACTERIZATIONS FOR DIRICHLET PROCESS MIXTURE MODELS

Although our interest in moments and density estimates is for the random effects in linear mixed models, it is easier notationally to outline the techniques in a simplified setting involving hidden variables without covariates. Thus consider the setting in which data $\mathbf{Y} = (Y_1, \dots, Y_n)$ are derived from a standard Dirichlet process mixture model,

$$\begin{aligned} (Y_i|\mathbf{V}_i) &\overset{\text{ind}}{\sim} f_0(Y_i|\mathbf{V}_i), \qquad i = 1, \dots, n \\ (\mathbf{V}_i|P) &\overset{\text{iid}}{\sim} P, \\ P &\sim \mathcal{P} = \mathrm{DP}(a_0 H), \end{aligned} \tag{5}$$

where f_0 is some given density. Model (5) can be thought of as a missing-data model, where $\mathbf{V}_i$ are the missing data that are assumed to be drawn from a Dirichlet process over a Borel space $\mathcal{U}$ (for our applications, $\mathcal{U} = \Re^s$). Ferguson (1973) has provided a general discussion of the Dirichlet process, Lo (1984) has given posterior characterizations of Dirichlet process mixture models, and Ferguson, Phadia, and Tiwari (1992) have published a survey paper discussing key properties for the Dirichlet process as well as some of its applications.

Now to go about estimating moments and so forth, we first need to characterize the posterior for (5). Consider an arbitrary functional $\psi(P)$ of a measure P. Then the posterior for P from (5) can be characterized as (Lo et al. 1996)

$$\begin{aligned} &\int \psi(P)\,\mathcal{P}(dP|\mathbf{Y}) \\ &\quad = \sum_{\mathbf{p}} \left[\int_{\mathcal{U}^{n(\mathbf{p})}} \int \psi(P)\,\mathcal{P}(dP|\mathbf{u},\mathbf{p}) \prod_{j=1}^{n(\mathbf{p})} \pi(d\mathbf{u}_j|C_j) \right] \pi(\mathbf{p}|\mathbf{Y}) \\ &\quad = \sum_{\mathbf{p}} \left[\int_{\mathcal{U}^{n(\mathbf{p})}} \int \psi(P)\,\mathcal{P}(dP|\mathbf{u},\mathbf{p}) \prod_{j=1}^{n(\mathbf{p})} \pi(d\mathbf{u}_j|C_j) \right] \\ &\qquad \times \frac{\Lambda(\mathbf{p})q(\mathbf{p})}{\sum_{\mathbf{p}} \Lambda(\mathbf{p})q(\mathbf{p})}, \end{aligned} \tag{6}$$

where $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_{n(\mathbf{p})})$ is the entire set of $n(\mathbf{p})$ unique values for $\mathbf{v}_1, \dots, \mathbf{v}_n$, $q(\mathbf{p})$ is the WCR density for (5) with importance weight $\Lambda(\mathbf{p})$, and $\pi(d\mathbf{u}_j|C_j)$ are the laws for the conditionally independent unique values $\mathbf{u}_j$. The measure $\mathcal{P}(dP|\mathbf{u},\mathbf{p})$ is the law for a Dirichlet process with finite measure $a_0 H(\cdot) + \sum_{j=1}^{n(\mathbf{p})} e_j \delta_{\mathbf{u}_j}(\cdot)$, where e_j is the cardinality of C_j.

3.1 Rao–Blackwellization: Moments and Density Estimation

Now suppose that we want to estimate $\boldsymbol{\mu}_k = E[\int \mathbf{v}^k P(d\mathbf{v})|\mathbf{Y}]$, the kth moment of the mean of the posterior random measure from (5). From the previous decomposition, with rearrangement by Fubini, we have

$$\begin{aligned} \boldsymbol{\mu}_k &= \iint \mathbf{v}^k\, P(d\mathbf{v})\, \mathcal{P}(dP|\mathbf{Y}) \\ &= \sum_{\mathbf{p}} \left[\frac{a_0}{a_0 + n} \int \mathbf{v}^k\, H(d\mathbf{v}) + \sum_{j=1}^{n(\mathbf{p})} \frac{e_j}{a_0 + n} E[\mathbf{U}_j^k|C_j] \right] \pi(\mathbf{p}|\mathbf{Y}) \\ &= \sum_{\mathbf{p}} \mathbf{g}_k(\mathbf{p})\, \pi(\mathbf{p}|\mathbf{Y}). \end{aligned}$$

Now reexpressing this in terms of the WCR density, we have

$$\boldsymbol{\mu}_k = \frac{\sum_{\mathbf{p}} \mathbf{g}_k(\mathbf{p}) \Lambda(\mathbf{p}) q(\mathbf{p})}{\sum_{\mathbf{p}} \Lambda(\mathbf{p}) q(\mathbf{p})},$$

which automatically suggests a Rao–Blackwell estimator for $\boldsymbol{\mu}_k$. Draw B iid values $\mathbf{p}$ from the WCR density, computing $\mathbf{g}_k(\mathbf{p})$ and $\Lambda(\mathbf{p})$ for each of these values. Estimate $\boldsymbol{\mu}_k$ with

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{b=1}^{B} \mathbf{g}_k(\mathbf{p}^{(b)}) \Lambda(\mathbf{p}^{(b)})}{\sum_{b=1}^{B} \Lambda(\mathbf{p}^{(b)})}.$$

We apply this method in Sections 5, 6, and 7 to estimate higher-order moments for random effects, such as skewness and kurtosis.

The same technique can also be used to derive a density estimate. This method allows for both smoothed and unsmoothed estimates, although here we focus on unsmoothed versions corresponding to discrete density estimates. Now suppose that $\mathcal{U} = \Re^s$. To estimate the posterior mean cumulative distribution function, let $\psi_{t,l}(P) = P\{U_l \le t\}$ be the cumulative distribution function for the lth marginal of P evaluated at some $t \in \Re$. By the previous decomposition, we have

$$\begin{aligned} &E[\psi_{t,l}(P)|\mathbf{Y}] \\ &\quad = \sum_{\mathbf{p}} \left[\frac{a_0}{a_0 + n} H\{U_l \le t\} + \sum_{j=1}^{n(\mathbf{p})} \frac{e_j}{a_0 + n} \pi(\{U_l \le t\}|C_j) \right] \pi(\mathbf{p}|\mathbf{Y}) \\ &\quad = \sum_{\mathbf{p}} h_{t,l}(\mathbf{p})\, \pi(\mathbf{p}|\mathbf{Y}). \end{aligned}$$

Hence, for a Rao–Blackwell estimate for $E[\psi_{t,l}(P)|\mathbf{Y}]$, we use

$$\widehat{\psi_{t,l}(P)} = \frac{\sum_{b=1}^{B} h_{t,l}(\mathbf{p}^{(b)}) \Lambda(\mathbf{p}^{(b)})}{\sum_{b=1}^{B} \Lambda(\mathbf{p}^{(b)})}.$$

Now choosing a grid of values $t_1 < \cdots < t_N$, we obtain a Rao–Blackwell estimate for the marginal cumulative distribution function, $(\widehat{\psi_{t_1,l}(P)}, \dots, \widehat{\psi_{t_N,l}(P)})$. This can be converted to a discrete density estimator or it can be smoothed using standard smoothing techniques; see Section 7 for an illustration.

Remark 1. So far we have only considered estimating functions of $\mathbf{p}$. However, (6) also points the way to estimating more general functions, such as those that can depend on both $\mathbf{p}$ and $\mathbf{u}$. We further elaborate this point in the following section.

3.2 Standard Errors

The Rao–Blackwell technique of integrating over the posterior gives stable Monte Carlo estimates for posterior mean values of functionals of P, such as for the mean posterior moments $\boldsymbol{\mu}_k = (\mu_{k,1}, \ldots, \mu_{k,s})$; however, the technique does not provide estimates for the variability of such estimators. One method for obtaining such estimates is to *expand the posterior* so as to sample values from the posterior random measure. For an estimator of a univariate functional, the estimates of variability derived in this way is what we call its "standard error." This use of terminology is, of course, not technically exact, but it is helpful in facilitating comparison to frequentist standard errors, such as those derived from REML that we explore later in this article.

