Bayesian Analysis (2006) **1**, Number 3, pp. 473–514

A comparison of Bayesian and likelihood-based methods for fitting multilevel models

William J. Browne*, and David Draper†

Abstract. We use simulation studies, whose design is realistic for educational and medical research (as well as other fields of inquiry), to compare Bayesian and likelihood-based methods for fitting variance-components (VC) and random-effects logistic regression (RELR) models. The likelihood (and approximate likelihood) approaches we examine are based on the methods most widely used in current applied multilevel (hierarchical) analyses: maximum likelihood (ML) and restricted ML (REML) for Gaussian outcomes, and marginal and penalized quasi-likelihood (MQL and PQL) for Bernoulli outcomes. Our Bayesian methods use Markov chain Monte Carlo (MCMC) estimation, with adaptive hybrid Metropolis-Gibbs sampling for RELR models, and several diffuse prior distributions ($\Gamma^{-1}(\epsilon, \epsilon)$ and $U(0, \frac{1}{\epsilon})$ priors for variance components). For evaluation criteria we consider bias of point estimates and nominal versus actual coverage of interval estimates in repeated sampling. In two-level VC models we find that (a) both likelihood-based and Bayesian approaches can be made to produce approximately unbiased estimates, although the automatic manner in which REML accomplishes this is an advantage, but (b) both approaches had difficulty achieving nominal coverage in small samples and with small values of the intraclass correlation. With the three-level RELR models we examine we find that (c) quasi-likelihood methods for estimating random-effects variances perform badly with respect to bias and coverage in the example we simulated, and (d) Bayesian diffuse-prior methods lead to well-calibrated point and interval RELR estimates. While it is true that the likelihood-based methods we study are considerably faster computationally than MCMC, (i) steady improvements in recent years in both hardware speed and efficiency of Monte Carlo algorithms and (ii) the lack of calibration of likelihood-based methods in some common hierarchical settings combine to make MCMC-based Bayesian fitting of multilevel models an attractive approach, even with rather large data sets. Other analytic strategies based on less approximate likelihood methods are also possible but would benefit from further study of the type summarized here.

Keywords: Adaptive MCMC, bias, calibration, diffuse priors, hierarchical modeling, hybrid Metropolis-Gibbs sampling, intraclass correlation, IGLS, interval coverage, MQL, mixed models, PQL, RIGLS, random-effects logistic regression, REML, variance-components models

*Division of Statistics, School of Mathematical Sciences, University of Nottingham, UK http://www.maths.nottingham.ac.uk/personal/pmzwjb/

†Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA, http://www.ams.ucsc.edu/~draper

 ba0003

1 Introduction

Multilevel models, for data possessing a nested hierarchy and—more generally—for the expression of uncertainty at several levels of aggregation, have gained dramatically in scope of application in the past 15 years, in fields as diverse as education and health policy (e.g., Goldstein et al. (1993), Draper (1995), Goldstein and Spiegelhalter (1996)). Statisticians and substantive researchers who use such models now have a variety of options in approaches to inference, with a corresponding variety of computer programs: to mention four, the maximum-likelihood (ML) Fisher-scoring approach in `VARCL` (Longford (1987)); ML via iterative generalized least squares (IGLS) and restricted IGLS (RIGLS, or REML) for Gaussian outcomes, and quasi-likelihood methods (MQL and PQL) for dichotomous outcomes, in `MLwiN` (Goldstein (1986, 1989), Rasbash et al. (2005)); empirical-Bayes estimation using the EM algorithm in `HLM` (Raudenbush et al. (2005)); and fully-Bayesian inference in `WinBUGS` (Spiegelhalter et al. (2003)) and `MLwiN`. This variety of fitting methods can lead to confusion, however: ML and Bayesian analyses of the same data can produce rather different point and interval estimates, and the applied multilevel modeler may well be left wondering what to report.

1.1 Example 1: The Junior School Project

The Junior School Project (JSP; Mortimore et al. (1988), Woodhouse et al. (1995)) was a longitudinal study of about 2,000 pupils from 50 primary schools chosen randomly from the 636 Inner London Education Authority (ILEA) schools in 1980. Here we will examine a random subsample of $N = 887$ students taken from $J = 48$ schools. A variety of measurements were made on the students during the four years of the study, including background variables (such as gender, age at entry, ethnicity, and social class) and measures of educational outcomes such as mathematics test scores (on a scale from 0 to 40) at year 3 (`math3`) and year 5 (`math5`). Both mathematics scores had distributions with negative skew due to a ceiling effect, with some students piling up at the maximum score, but transformations to normality produced results almost identical to those using the raw data (we report the latter). A principal goal of the study was to establish whether some schools were more effective than others in promoting pupils' learning and development, after adjusting for background differences.

Two simple baseline analyses that might be undertaken early on, before more complicated modeling, are as follows.

- Thinking (incorrectly) of the data as a simple random sample (SRS) from the population of ILEA pupils in the early 1980s, the mean mathematics score β_0 at year 5 would be estimated as 30.6 with a repeated-sampling standard error (SE) of 0.22, but this ignores the large estimated *intraclass* (intracluster; within-school) *correlation* of $\hat{\rho} = +0.12$ for this variable. The correct SE, from standard survey-sampling results (e.g., Cochran (1977)) or the Huber-White sandwich estimator (Huber (1967), White (1980), as implemented in the package `Stata`: StataCorp (2006)), is 0.43, almost double the SRS value. There is clearly scope for multilevel modeling here to account correctly for the nested structure of the data. (0.12 may

Table 1: *A comparison of ML, REML, and Bayesian fitting (with a diffuse prior) in model (1) applied to the JSP data. Figures in parentheses in the upper table are SEs (for the ML methods) or posterior SDs (for the Bayesian method). Bayesian point estimates are posterior means, and 95% central posterior intervals are reported.*

Point Estimates	Parameter		
Method	β_0	σ_u^2	σ_e^2
ML	30.6	5.16	39.3
	(0.400)	(1.55)	(1.92)
REML	30.6	5.32	39.3
	(0.404)	(1.59)	(1.92)
Bayesian with	30.6	6.09	39.5
diffuse priors	(0.427)	(1.91)	(1.94)

95% Interval Estimates	Parameter		
Method	β_0	σ_u^2	σ_e^2
REML (Gaussian)	(29.8, 31.4)	(2.22, 8.43)	(35.5, 43.0)
Bayesian	(29.8, 31.5)	(3.18, 10.6)	(35.9, 43.5)

not seem like a large value for ρ, but (a) despite its name the intraclass correlation is in fact comparable to a regression-style R^2 value (e.g., Donner (1986)), and (b) the *design effect* (e.g., Cochran (1977)) for estimating β_0 in this problem is $\left(\frac{0.43}{0.22}\right)^2 \doteq 3.8$, meaning that the cluster sample of 887 students was equivalent in information content for β_0 (because of the relatively high degree of within-school similarity of student achievement) to an SRS of only $\frac{887}{3.8} \doteq 230$ students.)

- Consider next a variance-components (VC) model,

$$\begin{array}{l} y_{ij} = \beta_0 + u_j + e_{ij}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J, \\ \sum_{j=1}^{J} n_j = N, \quad u_j \overset{\text{IID}}{\sim} N(0, \sigma_u^2), \quad e_{ij} \overset{\text{IID}}{\sim} N(0, \sigma_e^2), \end{array} \tag{1}$$

where y_{ij} is the math5 score for pupil i in school j; this model would generally be fit before a random-slopes regression model relating math5 to math3 is examined. (In our terminology i indexes level 1 of the model and j level 2. (1) is sometimes referred to as a *mixed linear model* for its combination of *fixed effects* (β_0) and *random effects* (the u_j and e_{ij}).) As noted above, the parameters in this model may be estimated in at least two ways: likelihood-based and Bayesian approaches. Maximum likelihood (ML) in turn may be based on iterative generalized least squares (IGLS, or some other equivalent method), or approximately unbiased estimation with restricted maximum likelihood (REML, based for instance on RIGLS) (Goldstein (1986, 1989)) may be preferred. Table 1 presents the results of ML, REML, and Bayesian fitting of model (1), in the latter case using a diffuse prior to be discussed in Section 2.3.1 ($U\left(0, \frac{1}{\epsilon}\right)$ on the variance scale).

While there is little difference in the three methods on point estimates for β_0 and σ_e^2 and on SEs/posterior standard deviations (SDs) for the latter quantity, (a) the posterior SD for β_0 is about 5% larger than the SE from ML and REML (note

that the Bayesian uncertainty assessment essentially coincides with the cluster-sampling SE 0.43 mentioned earlier), (b) the Bayesian estimate of σ_u^2 is 14–17% larger than the likelihood estimates, and (c) the posterior SD for σ_u^2 is 18–21% larger than the ML/REML SEs. Moreover, the default likelihood results (point estimates and estimated asymptotic SEs) in the ML computer programs in most widespread current use do not include interval estimates, encouraging investigators either to report no intervals at all (a practice to be frowned upon) or to use large-sample 95% Gaussian intervals of the form (estimate $\pm 1.96\,\widehat{SE}$). The bottom part of Table 1 compares Gaussian intervals based on REML estimates with Bayesian 95% posterior probability intervals, and it may be seen that in particular the two methods give quite different answers for σ_u^2. What should someone trying to arrive at substantive conclusions based on the JSP data report?

1.2 Example 2: The Guatemalan Child Health Study

The 1987 Guatemalan National Survey of Maternal and Child Health (Pebley and Goldman (1992)) was based on a multistage cluster sample of 5,160 women aged 15–44 years living in 240 communities, with the goal of increased understanding of the determinants of health for mothers and children in the period during and after pregnancy. The data have a three-level structure—births within mothers within communities—and one analysis of particular interest estimated the probability of receiving modern (physician or trained nurse) prenatal care as a function of covariates at all three levels. Rodríguez and Goldman (1995) studied a subsample of 2,449 births by 1,558 women who (a) lived in the 161 communities with accurate cluster-level information and (b) had some form of prenatal care during pregnancy. The random-effects logistic regression (RELR) model they examined is

$$\begin{gathered}(y_{ijk}\,|\,p_{ijk}) \overset{\text{indep}}{\sim} \text{Bernoulli}(p_{ijk}) \qquad \text{with} \\ \text{logit}(p_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + v_k,\end{gathered} \tag{2}$$

where y_{ijk} is a binary indicator of modern prenatal care or not and where $u_{jk} \sim N(0, \sigma_u^2)$ and $v_k \sim N(0, \sigma_v^2)$. In this formulation $i = 1, \ldots, I_{jk}$, $j = 1, \ldots, J_k$, and $k = 1, \ldots K$ index the level 1, 2, and 3 units, respectively, corresponding to births, mothers, and communities, and the variables x_1, x_2, and x_3 are composite scales, because the original Pebley-Goldman model contained many covariates at each level. The original Rodríguez-Goldman data set is not publicly available; however, these authors simulated 25 data sets with the same structure but with known parameter values, and they have kindly made these simulated data sets available to us.

As in Example 1, several likelihood-based and Bayesian fitting methods for model (2) are available: the main (approximate) likelihood alternatives (e.g., Goldstein (2002)) currently employed with greatest frequency by multilevel modelers in substantive fields of inquiry (based upon empirical usage in the recent literature) are marginal quasi-likelihood (MQL) and penalized (or predictive) quasi-likelihood (PQL), in both of which the investigator has to specify the order of the Taylor-series approximation, and a variety of prior distributions may be considered in the Bayesian approach. Table 2 summarizes

Table 2: *A comparison of first-order MQL, second-order PQL and Bayesian fitting (with a diffuse prior) in model (2) applied to the Rodríguez-Goldman simulated Guatemalan child health data set number 1. Figures in square brackets in the upper table are true parameter values; figures in parentheses in the upper table are SEs (for the ML methods) or posterior SDs (for the Bayesian method). Bayesian point estimates are posterior means, and 95% central posterior intervals are reported.*

Point Estimates	Parameter					
Method	β_0 [0.65]	β_1 [1.0]	β_2 [1.0]	β_3 [1.0]	σ_v^2 [1.0]	σ_u^2 [1.0]
MQL_1	0.491	0.791	0.631	0.806	0.546	0.000
	(0.149)	(0.172)	(0.081)	(0.189)	(0.102)	—
PQL_2	0.641	0.993	0.795	1.06	0.883	0.486
	(0.186)	(0.201)	(0.099)	(0.237)	(0.159)	(0.145)
Bayesian with	0.675	1.050	0.843	1.124	1.043	0.921
diffuse priors	(0.209)	(0.225)	(0.115)	(0.268)	(0.217)	(0.331)

95% Interval Estimates	Parameter					
Method	β_0	β_1	β_2	β_3	σ_v^2	σ_u^2
PQL_2 (Gaussian)	(0.276,	(0.599,	(0.601,	(0.593,	(0.571,	(0.202,
	1.01)	1.39)	0.989)	1.52)	1.19)	0.770)
Bayesian	(0.251,	(0.611,	(0.626,	(0.586,	(0.677,	(0.334,
	1.07)	1.50)	1.078)	1.62)	1.52)	1.63)

a comparison between first-order MQL, second-order PQL, and Bayesian fitting—again with a particular diffuse prior to be discussed in Section 2.3.1 ($U\left(0, \frac{1}{\epsilon}\right)$ on the variance scale for small ϵ)—on the Rodríguez-Goldman simulated data set number 1 (the true values of the parameters are given in the first row of this table). Here the differences are much more striking than those in Table 1: many MQL estimates are badly biased, and—although PQL does achieve some improvements—its estimates of β_2 and the variance components are still substantially too low, leading to dramatically different (and inferior) intervals for the variances. Because we have the luxury of knowing the right answer in this simulation context, it is easy to see which fitting method has produced better results on this one data set (and Section 4.2 will demonstrate that this table accurately reflects the superiority of Bayesian methods in models like (2) when compared with quasi-likelihood approaches, at least in settings similar to the Guatemalan Health study), but—if the data analyzed in Table 2 arose as the result of an actual sample survey—a researcher trying to draw substantive conclusions about variability within and between mothers and communities would certainly wonder which figures to publish.

1.3 Outline of the paper

Our interest is in comparing likelihood-based and Bayesian methods for fitting variance-components and random-effects logistic regression models, using bias and interval coverage behavior in repeated sampling as evaluation criteria. Following a brief literature

review below, Section 2 describes the fitting methods we compare; Sections 3 and 4 cover simulation study details and results for VC and RELR models, respectively; and Section 5 offers some conclusions and discussion. Browne and Draper (2000) and Browne et al. (2002) contain results that parallel those presented here for random-slopes regression models and multilevel models with heteroscedasticity at level 1, respectively.

We focus in this paper on the likelihood-based (and approximate likelihood) methods most readily available (given current usage patterns of existing software) to statisticians and substantive researchers making frequent use of multilevel models: ML and REML in VC models, and MQL and PQL in RELR models. Other promising likelihood-based approaches—including (a) methods based on Gaussian quadrature (e.g., Pinheiro and Bates (1995); see Section 5 for a software discussion); (b) the nonparametric maximum likelihood methods of Aitkin (1999a); (c) the Laplace-approximation approach of Raudenbush et al. (2000); (d) the work on hierarchical generalized linear models of Lee and Nelder (2001); and (e) interval estimation based on ranges of values of the parameters for which the log likelihood is within a certain distance of its maximum, for instance using profile likelihood (e.g., Longford (2000))—are not addressed here. It is evident from the recent applied literature that, from the point of view of multilevel analyses currently being conducted to inform educational and health policy choices and other substantive decisions, the use of methods (a–e) is not (yet) as widespread as REML and quasi-likelihood approaches. In particular, methods such as Gaussian quadrature may produce poor results in RELR models if not used carefully (see Lesaffre and Spiessens (2001) for a striking example); we intend to report elsewhere on a thorough comparison of quadrature with the methods examined here.

Statisticians are well aware that the highly skewed repeated-sampling distributions of ML estimators of random-effects variances in multilevel models with small sample sizes are not likely to lead to good coverage properties for large-sample Gaussian approximate interval estimates of the form $\hat{\sigma}^2 \pm 1.96\,\widehat{SE}\left(\hat{\sigma}^2\right)$, but with few exceptions the profession has not (yet) responded to this by making software for improved likelihood interval estimates for variance components widely available to multilevel modelers. In Sections 3 and 4 we document the extent of the poor coverage behavior of the Gaussian approach, and we offer several simple approximation methods, based only on information routinely output in multilevel software, which exhibit improved (although still not in many cases satisfactory) performance. Note that we are not advocating interval estimates for random-effects variances based on normal approximations in small samples; we are merely documenting how badly these intervals—which are all that will be readily available to many users of popular likelihood-based software packages—may behave, even with a variety of improvements to them.

The paper has been constructed so that readers interested in a fast path through it can proceed directly from this point to Section 5, where a summary of our conclusions is available.

1.4 Previous literature on comparisons between multilevel fitting methods

The literature on Bayesian and likelihood-based methods for fitting VC and RELR models is vast, e.g., Aitkin (1996, 1999b), Besag et al. (1995), Bryk and Raudenbush (1992), Corbeil and Searle (1976), Daniels and Gatsonis (1999), Gelfand and Smith (1990), Goldstein (2002), Harville and Zimmermann (1996), Kahn and Raftery (1996), Kass and Steffey (1989), Lee and Nelder (1996), Longford (1987, 1997), and Searle et al. (1992) (for a competing approach based on best linear unbiased prediction, see, e.g., Henderson (1950) and Robinson (1991)). Comparisons between multilevel fitting methods are less abundant, but some theoretical work has been done to demonstrate the equivalence of several of the leading approaches to fitting multilevel models: for instance, Raudenbush (1994) showed that Fisher scoring is equivalent to ML, and empirical-Bayes estimates based on the EM algorithm may be seen to coincide with maximum likelihood results in many Gaussian models (e.g., Goldstein (2002)). Fewer studies are available comparing the performance of the approaches in terms of bias of point estimates and calibration of interval estimates.

In the VC model (1), Box and Tiao (1973) reviewed results of Klotz et al. (1969) and Portnoy (1971) which contrast the mean squared error (MSE) behavior of the following estimators of σ_u^2: the classical unbiased estimator based on mean squares (e.g., Scheffé (1959)), the ML estimator, and the mean and mode of the marginal posterior distribution for σ_u^2 with several choices of relatively diffuse priors. They found, over all values of the intraclass (intracluster) correlation $\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$ they examined, that (a) the MSEs of the ML and posterior-mode estimators are comparable and much smaller than that of the unbiased estimator, and (b) the posterior mean is, by a substantial margin, the worst estimator on MSE grounds. Box and Tiao criticized MSE as an arbitrary criterion for performance assessment, and resisted the distillation of an entire posterior distribution down to a single point estimate. We are sympathetic with their position—from the Bayesian viewpoint the choice of posterior summaries should ideally be based on decision criteria arising from possible actions when using models like (1) and (2) to solve real-world problems—but we nevertheless find it relevant, particularly in the context of general-purpose multilevel modeling software (where the eventual use of the output is far from clear), to examine operating characteristics such as bias and interval coverage. See Rubin (1984) for a good discussion of the relevance of repeated-sampling properties in Bayesian inference, and Chapter 4 of Carlin and Louis (2001) for an evaluation in the spirit of the one presented here for some simpler non-hierarchical Gaussian and binary-outcome models.

Hulting and Harville (1991) compared frequentist and Bayesian methods of fitting the mixed-effects linear model

$$y = X\beta + Zs + e, \tag{3}$$

where y is an $n \times 1$ vector of quantitative outcomes, β is a $p \times 1$ vector of fixed effects, X and Z are known matrices, $s_i \sim N(0, \sigma_s^2)$, and $e_i \sim N(0, \sigma_e^2)$; the VC model (1) is a special case of (3). These authors obtained results which have points of contact with

some of our findings in Section 3.2 below, although Hulting and Harville focused on predictive inferences about quantities of the form $W = \lambda'\beta + \delta's$ and examined different frequentist estimators than the ones we consider. Chaloner (1987) carried out a similar frequentist/Bayesian comparison in model (1); however she used different diffuse prior distributions, focused on the variance ratio $\tau = \frac{\sigma_u^2}{\sigma_e^2} = \frac{\rho}{1-\rho}$ in her results on interval estimation, and conducted a less extensive simulation study than that reported here. See Swallow and Monahan (1984), Brown and Burgess (1984) and Huber et al. (1994) for additional simulation results comparing various non-Bayesian estimation methods in VC models, and Singh et al. (1998) for Bayesian and non-Bayesian comparisons in small-area estimation.

In model (2), Rodríguez and Goldman (1995) used the structure of the Guatemalan child health study to examine how well quasi-likelihood methods compare with fitting a standard logistic regression model and ignoring the multilevel structure. As noted in Section 1.2, their approach involved creating simulated data sets based on the original structure but with known true values for the fixed effects (the β_l in model (2)) and variance parameters. They considered the MQL method and showed that estimates of the fixed effects produced by MQL were even worse, in terms of bias, than estimates produced by standard logistic regression disregarding the hierarchical nature of the data. Goldstein and Rasbash (1996) considered the same problem but used the PQL method, and showed that the results produced by second-order PQL estimation were far better than for MQL, but still biased, in the Rodríguez-Goldman problem. Breslow and Clayton (1993) presented some brief comparisons between quasi-likelihood methods and a version of rejection Gibbs sampling in RELR models proposed by Zeger and Karim (1991); also see Natarajan and Kass (2000) for simulation results in a RELR model fit by the Zeger-Karim approach. Rodríguez and Goldman (2001) obtained results that parallel ours (in Section 4.2) with respect to bias of PQL random-effects variance estimates in REML models (and showed that a parametric bootstrap approach yields considerable improvement), but they have no corresponding findings on interval estimates.

