Sankhyā : The Indian Journal of Statistics
Special Issue on Bayesian Analysis
1998, Volume spl, Series , Pt. , pp. 19-30

A NOTE ON FIRST-STAGE APPROXIMATION IN TWO-STAGE HIERARCHICAL MODELS*

By MICHAEL J. DANIELS
Iowa State University, Ames
and
ROBERT E. KASS
Carnegie Mellon University, Pittsburgh

SUMMARY. We consider approximations to two-stage hierarchical models in which the second stage uses a Normal distribution to model the variation of the first-stage parameters. If we replace the first stage of the model with a Normal distribution based on first-stage maximum likelihood estimation, we obtain an alternative two-stage model that approximates the original model while allowing posterior simulation to become easy and efficient. We note that the MLE-based Normal approximation is not quite a special case of Laplace's method, but it does produce the same accuracy as Laplace's method in approximating the posterior of the second-stage parameters. In a previous paper we showed how draws from such approximate posteriors may be reweighted to produce importance samples from the original posterior. Here we show how the method extends to mixed models, and hierarchical nonlinear models. We demonstrate the possible utility of this kind of scheme by easily obtaining posterior inferences (without special-purpose MCMC code) for a model that could not, at the time of our writing, be fit by BUGS (Spiegelhalter, Best, Gilks, Inskip, 1996).

1. **Introduction**

In hierarchical models, it is very common to use a Normal distribution at the second stage to model variation of first-stage parameters. Thus, for example, we may have a model of the form :

Two-Stage Model

Stage one: $Y_i|\theta_i \sim f(y|\theta_i)$, $i = 1\ldots,k$, independently, with Y_i an $n_i \times 1$ dimensional vector and θ_i an m-dimensional vector;

Stage two: $\theta_i|\mu, D \sim N_m(\mu, D), i.i.d.$,

AMS (1991) *subject classification.* 62C10
Key words and phrases. Laplace's method, MCMC, importance sampling.
*This work was supported by grants from the National Science Foundation and the National Institutes of Mental Health.

for some given density f. If the density f is itself Normal with mean that depends linearly on θ_i then, using conjugate priors on μ and D, Gibbs sampling and other simulation methods are fast and easily implemented (Gilks, Wang, Yvonnet, and Coursaget 1993; Spiegelhalter, Best, Gilks, and Inskip, 1996; Everson and Morris, 1997). In this paper we are concerned with the situation in which either f is non-Normal, or the mean of Y_i is nonlinear in θ_i, or both. Such models arise frequently in practice. We investigate the use of approximation at the first stage, possibly in conjunction with an importance reweighting scheme that produces draws from the posterior for the original hierarchical model itself.

Our motivation is twofold. First, we have come across situations in which it might be too time-consuming to implement MCMC in a full-fledged two-stage model. For example, in applied work currently in progress, we have a multiparameter non-homogeneous Poisson process that is observed (as the first stage of the model) in a 4×2 design among each of roughly 300 individuals and a primary interest is to estimate the second-stage, or "population" parameters. In such problems, although we may not wish to spend the time implementing MCMC for the relevant two-stage model, it is not too hard to begin by using maximum likelihood estimation and then replace the original observations with their first-stage maximum likelihood estimates together with the information matrices; we may then assume Normality of the MLE's and obtain a familiar Normal conjugate model as an approximation:

MLE-Based Normal Approximation

Stage one: $\hat{\theta}_i|\theta_i \sim N_m(\theta_i, I(\hat{\theta}_i)^{-1})$, $i = 1 \ldots, k$, independently, where $I(\hat{\theta})$ is the observed information matrix (the negative Hessian of the loglikelihood function at the MLE);

Stage two: $\theta_i|\mu, D \sim N_m(\mu, D)$, i.i.d.

We consider the MLE-based Normal approximation to be of practical interest because posterior distributions based on it are so easy to generate (using conjugate priors on μ and D). Furthermore, it is plausible that the MLE-based Normal approximation would often be quite adequate for inferences about second-stage parameters because it essentially involves a version of Laplace's method applied to each of the k integrals in the likelihood function

$$L(\mu, D) = \prod_{i=1}^{k} \int f(y_i|\theta_i) n(\theta_i; \mu, D) d\theta_i$$

where $n(x; \mu, D)$ is the multivariate Normal density with mean μ and variance D evaluated at x. In Section 2 we examine this remark more closely. We show that the MLE-based Normal approximation is not quite an instance of Laplace's method, but it does have the same asymptotic accuracy. We also suggest that the adequacy of the Normal approximation to the first-stage likelihood can be checked using a diagnostic derived and discussed by Kass and Slate (1994). We assume throughout that the first-stage likelihood has a single, dominant peak

for all $i = 1, \ldots, k$. (That is, if any likelihood is multimodal, only the dominant mode contributes substantially to its integral.)