We focus specifically on estimates for the standard error of the moment estimator $\hat{\mu}_{k,l}$, for $l = 1, \ldots, s$, but the ideas can be generalized. A little thought shows that the standard error for $\hat{\mu}_{k,l}$ can be estimated by the square root of the lth diagonal element of the posterior variance-covariance matrix of the functional $\int \mathbf{v}^k P(d\mathbf{v})$, which in turn can be estimated using Monte Carlo if we can draw P from the posterior. The posterior characterization (6) shows how to do this. To draw P, we first draw $\mathbf{p}$ followed by a draw for $\mathbf{u}$ given $\mathbf{p}$. Finally, we draw P by conditioning on $\mathbf{u}$ and $\mathbf{p}$. The draw for P can now be used to produce a sampled value for our functional $\int \mathbf{v}^k P(d\mathbf{v})$. Thus, providing estimates for standard errors requires the following steps:

1. Draw $\mathbf{p}$ from the WCR density. Given $\mathbf{p}$, draw $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_{n(\mathbf{p})})$ from

$$\pi(d\mathbf{u}|\mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} \pi(d\mathbf{u}_j|C_j) \propto \prod_{j=1}^{n(\mathbf{p})} H(d\mathbf{u}_j) \prod_{i \in C_j} f_0(Y_i|\mathbf{u}_j),$$

and then draw P from $\mathcal{P}(\cdot|\mathbf{u}, \mathbf{p})$. The draw for $\mathbf{u}$ follows straightforwardly from parametric arguments, whereas the draw for P is facilitated by noting that (Ishwaran and James 2001)

$$\mathcal{P}(\cdot|\mathbf{u}, \mathbf{p}) \overset{\mathcal{D}}{=} \sum_{j=1}^{n(\mathbf{p})} p_j \, \delta_{\mathbf{u}_j}(\cdot) + p_{n(\mathbf{p})+1} \mathcal{P}(\cdot),$$

where $(p_1, \ldots, p_{n(\mathbf{p})}, p_{n(\mathbf{p})+1}) \sim \text{Dirichlet}(e_1, \ldots, e_{n(\mathbf{p})}, a_0)$ is independent of $\mathcal{P}$, which is a $\text{DP}(a_0 H)$ process. Although it is not possible to draw from $\mathcal{P}$ exactly, one can instead draw from a finite-dimensional approximation $\widehat{\mathcal{P}}$, which can be chosen to be arbitrarily accurate (Ishwaran and Zarepour 2002). Thus for a draw P, use $\sum_{j=1}^{n(\mathbf{p})} p_j \, \delta_{\mathbf{u}_j}(\cdot) + p_{n(\mathbf{p})+1} \widehat{\mathcal{P}}(\cdot)$. Compute $\tilde{\boldsymbol{\mu}}_k = \int \mathbf{v}^k P(d\mathbf{v})$, the kth moment of P.

2. Repeat B times independently. Now estimate the standard error of $\hat{\mu}_{k,l}$, for $l = 1, \ldots, s$, by

$$\sqrt{\frac{\sum_{b=1}^{B} (\tilde{\mu}_{k,l}^{(b)})^2 \Lambda(\mathbf{p}^{(b)})}{\sum_{b=1}^{B} \Lambda(\mathbf{p}^{(b)})} - \left(\frac{\sum_{b=1}^{B} \tilde{\mu}_{k,l}^{(b)} \Lambda(\mathbf{p}^{(b)})}{\sum_{b=1}^{B} \Lambda(\mathbf{p}^{(b)})}\right)^2}.$$

4. THE WEIGHTED CHINESE RESTAURANT ALGORITHM FOR DRAWING p

In this section we describe how to draw $\mathbf{p}$ from the WCR density under priors for $\boldsymbol{\beta}$ and σ^2. This serves as motivation for Section 5, in which we discuss a hybrid version involving external estimates for $\boldsymbol{\beta}$ and sequential updating for σ^2. In essence, the WCR method is a sequential method for generating a partition $\mathbf{p}$ of the set of integers $\{1, \ldots, n\}$, where the draw is designed in such a way to ensure that $\mathbf{p}$ has the WCR density $q(\mathbf{p})$. Specifically, the method works by creating a sequence of increasing partitions $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ formed by assigning $\{1, \ldots, n\}$ sequentially into sets using a random posterior partition rule. For $r > 1$, let $\mathbf{p}_r = \{C_{1,r}, \ldots, C_{n(\mathbf{p}_r),r}\}$ denote a partition of $\{1, \ldots, r\}$, where $C_{j,r}$ denotes the current jth set containing $e_{j,r}$ of the labels from $\{1, \ldots, r\}$. Applied to the linear mixed model (4), assuming priors for $(\boldsymbol{\beta}, \sigma^2)$, the sequential draw for $\mathbf{p}$ is as follows:

Step 1: Assign $\mathbf{p}_1 = \{1\}$. Let $\lambda(1) = \iint \phi(Y_{1,\boldsymbol{\beta}}|\mathbf{Z}_1'\mathbf{u}_1, \sigma^2) H(d\mathbf{u}_1) \, \pi(d\boldsymbol{\beta}, d\sigma^2)$.

Step r: Given $\mathbf{p}_{r-1}$, form $\mathbf{p}_r$ by assigning label r to a new set with probability

$$\frac{a_0}{(a_0 + r - 1)\lambda(r)} \times \iint \phi(Y_{r,\boldsymbol{\beta}}|\mathbf{Z}_r'\mathbf{u}_r, \sigma^2) \times H(d\mathbf{u}_r) \, \pi(d\boldsymbol{\beta}, d\sigma^2|\mathbf{p}_{r-1}, Y_1, \ldots, Y_{r-1}), \quad (7)$$

or to an existing set $C_{j,r-1}$ with probability $e_{j,r-1}/[(a_0 + r - 1)\lambda(r)]$ multiplied by

$$\iint \phi(Y_{r,\boldsymbol{\beta}}|\mathbf{Z}_r'\mathbf{u}_{j,r-1}, \sigma^2) \, \pi(d\mathbf{u}_{j,r-1}|C_{j,r-1}, \boldsymbol{\beta}, \sigma^2) \times \pi(d\boldsymbol{\beta}, d\sigma^2|\mathbf{p}_{r-1}, Y_1, \ldots, Y_{r-1}), \quad (8)$$

where $\lambda(r)$ is the appropriate normalizing constant.

Draw for $\mathbf{p}$: Run step 1 followed by step r for $r = 2, \ldots, n$. This gives a draw $\mathbf{p} = \{C_1, \ldots, C_{n(\mathbf{p})}\}$, which is a random partition of $\{1, \ldots, n\}$ with WCR density $q(\mathbf{p})$ from the posterior of (4), assuming priors for $(\boldsymbol{\beta}, \sigma^2)$. Its importance weight is $\Lambda(\mathbf{p}) = \lambda(1) \times \cdots \times \lambda(n)$.

In general, the integrals (7) and (8) in the update rules may be computed in closed form if H and $(\boldsymbol{\beta}, \sigma^2)$ are jointly conjugate. Relying on conjugacy thus allows for explicit update rules, which in turn provides a tractable method for full posterior inference for both fixed and random effects. However, in Section 5, we take a different approach by applying an approximation technique that uses a "plug-in" estimator, $\hat{\boldsymbol{\beta}}$, for $\boldsymbol{\beta}$. This simplifies computations and frees us from using conjugate priors and the tricky problem of choosing their hyperparameters. Of course, when using a plug-in method, we need to assess accuracy. This is studied in the simulations of Section 5.2 and also in the simulations of Section 6.2, in which we consider longitudinal data. What we find overall is that the hybrid WCR method is quite accurate in recovering moments for normal and nonnormal random effects and also provides robust estimates for standard errors. Section 6 illustrates another advantage of using external estimates, that they can be used to automate selection for hyperparameters for H

to further encourage accurate inference for random effects. Although one seeming disadvantage of a plug-in approach is that it does not allow for full posterior inference for fixed effects, this is a desirable trade-off in our view, because reliable and robust inference for $\boldsymbol{\beta}$ is readily available through frequentist estimates (see Sec. 5.2 and comment 3 in Sec. 6.2).