2 Methods for fitting multilevel models

2.1 Iterative generalized least squares (IGLS/ML) and restricted ML (RIGLS/ REML)

Iterative generalized least squares (IGLS/ML; Goldstein (1986)) is a sequential refinement procedure based on GLS estimation. The method can fit all Gaussian multilevel models, and has been described in detail elsewhere (e.g., Goldstein (2002)). Briefly, equations such as (1) are expressed in the usual general linear model form $Y = X\beta + e^*$ (for example, in (1) X is a vector of 1s, $\beta = \beta_0$, and $e_{ij}^* = u_j + e_{ij}$), in which the vector e^* has mean 0 and covariance matrix V; and then the observation is made that (i) if V were known, β could be estimated by GLS, yielding $\hat{\beta}$, and (ii) if β were known, one could form the residuals $\tilde{Y} = Y - X\beta$, calculate $Y^* = \tilde{Y}\tilde{Y}^T$, stack the columns of Y^* into one long column vector Y^{**}, and define a linear model $Y^{**} = Z^*\theta + \epsilon$, where Z^* is

the design matrix for the random-effects parameters θ (in (1) $\theta = \left(\sigma_u^2, \sigma_e^2\right)^T$). Another application of GLS then yields $\hat{\theta}$. Starting with an initial estimate of the fixed effect(s) β from ordinary least squares, IGLS iterates between steps (i) and (ii) to convergence, which is judged to occur when two successive sets of estimates differ by no more than a given tolerance (on a component-by-component basis). As with many ML procedures, IGLS produces biased estimates in small samples, often in particular underestimating random-effects variances because the sampling variation of $\hat{\beta}$ is not accounted for in the algorithm above. Defining the residuals instead as $\tilde{Y}^* = Y - X\hat{\beta}$ and $\hat{Y}^* = \tilde{Y}^* \left(\tilde{Y}^*\right)^T$, Goldstein (1989) showed that

$$E(\hat{Y}^*) = V - X\left(X^T V^{-1} X\right)^{-1} X^T, \tag{4}$$

so that the ML estimates can be bias-adjusted by adding an estimate of the second term on the right-hand side of (4) to $\hat{Y}^*$ at each iteration. This is restricted IGLS (RIGLS), which coincides with restricted maximum likelihood (REML) in Gaussian models such as (1). Estimated asymptotic standard errors of ML and REML estimates are based on the final values at convergence of the covariance matrices for $\hat{\beta}$ and $\hat{\theta}$, expressions for which are given by Goldstein (2002).

2.2 Marginal and penalized quasi-likelihood (MQL and PQL)

ML and REML are relevant to linear multilevel models with Gaussian outcomes; different likelihood-based methods are needed with models for dichotomous outcomes, such as (2). Following Goldstein (2002), in the simpler case of a two-level structure a reasonably general multilevel model for the binary outcome y_{ij} has the form

$$\begin{aligned} &(y_{ij} \,|\, p_{ij}) \sim \text{Bernoulli}(p_{ij}) \quad \text{with} \\ &p_{ij} = f\left(X_{ij}\beta + Z_{ij}^{(1)} e_{ij} + Z_{ij}^{(2)} u_j\right), \end{aligned} \tag{5}$$

where $f(l)$ has a nonlinear character such as $\text{logit}^{-1}(l) = \left(1 + e^{-l}\right)^{-1}$. One approach to the fitting of (5) is through quasi-likelihood methods, which proceed (e.g., Breslow and Clayton (1993)) by linearizing the model via Taylor series expansion; for instance, with H_t as a suitably chosen value around which to expand, the $f(\cdot)$ expression in (5) for the ijth unit at iteration $(t+1)$ may be approximated by

$$\begin{aligned} &f(H_t) + X_{ij}\left(\beta_{t+1} - \beta_t\right) f'(H_t) + \\ &\left(Z_{ij}^{(1)} e_{ij} + Z_{ij}^{(2)} u_j\right) f'(H_t) + \tfrac{1}{2}\left(Z_{ij}^{(1)} e_{ij} + Z_{ij}^{(2)} u_j\right)^2 f''(H_t) \end{aligned} \tag{6}$$

in terms of parameter values estimated at iteration t. The simplest choice, $H_t = X_{ij}\beta_t$, the fixed-part predicted value of the argument of f in (5), yields the marginal quasi-likelihood (MQL) algorithm. This can be improved upon by expanding around the entire current predicted value for the ijth unit, $H_t = X_{ij}\beta_t + Z_{ij}^{(1)}\hat{e}_{ij} + Z_{ij}^{(2)}\hat{u}_j$, where $\hat{e}_{ij}$ and $\hat{u}_j$ are the current estimated random effects; when this is combined with an

improved approximation obtained by replacing the second line in (6) with

$$\begin{aligned}&\left[Z_{ij}^{(1)}\left(e_{ij}-\hat{e}_{ij}\right)+Z_{ij}^{(2)}\left(u_j-\hat{u}_j\right)\right]f'(H_t)+\\&\tfrac{1}{2}\left[Z_{ij}^{(1)}\left(e_{ij}-\hat{e}_{ij}\right)+Z_{ij}^{(2)}\left(u_j-\hat{u}_j\right)\right]^2 f''(H_t)\,,\end{aligned}\tag{7}$$

the result is the penalized or predictive quasi-likelihood (PQL) algorithm. The order of an MQL or PQL algorithm refers to how many terms are used in the Taylor expansion underlying the linearization; for example, (6) is based on expansion up to second order and leads to MQL_2 and PQL_2 estimates. Estimated asymptotic standard errors for MQL/PQL estimates typically derive from a version of observed Fisher information based on the quasi-likelihood function underlying the estimation process; see Breslow and Clayton (1993) for details.

2.3 Markov chain Monte Carlo

The Bayesian fitting of both VC and RELR models involves, as usual in the Bayesian approach, the updating from prior to posterior distributions for the parameters via appropriate likelihood functions; but in both of these model classes closed-form exact expressions for most or all of the relevant joint and marginal posterior distributions are not available (see Chapter 5 of Box and Tiao (1973) for some limited analytical results in the VC model (1)). Instead we rely here on sampling-based approximations to the distributions of interest via Markov chain Monte Carlo (MCMC) methods (e.g., Gilks et al. (1996)): we use a Gibbs sampling approach in the VC model (cf. Seltzer (1993)) and an adaptive hybrid Metropolis-Gibbs method for random-effects logistic regression.

2.3.1 Diffuse priors for multilevel models

As with the Bayesian analysis of all statistical models, broadly speaking two classes of prior distributions are available for multilevel models: (a) diffuse and (b) informative, corresponding to situations in which (a) little is known about the quantities of interest *a priori* or (b) substantial prior information is available, for instance from previous studies judged relevant to the current data set. In situation (a), on which we focus in this paper, it seems natural to seek prior specifications that lead to well-calibrated inferences (e.g., Dawid (1985)), which we take here to mean point estimates with little bias and interval estimates whose actual coverage is close to the nominal level (in both cases in repeated sampling).

There is an extensive literature on the specification of diffuse priors (e.g., Bernardo and Smith (1994), Kass and Wasserman (1996), Spiegelhalter et al. (1997), Gelman et al. (2003)), leading in some models to more than one intuitively reasonable approach. It is sometimes asserted in this literature that the performance of the resulting Bayesian estimates is broadly insensitive, with moderate to large sample sizes, to how the diffuse prior is specified. In preliminary studies we found this to be the case for fixed effects in both the VC and RELR model classes, and in what follows we use (improper) priors that are uniform on the real line $\mathbb{R}$ for such parameters (these are functionally

equivalent to proper Gaussian priors with huge variances). As others (e.g., DuMouchel (1990)) have elsewhere noted, however, we found large differences in performance across plausible attempts to construct diffuse priors for random-effects variances in both model classes. Intuitively this is because the effective sample size for the level–2 variance in a two-level analysis with J level–2 units and N total level–1 units (typically $J << N$) is often much closer to J than to N; in other words, in the language of Example 1, even with data on hundreds of pupils the likelihood information about the between-school variance can be fairly weak when the number of schools is modest, so that prior specification can make a real difference in such cases.

The off-the-shelf (improper) choice for a diffuse prior on a variance in many Bayesian analyses is $p(\sigma^2) \propto \frac{1}{\sigma^2}$, which is equivalent to assuming that $\log\left(\sigma^2\right)$ is uniform on $\mathbb{R}$. This is typically justified by noting that the posterior for σ^2 will be proper even for very small sample sizes; but (e.g., DuMouchel and Waternaux (1992)) this choice can lead to improper posteriors in random-effects models. We avoid this problem by using two alternative diffuse (but proper) priors, both of which produce proper posteriors:

- A locally uniform prior for σ^2 on $(0, \frac{1}{\epsilon})$ for small positive ϵ (Gelman and Rubin (1992), Carlin (1992)), which is equivalent to a Pareto$(1, \epsilon)$ prior for the precision $\lambda = \frac{1}{\sigma^2}$ (Spiegelhalter et al. (1997)); and
- A $\Gamma^{-1}(\epsilon, \epsilon)$ prior for σ^2 (Spiegelhalter et al. (1997)), for small positive ϵ.

Both of these priors are members of the scaled inverse chi-squared $\chi^{-2}(\nu, s^2)$ family (e.g., Gelman et al. (2003)); this is equivalent to an inverse gamma $\Gamma^{-1}\left(\frac{\nu}{2}, \frac{\nu}{2}s^2\right)$ distribution, where ν is the prior effective sample size and s^2 is a prior estimate of σ^2. The $U(0, \frac{1}{\epsilon})$ and $\Gamma^{-1}(\epsilon, \epsilon)$ priors above are formally specified by the choices $(\nu, s^2) = (-2, 0)$ and $(2\epsilon, 1)$, respectively (in the former case in the limit as $\epsilon \to 0$). We have found that results are generally insensitive to the specific choice of ϵ in the region of 0.001 (the default setting in Spiegelhalter et al. (1997)); we report findings with this value. (We also studied the effects of a gently data-determined prior for σ^2—$\chi^{-2}(\epsilon, \hat{\sigma}^2)$ for small ϵ, with REML or PQL estimates used for $\hat{\sigma}^2$—but found that its results were indistinguishable from those of the $\Gamma^{-1}(\epsilon, \epsilon)$ prior.) See, e.g., Natarajan and Kass (2000, 2006) and Gelman (2006) for alternatives to the diffuse priors for variance parameters in hierarchical models which we examine here. Some of these priors (e.g., *approximate uniform shrinkage* and *default conjugate* priors (Natarajan and Kass), or uniform priors on standard deviations instead of random-effects variances (Gelman)) may have better repeated-sampling characteristics than the priors we study; our interest here is in reporting on the performance of two of the most widely-used approaches to diffuse-prior specification in current practice in multilevel modeling.

2.3.2 Gibbs sampling in the VC model

The unknown quantities in the VC model can be split into four groups: the fixed effect β_0, the level–2 residuals u_j, the level–2 variance σ_u^2, and the level–1 variance σ_e^2. Typically the parameters $(\beta_0, \sigma_u^2, \sigma_e^2)$ are of principal interest, but Gibbs sampling in

this model proceeds most smoothly by treating the level–2 residuals as latent variables and sampling in turn from the full conditional distributions $p(\beta_0|y,\sigma_u^2,\sigma_e^2,u)$, $p(u_j|y,\sigma_u^2,\sigma_e^2,\beta_0)$, $p(\sigma_u^2|y,\beta_0,u,\sigma_e^2)$, and $p(\sigma_e^2|y,\beta_0,u,\sigma_u^2)$ (here y and u are the N– and J–vectors of responses and residuals, respectively).

With $\chi^{-2}(\nu_u,s_u^2)$ and $\chi^{-2}(\nu_e,s_e^2)$ priors for σ_u^2 and σ_e^2, respectively, the full conditionals for model (1) have simple and intuitively reasonable Gaussian and inverse gamma forms (cf. Seltzer et al. (1996)):

$$\begin{aligned}
\left(\beta_0|y,\sigma_e^2,u\right) &\sim N\left[\tfrac{1}{N}\textstyle\sum_{ij}\left(y_{ij}-u_j\right),\tfrac{\sigma_e^2}{N}\right],\\
\left(u_j|y,\sigma_u^2,\sigma_e^2,\beta_0\right) &\sim N\left[\tfrac{\hat{D}_j}{\sigma_e^2}\textstyle\sum_{i=1}^{n_j}\left(y_{ij}-\beta_0\right),\hat{D}_j\right],\\
\left(\sigma_u^2|u\right) &\sim \Gamma^{-1}\left[\tfrac{J+\nu_u}{2},\tfrac{1}{2}\left(\nu_u s_u^2+\textstyle\sum_{j=1}^{J}u_j^2\right)\right], \quad \text{and}\\
\left(\sigma_e^2|y,\beta_0,u\right) &\sim \Gamma^{-1}\left[\tfrac{N+\nu_e}{2},\tfrac{1}{2}\left(\nu_e s_e^2+\textstyle\sum_{ij}e_{ij}^2\right)\right],
\end{aligned} \tag{8}$$

where $\hat{D}_j=\left(\frac{n_j}{\sigma_e^2}+\frac{1}{\sigma_u^2}\right)^{-1}$ and $e_{ij}=y_{ij}-\beta_0-u_j$.

It is possible to improve upon the Monte Carlo efficiency of the simple Gibbs sampler (8) in VC models with re-parameterization (e.g., Roberts and Sahu (1997)), and Metropolis-Gibbs hybrids based on block updating of the residuals (as in Browne and Draper (2000) for RELR models) may also lead to Monte Carlo acceleration; we do not pursue these possibilities here. It is worth noting in this context that the hierarchical centering parameterization introduced by Gelfand et al. (1995) only leads to better mixing in VC models if $\sigma_e^2<\sigma_u^2$, which rarely occurs with educational and medical data.

2.3.3 Adaptive hybrid Metropolis-Gibbs sampling in RELR models

Gibbs sampling in RELR models is not straightforward. For example, in the simple model

$$\begin{aligned}
\left(y_{ij}\,|\,p_{ij}\right) &\sim \text{Bernoulli}(p_{ij}), \quad \text{where}\\
\text{logit}(p_{ij}) &= \beta+u_j, \quad u_j\sim N(0,\sigma_u^2),
\end{aligned} \tag{9}$$

and assuming uniform priors for illustration, the full conditional for β is

$$p(\beta|y,u)\propto\prod_{ij}\left(1+e^{-\beta-u_j}\right)^{-y_{ij}}\left(1+e^{\beta+u_j}\right)^{y_{ij}-1}. \tag{10}$$

This distribution does not lend itself readily to direct sampling. Rejection sampling (Zeger and Karim (1991)) is possible, and the software package `WinBUGS` (Spiegelhalter et al. (2003)) employs adaptive rejection sampling (ARS; Gilks and Wild (1992)). In this paper we use a hybrid Metropolis-Gibbs approach in which (a) Gibbs sampling is employed for variances and (b) univariate-update random-walk Metropolis sampling with Gaussian proposal distributions is used for fixed effects and residuals; see Browne (1998) for details. As with VC models we take uniform priors on $\mathbb{R}$ for fixed effects and $\chi^{-2}(\nu,s^2)$ priors for the variances of random effects. The fixed effects and

residuals may also be block-updated using multivariate normal proposal distributions; Browne and Draper (2000) describes comparisons between these two Metropolis alternatives and documents the pronounced Monte Carlo efficiency advantage of the hybrid Metropolis-Gibbs approach over alternatives such as ARS in RELR models, where the former (with block updating) was 1.7 to 9.0 times faster than the latter in achieving the same accuracy of posterior summaries in the examples studied.

Metropolis sampling with univariate normal proposals requires specification of the variances of the proposal distributions. We use scaled versions of the estimated covariance matrices of REML or PQL estimates to set the initial values of the proposal distribution variances, but optimal scaling factors for many multilevel models are not known (Gelman et al. (1995) contains useful results in simple non-hierarchical settings). Our preferred method for specifying the proposal distribution variances is adaptive (see, e.g., Müller (1993) and Gilks et al. (1998) for other approaches to adaptive Metropolis sampling). From starting values based on the estimated covariance matrices, we first employ a sampling period of random length (but with an upper bound) during which the proposal distribution variances are adaptively tuned and eventually fixed for the remainder of the run; this is followed by the usual burn-in period (see Section 2.3.4); and then the main monitoring run from which posterior summaries are calculated occurs. The tuning of the proposal distribution variances is based on achieving an acceptance rate r for each parameter that lies within a specified tolerance interval $(r-\delta, r+\delta)$.

The algorithm examines empirical acceptance rates in batches of 100 iterations, comparing them for each parameter with the tolerance interval and modifying the proposal distribution appropriately before going on to the next batch of 100. With r^* as the acceptance rate in the most recent batch and σ_p as the proposal distribution SD for a given parameter, the modification performed at the end of each batch is as follows:

$$\text{If} \quad r^* \geq r, \quad \sigma_p \rightarrow \sigma_p \left[2 - \left(\frac{1-r^*}{1-r}\right)\right], \quad \text{else} \quad \sigma_p \rightarrow \frac{\sigma_p}{\left(2 - \frac{r^*}{r}\right)}. \tag{11}$$

This modifies the proposal standard deviation by a greater amount the farther the empirical acceptance rate is from the target r. If r^* is too low, the proposed moves are too big, so σ_p is decreased; if r^* is too high, the parameter space is being explored with moves that are too small, and σ_p is increased. If the r^* values are within the tolerance interval during three successive batches of 100 iterations, the parameter is marked as satisfying its tolerance condition, and once all parameters have been marked the overall tolerance condition is satisfied and adapting stops. After a parameter has been marked it is still modified as before until all parameters are marked, but each parameter only needs to be marked once for the algorithm to end. To limit the time spent in the adapting procedure an upper limit is set (we typically use 5,000 iterations) and after this time the adapting period ends regardless of whether the tolerance conditions are met (in practice this occurs rarely). Values of $(r, \delta) = (0.5, 0.1)$ appear to give near-optimal univariate-update Metropolis performance for a wide variety of multilevel models (Browne and Draper (2000)).

Table 3: *Summary of study designs for the VC model (1) simulations.*

Design (J)	Number of pupils per school (n_j)												Total number of pupils (N)
1 (6)	5	10	13	18	24	38							108
2 (6)	18	18	18	18	18	18							108
3 (12)	5	8	10	11	11	12	13	15	20	24	26	61	216
4 (12)	18	18	18	18	18	18	18	18	18	18	18	18	216
5 (24)	5	7	8	10	10	11	11	12	12	13	13	14	432
	15	16	18	19	20	21	23	24	26	29	34	61	
6 (24)	(18 for all schools)												432
7 (48)	5	6	7	8	8	10	10	10	11	11	11	11	864
	12	12	12	12	13	13	13	13	14	14	15	15	
	16	16	17	18	18	19	19	20	20	21	21	21	
	23	24	24	24	25	26	27	29	34	37	38	61	
8 (48)	(18 for all schools)												864

2.3.4 Starting values and burn-in strategy

In MCMC sampling with multilevel models it is natural to use as starting values the likelihood and quasi-likelihood results from ML/REML in VC models and from MQL/PQL in REML models. We have found that marginal posteriors in multilevel models of data sets with all but the tiniest sample sizes, even with diffuse priors, are almost invariably unimodal (but see Liu and Hodges (2003) for a cautionary note); this encourages a relatively short burn-in period without fear of missing significant posterior mass in all but the most unusual of situations. We have found burn-ins of 500 iterations to be more than adequate in both the VC and RELR model classes when likelihood-based starting values are used.

3 Variance-components models

3.1 Simulation study design

We have conducted a large simulation study of the properties of Bayesian and likelihood-based estimation methods in the VC model (1). The design of this study was based on the JSP data set introduced in Section 1.1. The numbers n_j of pupils per school in the subsample of $N = 887$ students described in that section averaged 18.5, with an SD of 10.3 and a range from 5 to 61 (i.e., the sampling design across the $J = 48$ schools was quite unbalanced). To examine the effects of J and the distribution of the n_j in the simulations, we removed one pupil at random from each of the 23 largest schools to yield $N = 864$ students, an average of 18 per school. We then varied the number of schools included in the study, with schools chosen so that the average number of pupils per school was maintained at 18 and the sizes of the individual schools were well spread

out. We considered four sizes of sampling experiment—6, 12, 24 and 48 schools—with a total of 108, 216, 432 and 864 pupils (respectively), and examined one balanced and one unbalanced design in each case. The resulting 8 study designs are given in Table 3. The school-level sample sizes in the cases with unequal n_j were chosen to resemble the actual (highly positively skewed) distribution of class size in the JSP data.

The other factors that varied in our simulations were the true values given to the parameters of model (1): β_0, σ_u^2, and σ_e^2. The fixed effect, β_0, is typically of lesser importance in VC models; we fixed it at 30 throughout all runs. The two variances are more interesting; we chose three possible values for each of these parameters. The between-schools variance, σ_u^2, took the values 1, 10 and 40, and we set the between-pupils variance σ_e^2 to 10, 40 and 80. For realism in the educational context of the JSP data we only examined cases in which $\sigma_e^2 \geq \sigma_u^2$.