Our second motivation here is that, having obtained posterior distributions based on an approximate model, we can easily reweight the sample to obtain a posterior sample from the original model. This reweighting was used by Daniels and Kass (1997) in posterior simulation with nonconjugate priors on covariance matrices. In Section 3 we describe this importance weighting scheme and extend it to generalized linear mixed models and nonlinear hierarchical models.

In Section 4 we refit a nonlinear hierarchical model to previously-analyzed data by (i) computing first-stage MLE's, (ii) using BUGS (Spiegelhalter, Best, Gilks, Inskip, 1996) to obtain draws from the approximating hierarchical model, and (iii) using S-PLUS to reweight those draws, thereby obtaining draws from the desired posterior. As of the writing of this paper, BUGS could not be used on its own to accomplish this fully Bayesian analysis. We add a few further comments in Section 5.

Although Daniels and Kass (1997) used the importance sampling method we discuss here in Section 3 for posterior simulation with non-conjugate priors, for simplicity, throughout this paper we assume we have a prior on (μ, D) that is conjugate to the Normal second stage of the model. Our interest here is in applying (via approximation and, possibly, importance reweighting) the easy and efficient Gibbs sampling engine that drives Normal conjugate posterior simulation.

2. First-Stage Normal Approximation

We now consider formally the MLE-based Normal approximation displayed in the Introduction: we would like to know how well the posterior distribution on (μ, D) obtained using the MLE-based Normal approximation approximates the posterior distribution obtained using the original two-stage model. This amounts to asking how well the likelihood function $L(\mu, D)$ in the Introduction is approximated by the corresponding likelihood from the MLE-based approximation, which we will write as $\hat{L}(\mu, D)$.

Intuitively, for fixed (μ, D), we are substituting a Normal distribution that approximates the posterior of θ, conditionally on (μ, D), for the exact conditional posterior. This is very similar to Laplace's method (see Kass, 1997, for references; also see Breslow and Clayton, 1993), but is not quite a special case. We now elucidate this distinction. The argument here is heuristic. Conditions for validity of such expansions may be found in Kass, Tierney, and Kadane (1990; see also Kass and Vos, 1997, Section 3.6).

Let us begin by writing the likelihood function as

$$L(\mu, D) = \prod I_i$$

where

$$I_i = \int L_i(\theta_i) n(\theta_i; \mu, D) d\theta_i,$$

and $L_i(\theta) = f(y_i|\theta)$. As shown in Theorem 1 below, making use of the Normality of the second-stage of the MLE-based Normal approximation, we obtain the approximation

$$\hat{I}_i = K(\hat{\theta}_i) \frac{1}{(2\pi)^{m/2}} \frac{1}{|D + I(\hat{\theta}_i)^{-1}|^{1/2}} \exp(-.5(\hat{\theta}_i - \mu)'(D + I(\hat{\theta}_i)^{-1})^{-1}(\hat{\theta}_i - \mu))$$

where $\hat{\theta}_i$ is the maximum likelihood estimate of θ_i and $K(\hat{\theta}_i) = L_i(\hat{\theta}_i)/n(\hat{\theta}_i; \hat{\theta}_i, I(\hat{\theta}_i)^{-1})$ does not depend on μ and D and thus will not affect the likelihood function $L(\mu, D)$. On the other hand, Laplace's method for approximating the integral I_i gives

$$I_i = \tilde{I}_i \left(1 + O(n_i^{-1})\right)$$

with

$$\tilde{I}_i = (2\pi)^{m/2} |\tilde{\Sigma}|^{1/2} L_i(\tilde{\theta}_i) n(\tilde{\theta}_i; \mu, D)$$

where $\tilde{\Sigma}^{-1}$ is the negative second-derivative matrix of $\log(L_i(\theta_i)n(\theta_i; \mu, D))$ evaluated at $\tilde{\theta}_i$, the maximizer of $L_i(\theta_i)n(\theta_i; \mu, D)$ conditionally on μ and D. Compared to $\tilde{I}_i$ the approximation $\hat{I}_i$ does two things: (i) it replaces $\tilde{\theta}_i$ in $\tilde{I}_i$ with the posterior mode obtained from the first-stage Normal approximation, and then (ii) it replaces the first-stage posterior evaluated at $\tilde{\theta}_i$ with the modal value of the posterior obtained from the first-stage Normal approximation. These two replacements, together, incur an error of order $O(n_i^{-1})$. Thus, as one might expect and as we next prove, the resulting approximation has the same asymptotic accuracy as Laplace's method itself; the virtue of the MLE-based Normal approximation is its simplicity, allowing immediate implementation of posterior simulation via Gibbs sampling.