5. FAST APPROXIMATE WEIGHTED CHINESE RESTAURANT ALGORITHMS

In the plug-in approach, the measure $\pi(d\boldsymbol{\beta}, d\sigma^2|\mathbf{p}_{r-1}, Y_1, \ldots, Y_{r-1})$ appearing in the update rule is replaced with some form of approximation. Our approach is to replace $\boldsymbol{\beta}$ throughout by the OLS estimate $\hat{\boldsymbol{\beta}}$ and to work with residuals $\widehat{Y}_i = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ in place of Y_i. Reliable estimates for σ^2 are not always available, and so we instead present a novel estimation scheme in which σ^2 is sequentially estimated within the algorithm. That is, at step r we compute a point estimate σ^2_{r-1}, now replacing $\pi(d\boldsymbol{\beta}, d\sigma^2|\mathbf{p}_{r-1}, Y_1, \ldots, Y_{r-1})$ with $(\hat{\boldsymbol{\beta}}, \sigma^2_{r-1})$. Thus (7) is replaced with

$$\frac{a_0}{(a_0 + r - 1)\lambda(r)} \times \int_{\Re^s} \phi\big(\widehat{Y}_r|\mathbf{Z}_r'\mathbf{u}_r, \sigma^2_{r-1}\big)\, H(d\mathbf{u}_r),$$

whereas (8) now becomes

$$\frac{e_{j,r-1}}{(a_0 + r - 1)\lambda(r)} \times \int_{\Re^s} \phi\big(\widehat{Y}_r|\mathbf{Z}_r'\mathbf{u}_{j,r-1}, \sigma^2_{r-1}\big) \times \pi\big(d\mathbf{u}_{j,r-1}|C_{j,r-1}, \hat{\boldsymbol{\beta}}, \sigma^2_{r-1}\big).$$

Assuming a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution for H, it is now straightforward to compute the update rules in closed form. Appendix A provides the details of the approximate WCR algorithm, as well as the details for computing $\hat{\boldsymbol{\beta}}$ and the r-step estimate σ^2_{r-1} for σ^2.

5.1 WCR Estimates for Parameters

As a method for estimating the posterior mean for a function $t(\boldsymbol{\alpha}, \sigma)$ from (4), run the following steps:

1. Use the approximate WCR algorithm (see App. A) to generate a partition $\mathbf{p} = \{C_1, \ldots, C_{n(\mathbf{p})}\}$. This is a draw from the approximate WCR density $q(\mathbf{p})$ from the posterior of (4). Compute its importance weight, $\Lambda(\mathbf{p})$.
2. Given the current draw $\mathbf{p}$, compute σ^2_n (see App. A) using all of the data. Draw $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_{n(\mathbf{p})})$ from

$$\prod_{j=1}^{n(\mathbf{p})} \pi\big(d\mathbf{u}_j|C_j, \hat{\boldsymbol{\beta}}, \sigma^2_n\big) \propto \prod_{j=1}^{n(\mathbf{p})} H(d\mathbf{u}_j) \prod_{i\in C_j} \phi\big(\widehat{Y}_i|\mathbf{Z}_i'\mathbf{u}_j, \sigma^2_n\big).$$

   In particular, if H is a $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, then the $\mathbf{u}_j$ are conditionally independent $\mathrm{N}(\mathbf{m}_j, \mathbf{S}_j)$ normal random vectors, where $\mathbf{S}_j = (\boldsymbol{\Sigma}^{-1} + \sum_{i\in C_j} \mathbf{Z}_i\mathbf{Z}_i'/\sigma^2_n)^{-1}$ and $\mathbf{m}_j = \mathbf{S}_j(\sum_{i\in C_j} \widehat{Y}_i\mathbf{Z}_i/\sigma^2_n + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$.
3. The importance draw for σ^2 is σ^2_n, whereas the importance draw for $\boldsymbol{\alpha}$ is defined by the $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_{n(\mathbf{p})})$ and $\mathbf{p}$ just drawn. That is, $\boldsymbol{\alpha}_i = \mathbf{u}_j$ if and only if $i \in C_j$.
4. Run steps 1–3 independently B times, getting draws $(\mathbf{u}^{(b)}, \mathbf{p}^{(b)}, \sigma_n^{(b)})$ and importance weights $\Lambda(\mathbf{p}^{(b)})$. Compute $\boldsymbol{\alpha}^{(b)}$ from $(\mathbf{u}^{(b)}, \mathbf{p}^{(b)})$. To estimate $E[t(\boldsymbol{\alpha}, \sigma)|\mathbf{Y}]$, use

$$\frac{\sum_{b=1}^{B} t(\boldsymbol{\alpha}^{(b)}, \sigma_n^{(b)})\Lambda(\mathbf{p}^{(b)})}{\sum_{b=1}^{B} \Lambda(\mathbf{p}^{(b)})}.$$

Remark 2. For example, to estimate moments of random effects, one could use the Rao–Blackwell estimator discussed in Section 3. Thus, to compute $E[\int \mathbf{v}^k P(d\mathbf{v})|\mathbf{Y}]$, use

$$t(\boldsymbol{\alpha}, \sigma_n) = \frac{a_0}{a_0 + n} \int_{\Re^s} \mathbf{v}^k\, H(d\mathbf{v}) + \sum_{j=1}^{n(\mathbf{p})} \frac{e_j}{a_0 + n} E\big[\mathbf{U}_j^k|C_j, \sigma^2_n\big],$$

where $E[\mathbf{U}_j^k|C_j, \sigma^2_n]$ is the kth moment of a $\mathrm{N}(\mathbf{m}_j, \mathbf{S}_j)$ distribution.

5.2 Two-Slope Problem

We return to the two-slope problem introduced earlier in Section 2 [see (2) and (3)]. To test our algorithm, we simulated data from this model using a format similar to that used in Figure 1. Here group sizes were randomly generated with m, the size of group 1, selected from a binomial$(n, 1/2)$ distribution where $n = 1,000$. The covariates X_i were drawn from a uniform distribution on $[-3, 3]$. Random intercepts were drawn from a uniform two-point mixture distribution with values $\{-5\sigma^2/3, 2.5\sigma^2/3\}$, whereas random slopes were drawn from an exponential distribution with a mean of 2. The value for β_1 was then set at 2 to coincide with this mean. We set $\sigma^2 = 1$, which produces a separation in the modes of the distribution for $\alpha_{0,i} + \epsilon_i$ (a two-point normal mixture).

For the fixed-effects estimator, we used the OLS with random effects centered by their means. As discussed in Appendix A, this gives us a consistent estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, as well as a consistent estimator $\hat{\boldsymbol{\alpha}}_0$ for the random-effects means $\boldsymbol{\alpha}_0 = E(\boldsymbol{\alpha}_i)$. Following our general strategy, the WCR algorithm was applied to the residuals $\widehat{Y}_i$ computed from $\hat{\boldsymbol{\beta}}$. The simulations were repeated independently a total of 100 times. In each simulation, the WCR algorithm was applied with $B = 2{,}500$ iterations. In all examples, we used a $\mathrm{DP}(a_0H)$ prior where $a_0 = 1$ and H is a bivariate $\mathrm{N}(\mathbf{0}, A\mathbf{I})$ distribution with $A = 10$. The initial value for σ^2_0 was drawn from a uniform distribution on $[0, 3]$ (see App. A).

Table 1 records the results of the simulations. Values recorded for the mean, variance, and skewness for random

Table 1. Parameter Estimates From the Two-Slope Problem

	Parameter	True	WCR mean ± std dev	OLS mean ± std dev	REML mean ± std dev
$\alpha_{i,0}$	μ	−.42	−.40 ± .07	−.41 ± .10	−.40 ± .07
	var	1.56	1.75 ± .22		2.06 ± .39
	κ_3	0	.01 ± .25		
$\alpha_{i,1}$	μ	2	1.91 ± .13	2.03 ± .13	2.02 ± .12
	var	4	3.41 ± .55		4.08 ± .69
	κ_3	2	1.74 ± .33		
ϵ_i	σ^2	1	1.02 ± .19		.47 ± .34

NOTE: Parameters μ, var, and κ_3 are mean, variance, and skewness for random effects. Skewness is defined to equal 0 for normal distributions.

effects were obtained using Rao–Blackwell estimates for posterior means of moments. Various estimates based on OLS and REML were also computed. The entries for the mean and standard deviation for a specific estimator are determined by taking the mean and standard deviation of estimated values over the 100 independent Monte Carlo experiments.

Table 1 shows that WCR estimates for the means are reasonably close to values from OLS and REML, although the latter two are somewhat more accurate. This suggests that we can improve estimation for means by substituting the OLS estimates $\hat{\boldsymbol{\alpha}}_0$ (which are simpler to compute than the REML estimates) for $\boldsymbol{\mu}$ in the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution for H. We apply an extension of this technique when we consider longitudinal data in the following section. As expected, variance components for intercepts and measurement errors under REML are not estimated consistently. The variance for the slope is estimated well. For the WCR, we find that variance and skewness estimates are reasonably accurate. It identifies the symmetry in the two-point mixture distribution for intercepts and the positive skewness in the exponential slope distribution. The variance σ^2 for measurement error is well estimated.