A full-factorial experiment varying both size/balance of the classroom samples and true parameter values was both computationally prohibitive and unnecessary (preliminary investigation revealed little or no interaction between these two factors), so we (a) made one set of runs varying the sample sizes as in Table 3, while holding the parameters fixed at values similar to those in the JSP data ($\beta_0 = 30, \sigma_u^2 = 10$, and $\sigma_e^2 = 40$), and (b) held the sample sizes constant at the values specified by design 7 in Table 3 (the layout most similar to the JSP data), and varied the parameters across seven settings—$(\sigma_u^2, \sigma_e^2) = (1, 10), (1, 40), (1, 80), (10, 10), (10, 40), (10, 80), (40, 80)$, giving rise to intraclass correlation values from 0.012 to 0.5—in all cases with $\beta_0 = 30$. We created 1,000 simulated data sets in each cell of the experimental grid; see the Appendix for additional simulation details.

3.2 VC results

3.2.1 Estimator bias

All methods of estimating β_0 we examined yielded negligible bias values; for brevity we omit details. Tables 4 and 5 present Monte Carlo estimates of the relative bias of eight methods of estimating σ_u^2 and σ_e^2 in the VC model (1), and Figures 1 and 2 graphically summarize some aspects of these tables. Two of the methods studied are likelihood-based (ML and REML), the other six Bayesian: two priors for the variances ($\Gamma^{-1}(\epsilon, \epsilon)$ and $U(0, \frac{1}{\epsilon})$) crossed with three methods of summarizing the posterior distribution for the purpose of point estimation (mean, median, mode). In Table 4 σ_u^2 and σ_e^2 were held constant at 10 and 40, respectively, with the results varying across the eight study designs in Table 3; in Table 5 study design 7 was maintained while (σ_u^2, σ_e^2) varied across seven settings. All methods were close to unbiased for the pupil-level variance σ_e^2, because—even in the smallest study designs—data on 108 or more pupils were available (in particular all relative bias estimates for σ_e^2 in the simulations summarized in Table 5 were less than 1%, and we omit these values for brevity). A number of clear conclusions emerge from these tables; we describe the results in the language of schools and pupils, with obvious extension to other settings.

- Bias for all methods drops steadily with increasing N, and tends to be somewhat

Table 4: *Estimates of relative bias for the variance parameters in VC model (1) with a variety of methods and study designs. The true values of σ_u^2 and σ_e^2 were 10 and 40, respectively. Figures in parentheses are Monte Carlo SEs.*

$\boxed{\sigma_u^2}$ Relative Bias (%)		Number of Level–2 Units J (U = unbalanced, B = balanced)							
Estimation Method		6–U	6–B	12–U	12–B	24–U	24–B	48–U	48–B
ML		–22.6 (2.1)	–20.1 (2.0)	–11.9 (1.6)	–9.8 (1.4)	–2.4 (1.1)	–4.1 (1.1)	–2.1 (0.9)	–2.0 (0.8)
REML		–1.0 (2.5)	0.0 (2.4)	–1.0 (1.7)	0.4 (1.5)	3.1 (1.2)	1.0 (1.2)	0.5 (0.9)	0.5 (0.8)
$\Gamma^{-1}(\epsilon,\epsilon)$ Prior	Mean	49.1 (4.1)	51.4 (4.0)	18.4 (2.2)	20.3 (2.1)	12.0 (1.3)	9.7 (1.3)	4.7 (0.9)	4.8 (0.9)
	Median	–6.7 (2.9)	–0.6 (2.8)	–1.9 (1.9)	1.1 (1.8)	3.5 (1.2)	1.7 (1.2)	0.9 (0.9)	1.0 (0.8)
	Mode	–33.6 (1.9)	–31.6 (1.9)	–27.3 (1.5)	–24.1 (1.4)	–12.8 (1.1)	–13.4 (1.1)	–7.7 (0.8)	–7.1 (0.8)
$U(0,\frac{1}{\epsilon})$ Prior	Mean	481 (10.2)	450 (9.7)	74.9 (2.7)	70.9 (2.6)	30.8 (1.4)	26.7 (1.4)	12.5 (1.0)	12.0 (0.9)
	Median	140 (5.1)	133 (4.9)	40.6 (2.3)	39.0 (2.2)	20.1 (1.5)	16.9 (1.3)	8.3 (0.9)	8.0 (0.9)
	Mode	107 (3.8)	94.3 (3.6)	1.2 (1.7)	0.4 (1.6)	0.8 (1.2)	–1.1 (1.2)	–1.0 (0.9)	–0.8 (0.8)

$\boxed{\sigma_e^2}$ Relative Bias (%)		Number of Level–2 Units J (U = unbalanced, B = balanced)							
Estimation Method		6–U	6–B	12–U	12–B	24–U	24–B	48–U	48–B
ML		–0.42	–0.45	–0.02	–0.16	–0.31	–0.15	–0.04	–0.09
REML		–0.42	–0.41	–0.03	–0.16	–0.31	–0.15	–0.04	–0.09
$\Gamma^{-1}(\epsilon,\epsilon)$	Mean	2.8	2.8	1.6	1.4	0.3	0.4	0.3	0.2
	Median	1.1	1.4	0.9	0.7	–0.0	0.1	0.1	0.1
	Mode	–1.2	–1.2	–0.4	–0.6	–0.7	–0.6	–0.2	–0.3
$U(0,\frac{1}{\epsilon})$	Mean	3.5	3.6	2.0	1.9	0.7	0.8	0.4	0.4
	Median	1.8	2.3	1.4	1.3	0.3	0.5	0.3	0.3
	Mode	–0.6	–0.5	0.0	–0.1	–0.3	–0.2	–0.1	–0.1

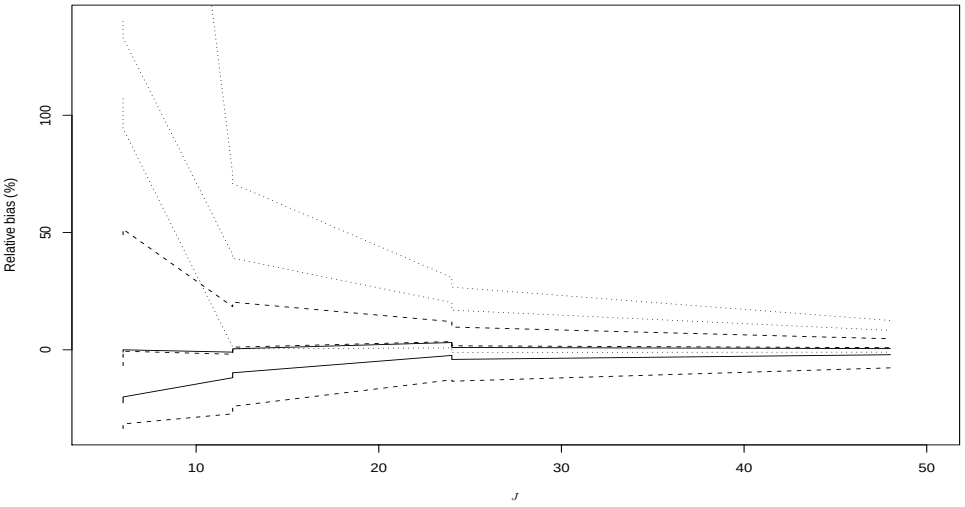
Note: The Monte Carlo SEs for all rows in the σ_e^2 portion of this table were 0.5 (designs 1 and 2), 0.3 (designs 3 and 4) and 0.2 (designs 5–8).

Table 5: *Estimates of relative bias for the variance parameter σ_u^2 in VC model (1) with a variety of methods and true parameter values. All runs use study design 7. Figures in parentheses are Monte Carlo SEs. Column headings record the true values of σ_u^2, σ_e^2, and the intraclass correlation ρ.*

$\boxed{\sigma_u^2}$ Relative Bias (%)		$\sigma_u^2; \sigma_e^2/\rho$			
Estimation Method		1; 80/ 0.012	1; 40/ 0.024	1; 10/ 0.091	10; 80/ 0.111
ML		–3.4	–6.5	–3.1	–2.8
		(3.0)	(2.1)	(1.1)	(1.0)
REML		7.2	0.3	0.4	0.4
		(3.2)	(2.1)	(1.1)	(1.0)
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior	Mean	–22.8	–18.5	3.2	3.7
		(2.5)	(2.1)	(1.2)	(1.1)
	Median	–47.9	–31.7	–1.7	–0.8
		(2.5)	(2.2)	(1.2)	(1.1)
	Mode	–60.0	–50.1	–15.1	–12.7
		(1.7)	(1.7)	(1.1)	(1.0)
$U(0, \frac{1}{\epsilon})$ Prior	Mean	84.5	39.6	15.9	14.8
		(3.1)	(2.2)	(1.2)	(1.1)
	Median	61.0	27.1	10.5	9.8
		(3.1)	(2.1)	(1.2)	(1.1)
	Mode	15.7	–4.2	–3.9	–2.9
		(2.4)	(1.8)	(1.1)	(1.0)

$\boxed{\sigma_u^2}$ Relative Bias (%)		$\sigma_u^2; \sigma_e^2/\rho$		
Estimation Method		10; 40/ 0.200	40; 80/ 0.333	10; 10/ 0.500
ML		–2.1	–1.9	–1.7
		(0.9)	(0.8)	(0.7)
REML		0.5	0.5	0.5
		(0.9)	(0.8)	(0.7)
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior	Mean	4.7	4.9	4.9
		(0.9)	(0.8)	(0.8)
	Median	0.9	1.4	1.6
		(0.9)	(0.8)	(0.7)
	Mode	–7.7	–5.5	–4.5
		(0.8)	(0.7)	(0.7)
$U(0, \frac{1}{\epsilon})$ Prior	Mean	12.5	11.2	10.6
		(1.0)	(0.9)	(0.8)
	Median	8.3	7.4	6.9
		(0.9)	(0.8)	(0.8)
	Mode	–1.0	–0.1	0.4
		(0.9)	(0.8)	(0.7)

Figure 1: *Relative bias (a visual analogue of Table 4) in estimating σ_u^2 as a function of the number of level–2 units J for each of eight estimation methods (likelihood-based estimates are plotted with solid lines, Bayesian estimates with $\Gamma^{-1}(\epsilon, \epsilon)$ priors appear as long dotted lines, and Bayesian estimates with with $U(0, \frac{1}{\epsilon})$ priors are indicated with short dotted lines). Vertical jumps at constant values of J indicate the effects of balanced versus unbalanced sampling designs.*



smaller with balanced designs than when substantial imbalance is present. In Table 5 the magnitude of the bias of estimates of σ_u^2 generally decreases as the intraclass correlation ρ increases from near 0 to 0.5.

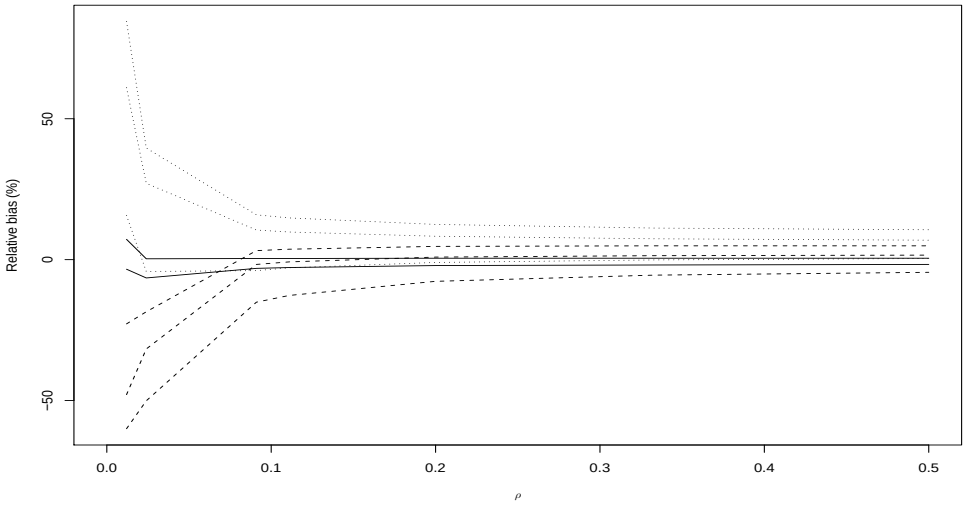
- ML estimates of σ_u^2 are biased low with the smallest designs; this is effectively remedied by the REML bias correction except when ρ is close to 0.

- Posterior means with the $\Gamma^{-1}(\epsilon, \epsilon)$ prior for the school-level variance σ_u^2 are sharply biased high with small sample sizes; this largely disappears when posterior medians are used with this prior. The exception to this pattern occurs when σ_e^2 is 40–80 times larger than σ_u^2, a situation which gave all of the methods trouble but which arguably casts doubt on the need for random effects at level 2 in the first place. Posterior modes with the $\Gamma^{-1}(\epsilon, \epsilon)$ prior are uniformly biased on the low side, sometimes substantially.

- The $U(0, \frac{1}{\epsilon})$ prior can produce huge positive biases when attention focuses on the posterior mean, but has good bias properties with all but the smallest sample sizes when the mode is used as a point estimate. There is clearly a trade-off between choice of prior distribution and choice of posterior summary; the need for these choices gives REML the advantage on bias grounds in small samples.

- The behavior of the two priors is understandable given their shape on the σ^2 scale: $\Gamma^{-1}(\epsilon, \epsilon)$ priors have a sharp spike near 0, which has no effect when the

Figure 2: *Relative bias (a visual analogue of Table 5) in estimating* σ_u^2 *as a function of the intraclass correlation* ρ *(in an unbalanced design with* $J = 48$ *level–2 units and a total of 864 level–1 units) for each of eight estimation methods (likelihood-based estimates are plotted with solid lines, Bayesian estimates with* $\Gamma^{-1}(\epsilon, \epsilon)$ *priors appear as long dotted lines, and Bayesian estimates with with* $U(0, \frac{1}{\epsilon})$ *priors are indicated with short dotted lines).*



likelihood is concentrated away from 0 but which can create appreciable negative bias when the data evidence for positive σ^2 is weak. By contrast $U(0, \frac{1}{\epsilon})$ priors do not have this defect, but claiming in the prior that σ^2 is as likely to be 500 (say) as it is to be 10 creates substantial positive bias when the true value is near 10 but sample sizes are small, leading to a relatively diffuse likelihood. Gelman (2006) makes similar criticisms of the $\Gamma^{-1}(\epsilon, \epsilon)$ prior and offers useful suggestions for alternatives.

3.2.2 Interval performance

We also monitored the coverage and length of interval estimates for the parameters in the VC model (1). To construct Bayesian $100(1-\gamma)\%$ intervals we simply used the $100\frac{\gamma}{2}\%$ and $100\left(1-\frac{\gamma}{2}\right)\%$ quantiles of the relevant posterior distributions (as estimated by MCMC). With the likelihood methods we examined six approaches: the first was intended (as in Examples 1 and 2) to reflect the behavior of many practitioners of multilevel modeling who are presented in the output of the standard computer programs with nothing more than an estimate and a standard error; the second through fifth are simple computationally inexpensive small-sample adjustments to the first for variance components; and the sixth is an idealized version of likelihood interval estimation for variances, assuming knowledge of the sampling distribution which would not be available with a single sample. For brevity we present ML results only for the first method.

- Method 1 used intervals of the form $\left[\hat{\sigma}^2 \pm \Phi^{-1}\left(1-\frac{\gamma}{2}\right)\widehat{SE}\left(\hat{\sigma}^2\right)\right]$ based on asymptotic normality of the MLE.

- In the case of variance parameters, method 2 approximates the sampling distribution of the likelihood estimate by a $\Gamma(\alpha, \beta)$ distribution (preliminary work suggested that this approximation was reasonable for moderate to large sample sizes). In this approach we equated the mean $\frac{\alpha}{\beta}$ of the gamma distribution to $\hat{\sigma}^2$ and the variance $\frac{\alpha}{\beta^2}$ to $\hat{V}\left(\hat{\sigma}^2\right)$, obtaining $[\hat{\alpha}, \hat{\beta}] = \left[\hat{\sigma}^4/\hat{V}\left(\hat{\sigma}^2\right), \hat{\sigma}^2/\hat{V}\left(\hat{\sigma}^2\right)\right]$, and then used quantiles of the corresponding gamma distribution to generate the interval endpoints. (In the smaller study designs the distribution of the REML estimate is a mixture of a point mass at 0 and an approximate gamma distribution conditional on being positive. Any attempt to achieve further improvement in a small-sample likelihood-based approximation would have to cope with the spike at 0.)

- Methods 3 and 4 use Taylor series and transformations to normality. Suppose that the sampling distribution of $g\left(\hat{\sigma}^2\right)$ is approximately Gaussian for some invertible function g, and $\hat{\sigma}^2$ is approximately unbiased. Then by the Δ–method $g\left(\hat{\sigma}^2\right)$ has approximate mean $g\left(\sigma^2\right)$ and variance $\left[g'\left(\sigma^2\right)\right]^2 V\left(\hat{\sigma}^2\right)$, and an approximate $100(1-\gamma)\%$ confidence interval for σ^2 is therefore of the form

$$g^{-1}\left[g\left(\hat{\sigma}^2\right) \pm \Phi^{-1}\left(1-\frac{\gamma}{2}\right)\left|g'\left(\hat{\sigma}^2\right)\right| \widehat{SE}\left(\hat{\sigma}^2\right)\right]. \tag{12}$$

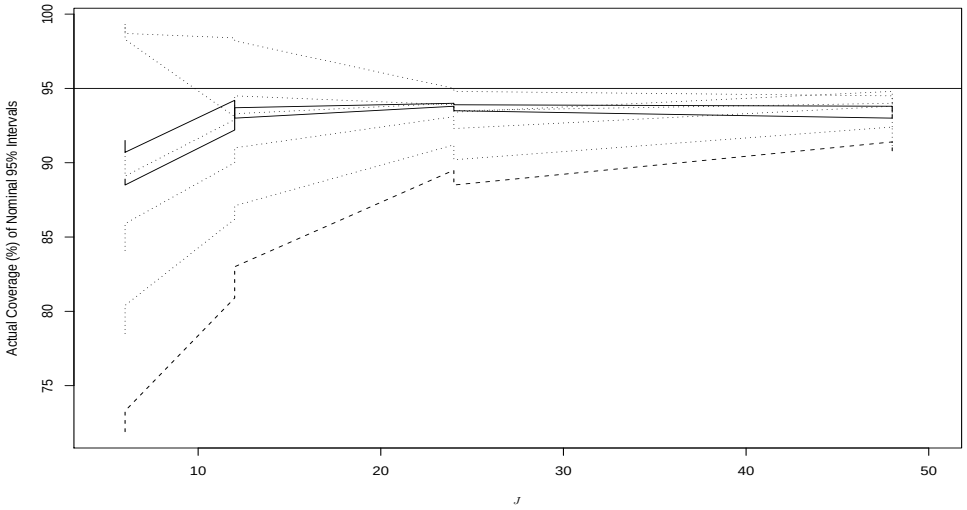
  Method 3 takes the sampling distribution of $g\left(\hat{\sigma}^2\right)$ to be approximately lognormal and uses $g(\cdot) = \ln(\cdot)$, and method 4 employs the Wilson and Hilferty (1931) optimal transformation to normality for gamma random variables, $g(\cdot) = (\cdot)^{\frac{1}{3}}$. Both of these methods fail when $\hat{\sigma}^2 = 0$ because of division by zero in the derivative calculation in (12).

- Method 5, which uses a variance-stabilizing (VS) transformation, is based on the observation by Longford (2000) that (a) the ML estimate of the variance ratio $\tau = \frac{\sigma_u^2}{\sigma_e^2}$ is highly correlated with its estimated asymptotic standard error $\widehat{SE}\left(\hat{\tau}\right)$, and (b) this dependence is removed asymptotically by working instead with $\eta = \ln\left(\bar{n}^{-1} + \tau\right)$, where $\bar{n}$ is a suitably chosen mean of the numbers n_j of level–1 units per level–2 unit (we found that harmonic means work best). This suggests building a Gaussian interval estimate on the η scale, relying on the large-sample result $V(\hat{\eta}_{\mathrm{ML}}) = \frac{2}{J}$, and back-transforming to obtain an interval for τ. We then convert this into an interval for σ_u^2 by using the REML estimate of σ_e^2 in place of σ_e^2, which should yield good performance in our context because the total number N of level–1 units in our simulations never drops below 108. The resulting intervals for σ_u^2 have the form

$$\hat{\sigma}_e^2\left\{\exp\left[\ln\left(\bar{n}^{-1} + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_e^2}\right) \pm \Phi^{-1}\left(1-\frac{\gamma}{2}\right)\sqrt{\frac{2}{J}}\right] - \bar{n}^{-1}\right\}. \tag{13}$$

  These intervals may have a negative left endpoint when σ_u^2 is small in relation to σ_e^2; in many uses of model (1) this is undesirable, but (as Longford points out) reformulations of the model exist in which negative values of τ are sensible subject to a positive-definite constraint on the implied covariance matrix of y.