THEOREM 1. *The MLE-based Normal approximation produces a product* $\hat{L}(\mu, D) = \prod_i \hat{I}_i$ *satisfying*

$$L(\mu, D) = \hat{L}(\mu, D) \left(1 + O(\sum \frac{1}{n_i})\right).$$

PROOF. For simplicity we consider the one-dimensional case and omit the subscript on θ_i. Let $nh(\theta)$ denote the log likelihood under the true model, $nh_N(\theta)$ denote the log likelihood under the Normal approximation, and $\rho(\theta)$ denote the logarithm of the Normal prior of θ with mean μ and covariance matrix D. We are interested in an integral of the form

$$I = \int \exp(nh(\theta) + \rho(\theta)) d\theta.$$

Let $\psi(\theta) = nh_N(\theta) + \rho(\theta)$ and consider an expansion about $\hat{\theta}$ of the exponential of $nh(\theta) + \rho(\theta) - \psi(\theta) - (nh(\hat{\theta}) + \rho(\hat{\theta}) - \psi(\hat{\theta}))$. The function $\exp(\psi(\theta))$, when rewritten, forms the product of two Normal densities, $n(\theta; \tilde{\theta}, \tilde{\Sigma})$ with mean $\tilde{\theta} = (D^{-1} + I(\hat{\theta}))^{-1}(D^{-1}\hat{\theta} + I(\hat{\theta})\mu)$ and covariance matrix, $\tilde{\Sigma} = (D^{-1} + I(\hat{\theta}))^{-1}$ and $n(\hat{\theta}; \mu, D + I(\hat{\theta})^{-1})$. Thus, combining this with the fact that $(nh(\hat{\theta}) + \rho(\hat{\theta}) - \psi(\hat{\theta})) = \log(K(\hat{\theta}))$,

$$I = \int \hat{I} \cdot n(\theta; \tilde{\theta}, \tilde{\Sigma})[1 + \frac{1}{6}(\theta - \hat{\theta})^3 nh^{(3)}(\hat{\theta}) + \cdots]d\theta$$

where $\cdots$ indicate higher-order terms. In the standard Laplace approximation, the odd powers would cancel out. However, here we are centering θ at the MLE $\hat{\theta}$, which is different than the mode $\tilde{\theta}$ of the Normal distribution. We now rewrite the term $(\theta - \hat{\theta})^3$,

$$\begin{aligned}(\theta - \hat{\theta})^3 &= (\theta - \tilde{\theta} + \tilde{\theta} - \hat{\theta})^3 \\ &= (\theta - \tilde{\theta})^3 + (\tilde{\theta} - \hat{\theta})^3 + 3(\theta - \tilde{\theta})^2(\tilde{\theta} - \hat{\theta}) + 3(\theta - \tilde{\theta})(\tilde{\theta} - \hat{\theta})^2.\end{aligned}$$

The integral of the first and last terms will drop out and the integral becomes,

$$\begin{aligned}I &= \int \hat{I} \cdot n(\theta; \tilde{\theta}, \tilde{\Sigma})\{1 + \frac{1}{6}h^{(3)}(\hat{\theta})[(\tilde{\theta} - \hat{\theta})^3 + 3(\theta - \tilde{\theta})^2(\tilde{\theta} - \hat{\theta})] + \cdots\}d\theta \\ &= \hat{I} \cdot \{1 + \frac{1}{6}nh^{(3)}(\hat{\theta})[(\tilde{\theta} - \hat{\theta})^3 + 3(\tilde{\theta} - \hat{\theta})\tilde{\Sigma}] + \cdots\}\end{aligned}$$

where $\hat{I}$ is the approximation to the integral and $\cdots$ indicate higher-order terms. Since $\tilde{\theta} - \hat{\theta} = O(1/n)$ (by Laplace's method, they each furnish order $O(1/n)$ approximations to posterior mean of θ, see, e.g., Kass, Tierney, and Kadane, 1990), the first term in square brackets will be $O(1/n^3)$ and the second term will be $O(1/n^2)$. Combining this with the term outside the brackets being $O(n)$, the overall error here will be $O(1/n)$. Finally, combining the approximations for each θ_i, the overall error will be $O(\sum 1/n_i)$. □