6. LINEAR MIXED MODELS FOR LONGITUDINAL DATA

The WCR algorithm and our methods for the single measurement case extend naturally to linear mixed models subject to repeated measurements. A key extension applies to longitudinal data

$$Y_{i,t} = \mathbf{X}_{i,t}'\boldsymbol{\beta} + \mathbf{Z}_{i,t}'\boldsymbol{\alpha}_i + \epsilon_{i,t}, \qquad i = 1, \ldots, n, \quad t = 1, \ldots, m, \tag{9}$$

where $\epsilon_{i,t}$ are iid $\mathrm{N}(0, \sigma^2)$, $\mathbf{X}_{i,t}$ are p-dimensional time-varying fixed-effects covariates, and $\mathbf{Z}_{i,t}$ are s-dimensional time-varying random-effects covariates. There are now m observations for each random-effects term $\boldsymbol{\alpha}_i$. (In what follows, m can be allowed to depend on i with straightforward modification.)

Let $\boldsymbol{\eta}_i(\boldsymbol{\beta}) = (\mathbf{X}_{i,1}'\boldsymbol{\beta}, \ldots, \mathbf{X}_{i,m}'\boldsymbol{\beta})'$ and $\boldsymbol{\gamma}_i(\boldsymbol{\alpha}_i) = (\mathbf{Z}_{i,1}'\boldsymbol{\alpha}_i, \ldots, \mathbf{Z}_{i,m}'\boldsymbol{\alpha}_i)'$. Then (9) can be written as

$$\mathbf{Y}_i = \boldsymbol{\eta}_i(\boldsymbol{\beta}) + \boldsymbol{\gamma}_i(\boldsymbol{\alpha}_i) + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n, \tag{10}$$

where $\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,m})'$ and $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,m})'$. Analogously to (4), the repeated-measurement model (10) is recast as a Bayesian semiparametric hierarchical model,

$$\begin{aligned} (\mathbf{Y}_i|\boldsymbol{\alpha}_i, \boldsymbol{\beta}, \sigma^2) &\overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{\eta}_i(\boldsymbol{\beta}) + \boldsymbol{\gamma}_i(\boldsymbol{\alpha}_i), \sigma^2\mathbf{I}), \qquad i = 1, \ldots, n, \\ (\boldsymbol{\alpha}_i|P) &\overset{\text{iid}}{\sim} P, \\ P &\sim \mathcal{P} = \mathrm{DP}(a_0 H). \end{aligned} \tag{11}$$

It follows automatically that to estimate posterior quantities from (11), we can simply apply the previous WCR algorithms with responses $\mathbf{Y}_i$ in place of Y_i. As before, we apply a fast approximate version, taking advantage of available external point estimates for $\boldsymbol{\beta}$. Thus, analogously, we replace $\mathbf{Y}_i$ by residuals $\widehat{\mathbf{Y}}_i = \mathbf{Y}_i - \boldsymbol{\eta}_i(\hat{\boldsymbol{\beta}})$ for some estimator $\hat{\boldsymbol{\beta}}$. Our strategy is to use REML, rewriting (9) by subtracting by the mean for the random effects $\boldsymbol{\alpha}_0 = E(\boldsymbol{\alpha}_i)$, and then computing the REML estimates for the fixed effects $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_0$. Such estimates are known to be $\sqrt{n}$ consistent under minimal assumptions (Jiang 1998). We also computed REML estimates for the variance components for the centered random effects $\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_0$ and the variance σ^2 for the measurement error. Learning from the earlier simulations, we used the estimates for $\boldsymbol{\alpha}_0$ and variances of random effects to specify the mean and variance for the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution H (see Sec. 6.3 for further details). We also used the REML estimate $\hat{\sigma}^2$ for σ^2, although it is possible to estimate σ^2 within the WCR analogously to Section 5. Appendix B provides the details of the approximate WCR algorithm.

6.1 WCR Estimates for Parameters

To estimate the posterior mean of a function $t(\boldsymbol{\alpha})$:

1. Use the approximate WCR algorithm (App. B) to generate a partition $\mathbf{p} = \{C_1, \ldots, C_{n(\mathbf{p})}\}$. Compute its importance weight, $\Lambda(\mathbf{p})$.
2. Given the current draw $\mathbf{p}$, draw $\mathbf{u}_j$ for $j = 1, \ldots, n(\mathbf{p})$ independently from a $\mathrm{N}(\mathbf{m}_j, \mathbf{S}_j)$ distribution, where $\mathbf{S}_j = (\boldsymbol{\Sigma}^{-1} + \sum_{i\in C_j} \mathbf{M}_i\mathbf{M}_i'/\hat{\sigma}^2)^{-1}$, $\mathbf{m}_j = \mathbf{S}_j(\sum_{i\in C_j} \mathbf{M}_i\widehat{\mathbf{Y}}_i/\hat{\sigma}^2 + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$, and $\mathbf{M}_i = [\mathbf{Z}_{i,1}, \ldots, \mathbf{Z}_{i,m}]$.
3. Run steps 1 and 2 independently B times, getting draws $(\mathbf{u}^{(b)}, \mathbf{p}^{(b)})$ and importance weights $\Lambda(\mathbf{p}^{(b)})$. Compute $\boldsymbol{\alpha}^{(b)}$ from $(\mathbf{u}^{(b)}, \mathbf{p}^{(b)})$. To estimate $E[t(\boldsymbol{\alpha})|\mathbf{Y}_1, \ldots, \mathbf{Y}_n]$, use
$$\frac{\sum_{b=1}^B t(\boldsymbol{\alpha}^{(b)})\Lambda(\mathbf{p}^{(b)})}{\sum_{b=1}^B \Lambda(\mathbf{p}^{(b)})}.$$

Remark 3. Although so far we have focused exclusively on estimation for the random effects, an ad hoc approach can also be used for inference for $\boldsymbol{\beta}$. In the step 2 just discussed, we could additionally add the following:

2*. Given the draw $\mathbf{p}$ and $\mathbf{u}$, calculate $\boldsymbol{\alpha}$ and the residual values $\mathbf{Y}_{i,\boldsymbol{\alpha}} = \mathbf{Y}_i - \boldsymbol{\gamma}_i(\boldsymbol{\alpha}_i)$. Now estimate $\boldsymbol{\beta}$ from the linear regression model

$$\mathbf{Y}_{i,\boldsymbol{\alpha}} = \boldsymbol{\eta}_i(\boldsymbol{\beta}) + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n,$$

using OLS. This is our estimator for $\boldsymbol{\beta}$.

6.2 Simulations for Longitudinal Data

To test our algorithm, we simulated data from the longitudinal model

$$Y_{i,t} = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \alpha_{i,1} + (t - (m_i + 1)/2)\alpha_{i,2} + \epsilon_{i,t}, \qquad i = 1, \ldots, n, \quad t = 1, \ldots, m_i,$$

where the number of repeated observations, m_i, was drawn randomly from $\{1, \ldots, 13\}$ and $n = 275$. The sample size and number of repeated observations were chosen to roughly match the applied example of the following section. Fixed-effects covariates $X_{i,1}$ and $X_{i,2}$ were independently sampled from a uniform discrete distribution on $\{-3, -2, -1, 0, 1, 2, 3\}$ and a N(0,1) distribution. For random effects, intercept values $\alpha_{i,1}$ were drawn from a $\mathrm{N}(-1, \tau^2)$ distribution, and four different distributions for $\alpha_{i,2}$ were considered: (i) an exponential distribution with mean ρ, (ii) a uniform distribution on $[0, \rho]$, (iii) a normal distribution, $\mathrm{N}(1, \rho^2)$, (iv) a mixture of two normals, $.5\mathrm{N}(-2\rho, \rho^2) + .5\mathrm{N}(\rho, \rho^2)$. In all cases the value for ρ was selected so that the variance for the random slope distribution was equal to $2\sigma^2$, twice the variance for the errors ϵ_i. This ensured that the signal-to-noise ratio was equal to two. Similarly, we set $\tau^2 = 2\sigma^2$ for a signal-to-noise ratio of 2 for $\alpha_{i,1}$. For convenience we chose $\sigma^2 = 1$.