Figure 3: *Actual coverage of nominal 95% intervals (a visual analogue of Table 6) for σ_u^2 as a function of the number of level–2 units J for each of eight estimation methods (ML intervals are plotted with long dotted lines, Bayesian intervals with $\Gamma^{-1}(\epsilon, \epsilon)$ priors appear as solid lines, and a variety of REML-based intervals are indicated with short dotted lines). Vertical jumps at constant values of J indicate the effects of balanced versus unbalanced sampling designs.*



- To estimate "best possible" (idealized) performance of the likelihood intervals for variances (method 6), we reasoned as follows. As in method 2, the sampling distribution for a likelihood estimate such as $\hat{\sigma}_u^2$ should be approximately gamma, with parameters $(\hat{\alpha}, \hat{\beta})$ which depend on the study design and underlying model parameters, and if these $(\hat{\alpha}, \hat{\beta})$ values were known an interval estimate for σ_u^2 could be formed by analogy with the usual result with an IID Gaussian sample of size n: $\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$, i.e., $\hat{\sigma}^2 \sim \Gamma\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right)$. In each of the cells of our simulation grid we therefore used maximum likelihood (e.g., Johnson et al. (1994)) to estimate $(\hat{\alpha}, \hat{\beta})$ from the 1,000 simulation replications, set $\hat{n} = 2\hat{\alpha} + 1$, and constructed 1,000 idealized interval estimates of the form

$$\left[\frac{(\hat{n}-1)}{\chi^2_{\hat{n}-1,1-\frac{\gamma}{2}}}\hat{\sigma}^2, \frac{(\hat{n}-1)}{\chi^2_{\hat{n}-1,\frac{\gamma}{2}}}\hat{\sigma}^2\right], \tag{14}$$

  where $\chi^2_{k,\gamma}$ is the γ quantile of the χ^2_k distribution. This method also fails when $\hat{\sigma}^2 = 0$ because the MLEs of the parameters of a gamma distribution are undefined if any of the data values are zero.

Tables 6–9 present the actual coverage and mean length of nominal 95% interval estimates for σ_u^2, and Figures 3 and 4 graphically summarize the interval-coverage information in these tables. The following conclusions are evident from the tables and figures, and from other simulation results not presented here (for more details see Browne (1998), which is available on the web at `www.ams.ucsc.edu/~draper`).

Table 6: *Performance of interval estimates of σ_u^2 in VC model (1) with a variety of methods and study designs: actual coverages of nominal 95% intervals. The true values of σ_u^2 and σ_e^2 were 10 and 40, respectively. Values in square brackets report the percentage of time REML yielded variance estimates of zero, and values in curly brackets record the percentage of time the VS intervals had a negative left endpoint (LE).*

$\boxed{\sigma_u^2}$ Coverage (%)		Number of Level–2 Units J (U = unbalanced, B = balanced)			
Estimation Method		6 (U)	6 (B)	12 (U)	12 (B)
ML	Gaussian	71.9	73.3	80.9	83.0
REML	Gaussian	78.5	80.4	86.2	87.1
	Gamma	84.1	85.9	90.0	91.0
	Lognormal*	99.1	98.7	98.4	98.2
	Cube Root*	99.3	98.3	93.1	94.5
	VS	90.7	89.1	92.9	93.3
	Idealized*	94.5	93.6	95.5	95.7
REML	% zero $\hat{\sigma}_u^2$	[4.8%]	[3.6%]	[0.4%]	[0%]
VS	% LE < 0	{43%}	{27%}	{12%}	{3.9%}
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior		88.9	88.5	92.2	93.7
Uniform$(0, \frac{1}{\epsilon})$ Prior		91.5	90.7	94.2	93.0

$\boxed{\sigma_u^2}$ Coverage (%)		Number of Level–2 Units J (U = unbalanced, B = balanced)			
Estimation Method		24 (U)	24 (B)	48 (U)	48 (B)
ML	Gaussian	89.5	88.5	91.4	90.7
REML	Gaussian	91.2	90.2	92.4	91.1
	Gamma	93.1	92.3	93.8	92.4
	Lognormal*	95.0	94.8	94.5	93.9
	Cube Root*	93.9	93.5	94.0	93.5
	VS	90.7	93.4	94.8	93.1
	Idealized*	93.7	94.5	94.7	94.8
REML	% zero $\hat{\sigma}_u^2$	[0%]	[0%]	[0%]	[0%]
VS	% LE < 0	{0.2%}	{0.1%}	{0%}	{0%}
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior		94.0	93.9	93.8	93.4
Uniform$(0, \frac{1}{\epsilon})$ Prior		93.8	93.5	93.0	93.2

Notes: Monte Carlo SEs for coverage rates ranged from 0.3% (for estimates near 99%) to 1.4% (for estimates near 70%). *In the lognormal, cube-root, and idealized cases, interval coverages were based only on the replications in which the estimates were nonzero.

Table 7: *Performance of interval estimates of σ_u^2 in VC model (1) with a variety of methods and study designs: mean interval lengths. The true values of σ_u^2 and σ_e^2 were 10 and 40, respectively. Figures in parentheses are Monte Carlo SEs.*

$\boxed{\sigma_u^2}$ Interval Length		Number of Level–2 Units J (U = unbalanced, B = balanced)			
Estimation Method		6 (U)	6 (B)	12 (U)	12 (B)
ML	Gaussian	23.2 (0.5)	22.9 (0.5)	18.6 (0.3)	18.0 (0.2)
REML	Gaussian	28.3 (0.6)	27.4 (0.6)	20.4 (0.3)	19.5 (0.3)
	Gamma	27.0 (0.6)	26.3 (0.5)	19.9 (0.3)	19.1 (0.3)
	Lognormal*	—	—	—	—
	Cube Root*	36.9 (3.8)	32.9 (1.2)	21.5 (0.3)	20.5 (0.3)
	VS	36.7 (0.7)	33.9 (0.7)	23.4 (0.3)	21.8 (0.3)
	Idealized*	131 (3.2)	112 (2.6)	43.6 (0.7)	39.7 (0.7)
$\Gamma^{-1}(\epsilon,\epsilon)$ Prior		59.5 (1.4)	58.1 (1.3)	29.5 (0.4)	28.4 (0.4)
Uniform$(0,\frac{1}{\epsilon})$ Prior		299 (5.0)	273 (4.7)	46.7 (0.6)	44.0 (0.6)

$\boxed{\sigma_u^2}$ Interval Length		Number of Level–2 Units J (U = unbalanced, B = balanced)			
Estimation Method		24 (U)	24 (B)	48 (U)	48 (B)
ML	Gaussian	14.1 (0.1)	13.4 (0.1)	9.9 (0.1)	9.6 (0.1)
REML	Gaussian	14.7 (0.1)	13.9 (0.1)	10.2 (0.1)	9.8 (0.1)
	Gamma	14.5 (0.1)	13.8 (0.1)	10.1 (0.1)	9.8 (0.1)
	Lognormal*	16.0 (0.1)	15.1 (0.1)	10.6 (0.1)	10.2 (0.1)
	Cube Root*	15.0 (0.1)	14.2 (0.1)	10.2 (0.1)	9.9 (0.1)
	VS	15.8 (0.1)	14.7 (0.1)	10.6 (0.1)	10.1 (0.1)
	Idealized*	19.7 (0.2)	19.4 (0.2)	12.4 (0.1)	11.8 (0.1)
$\Gamma^{-1}(\epsilon,\epsilon)$ Prior		17.5 (0.2)	16.6 (0.2)	11.0 (0.1)	10.6 (0.1)
Uniform$(0,\frac{1}{\epsilon})$ Prior		21.0 (0.2)	19.7 (0.2)	11.8 (0.1)	11.5 (0.1)

Notes: The dashes in the lognormal entries replace enormous numbers arising from division by near-zero values. *In the lognormal, cube-root, and idealized cases, interval lengths were based only on the replications in which the estimates were nonzero (see Table 6).

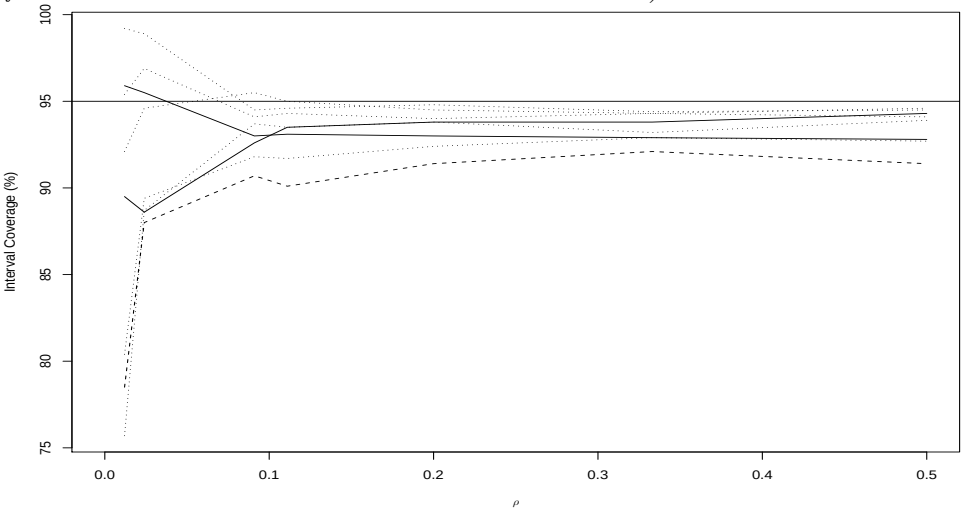
Table 8: *Performance of interval estimates of σ_u^2 in VC model (1) with a variety of methods and true parameter values: actual coverages of nominal 95% intervals. All runs use study design 7. Column headings record the true values of σ_u^2, σ_e^2, and the intraclass correlation ρ. Values in square brackets report the percentage of time REML yielded variance estimates of zero and values in curly brackets record the percentage of time the VS intervals had a negative left endpoint (LE).*

$\boxed{\sigma_u^2}$ Coverage (%)		$\sigma_u^2; \sigma_e^2/\rho$			
Estimation Method		1; 80/ 0.012	1; 40/ 0.024	1; 10/ 0.091	10; 80/ 0.111
ML	Gaussian	78.5	88.0	90.7	90.1
REML	Gaussian	80.4	89.4	91.8	91.7
	Gamma	75.7	88.7	93.7	93.5
	Lognormal*	92.1	94.6	95.5	95.0
	Cube Root*	95.4	96.9	94.1	94.3
	VS	99.2	98.9	94.5	94.6
	Idealized*	90.7	94.6	94.6	94.9
REML	% 0 Estimate	[19%]	[7.0%]	[0.1%]	[0%]
VS	% LE < 0	{92%}	{74%}	{1.2%}	{0.2%}
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior		89.5	88.6	92.6	93.5
Uniform$(0, \frac{1}{\epsilon})$ Prior		95.9	95.5	93.0	93.1

$\boxed{\sigma_u^2}$ Coverage (%)		$\sigma_u^2; \sigma_e^2/\rho$		
Estimation Method		10; 40/ 0.200	40; 80/ 0.333	10; 10/ 0.500
ML	Gaussian	91.4	92.1	91.4
REML	Gaussian	92.4	92.9	92.7
	Gamma	93.8	93.2	93.9
	Lognormal*	94.5	94.3	94.6
	Cube Root*	94.0	94.3	94.1
	VS	94.8	94.4	94.5
	Idealized*	94.7	95.5	95.8
REML	% 0 Estimate	[0%]	[0%]	[0%]
VS	% LE < 0	{0%}	{0%}	{0%}
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior		93.8	93.8	94.3
Uniform$(0, \frac{1}{\epsilon})$ Prior		93.0	92.9	92.8

Notes: See Table 6.

Figure 4: *Actual coverage of nominal 95% intervals (a visual analogue of Table 8) for σ_u^2 as a function of the intraclass correlation ρ (in an unbalanced design with $J = 48$ level–2 units and a total of 864 level–1 units) for each of eight estimation methods (ML intervals are plotted with long dotted lines, Bayesian intervals with $\Gamma^{-1}(\epsilon, \epsilon)$ priors appear as solid lines, and a variety of REML-based intervals are indicated with short dotted lines).*



- Intervals for σ_e^2 (not shown) had close to nominal coverage with all methods, and will not be discussed further. The coverage of ML/REML intervals for the fixed effect β_0 (also not shown) was below nominal with 6–12 schools and 108–216 pupils (study designs 1–4) but approached nominal levels with larger sample sizes. Bayesian interval coverage for β_0 with both $\Gamma^{-1}(\epsilon, \epsilon)$ and $U(0, \frac{1}{\epsilon})$ priors for the variance components was close to nominal in all designs examined (β_0 and σ_u^2 are correlated in the posterior, so the prior for σ_u^2 affects inferences about β_0).

- The effects of imbalance in the design were small but nonzero, and intuitively reasonable: holding the total number of pupils constant, balance yielded narrower intervals and generally better coverage.

- As was the case with bias in the estimation of σ_u^2 (Table 5), interval performance generally improved as the intraclass correlation ρ increased away from 0 (Table 8). Even with data on 48 schools and 864 pupils, both likelihood-based and Bayesian methods can have difficulty apportioning variation within and between schools when σ_e^2 is much larger than σ_u^2.

- ML produced Gaussian intervals for σ_u^2 that were consistently too narrow to achieve good coverage. REML improved on this but still fell below nominal coverage in all situations examined, using both the Gaussian and gamma intervals. In the two smallest designs the lognormal and cube root REML intervals failed to exist 4–5% of the time, and over-covered when they did not fail (and the lognormal intervals continued to over-cover in designs 3 and 4), but with 24 or more level–

Table 9: *Performance of interval estimates of σ_u^2 in VC model (1) with a variety of methods and true parameter values: mean interval lengths. All runs use study design 7; column headings record σ_u^2 and σ_e^2. Figures in parentheses are Monte Carlo SEs.*

σ_u^2 Interval Length		$\sigma_u^2; \sigma_e^2$						
Estimation Method		1; 10	1; 40	1; 80	10; 10	10; 40	10; 80	40; 80
ML	Gaussian	1.27	2.39	3.49	8.41	9.93	11.8	35.7
		(0.01)	(0.03)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)
REML	Gaussian	1.30	2.47	3.66	8.59	10.2	12.1	36.5
		(0.01)	(0.03)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)
	Gamma	1.29	2.34	3.31	8.56	10.1	12.0	36.2
		(0.01)	(0.03)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)
	Lognormal*	1.40	—	—	8.86	10.6	12.8	37.8
		(0.01)	—	—	(0.1)	(0.1)	(0.1)	(0.3)
	Cube Root*	1.32	40.3	14.1	8.65	10.2	12.2	36.8
		(0.01)	(25.5)	(4.0)	(0.1)	(0.1)	(0.1)	(0.2)
	VS	1.41	3.15	5.53	8.83	10.6	12.9	37.7
		(0.01)	(0.02)	(0.03)	(0.06)	(0.1)	(0.1)	(0.3)
	Idealized*	1.81	7.46	13.6	10.0	12.4	16.0	42.9
		(0.02)	(0.15)	(0.3)	(0.1)	(0.1)	(0.2)	(0.3)
$\Gamma^{-1}(\epsilon, \epsilon)$ Prior		1.40	2.28	2.95	9.34	11.0	13.0	39.6
		(0.01)	(0.03)	(0.1)	(0.1)	(0.1)	(0.1)	(0.3)
Uniform$(0, \frac{1}{\epsilon})$ Prior		1.53	2.99	4.71	9.97	11.8	14.2	42.5
		(0.01)	(0.03)	(0.05)	(0.1)	(0.1)	(0.1)	(0.3)

Notes: See Table 7.

2 units (schools) both transformation-based methods improved on the Gaussian and gamma intervals and achieved coverages close to nominal. The lognormal and cube root REML intervals failed to exist 7–19% of the time when $\rho \leq 0.024$, but—as mentioned earlier—in such situations the need for VC modeling is unclear. The VS intervals sharply over-covered when ρ was small and had a negative left endpoint 4–43% of the time in the smallest designs, but performed well otherwise.

- Bayesian intervals for σ_u^2 with the $U(0, \frac{1}{\epsilon})$ prior had actual coverages at or close to nominal levels in all study designs and parameter settings examined. The $\Gamma^{-1}(\epsilon, \epsilon)$ intervals undercovered to some extent (actual levels near 90% at nominal 95%) when the number of level–2 units or the variance ratio τ were small, but performed well in all other situations. Note, however, that the $U(0, \frac{1}{\epsilon})$ intervals were extremely wide with small samples (Table 7); further work is needed to see if other prior specifications might yield narrower but still well-calibrated intervals in such situations.

- In some cases the REML asymptotic standard errors underestimated the actual sampling variabilities they were meant to estimate. This may be seen from the substantially improved performance of the idealized intervals over the REML gamma intervals in small samples, and is also clear from a comparison of the

mean value of the REML squared standard errors for $\hat{\sigma}_u^2$ with the sample variance of the 1,000 simulated $\hat{\sigma}_u^2$ values: across studies 1–8 in Table 6, ratios of the form $\left\{\text{mean}\left[\widehat{SE}^2\left(\hat{\sigma}^2\right)\right] / \hat{V}\left(\hat{\sigma}^2\right)\right\}$ came out $(0.854, 0.837, 0.920, 0.910, 1.02, 0.933, 0.899, 0.925)$, respectively, i.e., the REML squared SEs underestimated the sampling variances on average by 15–16% in studies 1–2 (see Longford (2000) for a theoretical explanation of this phenomenon).

4 Random-effects logistic regression models

4.1 Simulation study design

We have also conducted a large simulation study of the properties of quasi-likelihood and Bayesian estimation methods in the RELR model (2). The design of this study was based on the Rodríguez-Goldman data set introduced in Section 1.2. Conditioning on both the covariates $(x_{1ijk}, x_{2jk}, x_{3k})$ and the true parameter values $(\beta_0 = 0.65, \beta_1 = \beta_2 = \beta_3 = \sigma_u^2 = \sigma_v^2 = 1.0)$ used by Rodríguez and Goldman (1995) in their likelihood-based simulation study, we used model (2) to create 500 simulation replications of the Rodríguez-Goldman data structure, each with 161 communities, 1,558 mothers, and 2,449 births.

For each simulated data set we estimated the six parameters using two quasi-likelihood methods—MQL_1 and PQL_2—and Bayesian fitting with two priors. In the quasi-likelihood estimation we used a convergence tolerance (maximum relative change in parameter values from one iteration to the next) of 0.01. For Bayesian estimation we used MCMC with (improper) uniform priors on $\mathbb{R}$ on the β_l and two prior distributions on the variance components: $\Gamma^{-1}(\epsilon, \epsilon)$ and (improper) uniform on $(0, \infty)$, functionally equivalent to a proper $U(0, \frac{1}{\epsilon})$ prior for small ϵ. We used the adaptive hybrid Metropolis-Gibbs method described in Section 2.3.3, with a maximum adaptation period of 5,000 iterations, a target acceptance rate of 44%, a burn-in from PQL_2 starting values of 500 iterations, and a monitoring run of 25,000 iterations (based on Raftery and Lewis (1992) default accuracy recommendations).

4.2 RELR results

Tables 10–11 and Figures 5 and 6 present our simulation findings. For each of the six parameters the tables contrast the mean estimate, and coverage and length of nominal 95% intervals, for the various estimation methods, using posterior means as Bayesian point estimates (medians and modes gave essentially the same results); the table also summarizes large-sample Gaussian intervals—and gamma, lognormal, and idealized intervals as in Section 3.2.2 for the variance parameters—based on the quasi-likelihood methods (the cube root results were inferior to those from the lognormal approximation, and the VS method is not readily adaptable to this setting since there is no direct estimate of the level–1 variance). Figures 5 and 6 give calibration plots for the six parameters (three in each figure), in which nominal and actual coverage of $100(1-\gamma)\%$ intervals for $\gamma = 0.01, 0.02, \ldots, 0.99$ are contrasted for the various estimation meth-

Table 10: *Mean estimates (top table) and coverage (bottom table) of nominal 95% intervals, for four estimation methods in RELR model (2) with the Rodríguez-Goldman data structure. True values of the parameters are given in square brackets in the top table. 95% central posterior Bayesian intervals are reported, and figures in parentheses are Monte Carlo SEs.*

Mean Estimate		Parameter					
Estimation Method		β_0 [0.65]	β_1 [1.0]	β_2 [1.0]	β_3 [1.0]	σ_v^2 [1.0]	σ_u^2 [1.0]
MQL_1		0.474	0.741	0.753	0.727	0.550	0.026
		(0.007)	(0.007)	(0.004)	(0.009)	(0.004)	(0.002)
PQL_2		0.612	0.945	0.958	0.942	0.888	0.568
		(0.009)	(0.009)	(0.005)	(0.011)	(0.009)	(0.010)
Bayesian	$\Gamma^{-1}(\epsilon,\epsilon)$ Priors	0.638	0.991	1.006	0.982	1.023	0.964
		(0.010)	(0.010)	(0.006)	(0.012)	(0.011)	(0.018)
	$U(0,\infty)$ Priors	0.655	1.015	1.031	1.007	1.108	1.130
		(0.010)	(0.010)	(0.005)	(0.013)	(0.011)	(0.016)

Actual Coverage (%)		Parameter					
Estimation Method		β_0	β_1	β_2	β_3	σ_v^2	σ_u^2
MQL_1	Gaussian	76.8	68.6	17.6	69.6	2.4	0.0
		(1.9)	(2.1)	(1.7)	(2.1)	(0.7)	(—)
PQL_2	Gaussian	92.0	96.2	90.8	89.8	77.6	26.8
		(1.2)	(0.9)	(1.3)	(1.4)	(1.9)	(2.0)
	Gamma	—	—	—	—	81.0	31.4
		—	—	—	—	(1.8)	(2.1)
	Lognormal	—	—	—	—	84.2	37.4
		—	—	—	—	(1.6)	(2.1)
	Idealized	—	—	—	—	93.6	83.4
		—	—	—	—	(1.1)	(1.7)
Bayesian	$\Gamma^{-1}(\epsilon,\epsilon)$ Priors	93.2	96.4	92.6	92.2	94.4	88.6
		(1.1)	(0.8)	(1.2)	(1.2)	(1.0)	(1.4)
	$U(0,\infty)$ Priors	93.6	96.4	92.8	93.6	92.2	93.0
		(1.1)	(0.8)	(1.2)	(1.1)	(1.2)	(1.1)

Table 11: *Mean length of nominal 95% intervals, for four estimation methods in RELR model (2) with the Rodríguez-Goldman data structure. True values of the parameters are given in square brackets. 95% central posterior Bayesian intervals are reported, and figures in parentheses are Monte Carlo SEs.*

Mean Interval Length		Parameter					
Estimation Method		β_0 [0.65]	β_1 [1.0]	β_2 [1.0]	β_3 [1.0]	σ_v^2 [1.0]	σ_u^2 [1.0]
MQL_1	Gaussian	0.589 (0.001)	0.681 (0.001)	0.327 (0.001)	0.746 (0.001)	0.404 (0.001)	0.177 (0.001)
PQL_2	Gaussian	0.735 (0.003)	0.796 (0.002)	0.400 (0.001)	0.930 (0.003)	0.638 (0.005)	0.591 (0.002)
	Gamma	— —	— —	— —	— —	0.636 (0.005)	0.586 (0.003)
	Lognormal	— —	— —	— —	— —	0.641 (0.005)	0.635 (0.004)
	Idealized	— —	— —	— —	— —	0.851 (0.009)	1.25 (0.022)
Bayesian	$\Gamma^{-1}(\epsilon,\epsilon)$ Priors	0.798 (0.004)	0.875 (0.003)	0.463 (0.002)	1.01 (0.004)	0.878 (0.009)	1.25 (0.015)
	$U(0,\infty)$ Priors	0.828 (0.003)	0.895 (0.002)	0.476 (0.002)	1.05 (0.004)	0.948 (0.008)	1.32 (0.011)

ods, using Gaussian intervals for MQL_1 and PQL_2 for the fixed effects and adding the PQL_2 lognormal intervals for the variance parameters. The following conclusions may be drawn from these summaries.