REMARK 1. Integrating out the Normal approximations for the θ_i's may provide an accurate approximation $\hat{L}(\mu, D)$ to the likelihood function $L(\mu, D)$. In particular, the error in the approximation in Theorem 1 is uniformly $O(\sum 1/n_i)$ on compact subsets of the maximizer of $L(\mu, D)$. Thus, to the extent that the individual n_i's are large enough for the approximations $\hat{I}_i$ to be accurate, we may expect the MLE-based Normal approximation to provide an accurate approximation to the posterior distribution of (μ, D) even in its tails.

REMARK 2. We noted above that the MLE-based Normal approximation may be obtained via two replacements in the Laplace approximation, which we labeled (i) and (ii). Theorem 1 shows that MLE-based Normal approximation, like the Laplace approximation, furnishes an approximation having accuracy

$O(\sum 1/n_i)$. That step (i) involves an error of order $O(\sum 1/n_i)$ is an immediate corollary of the following theorem. Taken together, Theorems 1 and 2 also show that step (ii) has accuracy $O(\sum 1/n_i)$ which, while not used in any way here, is of some interest on its own.

THEOREM 2. *If in the Laplace approximation $\tilde{I}_i$ to I_i we substitute for $\tilde{\theta}_i$ an alternative value θ_i^* for which $\tilde{\theta}_i - \theta_i^* = O(1/n_i^a)$, where $0 < a < 1$, then the resulting approximation I_i^* satisfies $I_i = I_i^*(1 + O(1/n^{1-2(1-a)}))$.*

The proof involves expansions of the kind used in proving Theorem 1. Because the MLE and the mode $\tilde{\theta}_i$ differ by order $O(1/n_i)$, it follows that the replacement in step (i) incurs an error of order $O(1/n_i)$.

The adequacy of the MLE-based Normal approximation can be assessed by computing a statistic suggested by Kass and Slate (1994). This diagnostic may be considered a multivariate generalization of Pearson skewness, which is the normalized difference between the posterior mode and posterior mean of a one-dimensional parameter: it is defined by $R^\star = (\tilde{\theta} - \bar{\theta}^\star)^T \tilde{G} (\tilde{\theta} - \bar{\theta}^\star)$ where $\tilde{G}$ is the observed information matrix for θ, $\tilde{\theta}$ is the posterior mode, and $\bar{\theta}^\star$ is the posterior mean, or any second-order approximation to it. A rough, conservative guideline for the interpretation of this quantity is that the posterior may be considered close to Normal when $R^\star < p/6$, where p is the dimension of θ. (See Kass and Slate (1994) for further details and references.) Here, it is most convenient to assess approximate Normality of the likelihood functions $L_i(\theta_i)$ by evaluating the corresponding $R^\star$ under the assumption of a flat prior on the θ_i. We obtain the values of $R^\star$ for each $i = 1, \ldots, k$. If a substantial percentage of these dramatically exceeded the cutoff $p/6$ then we would consider the MLE-based Normal approximation likely to be worrisomely inaccurate.

3. **Correcting via Importance weights**

In the previous section we discussed the theoretical accuracy of using the MLE-based Normal approximation, which simplifies sampling from the posterior distribution; we also mentioned a way to assess the approximation using an easily-computed diagnostic. We now consider correcting the approximation by reweighting the draws from its posterior.

In proposing a computational method for non-conjugate Bayesian inference about covariance matrices, Daniels and Kass (1997) noted that Gibbs sampling draws $(\mu^{(a)}, D^{(a)})$ from the posterior on (μ, D) based on any first-stage approximations $\hat{f}(y_i|\theta_i)$ may be corrected by the importance weights

$$w_a = \prod_{i=1}^{k} \frac{f(y_i|\theta_i^{(a)})}{\hat{f}(y_i|\theta_i^{(a)})}.$$