We followed the same strategy as in Section 5, running the WCR algorithm for $B = 2{,}500$ iterations. Means, variances, skewness, and kurtosis values for random effects, along with their standard errors, were computed using the methods outlined in Section 3. As before, we repeated the simulations (this time 250 times), recording the means and standard deviations of estimated values over the 250 experiments. In each of the examples we used a $\mathrm{DP}(a_0 H)$ prior, where $a_0 = 1$ and H was a bivariate $\mathrm{N}(\hat{\boldsymbol{\alpha}}_0, \widehat{\boldsymbol{\Sigma}})$ distribution with $\hat{\boldsymbol{\alpha}}_0$ taken to be the REML estimate for $\boldsymbol{\alpha}_0$ and $\widehat{\boldsymbol{\Sigma}}$ a diagonal matrix with diagonal values equal to three times the REML estimate of standard error for the corresponding random effect. For brevity, we report the results from only the simulations involving the exponential and two-point mixture slope distributions, because these are more difficult. Tables 2 and 3 record these values. We make the following general observations:

1. Point estimates for means and variances from WCR and REML are nearly identical in most cases. Standard errors for the mean were also similar, but standard errors for the variance were more accurate from WCR for the nonnormal slope distributions (including the uniform slope simulation, not shown).
2. Skewness and kurtosis were well estimated in all examples, including the two not shown. Corresponding standard errors were also good (agreeing quite closely with the standard deviations from the 250 replications), with only standard errors for kurtosis from the exponential slope not well estimated.
3. REML estimates for $\boldsymbol{\beta}$ are quite accurate and agree closely with the WCR ad hoc estimate. Standard errors from REML for fixed effects are also accurate—a decided advantage over the ad hoc method, which does not provide standard errors. Thus it seems questionable whether there can be any significant gain over using REML for fixed effects. From a practical perspective, a fully Bayesian inference for the fixed effects is not well motivated.

Table 2. Parameter Estimates Under an Exponential Random Slope Distribution for Longitudinal Data With $n = 275$ and $B = 2{,}500$ Iterations; Simulations Repeated 250 Times

	Parameter	*True*	*WCR mean ± std dev*	*REML mean ± std dev*
$\alpha_{i,1}$	μ	−1	−1.01 ± .11	−1.01 ± .11
	var	2	1.82 ± .23	2.01 ± .23
	κ_3	0	−.01 ± .22	
	κ_4	0	.29 ± .51	
	se(μ)		.10	.11
	se(var)		.22	.27
	se(κ_3)		.22	
	se(κ_4)		.53	
$\alpha_{i,2}$	μ	1.41	1.41 ± .08	1.41 ± .08
	var	2	2.00 ± .32	2.01 ± .32
	κ_3	2	1.91 ± .44	
	κ_4	6	5.29 ± 4.03	
	se(μ)		.08	.09
	se(var)		.30	.19
	se(κ_3)		.25	
	se(κ_4)		1.57	
ϵ_i	σ^2	1	.99 ± .04	
	se(σ^2)			.04
Fixed effects	β_1	1	1.00 ± .06	1.00 ± .06
	β_2	3	3.00 ± .11	3.00 ± .11
	se(β_1)			.05
	se(β_2)			.11

NOTE: Mean, variance, skewness and kurtosis are denoted by μ, var, κ_3, and κ_4. Skewness and kurtosis are defined to equal 0 for normal distributions. Variance estimate for ϵ_i are obtained from REML. Standard errors are denoted by "se."

Table 3. Parameter Estimates (similar to Table 2) Under a Two-Point Normal Mixture Slope Distribution

	Parameter	*True*	*WCR mean ± std dev*	*REML mean ± std dev*
$\alpha_{i,1}$	μ	−1	−.99 ± .12	−.99 ± .12
	var	2	1.77 ± .23	1.98 ± .23
	κ_3	0	−.01 ± .21	
	κ_4	0	.28 ± .46	
	se(μ)		.10	.11
	se(var)		.21	.27
	se(κ_3)		.22	
	se(κ_4)		.49	
$\alpha_{i,2}$	μ	−.39	−.39 ± .09	−.38 ± .09
	var	2	1.98 ± .13	1.99 ± .13
	κ_3	0	−.01 ± .10	
	κ_4	−.95	−.88 ± .13	
	se(μ)		.09	.09
	se(var)		.12	.19
	se(κ_3)		.10	
	se(κ_4)		.14	
ϵ_i	σ^2	1	.99 ± .03	
	se(σ^2)			.04
Fixed effects	β_1	1	.99 ± .05	.99 ± .05
	β_2	3	3.00 ± .09	3.00 ± .09
	se(β_1)			.05
	se(β_2)			.11

7. CHRONIC RENAL DISEASE

Here we illustrate the WCR method applied to data collected from the Modification of Diet in Renal Disease (MDRD) study, a longitudinal study investigating the effects of dietary protein restriction and strict blood pressure control on the rate of renal disease progression (Klahr et al. 1994). A key measurement in the study for evaluating renal disease progression are sequential measurements of glomerular filtration rate (GFR). We consider data on GFR measurements collected over 3.7 years starting from measurements collected 4 months after follow-up. The data that we considered consisted of $n = 571$ patients, with the treatment group (i.e., those

Table 4. Parameter Estimates From the Approximate WCR Algorithm Applied to the MDRD Data

		μ	var	κ_3	κ_4	se(μ)	se(var)	se(κ_3)	se(κ_4)
Normal protein	Intercept	42.20	156.75	.53	−.30	.79	13.60	.13	.32
	Slope	−.31	.15	−1.51	4.71	.03	.03	.38	1.85
	ϵ_i	0	17.22				.68		
Low protein	Intercept	40.76	143.55	.55	.74	.67	15.87	.19	.64
	Slope	−.22	.07	−1.40	3.94	.02	.02	.05	2.42
	ϵ_i	0	19.06				.77		

NOTE: Estimates for ϵ_i are obtained from REML. GFR measurements are divided by 10 here.

subject to a protein-restricted diet) comprising 284 patients (an average of 6.73 observations per subject) and the control group (i.e., those subject to a normal protein diet) comprising 287 patients (an average of 6.84 observations per subject).

Analysis of the MDRD data conducted after the study completion indicated that the distribution of the GFR slopes deviated from normality due to negative skewness and positive kurtosis. Such a conclusion was shared by Greene (2001), who analyzed the same data using a proportional-effects model. A comprehensive discussion of the data and detailed analyses can be found there. Here we consider a somewhat simpler scenario, fitting separate random slope-intercept models for each of the experimental groups. The results of the analysis are presented in Table 4 and Figures 4 and 5. Estimates for each experimental group are based on 10,000 sampled values from the WCR algorithm using the same strategy for fitting used in the previous section.

Table 4 confirms the negative skewness and positive kurtosis for the slope distributions. We find that the slope distribution for the treatment group is less skewed and has smaller kurtosis, signifying a complex effect of treatment. This same effect can also be seen in Figure 4 by comparing the left tails of the posterior slope densities (with density estimates based on the method of Sec. 3). Clearly the protein-restricted group has a thinner left tail, confirming that a restricted diet slows disease progression overall, but the extreme left-tail behavior also exhibits evidence of a substantial reduction in disease for the "progressors," that is, patients with large negative GFR slope and hence with severe disease progression. (See also the QQ plot in Figure 5, which clearly shows a difference in left-tail behavior.) It is interesting to note that an analysis based on only the first two moments, such as in a standard REML approach, would miss this important complex interaction occurring for progressors. Based on only the first two moments from Table 4, we could only conclude from the smaller mean value for the slope that there is an overall treatment effect due to diet.

Remark 4. To assess Monte Carlo efficiency of the WCR method, one can use the importance weights $\Lambda(\mathbf{p})$ to estimate the effective sample size, a measure of efficiency discussed by Kong et al. (1994). A simple calculation reveals that the effective sample sizes here are $B^* = 8{,}938$ for the control group and $B^* = 9{,}088$ for the treatment group (89% and 91% of the total Monte Carlo sample size $B = 10{,}000$). This high efficiency, reflected by the large effective sample sizes, means that the importance weights must be fairly evenly distributed. Some evidence of this can be seen in Figure 6, which plots the density of the log of the renormalized weights for both groups combined.

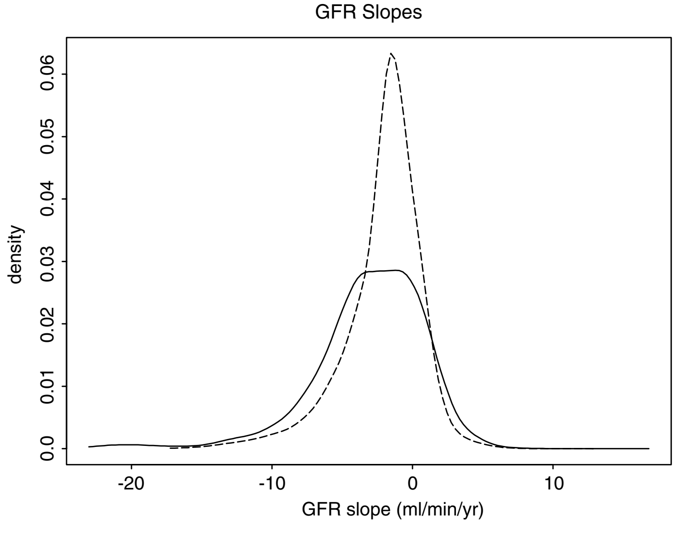

Figure 4. Posterior Mean Densities for Slopes of Normal (——) and Low (- - - -) Protein Groups (expressed in original measurements). Densities were evaluated over 250 grid points.