- MQL_1 yielded sharply biased estimates and very poor coverage properties, especially for the random-effects variances (e.g., the MQL_1 point estimate of the level–2 variance σ_u^2 was 0 in 58% of the simulated data sets). PQL_2 produced a considerable improvement, but bias and undercoverage with the Gaussian intervals were still noticeable, especially for σ_u^2. The lognormal intervals offered some improvement but still exhibited substantial undercoverage.

- PQL_2 underperformed for the variance estimates both because the PQL estimates are biased low and because the PQL standard errors are too small (see Engel (1998) and Lee and Nelder (2001) for theoretical results that support this conclusion). As was the case with the VC model, this may be seen in two ways: (a) by the improved performance of the idealized interval estimates and (b) through the ratios $\left\{\text{mean}\left[\widehat{SE}^2\left(\hat{\sigma}^2\right)\right] / \hat{V}\left(\hat{\sigma}^2\right)\right\}$, which were 0.447 and 0.672 for σ_u^2 and σ_v^2, respectively, i.e., the typical estimated variance of $\hat{\sigma}_u^2$ in any given simulated data set was only about 45% of the actual sampling variance across the 500 data sets. This seems to be largely a small-sample problem for PQL—even though each simulated data set had 2,449 births, the average number of women per community

(the most important determinant of the accuracy of $\hat{\sigma}_u^2$) was only $\frac{1{,}558}{161} \doteq 9.7$—but note that even with 161 communities the PQL performance for σ_v^2 was also unsatisfactory.

- Bayesian estimates with both priors were close to unbiased and well calibrated for all parameters, with actual coverage values close to nominal at all levels in Figures 1 and 2.

5 Summary and conclusions

In two large simulations studies whose design is realistic for educational and medical research (as well as other fields of inquiry), we have examined the performance of likelihood-based and Bayesian methods of fitting variance-components (VC) and random-effects logistic regression (RELR) models, focusing on the likelihood-based approaches in most frequent current use in the applied multilevel-modeling literature. Our main findings are as follows.

- In two-level VC models with a wide variety of sample sizes and true parameter values,
  - Both likelihood-based (ML and REML) and Bayesian (diffuse-prior) methods can be made to yield approximately unbiased point estimates, in the likelihood case by using REML rather than ML estimates, and in the Bayesian case by choosing one of several combinations of diffuse priors and posterior point summaries (specifically, for random-effects variances: posterior medians for $\Gamma^{-1}(\epsilon, \epsilon)$ priors and posterior modes for $U(0, \frac{1}{\epsilon})$ priors, in both cases for small ϵ; these combinations produce approximate unbiasedness in all but the smallest designs (i.e., those with fewer than about a dozen cluster units in the hierarchy)). The automatic nature of REML's bias correction represents an advantage for the likelihood-based approach as far as bias is concerned with small samples (see Section 3.2.1, Tables 4 and 5, and Figures 1 and 2 for details);
  - However, both approaches experienced difficulty in attaining nominal coverage of interval estimates in two situations: when (i) the number J of level–2 (cluster) units and/or (ii) the variance ratio $\tau = \frac{\sigma_u^2}{\sigma_e^2}$ between levels 2 and 1 (or equivalently the intraclass correlation $\rho = \frac{\tau}{\tau+1}$) are small (see Section 3.2.2, Tables 6–9, and Figures 3 and 4 for details on the magnitude of these effects).

- In the three-level RELR model we studied (which had 161 units at level 3, an average of 9.7 level–2 units per level–3 unit, and a total of 2,449 level–1 units),
  - quasi-likelihood methods performed badly in terms of bias of point estimates and coverage of interval estimates for random-effects variances (see Section 4.2 and Tables 10 and 11 for details); and

Figure 5: *Actual versus nominal coverage of four estimation methods for the parameters β_0, β_1 and β_2 in the RELR model (2).*

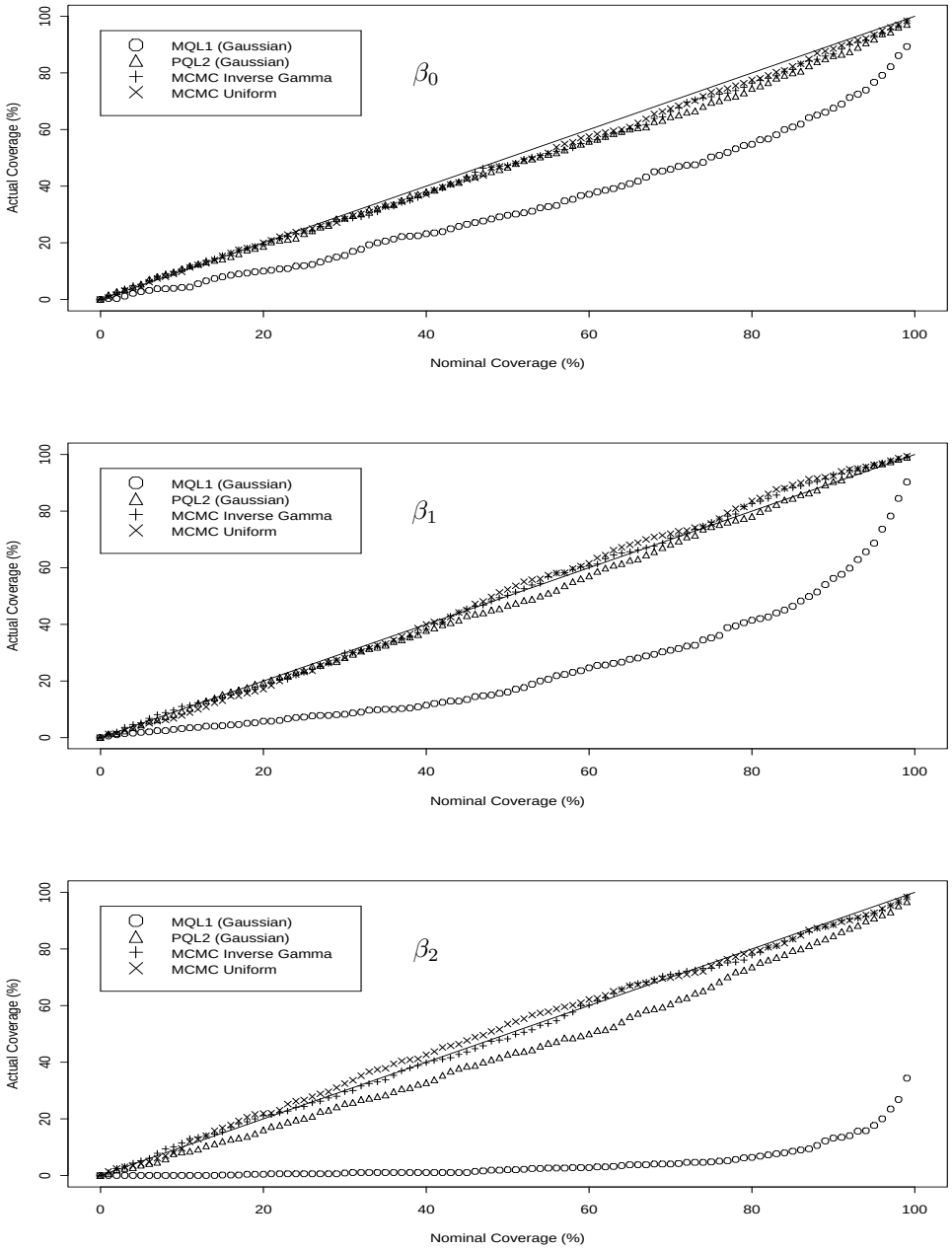
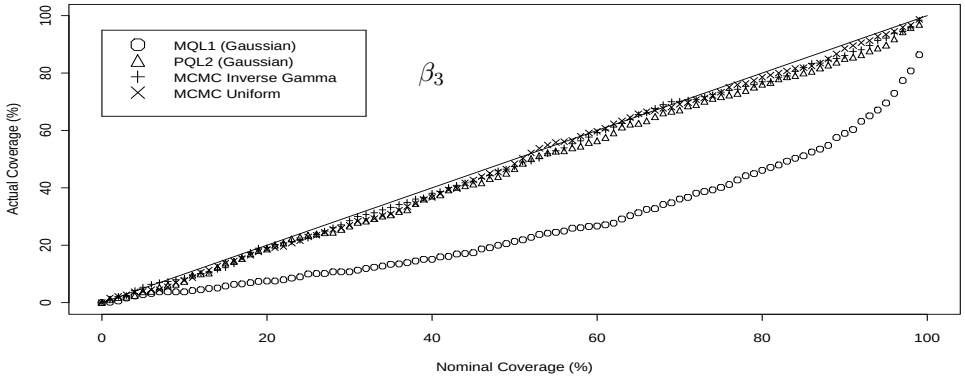
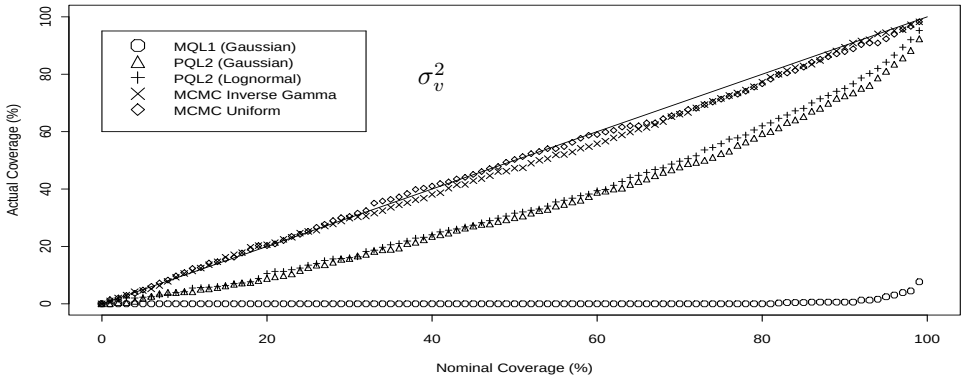
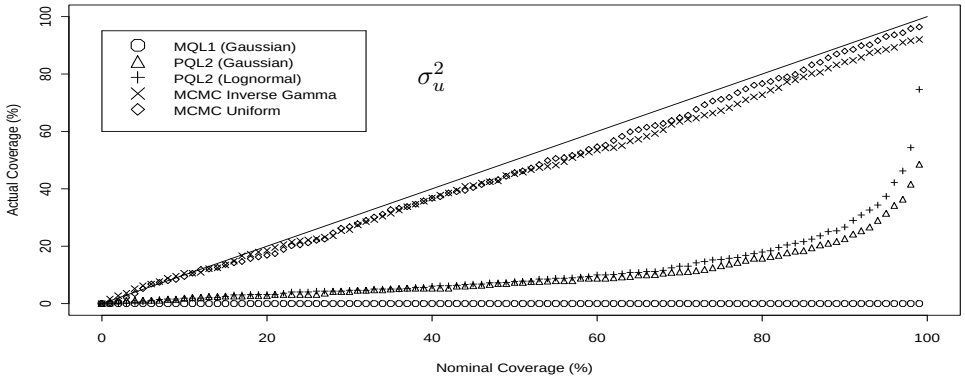


Figure 6: *Actual versus nominal coverage of four estimation methods for the parameters* β_3, σ_v^2 *and* σ_u^2 *in the RELR model (2).*







  - Bayesian methods with diffuse priors were well-calibrated in both point and interval estimation for all parameters of the model (see Figures 5 and 6 for a dramatic summary of the calibration picture in RELR models).

Our RELR results, narrowly construed, apply only to the 3–level model (2) with sample sizes like those in Example 2 of Section 1.2, but our quasi-likelihood conclusions are consistent with broad theoretical predictions made by Engel (1998) and Lee and Nelder (2001), and our Bayesian calibration findings are in line with those in other multilevel settings we have examined (e.g., Browne and Draper (2000), Browne et al. (2002)).

These results bear comment both methodologically and in their practical implications for applied multilevel modeling in health care, education, and other fields. On the methodological side,

- Further study is needed to see if alternative diffuse priors (e.g., Daniels (1999), Natarajan and Kass (2000, 2006), Gelman (2006)) can remedy the undercoverage of Bayesian intervals (and achieve approximate unbiasedness without the need to select a method of posterior summary depending on the problem) with small numbers of level–2 units in 2–level VC models; we intend to report on this elsewhere. Likelihood-based intervals of the kind we have studied here underperform in that situation for a fundamental reason that would be harder to remedy: the insistence on maximization (rather than integration) over the parameters of a highly-skewed likelihood surface with its marginal maximum at $\sigma_u^2 = 0$ leads to zero point estimates in small samples with some frequency when the true value is well away from 0;

- The usual quasi-likelihood computer output in RELR models may not be trustworthy either for point estimation or uncertainty assessment, in the latter case because the estimated asymptotic standard errors can be systematically too small when the mean numbers of level–k units per level–$(k+1)$ unit (and/or the number of level–M units in an M–level model) are small for $k \geq 1$; and

- There is an expectation, expressed formally in the Bernstein-von Mises theorem (e.g., Freedman (1999); also see Samaniego and Reneau (1994) and Severini (1994)), that likelihood and diffuse-prior Bayesian results will be close in large samples, and this will typically occur when parametric models with a modest number of parameters are fit to data not possessing a hierarchical structure. However,

  - what looks like a large sample in multilevel modeling may not be so large in reality, because the effective sample sizes for variances of random effects at levels greater than 1 in the hierarchy are mainly governed not by the total number of level–1 units (which will often be large) but by the numbers of units at the other levels, which are often much smaller; and
  - exact-likelihood methods for non-Gaussian multilevel models have until fairly recently been difficult to implement (because evaluation of the likelihood

function involves integrating over the random effects), with the result that approximate methods such as quasi-likelihood techniques in RELR models have gained widespread use, and the Bernstein-von Mises theorem says nothing about agreement between Bayesian and *approximate* likelihood approaches unless the approximation is good.

On the practical side, as mentioned in Section 1.3, likelihood methods that may prove superior to quasi-likelihood have recently been under development, based on (a) Gaussian quadrature (e.g., Pinheiro and Bates (1995); see the `SAS` (SAS-Institute (2006)) procedure `MIXED` for VC model fitting and the packages `EGRET`, `MIXOR`, and `LIMDEP`, the `SAS` procedure `NLMIXED`, and the `SAS` macro `NLINMIX` for examples of quadrature implementations in RELR models; note however that, since these programs are only applicable to 2–level designs, they could not be used on the RELR models in this paper), (b) nonparametric maximum likelihood (Aitkin (1999a), supported by `GLIM4` macros written by the author), (c) Laplace approximations (Raudenbush et al. (2000), available in `HLM`), (d) hierarchical generalized linear models (Lee and Nelder (2001), as implemented in `GENSTAT` macros), and (e) profile likelihood (e.g., Longford (2000)); parametric bootstrapping of PQL estimates (e.g., Rodríguez and Goldman (2001), for instance using `MLwiN`) may well lead to significant improvement in RELR models as well. We are not aware of large-scale simulation results on the calibration of these approaches in small samples; the literature seems particularly silent on the quality of interval estimates produced by these methods.

One important likelihood-Bayesian comparison we have not addressed is computational speed, where ML/REML and MQL/PQL approaches have a distinct advantage (for example, PQL_2 fitting of model (2) to the Rodríguez-Goldman data set in Table 2 takes less than 3 seconds on a 3GHz PC versus 1.8 minutes using MCMC with 25,000 monitoring iterations). However, (i) steady improvements in recent years in both hardware speed and efficiency of Monte Carlo algorithms and (ii) the lack of calibration of likelihood-based methods in some common hierarchical settings combine to make MCMC-based Bayesian fitting of multilevel models an attractive approach, even with rather large data sets. Other analytic strategies based on less approximate likelihood methods are also possible but would benefit from further study of the type summarized here.

Acknowledgments

We are grateful to the U.K. EPSRC and ESRC and the University of Bath (U.K.) for financial support, and to Murray Aitkin, David Clayton, Constantine Gatsonis, Andrew Gelman, Wally Gilks, Harvey Goldstein, Sander Greenland, Jim Hodges, Rob Kass, Youngjo Lee, Dennis Lindley, Nick Longford, John Nelder, Jon Rasbash, Steve Raudenbush, Tony Robinson, Michael Seltzer, David Spiegelhalter and a referee for references and valuable comments on this and related papers and presentations. Membership on this list does not imply agreement with the ideas expressed here, nor are any of these people responsible for any errors that may be present.

Appendix: Computing details

In the VC simulations, to decide how long to monitor the Gibbs-sampling output we estimated time per iteration and calculated Raftery and Lewis (1992) diagnostics as a function of the total number of pupils N. This revealed that the smaller designs in Table 3 needed longer monitoring runs to satisfy Raftery-Lewis default accuracy constraints but took less time per iteration, leading to the following monitoring run lengths M: 50,000 in studies 1 and 2, 30,000 in 3 and 4, 20,000 in 5 and 6, and 10,000 in studies 7 and 8. The full set of VC simulations took 1.8 GHz-months of CPU time on 3 Sun SPARCstations and a Pentium-based PC.

The data sets in Examples 1 and 2, and `WinBUGS` and `MLwiN` programs to fit models (1) and (2) to those examples, are available on the web at `www.ams.ucsc.edu/~draper`.