That is, given a conjugate prior on (μ, D), if we take Gibbs sample draws $(\theta^{(a)}, \mu^{(a)}, D^{(a)})$ for $a = 1, \ldots, A$ from the joint posterior of (θ, μ, D) based on the approximation and then attach to each $(\mu^{(a)}, D^{(a)})$ the corresponding weight w_a, we obtain a valid importance sampling scheme in the sense that $E(g(\mu, D)|y) = \frac{E_{\hat{p}}(wg(\mu,D)|y)}{E_{\hat{p}}(w|y)}$, where the expectation, $E_{\hat{p}}$, is taken over the approximate joint posterior distribution of (θ, μ, D). Thus, if the approximation is well-behaved, weighted means of the form $\sum_{a=1}^{A} w_a g(\mu^{(a)}, D^{(a)}) / \sum_a w_a$ will converge (as $A \to \infty$) to posterior means $E(g(\mu, D)|y)$ (and variances may similarly be estimated; see, e.g., Tanner, 1993, section 3.3.3). Alternatively, draws from the posterior may be obtained via importance resampling (Tanner, 1993, section 5.7). Here and elsewhere, we are assuming that the $E(g(\mu, D)|y)$ exists and is finite.

In the context of the MLE-based Normal approximation, we may obtain draws from the posterior based on the approximation, for example, from convenient software such as BUGS and may then quite easily use a high-level general language such as S-PLUS to reweight and obtain corrected posterior inferences.

In the remainder of this section we generalize the result of Daniels and Kass (1997) to two important classes of models.

3.1 *Importance reweighting for mixed effects models.* We now consider the class of generalized linear mixed models (GLMMs) having the following form.

Stage one: $Y_i|\alpha, \beta_i \sim f(y|\alpha, \beta_i, X_i, Z_i)$, $i = 1 \ldots, k$, where f denotes the density of a one-parameter exponential family, Y_i is an $n_i \times 1$ dimensional vector, α a p-dimensional vector, and β_i an m-dimensional vector; the data are connected to the covariates (X_i, Z_i) through a link function $\eta_i = X_i\alpha + Z_i\beta_i$, which is itself a transformation of the expected value of Y_i;

Stage two: $\beta_i|\mu, D \sim N_m(\mu, D)$.

Here, α is often called a "fixed effect" while the β_i's are the "random effects (coefficients)." For details on such models and the standard GLMM form, see for example Breslow and Clayton (1993).

Here are steps that lead to importance weights for mixed models starting with draws from the posterior based on the MLE-based Normal approximation:

1. Ignoring the distinctions among the individual units $i = 1, \ldots, n$, pool the data and fit a single-stage generalized linear model with $\eta_i = X_i\alpha + Z_i\mu$ to obtain MLEs $\hat{\alpha}^{(1)}, \hat{\mu}^{(1)}$.

2. Conditioning on $\hat{\alpha}^{(1)}$, for each i compute the MLE of β_i using the individual likelihoods $p(y_i|\hat{\alpha}^{(1)}, \beta_i)$.

3. For each i, replace the likelihood $p(y_i|\hat{\alpha}^{(1)}, \beta_i)$ with its Normal approximation based on the MLE of β_i found in step 2.

4a. For the conjugate Normal two-stage approximate model based on Step

3, involving the parameters $\beta_1, \ldots, \beta_m, \mu, D$, use Gibbs sampling to obtain a sample, $(\beta_i^{(a)}, \mu^{(a)}, D^{(a)})$, $a = 1, \ldots, A$ from the approximate joint posterior, $\hat{p}(\beta_i, \mu, D|y, \hat{\alpha}^{(1)})$.

4b. We now need to sample from an approximation to $\hat{p}(\alpha|\beta_i, \mu, y)$. We begin with the joint Normal approximation based on the MLE's in Step 1, then condition on the current value of μ, $\mu^{(a)}$, to sample $\alpha^{(a)}$. [There are several possible alternatives. For instance, one might instead compute a Normal approximation to α from the original likelihood, conditional on the current values of the β_i, $\beta_i^{(a)}$.]

5. We now have A draws from the approximate posterior. We attach to the draw $(\mu^{(a)}, D^{(a)})$ the importance weight $w_a = \prod_i \frac{p(y_i|\beta_i^{(a)}, \alpha^{(a)})}{\hat{p}(y_i|\beta_i^{(a)}, \hat{\alpha}^{(1)})} \frac{p(\alpha^{(a)})}{\hat{p}(\alpha^{(a)}|\beta_i^{(a)}, \mu^{(a)}, y)}$.