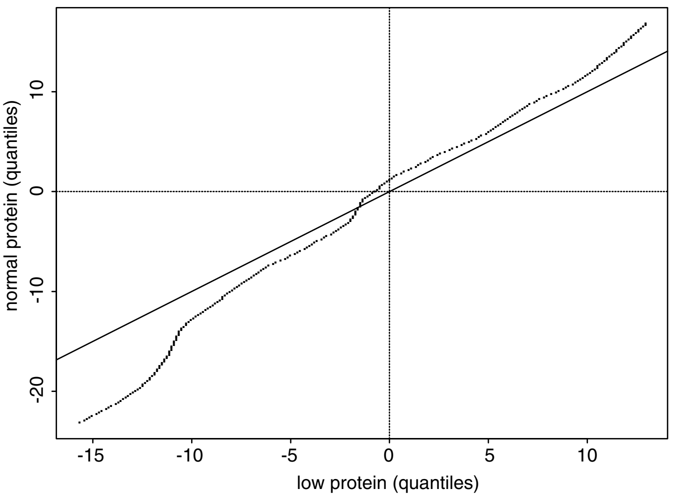

Figure 5. QQ Plot of Quantiles for Slope Distributions of the Normal and Low Protein Groups.

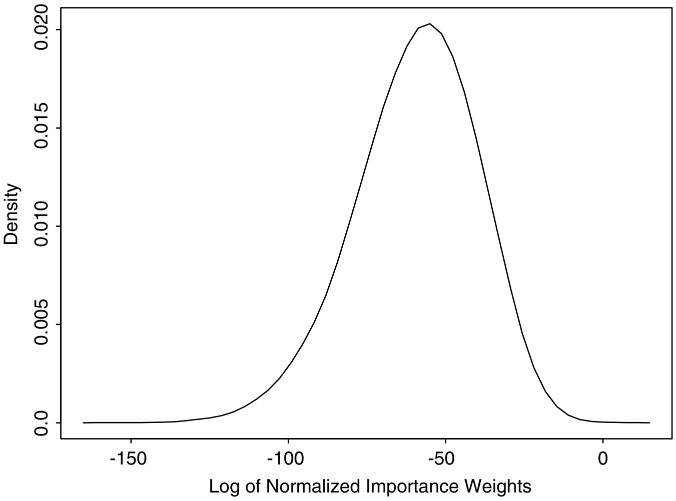

Figure 6. Log of Renormalized Importance Weights $\Lambda(\mathbf{p})$.

8. COMPUTATIONAL FEATURES

We end by summarizing several key computational features of our methods.

1. A key aspect forming the basis of the WCR's numerical stability, is the use of the partition structure underlying our sampling schemes. This naturally leads to lower Monte Carlo error due to a form of Rao–Blackwellization (see Sec. 3), thus improving on usual methods that work with hidden variables.
2. Sequential generation of a candidate partition in the WCR algorithm is simple to program and also is computationally efficient, because each step relies only on the current partition structure, which can be summarized in the form of simple linear and quadratic sufficient statistics (see Remarks A.1 in Appendix A and B.1 in Appendix B). This algorithm should be easy to implement in standard software packages, our model fitting was done in S-PLUS.
3. The use of external plug-in estimates reduces computations and minimizes the need for specifying hyperparameters, thus making programs fast and more easily automated. At the same time, their use leads to accurate and flexible inference for random effects.
4. Another nice feature is the iid nature of our algorithms. This avoids the obvious problem of convergence associated with MCMC methods (see Kleinman and Ibrahim 1998 for problems in linear mixed models) and also some side effects, such as the need to reparameterize and work with hierarchical centerings to improve Markov chain mixing (Gilks and Roberts 1996), practices that hinder overall automation.
5. The sequential and iid feature of our algorithms means that they are "interruptible," making it possible to easily update models based on new data. If a new data value Y_{n+1} arrives at some point in the future, then we simply run an $n+1$ step for each of the current partitions—where partition information is encoded using only simple sufficient statistics—thus allowing models to be updated without the need to rerun past data. Note that the number of sufficient statistics for each partition is a linear function of its cardinality, which is typically a small fraction of the sample size. Thus the notion of interruptibility can be applied when n is large.

APPENDIX A: APPROXIMATE WEIGHTED CHINESE RESTAURANT ALGORITHM FOR SINGLE-MEASUREMENT DATA

Plug-In Estimates

Before presenting the approximate WCR algorithm for (4), we first discuss how to compute our OLS plug-in estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$. We also give the details for computing our estimator σ^2_{r-1} for σ^2 used in step r of the algorithm. To compute $\hat{\boldsymbol{\beta}}$, rewrite the linear mixed model (1) as

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{Z}_i'\boldsymbol{\alpha}_0 + \xi_i, \tag{A.1}$$

where $\boldsymbol{\alpha}_0 = E(\boldsymbol{\alpha}_i)$ is the (unknown) mean for $\boldsymbol{\alpha}_i$ and $\xi_i = \mathbf{Z}_i'(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_0) + \epsilon_i$ are independent (but not identically distributed) errors with mean 0. Now compute the OLS estimate $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_0)$ for $(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)$ from (A.1). Both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}_0$ can be shown to be $\sqrt{n}$ consistent under mild assumptions for covariates by appealing to a triangular central limit theorem.

Our estimator for σ^2_{r-1} is also obtained using OLS. Recall that the current partition $\mathbf{p}_{r-1}$ at step r identifies $n(\mathbf{p}_{r-1})$ distinct random-effects values over the first $r-1$ observations, $\widehat{Y}_1, \ldots, \widehat{Y}_{r-1}$. We estimate these random effects using least squares and use the resulting mean squared error for σ^2_{r-1}. Such an estimator can also be interpreted as a maximum likelihood estimator under diffuse priors. For convenience, consider the case where $r = n+1$ (estimate based on all of the data). Then

$$\sigma_n^2 = \left[\sum_{\{j:e_j>s\}} (e_j - s)\right]^{-1} \sum_{\{j:e_j>s\}} \left[\left(\sum_{\{i:j\in C_j\}} \widehat{Y}_i\right)^2 - \mathbf{c}_j'\mathbf{B}_j^{-1}\mathbf{c}_j\right],$$

where $\mathbf{c}_j = \sum_{i\in C_j} \widehat{Y}_i\mathbf{Z}_i'$ and $\mathbf{B}_j = \sum_{i\in C_j} \mathbf{Z}_i\mathbf{Z}_i'$. Note the correction to the degrees of freedom, because σ_n^2 is based only on those clusters with at least $s+1$ members. This handles numerical problems that can arise for small clusters C_j.

Approximate WCR Algorithm

Assume a $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution for H. To get an approximate draw for $\mathbf{p}$ from the WCR density for (4), run the following steps:

Step 1: Compute $\hat{\boldsymbol{\beta}}$ and let σ_0^2 be an initial estimate for σ^2. Assign $\mathbf{p}_1 = \{1\}$ and let

$$\begin{aligned}\lambda(1) &= \int_{\Re^s} \phi(\widehat{Y}_1 | \mathbf{Z}_1'\mathbf{u}_1, \sigma_0^2) H(d\mathbf{u}_1)\\ &= \frac{\sigma_0^{s-1}|\boldsymbol{\Sigma}\boldsymbol{\Sigma}_1|^{-1/2}}{\sqrt{2\pi}}\\ &\quad \times \exp\left[-\frac{1}{2\sigma_0^2}\widehat{Y}_1^2 + \frac{1}{2\sigma_0^2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right],\end{aligned}$$

where $\boldsymbol{\Sigma}_1 = \sigma_0^2\boldsymbol{\Sigma}^{-1} + \mathbf{Z}_1\mathbf{Z}_1'$ and $\boldsymbol{\mu}_1 = \sigma_0^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \widehat{Y}_1\mathbf{Z}_1$.