References

Aitkin, M. 1996. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**: 251–262. 479

—. 1999a. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**: 117–128. 478, 506

—. 1999b. Meta-analysis by random-effects modelling in generalized linear models. *Statistics in Medicine* **18**: 2343–2351. 479

Bernardo, J. M. and A. F. M. Smith. 1994. *Bayesian Theory*. New York: Wiley. 482

Besag, J., P. Green, D. Higdon, and K. Mengersen. 1995. Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**: 3–41. 479

Box, G. E. P. and G. C. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. New York: Wiley. 479, 482

Breslow, N. E. and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *Journal of Statistical Computation and Simulation* **88**: 9–25. 480, 481, 482

Brown, K. G. and M. A. Burgess. 1984. On maximum likelihood and restricted maximum likelihood approaches to estimation of variance components. *Journal of Statistical Computation and Simulation* **19**: 59–77. 480

Browne, W. J. 1998. *Applying MCMC Methods to Multilevel Models*. Ph.D. dissertation, Department of Mathematical Sciences, University of Bath, U.K. 484, 493

Browne, W. J. and D. Draper. 2000. Implementation and performance issues in the Bayesian fitting of multilevel models. *Computational Statistics* **15**: 391–420. 478, 484, 485, 505

Browne, W. J., D. Draper, H. Goldstein, and J. Rasbash. 2002. Bayesian and likelihood methods for fitting multilevel models with complex level–1 variation. *Computational Statistics and Data Analysis* **39**: 203–225. 478, 505

Bryk, A. S. and S. W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. London: Sage. 479

Carlin, B. 1992. Discussion of "Hierarchical models for combining information and for meta-analysis," by C. N. Morris and S. L. Normand. In *Bayesian Statistics*, vol. 4, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (editors), 336–338. Oxford: Clarendon Press. 483

Carlin, B. and T. A. Louis. 2001. *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd ed. London: Chapman & Hall. 479

Chaloner, K. 1987. A Bayesian approach to the estimation of variance components in the unbalanced one-way random-effects model. *Technometrics* **29**: 323–337. 480

Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley. 474, 475

Corbeil, R. R. and S. R. Searle. 1976. Restricted maximum likelihood (REML) estimation of variance components in mixed models. *Technometrics* **18**: 31–38. 479

Daniels, M. 1999. A prior for the variance in hierarchical models. *Canadian Journal of Statistics* **27**: 569–580. 505

Daniels, M. J. and C. Gatsonis. 1999. Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association* **94**: 29–42. 479

Dawid, A. P. 1985. Calibration-based empirical probability. *Annals of Statistics* **13**: 1251–1274. 482

Donner, A. 1986. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* **54**: 67–82. 475

Draper, D. 1995. Inference and hierarchical modeling in the social sciences (with discussion). *Journal of Educational and Behavioral Statistics* **20**: 115–147, 233–239. 474

DuMouchel, W. 1990. Bayesian meta-analysis. In *Statistical Methodology in the Pharmaceutical Sciences*, D. Berry (editor), 509–529. New York: Marcel Dekker. 483

DuMouchel, W. and C. Waternaux. 1992. Discussion of "Hierarchical models for combining information and for meta-analysis," by C. N. Morris and S. L. Normand. In *Bayesian Statistics*, vol. 4, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (editors), 338–341. Oxford: Clarendon Press. 483

Engel, B. 1998. A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal* **40**: 141–154. 501, 505

Freedman, D. 1999. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics* **27**: 1119–1140. 505

Gelfand, A. and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409. 479

Gelfand, A. E., S. K. Sahu, and B. P. Carlin. 1995. Efficient parameterizations for generalized linear mixed models (with discussion). In *Bayesian Statistics*, vol. 5, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (editors), 165–180. Oxford: Clarendon Press. 484

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* (this issue). 483, 491, 505

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. *Bayesian Data Analysis*. 2nd ed. London: Chapman & Hall. 482, 483

Gelman, A., G. O. Roberts, and W. R. Gilks. 1995. Efficient Metropolis jumping rules. In *Bayesian Statistics*, vol. 5, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (editors), 599–607. Oxford: Clarendon Press. 485

Gelman, A. and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**: 457–511. 483

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall. 482

Gilks, W. R., G. O. Roberts, and S. K. Sahu. 1998. Adaptive Markov chain Monte Carlo sampling through regeneration. *Journal of the American Statistical Association* **93**: 1045–1054. 485

Gilks, W. R. and P. Wild. 1992. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**: 337–348. 484

Goldstein, H. 1986. Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika* **73**: 43–56. 474, 475, 480

—. 1989. Restricted unbiased iterative generalised least squares estimation. *Biometrika* **76**: 622–623. 474, 475, 481

—. 2002. *Multilevel Statistical Models*. 3rd ed. London: Hodder Arnold. 476, 479, 480, 481

Goldstein, H. and J. Rasbash. 1996. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society (Series A)* **159**: 505–513. 480

Goldstein, H., J. Rasbash, M. Yang, G. Woodhouse, H. Pan, D. Nutall, and S. Thomas. 1993. A multilevel analysis of school examination results. *Oxford Review of Education* **19**: 425–433. 474

Goldstein, H. and D. J. Spiegelhalter. 1996. League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society (Series A)* **159**: 385–444. 474

Harville, D. A. and A. G. Zimmermann. 1996. The posterior distribution of the fixed and random effects in a mixed-effects linear model. *Journal of Statistical Computation and Simulation* **54**: 211–229. 479

Henderson, C. R. 1950. Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics* **21**: 309–310. 479

Huber, D. A., T. L. White, and G. R. Hodge. 1994. Variance-component estimation techniques compared for two mating designs with forest genetic architecture through computer simulation. *Theoretical and Applied Genetics* **88**: 236–242. 480

Huber, P. J. 1967. The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, vol. 1, 221–233. Berkeley CA: University of California Press. 474

Hulting, F. L. and D. A. Harville. 1991. Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association* **86**: 557–568. 479

Johnson, N. L., S. Kotz, and N. Balakrishnan. 1994. *Continuous Univariate Distributions (Volume 1)*. 2nd ed. New York: Wiley. 493

Kahn, M. J. and A. E. Raftery. 1996. Discharge rates of Medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. *Journal of the American Statistical Association* **91**: 29–41. 479

Kass, R. E. and D. Steffey. 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association* **84**: 717–726. 479

Kass, R. E. and L. Wasserman. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**: 1343–1370. 482

Klotz, J. H., R. C. Milton, and S. Zacks. 1969. Mean square efficiency of estimators of variance components. *Journal of the American Statistical Association* **64**: 1383–1394. 479

Lee, Y. and J. A. Nelder. 1996. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society (Series B)* **58**: 619–678. 479

—. 2001. Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models, and structured dispersion. *Biometrika* **88**: 987–1006. 478, 501, 505, 506

Lesaffre, E. and B. Spiessens. 2001. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Journal of the Royal Statistical Society (Series C)* **50**: 325–335. 478

Liu, J. N. and J. S. Hodges. 2003. Posterior bimodality in the balanced one-way random-effects model. *Journal of the Royal Statistical Society (Series B)* **65**: 247–255. 486

Longford, N. T. 1987. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**: 817–827. 474, 479

—. 1997. Comment on "Improved approximations for multilevel models with binary responses," by H. Goldstein and J. Rasbash. *Journal of the Royal Statistical Society (Series A)* **160**: 593. 479

—. 2000. On estimating standard errors in multilevel analysis. *The Statistician* **49**: 389–398. 478, 492, 499, 506

Mortimore, P., P. Sammons, L. Stoll, D. Lewis, and R. Ecob. 1988. *School Matters*. Wells: Open Books. 474

Müller, P. 1993. A generic approach to posterior integration and Gibbs sampling. Technical report, Institute of Statistics and Decision Sciences, Duke University. 485

Natarajan, R. and R. E. Kass. 2000. Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association* **95**: 227–237. 480, 483, 505

—. 2006. A defaulty conjugate prior for variance components in generalized linear mixed models. *Bayesian Analysis* (this issue). 483, 505

Pebley, A. R. and N. Goldman. 1992. *Family, community, ethnic identity, and the use of formal health care services in Guatemala*. Working Paper 92–12. Princeton NJ: Office of Population Research. 476

Pinheiro, J. C. and D. M. Bates. 1995. Approximations to the log-likelihood function in the non-linear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**: 12–35. 478, 506

Portnoy, S. 1971. Formal Bayes estimation with application to a random-effects model. *Annals of Mathematical Statistics* **42**: 1379–1388. 479

Raftery, A. L. and S. Lewis. 1992. How many iterations in the Gibbs sampler? In *Bayesian Statistics*, vol. 4, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (editors), 763–774. Oxford: Clarendon Press. 499, 507

Rasbash, J., F. Steele, W. Browne, and B. Prosser. 2005. *A User's Guide To `MLwiN`, Version 2.0*. University of Bristol, U.K. 474

Raudenbush, S., A. Bryk, and R. Congdon. 2005. *HLM: Hierarchical Linear and Nonlinear Modeling*. Available at `www.ssicentral.com/hlm`. 474

Raudenbush, S. W. 1994. Equivalence of Fisher scoring to iterative generalized least squares in the normal case, with application to hierarchical linear models. Technical Report, College of Education, Michigan State University. 479

Raudenbush, S. W., M.-L. Yang, and M. Yosef. 2000. Maximum likelihood for hierarchical models via high-order multivariate Laplace approximations. *Journal of Computational and Graphical Statistics* **9**: 141–157. 478, 506

Roberts, G. O. and S. K. Sahu. 1997. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society (Series B)* **59**: 291–318. 484

Robinson, G. K. 1991. That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science* **6**: 15–51. 479

Rodríguez, G. and N. Goldman. 1995. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society (Series A)* **158**: 73–89. 476, 480, 499

—. 2001. Improved estimation procedures for multilevel models with binary responses: a case study. *Journal of the Royal Statistical Society (Series A)* **164**: 339–355. 480, 506

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**: 1151–1172. 479

Samaniego, F. J. and D. M. Reneau. 1994. Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association* **89**: 947–957. 505

SAS-Institute. 2006. *SAS Documentation, Release 9*. Cary, NC: SAS Institute. 506

Scheffé, H. 1959. *The Analysis of Variance*. New York: Wiley. 479

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. *Variance Components*. New York: Wiley. 479

Seltzer, M. H. 1993. Sensitivity analysis for fixed effects in the hierarchical model: a Gibbs sampling approach. *Journal of Educational Statistics* **18**: 207–235. 482

Seltzer, M. H., W. H. Wong, and A. S. Bryk. 1996. Bayesian analysis in applications of hierarchical models: issues and methods. *Journal of Educational and Behavioral Statistics* **21**: 131–167. 484

Severini, T. A. 1994. On the relationship between Bayesian and non-Bayesian interval estimates. *Journal of the Royal Statistical Society (Series B)* **53**: 611–618. 505

Singh, A. C., D. M. Stukel, and D. Pfefferman. 1998. Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society (Series B)* **60**: 377–396. 480

Spiegelhalter, D. J., A. Thomas, N. Best, and W. R. Gilks. 1997. *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.60*. Cambridge: Medical Research Council Biostatistics Unit. 482, 483

Spiegelhalter, D. J., A. Thomas, N. Best, and D. Lunn. 2003. *WinBUGS User Manual, Version 1.4*. Available at `www.mrc-bsu.cam.ac.uk/bugs`. 474, 484

StataCorp. 2006. *Stata Statistical Software: Release 9*. College Station TX: Stata Corporation. 474

Swallow, W. H. and J. F. Monahan. 1984. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics* **26**: 47–57. 480

White, H. 1980. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* **48**: 817–830. 474

Wilson, E. B. and M. M. Hilferty. 1931. The distribution of chi-square. *Proceedings of the US National Academy of Sciences* **17**: 684. 492

Woodhouse, G., J. Rasbash, H. Goldstein, M. Yang, J. Howarth, and I. Plewis. 1995. *A Guide to MLn for New Users*. London: Institute of Education, University of London. 474

Zeger, S. L. and M. R. Karim. 1991. Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**: 79–86. 480, 484

Bayesian Analysis (2006) **1**, Number 3, pp. 515–534

Prior distributions for variance parameters in hierarchical models(Comment on Article by Browne and Draper)

Andrew Gelman*

Abstract. Various noninformative prior distributions have been suggested for scale parameters in hierarchical models. We construct a new folded-noncentral-t family of conditionally conjugate priors for hierarchical standard deviation parameters, and then consider noninformative and weakly informative priors in this family. We use an example to illustrate serious problems with the inverse-gamma family of "noninformative" prior distributions. We suggest instead to use a uniform prior on the hierarchical standard deviation, using the half-t family when the number of groups is small and in other settings where a weakly informative prior is desired. We also illustrate the use of the half-t family for hierarchical modeling of multiple variance parameters such as arise in the analysis of variance.

Keywords: Bayesian inference, conditional conjugacy, folded-noncentral-t distribution, half-t distribution, hierarchical model, multilevel model, noninformative prior distribution, weakly informative prior distribution

1 Introduction

Fully-Bayesian analyses of hierarchical linear models have been considered for at least forty years (Hill, 1965, Tiao and Tan, 1965, and Stone and Springer, 1965) and have remained a topic of theoretical and applied interest (see, e.g., Portnoy, 1971, Box and Tiao, 1973, Gelman et al., 2003, Carlin and Louis, 1996, and Meng and van Dyk, 2001). Browne and Draper (2005) review much of the extensive literature in the course of comparing Bayesian and non-Bayesian inference for hierarchical models. As part of their article, Browne and Draper consider some different prior distributions for variance parameters; here, we explore the principles of hierarchical prior distributions in the context of a specific class of models.

Hierarchical (multilevel) models are central to modern Bayesian statistics for both conceptual and practical reasons. On the theoretical side, hierarchical models allow a more "objective" approach to inference by estimating the parameters of prior distributions from data rather than requiring them to be specified using subjective information (see James and Stein, 1960, Efron and Morris, 1975, and Morris, 1983). At a practical level, hierarchical models are flexible tools for combining information and partial pooling of inferences (see, for example, Kreft and De Leeuw, 1998, Snijders and Bosker, 1999, Carlin and Louis, 2001, Raudenbush and Bryk, 2002, Gelman et al., 2003).

*Department of Statistics, Columbia University, New York, NY, http://www.stat.columbia.edu/~gelman/

A hierarchical model requires hyperparameters, however, and these must be given their own prior distribution. In this paper, we discuss the prior distribution for hierarchical variance parameters. We consider some proposed noninformative prior distributions, including uniform and inverse-gamma families, in the context of an expanded conditionally-conjugate family. We propose a half-t model and demonstrate its use as a weakly-informative prior distribution and as a component in a hierarchical model of variance parameters.

1.1 The basic hierarchical model

We shall work with a simple two-level normal model of data y_{ij} with group-level effects α_j:

$$\begin{aligned} y_{ij} &\sim \text{N}(\mu + \alpha_j, \sigma_y^2), \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J \\ \alpha_j &\sim \text{N}(0, \sigma_\alpha^2), \quad j = 1, \ldots, J. \end{aligned} \tag{1}$$

We briefly discuss other hierarchical models in Section 7.2.

Model (1) has three hyperparameters—μ, σ_y, and σ_α—but in this paper we concern ourselves only with the last of these. Typically, enough data will be available to estimate μ and σ_y that one can use any reasonable noninformative prior distribution—for example, $p(\mu, \sigma_y) \propto 1$ or $p(\mu, \log \sigma_y) \propto 1$.

Various noninformative prior distributions for σ_α have been suggested in Bayesian literature and software, including an improper uniform density on σ_α (Gelman et al., 2003), proper distributions such as $p(\sigma_\alpha^2) \sim$ inverse-gamma$(0.001, 0.001)$ (Spiegelhalter et al., 1994, 2003), and distributions that depend on the data-level variance (Box and Tiao, 1973). In this paper, we explore and make recommendations for prior distributions for σ_α, beginning in Section 3 with conjugate families of proper prior distributions and then considering noninformative prior densities in Section 4.

As we illustrate in Section 5, the choice of "noninformative" prior distribution can have a big effect on inferences, especially for problems where the number of groups J is small or the group-level variance σ_α^2 is close to zero. We conclude with recommendations in Section 7.

2 Concepts relating to the choice of prior distribution

2.1 Conditionally-conjugate families

Consider a model with parameters θ, for which ϕ represents one element or a subset of elements of θ. A family of prior distributions $p(\phi)$ is *conditionally conjugate* for ϕ if the conditional posterior distribution, $p(\phi|y)$ is also in that class. In computational terms, conditional conjugacy means that, if it is possible to draw ϕ from this class of prior distributions, then it is also possible to perform a Gibbs sampler draw of ϕ in the posterior distribution. Perhaps more important for understanding the model,

conditional conjugacy allows a prior distribution to be interpreted in terms of equivalent data (see, for example, Box and Tiao, 1973).

Conditional conjugacy is a useful idea because it is preserved when a model is expanded hierarchically, while the usual concept of conjugacy is not. For example, in the basic hierarchical normal model, the normal prior distributions on the α_j's are conditionally conjugate but not conjugate; the α_j's have normal posterior distributions, conditional on all other parameters in the model, but their marginal posterior distributions are not normal.

As we shall see, by judicious model expansion we can expand the class of conditionally conjugate prior distributions for the hierarchical variance parameter.

2.2 Improper limit of a prior distribution

Improper prior densities can, but do not necessarily, lead to proper posterior distributions. To avoid confusion it is useful to define improper distributions as particular limits of proper distributions. For the variance parameter σ_α, two commonly-considered improper densities are uniform$(0, A)$, as $A \to \infty$, and inverse-gamma(ϵ, ϵ), as $\epsilon \to 0$.

As we shall see, the uniform$(0, A)$ model yields a limiting proper posterior density as $A \to \infty$, as long as the number of groups J is at least 3. Thus, for a finite but sufficiently large A, inferences are not sensitive to the choice of A.

In contrast, the inverse-gamma(ϵ, ϵ) model does *not* have any proper limiting posterior distribution. As a result, posterior inferences are sensitive to ϵ—it cannot simply be comfortably set to a low value such as 0.001.

2.3 Weakly-informative prior distribution

We characterize a prior distribution as *weakly informative* if it is proper but is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available. We will discuss this further in the context of a specific example, but in general any problem has some natural constraints that would allow a weakly-informative model. For example, for regression models on the logarithmic or logit scale, with predictors that are binary or scaled to have standard deviation 1, we can be sure for most applications that effect sizes will be less than 10, or certainly less than 100.

Weakly-informative distributions are useful for their own sake and also as necessary limiting steps in noninformative distributions, as discussed in Section 2.2 above.

2.4 Calibration

Posterior inferences can be evaluated using the concept of *calibration* of the posterior mean, the Bayesian analogue to the classical notion of "bias." For any parameter θ, we

label the posterior mean as $\hat{\theta} = \mathrm{E}(\theta|y)$ and define the *miscalibration* of the posterior mean as $\mathrm{E}(\theta|\hat{\theta}, y) - \hat{\theta}$, for any value of $\hat{\theta}$. If the prior distribution is true—that is, if the data are constructed by first drawing θ from $p(\theta)$, then drawing y from $p(y|\theta)$—then the posterior mean is automatically calibrated; that is its miscalibration is 0 for all values of $\hat{\theta}$.

For improper prior distributions, however, things are not so simple, since it is impossible for θ to be drawn from an unnormalized density. To evaluate calibration in this context, it is necessary to posit a "true prior distribution" from which θ is drawn along with the "inferential prior distribution" that is used in the Bayesian inference.

For the hierarchical model discussed in this paper, we can consider the improper uniform density on σ_α as a limit of uniform prior densities on the range $(0, A)$, with $A \to \infty$. For any finite value of A, we can then see that the improper uniform density leads to inferences with a positive miscalibration—that is, overestimates (on average) of σ_α.

We demonstrate this miscalibration in two steps. First, suppose that both the true and inferential prior distributions for σ_α are uniform on $(0, A)$. Then the miscalibration is trivially zero. Now keep the true prior distribution at $\mathrm{U}(0, A)$ and let the inferential prior distribution go to $\mathrm{U}(0, \infty)$. This will necessarily increase $\hat{\theta}$ for any data y (since we are now averaging over values of θ in the range $[A, \infty)$) without changing the true θ, thus causing the average value of the miscalibration to become positive.

This miscalibration is an unavoidable consequence of the asymmetry in the parameter space, with variance parameters restricted to be positive. Similarly, there are no always-nonnegative classical unbiased estimators of σ_α or σ_α^2 in the hierarchical model. Similar issues are discussed by Bickel and Blackwell (1967) and Meng and Zaslavsky (2002).

3 Conditionally-conjugate prior distributions for hierarchical variance parameters

3.1 Inverse-gamma prior distribution for σ_α^2

The parameter σ_α^2 in model (1) does not have any simple family of conjugate prior distributions because its marginal likelihood depends in a complex way on the data from all J groups (Hill, 1965, Tiao and Tan, 1965). However, the inverse-gamma family is conditionally conjugate, in the sense defined in Section 2.1: if σ_α^2 has an inverse-gamma prior distribution, then the conditional posterior distribution $p(\sigma_\alpha^2 \,|\, \alpha, \mu, \sigma_y, y)$ is also inverse-gamma.

The inverse-gamma(α, β) model for σ_α^2 can also be expressed as an inverse-χ^2 distribution with scale $s_\alpha^2 = \beta/\alpha$ and degrees of freedom $\nu_\alpha = 2\alpha$ (Gelman et al., 2003). The inverse-χ^2 parameterization can be helpful in understanding the information underlying various choices of proper prior distributions, as we discuss in Section 4.

3.2 Folded-noncentral-t prior distribution for σ_α

We can expand the family of conditionally-conjugate prior distributions by applying a redundant multiplicative reparameterization to model (1):

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\mu + \xi\eta_j, \sigma_y^2) \\ \eta_j &\sim \mathrm{N}(0, \sigma_\eta^2). \end{aligned} \tag{2}$$

The parameters α_j in (1) correspond to the products $\xi\eta_j$ in (2), and the hierarchical standard deviation σ_α in (1) corresponds to $|\xi|\sigma_\eta$ in (2). This "parameter expanded" model was originally constructed to speed up EM and Gibbs sampler computations. The overparameterization reduces dependence among the parameters in a hierarchical model and improves MCMC convergence (Liu, Rubin, and Wu, 1998, Liu and Wu, 1999, van Dyk and Meng, 2001, Gelman et al., 2005). It has also been suggested that the additional parameter can increase the flexibility of applied modeling, especially in hierarchical regression models with several batches of varying coefficients (Gelman, 2004). Here we merely note that this expanded model form allows conditionally conjugate prior distributions for both ξ and σ_η, and these parameters are independent in the conditional posterior distribution. There is thus an implicit conditionally conjugate prior distribution for $\sigma_\alpha = |\xi|\sigma_\eta$.

For simplicity we restrict ourselves to independent prior distributions on ξ and σ_η. In model (2), the conditionally-conjugate prior family for ξ is normal—given the data and all the other parameters in the model, the likelihood for ξ has the form of a normal distribution, derived from $\sum_{j=1}^{J} n_j$ factors of the form $(y_{ij} - \mu)/\eta_j \sim \mathrm{N}(\xi, \sigma_y^2/\eta_j^2)$. The conditionally-conjugate prior family for σ_η^2 is inverse-gamma, as discussed in Section 3.1.