THEOREM 3. *Letting E_p be the expectation under the marginal posterior of (μ, D) and $E_{\hat{p}}$ the expectation under the approximation to the joint posterior created in the steps outlined above, for any function of the parameters $g(\mu, D)$ we have $E_p(g(\mu, D)) = E_{\hat{p}}(w_a g(\mu, D))/E_{\hat{p}}(w_a)$.*

PROOF. Consider computing the expectation of some function $g(\cdot)$ of (μ, D), i.e., $E[g(\mu, D|y)] = \frac{A}{B}$, where the numerator integral, A, is defined as

$$A = \int\int \cdots \int g(\mu, D) \prod_{i=1}^{n} p(\beta_i|\alpha, \mu, D, y) p(\alpha|\mu, D, y) p(\mu, D|y) d\mu dD d\alpha d\beta_1 \ldots d\beta_n$$

and denominator integral is

$$B = \int\int \cdots \int \prod_{i=1}^{n} p(\beta_i|\alpha, \mu, D, y) p(\alpha|\mu, D, y) p(\mu, D|y) d\mu dD d\alpha d\beta_1 \ldots d\beta_n.$$

We begin by rewriting A:

$$\begin{aligned} A &= \int\int \cdots \int g(\mu, D) \prod_i \frac{p(\beta_i, \alpha, \mu, D|y)}{\hat{p}(\beta_i, \alpha, \mu, D|y)} \hat{p}(\beta_i, \alpha, \mu, D|y) d\mu dD d\alpha d\beta_1 \ldots d\beta_n \\ &= \int\int \cdots \int g(\mu, D) \frac{\prod_i \{p(y_i|\beta_i, \alpha) p(\beta_i|\mu, D)\} p(\mu, D) p(\alpha)/m(y)}{\prod_i \hat{p}(\beta_i|\mu, D, y) \hat{p}(\mu|D, y) \hat{p}(D|y) \hat{p}(\alpha|\mu, \beta_i, y)} \\ &\qquad \times \prod_i \hat{p}(\beta_i|\mu, D, y) \hat{p}(\mu|D, y) \hat{p}(D|y) \hat{p}(\alpha|\mu, \beta_i, y) d\mu dD d\beta_1 \ldots d\beta_n \end{aligned}$$

where $m(y)$ denotes the marginal distribution of the data under the true model. Then, with $\hat{m}(y)$ denoting the marginal distribution of the data under the approximate model we obtain

$$
\begin{aligned}
A &= \int\int\cdots\int g(\mu,D)\frac{\prod_i\{p(y_i|\beta_i,\alpha)p(\beta_i|\mu,D)\}p(\mu,D)p(\alpha)/m(y)}{\prod_i\{\hat{p}(y_i|\beta_i,\hat{\alpha})p(\beta_i|\mu,D)\}p(\mu,D)\hat{p}(\alpha|\mu,\beta_i,y)/\hat{m}(y)} \\
&\qquad \times\prod_i \hat{p}(\beta_i|\mu,D,y)\hat{p}(\mu|D,y)\hat{p}(D|y)\hat{p}(\alpha|\mu,\beta_i,y)d\mu dD d\beta_1\ldots d\beta_n \\
&= \int\int\cdots\int g(\mu,D)\frac{\prod_i p(y_i|\beta_i,\alpha)p(\alpha)/m(y)}{\prod_i \hat{p}(y_i|\beta_i,\hat{\alpha})\hat{p}(\alpha|\mu,\beta_i,y)/\hat{m}(y)} \\
&\qquad \times\prod_i \hat{p}(\beta_i|\mu,D,y)\hat{p}(\mu|D,y)\hat{p}(D|y)\hat{p}(\alpha|\mu,\beta_i,y)d\mu dD d\beta_1\ldots d\beta_n.
\end{aligned}
$$

Thus, writing the importance weight attached to a random draw (μ, D) as $w(\mu, D)$

$$A = \int g(\mu,D)w(\mu,D)\hat{p}(\mu,D)d\mu dD$$

as required. By a similar development, the denominator integral, B, will be

$$
\begin{aligned}
B &= \int\int\cdots\int \frac{\prod_i p(y_i|\beta_i,\alpha)p(\alpha)/m(y)}{\prod_i \hat{p}(y_i|\beta_i,\hat{\alpha})\hat{p}(\alpha|\mu,\beta_i,y)/\hat{m}(y)} \\
&\qquad \times\prod_i \hat{p}(\beta_i|\mu,D,y)\hat{p}(\mu|D,y)\hat{p}(D|y)\hat{p}(\alpha|\mu,\beta_i,y)d\mu dD d\beta_1\ldots d\beta_n
\end{aligned}
$$

and

$$B = \int w(\mu,D)\hat{p}(\mu,D)d\mu dD.$$

□

3.2 *Importance reweighting for nonlinear hierarchical models.* We now consider a general Normal nonlinear hierachical model.