Step r: Compute σ^2_{r-1} from $\widehat{Y}_1, \ldots, \widehat{Y}_{r-1}$ and $\mathbf{p}_{r-1}$. Create $\mathbf{p}_r$ by assigning label r to a new set with probability

$$\begin{aligned}&\frac{a_0}{(a_0 + r - 1)\lambda(r)} \times \frac{\sigma_{r-1}^{s-1}|\boldsymbol{\Sigma}\boldsymbol{\Sigma}_r|^{-1/2}}{\sqrt{2\pi}}\\ &\qquad \times \exp\left[-\frac{1}{2\sigma_{r-1}^2}\widehat{Y}_r^2 + \frac{1}{2\sigma_{r-1}^2}\boldsymbol{\mu}_r'\boldsymbol{\Sigma}_r^{-1}\boldsymbol{\mu}_r - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right],\end{aligned}$$

where $\boldsymbol{\Sigma}_r = \sigma_{r-1}^2\boldsymbol{\Sigma}^{-1} + \mathbf{Z}_r\mathbf{Z}_r'$ and $\boldsymbol{\mu}_r = \sigma_{r-1}^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \widehat{Y}_r\mathbf{Z}_r$. Otherwise, assign label r to an existing set $C_{j,\,r-1}$ with probability

$$\frac{e_{j,r-1}}{(a_0+r-1)\lambda(r)} \times \frac{|\boldsymbol{\Sigma}_{r,j}|^{1/2}}{\sigma_{r-1}\sqrt{2\pi}|\boldsymbol{\Sigma}_{r,j}^*|^{1/2}} \times \exp\left[-\frac{1}{2\sigma_{r-1}^2}\left(\widehat{Y}_r^2 - (\boldsymbol{\mu}_{r,j}^*)'(\boldsymbol{\Sigma}_{r,j}^*)^{-1}\boldsymbol{\mu}_{r,j}^* + \boldsymbol{\mu}_{r,j}'\boldsymbol{\Sigma}_{r,j}^{-1}\boldsymbol{\mu}_{r,j}\right)\right],$$

where

$$\boldsymbol{\Sigma}_{r,\,j} = \sigma_{r-1}^2\boldsymbol{\Sigma}^{-1} + \sum_{i\in C_{j,\,r-1}} \mathbf{Z}_i\mathbf{Z}_i'$$

and

$$\boldsymbol{\mu}_{r,\,j} = \sigma_{r-1}^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \sum_{i\in C_{j,\,r-1}} \widehat{Y}_i\mathbf{Z}_i,$$

whereas $\boldsymbol{\Sigma}_{r,\,j}^* = \boldsymbol{\Sigma}_{r,\,j} + \mathbf{Z}_r\mathbf{Z}_r'$ and $\boldsymbol{\mu}_{r,\,j}^* = \boldsymbol{\mu}_{r,\,j} + \widehat{Y}_r\mathbf{Z}_r$. As before, $\lambda(r)$ is the appropriate normalizing constant.

Approximate draw for $\mathbf{p}$: Run step 1 followed by step r for $r = 2, \ldots, n$. This gives a draw $\mathbf{p} = \{C_1, \ldots, C_{n(\mathbf{p})}\}$, which is a random partition of $\{1, \ldots, n\}$ with approximate WCR density $q(\mathbf{p})$ from the posterior of (4). Its importance weight is $\Lambda(\mathbf{p}) = \lambda(1) \times \cdots \times \lambda(n)$.

Remark A.1. A key feature of the WCR algorithm is that it depends only on simple linear and quadratic expressions of the data such as $\sum_{i\in C_{j,\,r-1}} \widehat{Y}_i\mathbf{Z}_i$ and $\sum_{i\in C_{j,\,r-1}} \mathbf{Z}_i\mathbf{Z}_i'$. By *building these values up as the algorithm proceeds*, we can considerably reduce overall computations. Consider, for example, what happens at the end of the rth iteration of the algorithm: We assign label r to either a new set or a previous set $C_{j,\,r-1}$. In the second case, the cluster associated with $C_{j,\,r-1}$ (which now becomes $C_{j,\,r}$) is expanded to include the new value $\widehat{Y}_r$ and its covariate $\mathbf{Z}_r$. Thus, to move to step $r+1$, the only bookkeeping required is in updating the sufficient statistics $\sum_{i\in C_{j,\,r-1}} \widehat{Y}_i\mathbf{Z}_i$ and $\sum_{i\in C_{j,\,r-1}} \mathbf{Z}_i\mathbf{Z}_i'$ for this set. Similarly, if r is assigned to a new set, then the only bookkeeping required is to introduce a new set of sufficient statistics for the newly created cluster corresponding to $\widehat{Y}_r$ and its covariate $\mathbf{Z}_r$.

Remark A.2. We recommend the use of a "shuffle" step. This step is introduced at the start of the WCR algorithm and involves permuting the data and covariates randomly. Thus, draw a random permutation $(i_1, \ldots, i_n)$ of $\{1, \ldots, n\}$ and apply the algorithm to the data $\widehat{Y}_{i_1}, \ldots, \widehat{Y}_{i_n}$ and covariates $\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_n}$ in the order of the permutation. This reduces the data order dependence of the WCR algorithm.

APPENDIX B: APPROXIMATE WEIGHTED CHINESE RESTAURANT ALGORITHM FOR LONGITUDINAL DATA

Let $\phi(\cdot|\boldsymbol{\gamma}_i(\mathbf{u}_i), \sigma^2)$ denote a multivariate normal density with variance $\sigma^2\mathbf{I}$ and mean $\boldsymbol{\gamma}_i(\mathbf{u}_i) = (\mathbf{Z}_{i,1}'\mathbf{u}_i, \ldots, \mathbf{Z}_{i,m}'\mathbf{u}_i)'$. For an approximate draw for $\mathbf{p}$ from the WCR density for (11), run the following steps:

Step 1: Compute $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. Assign $\mathbf{p}_1 = \{1\}$ and let

$$\begin{aligned}\lambda(1) &= \int_{\Re^s} \phi\left(\widehat{\mathbf{Y}}_1|\boldsymbol{\gamma}_1(\mathbf{u}_1), \hat{\sigma}^2\right)H(d\mathbf{u}_1)\\ &= \frac{\hat{\sigma}^{s-m}|\boldsymbol{\Sigma}\boldsymbol{\Sigma}_1|^{-1/2}}{(2\pi)^{m/2}}\\ &\quad\times\exp\left[-\frac{1}{2\hat{\sigma}^2}\widehat{\mathbf{Y}}_1'\widehat{\mathbf{Y}}_1 + \frac{1}{2\hat{\sigma}^2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right],\end{aligned}$$

where $\boldsymbol{\Sigma}_1 = \hat{\sigma}^2\boldsymbol{\Sigma}^{-1} + \mathbf{M}_1\mathbf{M}_1'$, $\boldsymbol{\mu}_1 = \hat{\sigma}^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{M}_1\widehat{\mathbf{Y}}_1$, and $\mathbf{M}_1 = [\mathbf{Z}_{1,\,1}, \ldots, \mathbf{Z}_{1,\,m}]$ is the $s \times m$ matrix of random-effects covariates for $i = 1$.

Step r: Create $\mathbf{p}_r$ by assigning label r to a new set with probability

$$\begin{aligned}&\frac{a_0}{(a_0+r-1)\lambda(r)} \times \int_{\Re^s} \phi\left(\widehat{\mathbf{Y}}_r|\boldsymbol{\gamma}_r(\mathbf{u}_r), \hat{\sigma}^2\right)H(d\mathbf{u}_r)\\ &\quad= \frac{a_0}{(a_0+r-1)\lambda(r)} \times \frac{\hat{\sigma}^{s-m}|\boldsymbol{\Sigma}\boldsymbol{\Sigma}_r|^{-1/2}}{(2\pi)^{m/2}}\\ &\quad\quad\times\exp\left[-\frac{1}{2\hat{\sigma}^2}\widehat{\mathbf{Y}}_r'\widehat{\mathbf{Y}}_r + \frac{1}{2\hat{\sigma}^2}\boldsymbol{\mu}_r'\boldsymbol{\Sigma}_r^{-1}\boldsymbol{\mu}_r - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right],\end{aligned}$$

where $\boldsymbol{\Sigma}_r = \hat{\sigma}^2\boldsymbol{\Sigma}^{-1} + \mathbf{M}_r\mathbf{M}_r'$, $\boldsymbol{\mu}_r = \hat{\sigma}^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{M}_r\widehat{\mathbf{Y}}_r$, and $\mathbf{M}_r = [\mathbf{Z}_{r,\,1}, \ldots, \mathbf{Z}_{r,\,m}]$. Alternatively, assign label r to a previous set $C_{j,\,r-1}$ with probability