The implicit conditionally-conjugate family for σ_α is then the set of distributions corresponding to the absolute value of a normal random variable, divided by the square root of a gamma random variable. That is, σ_α has the distribution of the absolute value of a noncentral-t variate (see, for example, Johnson and Kotz, 1972). We shall call this the *folded noncentral t distribution*, with the "folding" corresponding to the absolute value operator. The noncentral t in this context has three parameters, which can be identified with the mean of the normal distribution for ξ, and the scale and degrees of freedom for σ_η^2. (Without loss of generality, the scale of the normal distribution for ξ can be set to 1 since it cannot be separated from the scale for σ_η.)

The folded noncentral t distribution is not commonly used in statistics, and we find it convenient to understand it through various special and limiting cases. In the limit that the denominator is specified exactly, we have a folded normal distribution; conversely, specifying the numerator exactly yields the square-root-inverse-χ^2 distribution for σ_α, as in Section 3.1.

An appealing two-parameter family of prior distributions is determined by restricting the prior mean of the numerator to zero, so that the folded noncentral t distribution for σ_α becomes simply a half-t—that is, the absolute value of a Student-t distribution centered at zero. We can parameterize this in terms of scale A and degrees of freedom

ν:

$$p(\sigma_\alpha) \propto \left(1 + \frac{1}{\nu}\left(\frac{\sigma_\alpha}{A}\right)^2\right)^{-(\nu+1)/2}.$$

This family includes, as special cases, the improper uniform density (if $\nu = -1$) and the proper half-Cauchy, $p(\sigma_\alpha) \propto \left(\sigma_\alpha^2 + s_\alpha^2\right)^{-1}$ (if $\nu = 1$).

The half-t family is not itself conditionally-conjugate—starting with a half-t prior distribution, you will still end up with a more general folded noncentral t conditional posterior—but it is a natural subclass of prior densities in which the distribution of the multiplicative parameter ξ is symmetric about zero.

4 Noninformative and weakly-informative prior distributions for hierarchical variance parameters

4.1 General considerations

Noninformative prior distributions are intended to allow Bayesian inference for parameters about which not much is known beyond the data included in the analysis at hand. Various justifications and interpretations of noninformative priors have been proposed over the years, including invariance (Jeffreys, 1961), maximum entropy (Jaynes, 1983), and agreement with classical estimators (Box and Tiao, 1973, Meng and Zaslavsky, 2002). In this paper, we follow the approach of Bernardo (1979) and consider so-called noninformative priors as "reference models" to be used as a standard of comparison or starting point in place of the proper, informative prior distributions that would be appropriate for a full Bayesian analysis (see also Kass and Wasserman, 1996).

We view any noninformative or weakly-informative prior distribution as inherently provisional—after the model has been fit, one should look at the posterior distribution and see if it makes sense. If the posterior distribution does not make sense, this implies that additional prior knowledge is available that has not been included in the model, and that contradicts the assumptions of the prior distribution that has been used. It is then appropriate to go back and alter the prior distribution to be more consistent with this external knowledge.

4.2 Uniform prior distributions

We first consider uniform prior distributions while recognizing that we must be explicit about the scale on which the distribution is defined. Various choices have been proposed for modeling variance parameters. A uniform prior distribution on $\log \sigma_\alpha$ would seem natural—working with the logarithm of a parameter that must be positive—but it results in an improper posterior distribution. An alternative would be to define the prior distribution on a compact set (e.g., in the range $[-A, A]$ for some large value of A), but then the posterior distribution would depend strongly on the lower bound $-A$

of the prior support.

The problem arises because the marginal likelihood, $p(y|\sigma_\alpha)$—after integrating over α, μ, σ_y in (1)—approaches a finite nonzero value as $\sigma_\alpha \to 0$. Thus, if the prior density for $\log \sigma_\alpha$ is uniform, the posterior distribution will have infinite mass integrating to the limit $\log \sigma_\alpha \to -\infty$. To put it another way, in a hierarchical model the data can never rule out a group-level variance of zero, and so the prior distribution cannot put an infinite mass in this area.

Another option is a uniform prior distribution on σ_α itself, which has a finite integral near $\sigma_\alpha = 0$ and thus avoids the above problem. We have generally used this noninformative density in our applied work (see Gelman et al., 2003), but it has a slightly disagreeable miscalibration toward positive values (see Section 2.4), with its infinite prior mass in the range $\sigma_\alpha \to \infty$. With $J = 1$ or 2 groups, this actually results in an improper posterior density, essentially concluding $\sigma_\alpha = \infty$ and doing no shrinkage (see Gelman et al., 2003, Exercise 5.8). In a sense this is reasonable behavior, since it would seem difficult from the data alone to decide how much, if any, shrinkage should be done with data from only one or two groups—and in fact this would seem consistent with the work of Stein (1955) and James and Stein (1960) that unshrunken estimators are admissible if $J < 3$. However, from a Bayesian perspective it is awkward for the decision to be made ahead of time, as it were, with the data having no say in the matter. In addition, for small J, such as 4 or 5, we worry that the heavy right tail of the posterior distribution would lead to overestimates of σ_α and thus result in shrinkage that is less than optimal for estimating the individual α_j's.

We can interpret the various improper uniform prior densities as limits of weakly-informative conditionally-conjugate priors. The uniform prior distribution on $\log \sigma_\alpha$ is equivalent to $p(\sigma_\alpha) \propto \sigma_\alpha^{-1}$ or $p(\sigma_\alpha^2) \propto \sigma_\alpha^{-2}$, which has the form of an inverse-χ^2 density with 0 degrees of freedom and can be taken as a limit of proper conditionally-conjugate inverse-gamma priors.

The uniform density on σ_α is equivalent to $p(\sigma_\alpha^2) \propto \sigma_\alpha^{-1}$, an inverse-$\chi^2$ density with -1 degrees of freedom. This density cannot easily be seen as a limit of proper inverse-χ^2 densities (since these must have positive degrees of freedom), but it can be interpreted as a limit of the half-t family on σ_α, where the scale approaches ∞ (and any value of ν). Or, in the expanded notation of (2), one could assign any prior distribution to σ_η and a normal to ξ, and let the prior variance for ξ approach ∞.

Another noninformative prior distribution sometimes proposed in the Bayesian literature is uniform on σ_α^2. We do not recommend this, as it seems to have the miscalibration toward higher values as described above, but more so, and also requires $J \geq 4$ groups for a proper posterior distribution.

4.3 Inverse-gamma(ϵ, ϵ) prior distributions

The inverse-gamma(ϵ, ϵ) prior distribution is an attempt at noninformativeness within the conditionally conjugate family, with ϵ set to a low value such as 1 or 0.01 or 0.001

(the latter value being used in the examples in Bugs; see Spiegelhalter et al., 1994, 2003). A difficulty of this prior distribution is that in the limit of $\epsilon \to 0$ it yields an improper posterior density, and thus ϵ must be set to a reasonable value. Unfortunately, for datasets in which low values of σ_α are possible, inferences become very sensitive to ϵ in this model, and the prior distribution hardly looks noninformative, as we illustrate in Section 5.

4.4 Half-Cauchy prior distributions

The half-Cauchy is a special case of the conditionally-conjugate folded-noncentral-t family of prior distributions for σ_α; see Section 3.2, which has a broad peak at zero and a scale parameter A. In the limit $A \to \infty$ this becomes a uniform prior density on $p(\sigma_\alpha)$. Large but finite values of A represent prior distributions which we call "weakly informative" because, even in the tail, they have a gentle slope (unlike, for example, a half-normal distribution) and can let the data dominate if the likelihood is strong in that region. In Sections 5.2 and 6, we consider half-Cauchy models for variance parameters which are estimated from a small number of groups (so that inferences are sensitive to the choice of weakly-informative prior distribution).

5 Application to the 8-schools example

We demonstrate the properties of some proposed noninformative prior densities with a simple example of data from $J = 8$ educational testing experiments described in Rubin (1981) and Gelman et al. (2003, Chapter 5 and Appendix C). Here, the parameters $\alpha_1, \ldots, \alpha_8$ represent the relative effects of Scholastic Aptitude Test coaching programs in eight different schools, and σ_α represents the between-school standard deviations of these effects. The effects are measured as points on the test, which was scored from 200 to 800 with an average of about 500; thus the largest possible range of effects could be about 300 points, with a realistic upper limit on σ_α of 100, say.

5.1 Noninformative prior distributions for the 8-schools problem

Figure 1 shows the posterior distributions for the 8-schools model resulting from three different choices of prior distributions that are intended to be noninformative.

The leftmost histogram shows the posterior inference for σ_α (as represented by 6000 simulation draws from a model fit using Bugs) for the model with uniform prior density. The data show support for a range of values below $\sigma_\alpha = 20$, with a slight tail after that, reflecting the possibility of larger values, which are difficult to rule out given that the number of groups J is only 8—that is, not much more than the $J = 3$ required to ensure a proper posterior density with finite mass in the right tail.

In contrast, the middle histogram in Figure 1 shows the result with an inverse-gamma$(1, 1)$ prior distribution for σ_α^2. This new prior distribution leads to changed

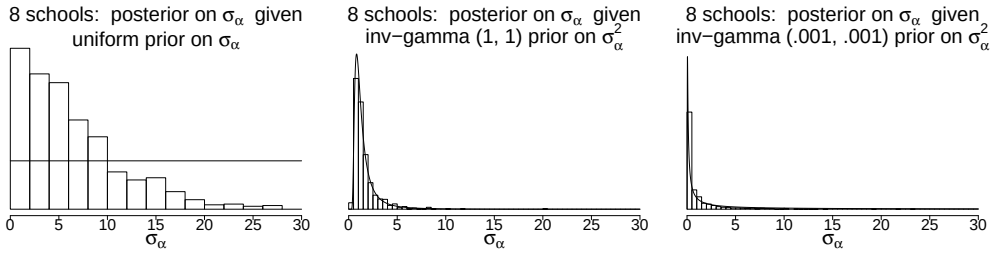


Figure 1: Histograms of posterior simulations of the between-school standard deviation, σ_α, from models with three different prior distributions: (a) uniform prior distribution on σ_α, (b) inverse-gamma$(1, 1)$ prior distribution on σ_α^2, (c) inverse-gamma$(0.001, 0.001)$ prior distribution on σ_α^2. Overlain on each is the corresponding prior density function for σ_α. (For models (b) and (c), the density for σ_α is calculated using the gamma density function multiplied by the Jacobian of the $1/\sigma_\alpha^2$ transformation.) In models (b) and (c), posterior inferences are strongly constrained by the prior distribution. Adapted from Gelman et al. (2003, Appendix C).

inferences. In particular, the posterior mean and median of σ_α are lower and shrinkage of the α_j's is greater than in the previously-fitted model with a uniform prior distribution on σ_α. To understand this, it helps to graph the prior distribution in the range for which the posterior distribution is substantial. The graph shows that the prior distribution is concentrated in the range $[0.5, 5]$, a narrow zone in which the likelihood is close to flat compared to this prior (as we can see because the distribution of the posterior simulations of σ_α closely matches the prior distribution, $p(\sigma_\alpha)$). By comparison, in the left graph, the uniform prior distribution on σ_α seems closer to "noninformative" for this problem, in the sense that it does not appear to be constraining the posterior inference.

Finally, the rightmost histogram in Figure 1 shows the corresponding result with an inverse-gamma$(0.001, 0.001)$ prior distribution for σ_α^2. This prior distribution is even more sharply peaked near zero and further distorts posterior inferences, with the problem arising because the marginal likelihood for σ_α remains high near zero.

In this example, we do not consider a uniform prior density on $\log \sigma_\alpha$, which would yield an improper posterior density with a spike at $\sigma_\alpha = 0$, like the rightmost graph in Figure 1, but more so. We also do not consider a uniform prior density on σ_α^2, which would yield a posterior distribution similar to the leftmost graph in Figure 1, but with a slightly higher right tail.

This example is a gratifying case in which the simplest approach—the uniform prior density on σ_α—seems to perform well. As detailed in Gelman et al. (2003, Appendix C), this model is also straightforward to program directly using the Gibbs sampler or in Bugs, using either the basic model (1) or slightly faster using the expanded parameterization (2).

The appearance of the histograms and density plots in Figure 1 is crucially affected

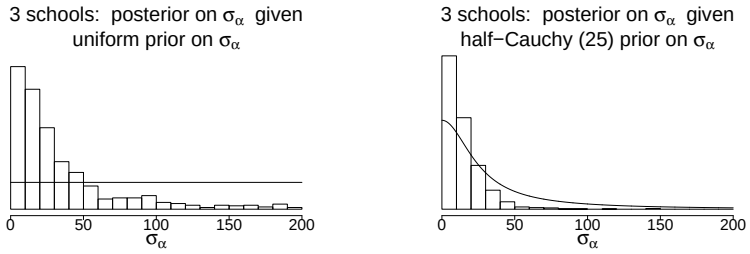


Figure 2: Histograms of posterior simulations of the between-school standard deviation, σ_α, from models for the 3-schools data with two different prior distributions on σ_α: (a) uniform $(0, \infty)$, (b) half-Cauchy with scale 25, set as a weakly informative prior distribution given that σ_α was expected to be well below 100. The histograms are not on the same scales. Overlain on each histogram is the corresponding prior density function. With only $J = 3$ groups, the noninformative uniform prior distribution is too weak, and the proper Cauchy distribution works better, without appearing to distort inferences in the area of high likelihood.

by the choice to plot them on the scale of σ_α. If instead they were plotted on the scale of $\log \sigma_\alpha$, the inverse-gamma$(0.001, 0.001)$ prior density would appear to be the flattest. However, the inverse-gamma(ϵ, ϵ) prior is not at all "noninformative" for this problem since the resulting posterior distribution remains highly sensitive to the choice of ϵ. As explained in Section 4.2, the hierarchical model likelihood does not constrain $\log \sigma_\alpha$ in the limit $\log \sigma_\alpha \to -\infty$, and so a prior distribution that is noninformative on the log scale will not work.

5.2 Weakly informative prior distribution for the 3-schools problem

The uniform prior distribution seems fine for the 8-school analysis, but problems arise if the number of groups J is much smaller, in which case the data supply little information about the group-level variance, and a noninformative prior distribution can lead to a posterior distribution that is improper or is proper but unrealistically broad. We demonstrate by reanalyzing the 8-schools example using just the data from the first 3 of the schools.

Figure 2 displays the inferences for σ_α from two different prior distributions. First we continue with the default uniform distribution that worked well with $J = 8$ (as seen in Figure 1). Unfortunately, as the left histogram of Figure 2 shows, the resulting posterior distribution for the 3-schools dataset has an extremely long right tail, containing values of σ_α that are too high to be reasonable. This heavy tail is expected since J is so low (if J were any lower, the right tail would have an infinite integral), and using this as a posterior distribution will have the effect of undershrinking the estimates of the school effects α_j, as explained in Section 4.2.

The right histogram of Figure 2 shows the posterior inference for σ_α resulting from a half-Cauchy prior distribution of the sort described at the end of Section 3.2, with scale

parameter $A = 25$ (a value chosen to be a bit higher than we expect for the standard deviation of the underlying θ_j's in the context of this educational testing example, so that the model will constrain σ_α only weakly). As the line on the graph shows, this prior distribution is high over the plausible range of $\sigma_\alpha < 50$, falling off gradually beyond this point. This prior distribution appears to perform well in this example, reflecting the marginal likelihood for σ_α at its low end but removing much of the unrealistic upper tail.

This half-Cauchy prior distribution would also perform well in the 8-schools problem; however it was unnecessary because the default uniform prior gave reasonable results. With only 3 schools, we went to the trouble of using a weakly informative prior, a distribution that was not intended to represent our actual prior state of knowledge about σ_α but rather to constrain the posterior distribution, to an extent allowed by the data.

6 Modeling variance components hierarchically

6.1 Application to a latin square Anova

We next consider an analysis of variance problem which has several variance components, one for each source of variation. Gelman (2005) analyzes data from a $5 \times 5 \times 2$ split-plot latin square with five full-plot treatments (labeled A, B, C, D, E), and with each plot divided into two subplots (labeled 1 and 2).

Source	df
row	4
column	4
(A,B,C,D,E)	4
plot	12
(1,2)	1
row × (1,2)	4
column × (1,2)	4
(A,B,C,D,E) × (1,2)	4
plot × (1,2)	12

Each row of the table corresponds to a different variance component, and the split-plot Anova can be understood as a linear model with nine variance components, $\sigma_1^2, \ldots, \sigma_9^2$—one for each row of the table. A default Bayesian analysis assigns a uniform prior distribution, $p(\sigma_1, \ldots, \sigma_9) \propto 1$ (Gelman, 2005).

More generally, we can set up a hierarchical model, where the variance parameters have a common distribution with hyperparameters estimated from the data. Based on the analyses given above, we consider a half-Cauchy prior distribution with peak 0 and scale A, and with a uniform prior distribution on A. The hierarchical half-Cauchy model allows most of the variance parameters to be small but with the occasionally large σ_α,

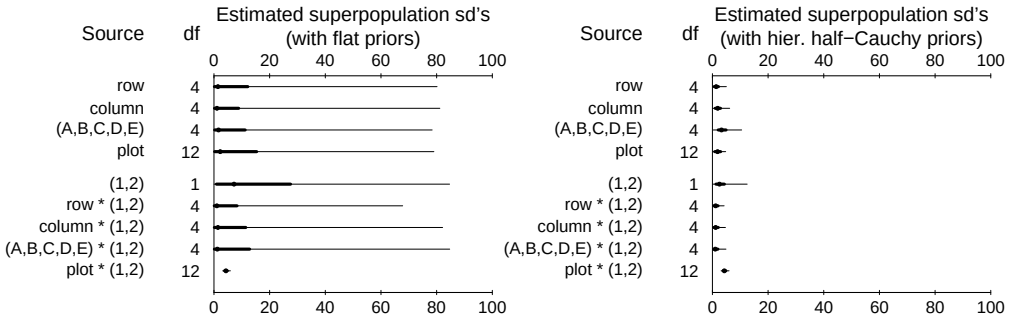


Figure 3: Posterior medians, 50%, and 95% intervals for standard deviation parameters σ_k estimated from a split-plot latin square experiment. The left plot shows inferences given uniform prior distributions on the σ_k's, and the right plot shows inferences given a hierarchical half-Cauchy model with scale fit to the data. The half-Cauchy model gives much sharper inferences, using the partial pooling that comes with fitting a hierarchical model.

which seems reasonable in the typical settings of analysis of variance, in which most sources of variation are small but some are large (Daniel, 1959, Gelman, 2005).

6.2 Superpopulation and finite-population standard deviations

Figure 3 shows the inferences in the latin square example, given uniform and hierarchical half-Cauchy prior distributions for the standard deviation parameters σ_k. As the left plot shows, the uniform prior distribution does not rule out the potential for some extremely high values of the variance components—the degrees of freedom are low, and the interlocking of the linear parameters in the latin square model results in difficulty in estimating any single variance parameter. In contrast, the hierarchical half-Cauchy model performs a great deal of shrinkage, especially of the high ranges of the intervals. (For most of the variance parameters, the posterior medians are similar under the two models; it is the 75th and 97.5th percentiles that are shrunk by the hierarchical model.) This is an ideal setting for hierarchical modeling of variance parameters in that it combines separately imprecise estimates of each of the individual σ_k's.

As discussed in Gelman (2005, Section 3.5), the σ_k's are *superpopulation* parameters in that each represents the standard deviation of an entire population of effects, of which only a few of which were sampled for the experiment at hand. In estimating variance parameters estimated from few degrees of freedom, it can be helpful also to look at the *finite-population* standard deviation s_α of the corresponding linear parameters α_j.

For a simple hierarchical model of the form (1), s_α is simply the standard deviation of the J values of α_j. More generally, for more complicated linear models such as

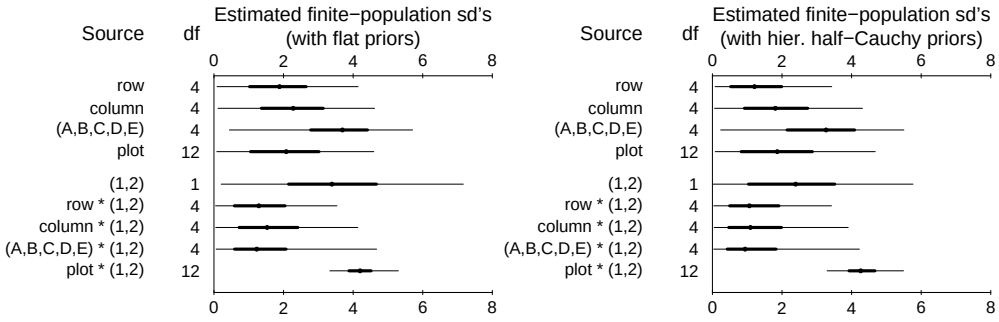


Figure 4: Posterior medians, 50%, and 95% intervals for finite-population standard deviations s_k estimated from a split-plot latin square experiment. The left plot shows inferences given uniform prior distributions on the σ_k's, and the right plot shows inferences given a hierarchical half-Cauchy model with scale fit to the data. The half-Cauchy model gives sharper estimates even for these finite-population standard deviations, indicating the power of hierarchical modeling for these highly uncertain quantities. Compare to Figure 3 (which is on a different scale).

the split-plot latin square, s_α for any variance component is the root mean square of the coefficients' residuals after projection to their constraint space (see Gelman, 2005, Section 3.1). In any case, this finite-population standard deviation s can be calculated from its posterior simulations and, especially when degrees of freedom are low, is more precisely estimated than the superpopulation standard deviation σ.