Stage one: $Y_{ij}|\alpha,\beta_i,\tau^2,X_{ij} \sim N(h(\beta_i,\alpha;X_{ij}),\tau^2)$, $j=1\ldots,n_i$, $i=1\ldots,n$, where α is a p-dimensional vector, β_i an m-dimensional vector, and h a nonlinear function of the parameters that depends on some explanatory variables X_{ij};

Stage two: $\beta_i|\mu,D \sim N_m(\mu,D)$.

We assume there is either a conjugate or flat prior on τ^2.

The steps for fitting the approximate model for the generalized linear mixed model in Section 3.1 only need to be slightly modified. Step 1 now includes the pooled MLEs $\hat{\alpha},\hat{\mu},\hat{\tau}^2$ and the steps 2-4 now are conditional on both $\hat{\alpha}$ and $\hat{\tau}^2$. In Step 5 we sample from the conditional approximate posterior on α as before and then also sample from τ^2, conditional on β_i,μ,α,y, using an Inverse gamma distribution. In Step 6 the importance weight becomes

$$w_a = \frac{\prod_i p(y_i|\beta_i^{(a)},\alpha^{(a)})}{\prod_i \hat{p}(y_i|\beta_i^{(a)},\hat{\alpha})}\frac{p(\alpha^{(a)})p(\tau^{2(a)})}{\hat{p}(\alpha^{(a)}|\mu^{(a)},\beta_i^{(a)},y)p(\tau^{2(a)}|\beta_i^{(a)},\mu^{(a)},\alpha^{(a)},y)}$$

THEOREM 4. *Letting E_p be the expectation under the marginal posterior of (μ, D) and $E_{\hat{p}}$ the expectation under the approximation to the joint posterior created in the steps outlined above, for any function of the parameters $g(\mu, D)$ we have $E_p(g(\mu, D)) = E_{\hat{p}}(w_a g(\mu, D))/E_{\hat{p}}(w_a)$.*

The proof is similar to Theorem 3.

4. **Example**

The general procedure we envision is summarized by the steps (i) find first-stage MLE's, (ii) obtain posterior draws from the MLE-based approximate Normal model, (iii) possibly, assess adequacy of MLE-based Normal approximation via $R^\star$, and, if desired, (iv) reweight the draws to obtain accurate posterior inferences.

As an example of our strategy, we consider the nonlinear hierarchical model most recently analyzed by Bennet, Racine-Poon, and Wakefield (1996). This application deals with prediction of uptake volume of guinea pig tissue by concentration of β-methylglucoside. The response, y_{ij} represents the uptake volume for the jth concentration ($j = 1, \ldots, 10$) of the ith guinea pig ($i = 1, \ldots, 8$) and X_{ij} the corresponding concentration of β-methylglucoside. The following nonlinear hierarchical model was fit to the data

$$\log(y_{ij}) = \log(\frac{\exp(\beta_{1i})X_{ij}}{\exp(\beta_{2i}) + X_{ij}} + \exp(\beta_{3i})X_{ij}) + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \tau^2)$$

$$\beta_i \sim N(\mu, D)$$

$$\mu \sim d\mu,\ D \sim \pi(D)$$

with $\pi(D)$ an inverse Wishart prior with 3 degrees of freedom and known scale matrix. This example is of interest not only because Bennet et al. used it to discuss alternative MCMC strategies but also because the within-pig sample size of $n_i = 10$ for all i is sufficiently small for a three-parameter nonlinear model that one might be dubious about the adequacy of the MLE-based Normal approximation for estimation within pigs. Here, of course, we are interested in the use of this approximation for inference about μ and D.

Following the steps in Section 3.2, we began by fitting the MLE-based approximate Normal model with τ^2 fixed at its MLE ($\hat{\tau}^2$), using BUGS. In S-PLUS we then computed the appropriate weights and sampled from the full conditional

distribution for τ^2. The importance weights now become $\prod_i \frac{p(y_i|\theta_i,\tau^2)}{\hat{p}(y_i|\theta_i,\tau^2)} \frac{p(\tau^2)}{p(\tau^2|\theta_i,y)}$ where $\hat{p}$ is the Normal approximation to the distribution of the MLE conditional on $\hat{\tau}^2$.