$$\begin{aligned}&\frac{e_{j,\,r-1}}{(a_0+r-1)\lambda(r)}\\ &\quad\times\int_{\Re^s} \phi\left(\widehat{\mathbf{Y}}_r|\boldsymbol{\gamma}_r(\mathbf{u}_{j,\,r-1}), \hat{\sigma}^2\right)\pi\left(d\mathbf{u}_{j,\,r-1}|C_{j,\,r-1}, \hat{\sigma}^2\right)\\ &\quad= \frac{e_{j,\,r-1}}{(a_0+r-1)\lambda(r)} \times \frac{|\boldsymbol{\Sigma}_{r,\,j}|^{1/2}}{\hat{\sigma}^m(2\pi)^{m/2}|\boldsymbol{\Sigma}_{r,\,j}^*|^{1/2}}\\ &\quad\quad\times\exp\left[-\frac{1}{2\hat{\sigma}^2}\left(\widehat{\mathbf{Y}}_r'\widehat{\mathbf{Y}}_r - (\boldsymbol{\mu}_{r,\,j}^*)'(\boldsymbol{\Sigma}_{r,\,j}^*)^{-1}\boldsymbol{\mu}_{r,\,j}^*\right.\right.\\ &\quad\quad\quad\left.\left.+\boldsymbol{\mu}_{r,\,j}'\boldsymbol{\Sigma}_{r,\,j}^{-1}\boldsymbol{\mu}_{r,\,j}\right)\right],\end{aligned}$$

where

$$\boldsymbol{\Sigma}_{r,\,j} = \hat{\sigma}^2\boldsymbol{\Sigma}^{-1} + \sum_{i\in C_{j,\,r-1}} \mathbf{M}_i\mathbf{M}_i',$$

$$\boldsymbol{\mu}_{r,\,j} = \hat{\sigma}^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \sum_{i\in C_{j,\,r-1}} \mathbf{M}_i\widehat{\mathbf{Y}}_i,$$

$\boldsymbol{\Sigma}_{r,\,j}^* = \boldsymbol{\Sigma}_{r,\,j} + \mathbf{M}_r\mathbf{M}_r'$, and $\boldsymbol{\mu}_{r,\,j}^* = \boldsymbol{\mu}_{r,\,j} + \mathbf{M}_r\widehat{\mathbf{Y}}_r$. As before, $\lambda(r)$ is the appropriate normalizing constant.

Approximate draw for $\mathbf{p}$: Run step 1 followed by step r, for $r = 2, \ldots, n$. This gives an approximate draw $\mathbf{p} = \{C_1, \ldots, C_{n(\mathbf{p})}\}$ from the WCR density $q(\mathbf{p})$ for the posterior of (11). Its importance weight is $\Lambda(\mathbf{p}) = \lambda(1) \times \cdots \times \lambda(n)$.

Remark B.1. Note again that the update rule is based on simple linear and quadratic expressions of the data. As before, we can reduce overall computations considerably by keeping track of these values. Here it suffices to update the values $\sum_{i\in C_{j,\,r-1}} \mathbf{M}_i\mathbf{M}_i'$ and $\sum_{i\in C_{j,\,r-1}} \mathbf{M}_i\widehat{\mathbf{Y}}_i$ at the end of each iteration r for only the set $C_{j,\,r-1}$ corresponding to label r or, if a new set is created, doing the necessary initial bookkeeping for it.

Remark B.2. A shuffle step similar to that in Remark A.2 can be used to reduce data dependence.

[Received November 2001. Revised March 2002.]

REFERENCES

Aitkin, M. (1999), "A general Maximum Likelihood Analysis of Variance Components in Generalized Linear Models," *Biometrics*, 55, 117–128.

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

Brunner, L. J., Chan, A. T., James, L. F., and Lo, A. Y. (2001), "Weighted Chinese Restaurant Processes and Bayesian Mixture Models," unpublished manuscript.

Bush, C. A., and MacEachern, S. N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275–285.

Butler, S. M., and Louis, T. A. (1992), "Random Effects Models With Non-Parametric Priors," *Statistics in Medicine*, 11, 1981–2000.

Escobar, M. D. (1988), "Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means," unpublished doctoral dissertation, Yale University.

——— (1994), "Estimating Normal Means With a Dirichlet Process Prior," *J. Amer. Stat. Assoc.*, 89, 268–277.

Escobar, M. D., and West, M. (1998), "Computing Nonparametric Hierarchical Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Mueller, and D. Sinha, New York: Springer, pp. 1–22.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Ann. Statist.*, 1, 209–230.

——— (1974), "Prior Distributions on Spaces of Probability Measures," *Ann. Statist.*, 2, 615–629.

Ferguson, T. S., Phadia, E. G., and Tiwari, R. C. (1992), "Bayesian Nonparametric Inference," in *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, Hayward, CA: Institute of Mathematical Statistics, pp. 127–150.

Gilks, W. R., and Roberts, G. O. (1996), "Strategies for Improving MCMC," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 89–114.

Greene, T. (2001), "A Model for a Proportional Treatment Effect on Disease Progression," *Biometrics*, 57, 354–360.

Ishwaran, H., and James, L. F. (2000), "Generalized Weighted Chinese Restaurant Processes for Species Sampling Mixture Models," unpublished manuscript.

——— (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *J. Amer. Stat. Assoc.*, 96, 161–173.

Ishwaran, H., James, L. F., and Lo, A. Y. (2001), "Generalized Weighted Chinese Restaurant and SIS Stick-Breaking Algorithms for Semiparametric Models," unpublished manuscript.

Ishwaran, H., James, L. F., and Sun, J. (2001), "Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions," *J. Amer. Stat. Assoc.*, 96, 1316–1332.

Ishwaran, H., and Zarepour, M. (2002), "Exact and Approximate Sum-Representations for the Dirichlet Process," *Canad. J. of Statist*, to appear.

Jiang, J. (1996), "REML Estimation: Asymptotic Behavior and Related Topics," *The Annals of Statistics*, 24, 255–286.

——— (1998), "Asymptotic Properties of the Empirical BLUP and BLUE in Mixed Linear Models," *Statistica Sinica*, 8, 861–885.

Kong, A., Liu, J. S., and Wong, W. H. (1994), "Sequential Imputations and Bayesian Missing Data Problems," *J. Amer. Stat. Assoc.*, 89, 278–288.

Klahr, S., Levey, A., Beck, G., Caggiula, A., Hunsikcer, L., Kusek J., and Striker, G. (1994), "The Effects of Dietary Protein Restriction and Blood Pressure Control on the Progression of Chronic Renal Disease," *New England Journal of Medicine*, 330, 877–884.

Kleinman, K. P., and Ibrahim, J. G. (1998), "A Semiparametric Bayesian Approach to the Random Effects Model," *Biometrics*, 54, 921–938.

Laird, N. M., and Ware, J. H. (1982), "Random Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

Liu, J. S. (1996), "Nonparametric Hierarchical Bayes via Sequential Imputations," *Ann. Statist.*, 24, 911–930.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *Ann. Statist.*, 12, 351–357.

Lo, A. Y., Brunner, L. J., and Chan, A. T. (1996), "Weighted Chinese Restaurant Processes and Bayesian Mixture Models," Research Report 1, Hong Kong University of Science and Technology.

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999), "Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation," *Canadian J. Statist.*, 27, 251–267.

Quintana, F. A. (1998), "Nonparametric Bayesian Analysis for Assessing Homogeneity in $k \times l$ Contingency Tables With Fixed Right Margin Totals," *J. Amer. Stat. Assoc.*, 93, 1140–1149.

Quintana, F. A., and Newton, M. A. (2000), "Computational Aspects of Nonparametric Bayesian Analysis With Applications to the Modeling of Multiple Binary Sequences," *J. Comp. Graph. Statist.*, 9, 711–737.

Richardson, A. M., and Welsh, A. H. (1994), "Asymptotic Properties of Restricted Maximum Likelihood (REML) Estimates for Hierarchical Mixed Linear Models," *Austral. J. Statist.*, 36, 31–43.

Tao, H., Palta, M., Yandell, B. S., and Newton, M. A. (1999), "An Estimation Method for the Semiparametric Mixed Effects Model," *Biometrics*, 55, 102–110.

Verbeke, G., and Lesaffre, E. (1996), "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population," *J. Amer. Stat. Assoc.*, 91, 217–221.

Wasserman, L. (2000), "Asymptotic Inference for Mixture Models by Using Data-Dependent Priors," *J. Royal Statist. Soc.*, Ser. B, 62, 159–180.

West, M., Müller, P., and Escobar, M. D. (1994), "Hierarchical Priors and Mixture Models, With Applications in Regression and Density Estimation," in *A Tribute to D. V. Lindley*, eds. A. F. M. Smith and P. R. Freeman, New York: Wiley and Sons.

Zhang, D., and Davidian, M. (2002), "Linear Mixed Models With Flexible Distributions of Random Effects for Longitudinal Data," *Biometrics*, to appear.