Assessment of the Normal approximation showed that it appeared reasonably accurate in 7 of the 8 pigs: the conservative guideline was $p/6 = .5$ and the respective values of $R^\star$ were .09, .58, .58, .54, .57, 3.09, .56, .48. Both the MLE-based approximate Normal results and the importance reweighted results are given in Table 1 (where simulation error is small enough that all digits listed are very likely accurate). It may be seen that the approximate method is indeed quite good in this case.

Table 1: Posterior means of components of μ (with posterior standard deviations) and D using the MLE-based Normal approximation (Approx) and the reweighting (Exact).

Parameter	Approx	Exact
μ_1	-1.61 (.07)	-1.59 (.07)
μ_2	.87 (.10)	.89 (.10)
μ_3	-5.44 (.17)	-5.48 (.17)
D_{11}	.022	.022
D_{21}	.0075	.0077
D_{22}	.030	.031
D_{31}	-.0056	-.0070
D_{32}	-.023	-.025
D_{33}	.157	.168

5. **Discussion**

The purpose of this note was to indicate the potential usefulness of the MLE-based Normal approximation specified in the Introduction. It may often be quite adequate on its own, and its accuracy may be assessed using the $R^\star$ diagnostic. Generally (based on a limited simulation study not described here), we feel that when the large majority of individual approximations are judged roughly adequate in the sense that they are close to satisfying $R^\star < p/6$, the posterior approximation itself is likely to be adequate for most practical purposes in making inferences about (μ, D). However, there is no guarantee of this, and it is possible to construct cases in which systematic bias occurs in the approximations, so that even a small number of poor individual approximations may seriously affect the posterior on (μ, D). It is thus clearly desirable to reweight.

The MLE-based Normal approximation is worth considering whenever implementation of a comprehensive MCMC method for a two-stage model seems daunting. In addition, because the underlying Gibbs sampling routine for the MLE-based Normal approximation may be efficiently coded, we believe the

reweighting scheme we have articulated may be attractive for implementing Bayesian analysis of generalized linear and nonlinear mixed models in software such as S-PLUS.

Acknowledgement. The authors are grateful for comments from a referee. This work was supported by grants from the National Science Foundation and the National Institutes of Mental Health.

References

Bennet JE, Racine-Poon A, and Wakefield JC (1996). MCMC for nonlinear hierarchical models. in *Markov Chain Monte Carlo in Practice*, eds. Gilks WR, Richardson S, Spiegelhalter DJ, Chapman and Hall, pp. 339-358.

Breslow NE and Clayton DG (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9-25.

Daniels MJ and Kass RE (1997). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Carnegie Mellon University Technical Report* # 659.

Everson P and Morris C (1997). Inference for Multivariate Normal hierarchical models. *Harvard University Technical Report.*

Gilks WR, Wang CC, Yvonnet B, and Coursaget P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics*, 49:441-453.

Kass RE and Slate (1994). Some diagnostics for maximum likelihood and posterior non-normality. *Annals of Statistics*, 22:668-695.

Kass RE (1997). Laplace's method, in *Encyclopedia of Statistical Sciences, Update Volume 1*, eds. Kotz S, Read CB, and Banks DB, Wiley.

Kass RE and Vos (1997). *Geometrical Foundations of Asymptotic Inference*, Wiley.

Kass RE, Tierney L, and Kadane JB (1990). The validity of posterior expansions based on Laplace's method, in *Essays in Honor of George Barnard*, eds. Geisser S, Hodges JS, Press SJ, and Zellner A, North Holland, pp. 473-488.

Smith AFM and Roberts GO (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of Royal Statistical Society B*, 55:3-23.

Spiegelhalter DJ, Best NG, Gilks WR, and Inskip H (1996). Hepatitis B: a case study in MCMC methods. in *Markov Chain Monte Carlo in Practice*, eds. Gilks WR, Richardson S, Spiegelhalter DJ, Chapman and Hall, pp. 339-358.

Tanner MA (1993). *Tools for Statistical Inference: Methods for Exporation of Posterior Distributions and Likelihood Functions*, Springer-Verlag.

Michael J. Daniels
Department of Statistics
Iowa State University
102G Snedecor Hall
Ames, IA 50011
e-mail : mdaniels@iastate.edu

Robert E. Kass
Department of Statistics
Carnegie Mellon University
229 Baker Hall
Pittsburgh, PA 15213
e-mail : kass@stat.cmu.edu