Figure 4 shows posterior inferences for the finite-population standard deviation parameters s_α for each row of the latin square split-plot Anova, showing inferences given the uniform and hierarchical half-Cauchy prior distributions for the variance parameters σ_α. The half-Cauchy prior distribution does slightly better than the uniform, with the largest shrinkage occurring for the variance component that has just one degree of freedom. The Cauchy scale parameter A was estimated at 1.8, with a 95% posterior interval of $[0.5, 5.1]$.

7 Recommendations

7.1 Prior distributions for variance parameters

In fitting hierarchical models, we recommend starting with a noninformative uniform prior density on standard deviation parameters σ_α. We expect this will generally work well unless the number of groups J is low (below 5, say). If J is low, the uniform prior density tends to lead to high estimates of σ_α, as discussed in Section 5.2. This miscalibration is an unavoidable consequence of the asymmetry in the parameter space,

with variance parameters restricted to be positive. Similarly, there are no always-nonnegative classical unbiased estimators of σ_α or σ_α^2 in the hierarchical model.

A user of a noninformative prior density might still like to use a proper distribution—reasons could include Bayesian scruple, the desire to perform prior predictive checks (see Box, 1980, Gelman, Meng, and Stern, 1996, and Bayarri and Berger, 2000) or Bayes factors (see Kass and Raftery, 1995, O'Hagan, 1995, and Pauler, Wakefield, and Kass, 1999), or because computation is performed in Bugs, which requires proper distributions. For a noninformative but proper prior distribution, we recommend approximating the uniform density on σ_α by a uniform on a wide range (for example, U$(0, 100)$ in the SAT coaching example) or a half-normal centered at 0 with standard deviation set to a high value such as 100. The latter approach is particularly easy to program as a $N(0, 100^2)$ prior distribution for ξ in (2).

When more prior information is desired, for instance to restrict σ_α away from very large values, we recommend working within the half-t family of prior distributions, which are more flexible and have better behavior near 0, compared to the inverse-gamma family. A reasonable starting point is the half-Cauchy family, with scale set to a value that is high but not off the scale; for example, 25 in the example in Section 5.2. When several variance parameters are present, we recommend a hierarchical model such as the half-Cauchy, with hyperparameter estimated from data.

We do *not* recommend the inverse-gamma(ϵ, ϵ) family of noninformative prior distributions because, as discussed in Sections 4.3 and 5.1, in cases where σ_α is estimated to be near zero, the resulting inferences will be sensitive to ϵ. The setting of near-zero variance parameters is important partly because this is where classical and Bayesian inferences for hierarchical models will differ the most (see Draper and Browne, 2005, and Section 3.4 of Gelman, 2005).

Figure 1 illustrates the generally robust properties of the uniform prior density on σ_α. Many Bayesians have preferred the inverse-gamma prior family, possibly because its conditional conjugacy suggested clean mathematical properties. However, by writing the hierarchical model in the form (2), we see conditional conjugacy in the wider class of half-t distributions on σ_α, which include the uniform and half-Cauchy densities on σ_α (as well as inverse-gamma on σ_α^2) as special cases. From this perspective, the inverse-gamma family has nothing special to offer, and we prefer to work on the scale of the standard deviation parameter σ_α, which is typically directly interpretable in the original model.

7.2 Generalizations

The reasoning in this paper should apply to hierarchical regression models (including predictors at the individual or group levels), hierarchical generalized linear models (as discussed by Christiansen and Morris, 1997, and Natarajan and Kass, 2000), and more complicated nonlinear models with hierarchical structure. The key idea is that parameters α_j—in general, group-level exchangeable parameters—have a common distribution with some scale parameter which we label σ_α. Some of the details will change—in

particular, if the model is nonlinear, then the normal prior distribution for the multiplicative parameter ξ in (2) will not be conditionally conjugate, however ξ can still be updated using the Metropolis algorithm. In addition, when regression predictors must be estimated, more than $J = 3$ groups may be necessary to estimate σ_α from a noninformative prior distribution, thus requiring at least weakly informative prior distributions for the regression coefficients, the variance parameters, or both.

There is also room to generalize these distributions to variance matrices in multivariate hierarchical models, going beyond the commonly-used inverse-Wishart family of prior distributions (Box and Tiao, 1973), which has problems similar to the inverse-gamma for scalar variances. Noninformative or weakly informative conditionally-conjugate priors could be applied to structured models such as described by Barnard, McCulloch, and Meng (2000) and Daniels and Kass (1999, 2001), expanded using multiplicative parameters as in Liu (2001) to give the models more flexibility.

Further work needs to be done in developing the next level of hierarchical models, in which there are several batches of exchangeable parameters, each with their own variance parameter—the Bayesian counterpart to the analysis of variance (Sargent and Hodges, 1997, Gelman, 2005). Specifying a prior distribution jointly on variance components at different levels of the model could be seen as a generalization of priors on the shrinkage factor, which is a function of both σ_y and σ_α (see Daniels, 1999, Natarajan and Kass, 2000, and Spiegelhalter, Abrams, and Myles, 2004, for an overview). In a model with several levels, it would make sense to give the variance parameters a parametric model with hyper-hyperparameters. This could be the ultimate solution to the difficulties of estimating σ_α for batches of parameters α_j where J is small, and we suppose that the folded-noncentral-t family could be useful here, as illustrated in Section 6.

Appendix: R and Bugs code for the hierarchical model with half-Cauchy prior density

Computations for the hierarchical normal model are most conveniently performed using Bugs (Spiegelhalter et al., 1994, 2003) as called from R (R Development Core Team, 2003), or by programming the Gibbs sampler directly in R. Both these strategies are described in detail in Gelman et al. (2003, Appendix C). Here we give an Bugs implementation of the 8-schools model with the half-Cauchy prior distribution (that is, the half-t with degrees-of-freedom parameter $\nu = 1$).

We put the following Bugs code in the file `schools.halfcauchy.bug`:

```
# Bugs model:  a half-Cauchy prior distribution on sigma.theta is induced
# using a normal prior on xi and an inverse-gamma on tau.eta

model {
  for (j in 1:J){                          # J = the number of schools
    y[j] ~ dnorm (theta[j], tau.y[j])      # data model:  the likelihood
    theta[j] <- mu.theta + xi*eta[j]
```

```
    tau.y[j] <- pow(sigma.y[j], -2)
  }
  xi ~ dnorm (0, tau.xi)
  tau.xi <- pow(prior.scale, -2)
  for (j in 1:J){
    eta[j] ~ dnorm (0, tau.eta)            # hierarchical model for theta
  }
  tau.eta ~ dgamma (.5, .5)                # chi^2 with 1 d.f.
  sigma.theta <- abs(xi)/sqrt(tau.eta)     # cauchy = normal/sqrt(chi^2)
  mu.theta ~ dnorm (0.0, 1.0E-6)           # noninformative prior on mu
}
```

We can then set up the data and call the Bugs model from R (using the `bugs.R` routines at Gelman, 2003). The scale parameter in the half-Cauchy distribution is `prior.scale`, which we set to the value 25 in the R code.

```
# R code for calling the Bugs 8-schools model with half-Cauchy prior dist

schools <- read.table ("schools.dat", header=T)
J <- nrow (schools)
y <- schools$estimate
sigma.y <- schools$sd
prior.scale <- 25
data <- list ("J", "y", "sigma.y", "prior.scale")
inits <- function (){
  list (eta=rnorm(J), mu.theta=rnorm(1), xi=rnorm(1), tau.eta=runif(1))}
parameters <- c ("theta", "mu.theta", "sigma.theta")
schools.sim <- bugs (data, inits, parameters, "schools.halfcauchy.bug",
  n.chains=3, n.iter=1000)
```

References

Barnard, J., McCulloch, R. E., and Meng, X. L. (2000). "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica*, 10: 1281–1311.

Bayarri, M. J. and Berger, J. (2000). "P-values for composite null models." *Journal of the American Statistical Association*, 95: 1127–1142. (with discussion).

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society B*, 41: 113–147. (with discussion).

Bickel, P. and Blackwell, D. (1967). "A note on Bayes estimates." *Annals of Mathematical Statistics*, 38: 1907–1911.

Box, G. E. P. (1980). "Sampling and Bayes inference in scientific modelling and robustness." *Journal of the Royal Statistical Society A*, 143: 383–430.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, Mass.: Addison-Wesley.

Browne, W. J. and Draper, D. (2005). "A comparison of Bayesian and likelihood-based methods for fitting multilevel models." *Bayesian Analysis*, This issue.

Carlin, B. P. and Louis, T. A. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, second edition edition.

Christiansen, C. and Morris, C. (1997). "Hierarchical Poisson regression models." *Journal of the American Statistical Association*, 92: 618–632.

Daniel, C. (1959). "Use of half-normal plots in interpreting factorial two-level experiments." *Technometrics*, 1: 311–341.

Daniels, M. J. (1999). "A prior for the variance in hierarchical models." *Canadian Journal of Statistics*, 27: 569–580.

Daniels, M. J. and Kass, R. E. (1999)). "Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models." *Journal of the American Statistical Association*, 94: 1254–1263.

— (2001). "Shrinkage estimators for covariance matrices." *Biometrics*, 57: 1173–1184.

Efron, B. and Morris, C. (1975). "Data analysis using Stein's estimator and its generalizations." *Journal of the American Statistical Association*, 70: 311–319.

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85: 398–409.

Gelman, A. (2003). "Bugs.R: functions for calling Bugs from R." `http://www.stat.columbia.edu/~gelman/bugsR/`.

— (2004). "Parameterization and Bayesian modeling." *Journal of the American Statistical Association*, 99: 537 – 545.

— (2005). "Analysis of variance: why it is more important than ever." *Annals of Statistics*, 33: 1 – 53. With discussion.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. London: Chapman and Hall, second edition edition.

Gelman, A., Huang, Z., van Dyk, D., and Boscardin, W. J. (2005). "Transformed and parameter-expanded Gibbs samplers for multilevel linear and generalized linear models." Technical report, Department of Statistics, Columbia University.

Gelman, A., Meng, X. L., and Stern, H. S. (1996). "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica Sinica*, 6: 733–807. (with discussion).

Hill, B. M. (1965). "Inference about variance components in the one-way model." *Journal of the American Statistical Association*, 60: 806–825.

James, W. and Stein, C. (1960). "Estimation with quadratic loss." In Neyman, J. (ed.), *Proceedings of the Fourth Berkeley Symposium*, volume 1, 361–380. Berkeley: University of California Press.

Jaynes, E. T. (1983). *Papers on Probability, Statistics, and Statistical Physics*. Dordrecht, Netherlands: Reidel.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, third edition edition.

Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics*. New York: Wiley. 4 vols.

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors and model uncertainty." *Journal of the American Statistical Association*, 90: 773–795.

Kass, R. E. and Wasserman, L. (1996). "The selection of prior distributions by formal rules." *Journal of the American Statistical Association*, 91: 1343–1370.

Kreft, I. and De Leeuw, J. (1998). *Introducing Multilevel Modeling*. Sage.

Liu, C. (2001). "Bayesian analysis of multivariate probit models. Discussion of "The art of data augmentation" by D. A. van Dyk and X. L. Meng." *Journal of Computational and Graphical Statistics*, 10: 75–81.

Liu, C., Rubin, D. B., , and Wu, Y. N. (1998). "Parameter expansion to accelerate EM: the PX-EM algorithm." *Biometrika*, 85: 755–770.

Liu, J. and Wu, Y. N. (1999). "Parameter expansion for data augmentation." *Journal of the American Statistical Association*, 94: 1264–1274.

Meng, X. L. and Zaslavsky, A. M. (2002). "Single observation unbiased priors." *Annals of Statistics*, 30: 1345–1375.

Morris, C. (1983). "Parametric empirical Bayes inference: theory and applications (with discussion)." *Journal of the American Statistical Association*, 78: 47–65.

Natarajan, R. and Kass, R. E. (2000). "Reference Bayesian methods for generalized linear mixed models." *Journal of the American Statistical Association*, 95: 227–237.

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison (with discussion)." *Journal of the Royal Statistical Society B*, 57: 99–138.

Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999). "Bayes factors for variance component models." *Journal of the American Statistical Association*, 94: 1242–1253.

Portnoy, S. (1971). "Formal Bayes estimation with applications to a random effects model." *Annals of Mathematical Statistics*, 42: 1379–1402.

R Development Core Team (2003). "R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing." `http://www.r-project.org`.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models*. Thousand Oaks, Calif.: Sage., second edition.

Rubin, D. B. (1981). "Estimation in parallel randomized experiments." *Journal of Educational Statistics*, 6: 377–401.

Sargent, D. J. and Hodges, J. S. (1997). "Smoothed ANOVA with application to subgroup analysis." Technical report, Department of Biostatistics, University of Minnesota.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Dover.

Snijders, T. A. B. and Bosker, R. J. (1999). *Multilevel Analysis*. London: Sage.

Spiegelhalter, D. J., Abrams, K. R., , and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, chapter section 5.7.3. Chichester: Wiley.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., , and Lunn, D. (1994, 2003). "BUGS: Bayesian inference using Gibbs sampling." MRC Biostatistics Unit, Cambridge, England,`http://www.mrc-bsu.cam.ac.uk/bugs/`.

Stein, C. (1955). "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." In Neyman, J. (ed.), *Proceedings of the Third Berkeley Symposium*, volume 1, 197–206. Berkeley: University of California Press.

Stone, M. and Springer, B. G. F. (1965). "A paradox involving quasi-prior distributions." *Biometrika*, 52: 623–627.

Tiao, G. C. and Tan, W. Y. (1965). "Bayesian analysis of random-effect models in the analysis of variance. I: Posterior distribution of variance components." *Biometrika*, 52: 37–53.

van Dyk, D. A. and Meng, X. L. (2001). "The art of data augmentation (with discussion)." *Journal of Computational and Graphical Statistics*, 10: 1–111.

Acknowledgments

We thank Rob Kass for inviting this paper, John Boscardin, John Carlin, Samantha Cook, Chuanhai Liu, Iain Pardoe, Hal Stern, Francis Tuerlinckx, Aki Vehtari, Phil Woodward, Shouhao Zhao, and reviewers for helpful suggestions, and the National Science Foundation for financial support.

Bayesian Analysis (2006) **1**, Number 3, pp. 535–542

A Default Conjugate Prior for Variance Components in Generalized Linear Mixed Models (Comment on Article by Browne and Draper)

Robert E. Kass* and Ranjini Natarajan†

Abstract. For a scalar random-effect variance, Browne and Draper (2005) have found that the uniform prior works well. It would be valuable to know more about the vector case, in which a second-stage prior on the random effects variance matrix **D** is needed. We suggest consideration of an inverse Wishart prior for **D** where the scale matrix is determined from the first-stage variance.

Keywords: Choice of prior, hierarchical models, noninformative priors, random effects

1 Comments

There is no standard solution to the problem of choosing a prior on the random-effects variance in random-effects models, or mixed models, or what Bayesian analysts usually call "hierarchical models." In the case of a scalar random effect, Browne and Draper (2005) investigated the frequentist behavior of posterior estimates based on a uniform prior and an inverted-gamma prior. They also compared the Bayesian methods to likelihood and quasi-likelihood alternatives.

The main Bayesian messages we take home from Browne and Draper's study are that, in the case of a scalar random effect, (1) a uniform prior on the variance produces posterior distributions with very good operating characteristics: the coverage probabilities remain close to .95 for all of their simulations; and (2) the uniform prior is a bit better than a quasi-uniform inverted-gamma prior. Though the situations for Normal and non-Normal models seem to us different in principle, with some kind of correction seeming necessary before prior rules for non-Normal models match those for the Normal models, the work by Browne and Draper strengthens an already strong case for the uniform prior becoming the "standard solution." The main general statistical message seems to be that this Bayesian method works well. We would underscore the additional general comment made by Browne and Draper, and many before them, that estimates of fixed effects remain very good in the presence of modest errors in estimation of the variance components. This is part of what makes generalized estimating equation estimators so effective (Diggle et al., 2002).

*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, http://www.stat.cmu.edu/~kass

†Unaffiliated

What happens in the vector case? As the dimensionality increases, one anticipates degradation of performance: the choice of prior is likely to matter much more, and one may expect trouble in estimating fixed effects, as well. It would be good to have results like those of Browne and Draper's so that we would know more precisely when to worry, and it would also be very valuable if the field could settle on a reasonable default prior for the non-worrisome and not-very-worrisome situations. The tradition in statistical research is to report results of the form "method A (often the authors' method) works better than method B." This is useful, but statisticians too rarely give practical guidance as to when a method breaks down.

Perhaps future studies of priors for random effects in the vector case will be undertaken. If so, we would like to make one more suggestion: it may be worthwhile to evaluate yet another prior, one we call a "default conjugate prior." In the remainder of our commentary we will describe this prior and indicate why we think it may be of use.

2 A Default Conjugate Prior

In the vector case, under the assumption of a Normal distribution for the random effects (the second stage of the hierarchical model), the uniform prior remains a reasonable candidate. It is also possible to use an inverted Wishart prior on the random effects variance matrix $\mathbf{D}$, which requires the specification of a scale matrix typically considered to be a guess at the value of $\mathbf{D}$. There is, however, rarely good scientific information on which to base this guess. A frequently-applied procedure is to set the scale matrix equal to the maximum likelihood estimator (MLE) of $\mathbf{D}$. Natarajan and Kass (2000) reported simulations indicating that posterior distributions based on this procedure can lead to poor estimates of $\mathbf{D}$, and we also gave a real-data example where scientific inferences are seriously affected. In that paper we also proposed an alternative — the "approximate uniform shrinkage" prior — and showed it to lead to better-behaved posteriors. That prior is easy enough to use, but has not caught on. We here draw attention to yet another alternative, namely the "default conjugate prior." Rather than using the MLE as the scale matrix of the inverse Wishart prior, it may be preferable to base a "guess" at the value of $\mathbf{D}$ on the first-stage data variability. Although the method uses first-stage data both for formulation of the second-stage prior and for computation of the posterior, we note that this particular re-use of the data has asymptotically negligible effects on the posterior.

2.1 The Two-Stage Hierarchical Model

Let us consider the following class of two-stage models:

$$\begin{aligned} \mathbf{Y}_i|\mathbf{b}_i &\sim \prod_{j=1}^{n_i} f\left(Y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}\right), \qquad i = 1, \ldots, k,\ j = 1, \ldots, n_i, \\ \mathbf{b}_i &\sim \mathrm{N}_q\left(\mathbf{0},\ \mathbf{D}\right), \end{aligned} \tag{1}$$

where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^{\mathrm{T}}$ is a vector of observed responses for the i^{th} experimental unit (cluster), $\mathbf{b}_i$ is a $q \times 1$ vector of unobserved cluster-specific random effects and $f(.)$ is an exponential family density function with dispersion parameter ϕ assumed known. The conditional mean of Y_{ij} is assumed to satisfy $\mu_{ij}^{\mathbf{b}} = h\left(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i\right)$, where $\mathbf{x}_{ij}\,(p \times 1)$ and $\mathbf{z}_{ij}\,(q \times 1)$ are design vectors corresponding to the fixed effects $\boldsymbol{\beta}$ and the random effects $\mathbf{b}_i$ respectively and $h(.)$ is a known link function with inverse $g(.)$. Such models belong to the family of generalized linear mixed models (GLMMs). By way of notation we let $\mathbf{X}_i\,(n_i \times p)$ and $\mathbf{Z}_i\,(n_i \times q)$ denote full-rank matrices with rows $\mathbf{x}_{ij}^{\mathrm{T}}$ and $\mathbf{z}_{ij}^{\mathrm{T}}$, respectively.

2.2 Definition and motivation

In this section we assume the prior on $\boldsymbol{\beta}$ will be diffuse (in implementation, typically a multivariate Normal with large variances), and consider the problem of specifying the $q \times q$ scale matrix $\mathbf{R}$ of an inverted Wishart prior for $\mathbf{D}$. Specifically, a random positive-definite symmetric matrix $\mathbf{D}$ is distributed according to an inverted Wishart distribution with $\rho(> q-1)$ degrees of freedom and scale matrix $\mathbf{R}$ if its probability density function is proportional to $\det(\mathbf{D})^{-(\rho+q+1)/2}\, exp\left(-\frac{\rho}{2}\mathrm{tr}\left(\mathbf{R}\mathbf{D}^{-1}\right)\right)$. We denote this inverted Wishart distribution by $\mathrm{IW}(\rho, \rho\mathbf{R})$. Note that when $q = 1$, the inverted Wishart reduces to an inverted gamma distribution and ρ is typically referred to as the shape parameter. We will denote the inverted gamma by IG. Conventional wisdom dictates that a good default specification is one for which ρ is taken to be small and $\mathbf{R}$ is a "minimally informative" prior guess of $\mathbf{D}$.

We now define a default Wishart prior for $\mathbf{D}$ with

$$\begin{aligned} \rho &= q, \\ \tilde{\mathbf{R}} &= c \cdot \left(\frac{1}{k} \sum_{i=1}^{k} \mathbf{Z}_i^{\mathrm{T}} \mathbf{W}_i(\boldsymbol{\beta}) \mathbf{Z}_i \right)^{-1}, \end{aligned}$$

where $\mathbf{W}_i(\boldsymbol{\beta})\,(n_i \times n_i)$ denotes the usual diagonal GLM weight matrix with diagonal elements $\left\{\phi v\left(\mu_{ij}^{\mathbf{0}}\right)\left[\partial g\left(\mu_{ij}^{\mathbf{0}}\right)/\partial \mu_{ij}^{\mathbf{0}}\right]^2\right\}^{-1}$, $v(.)$ is the known variance function based on the density $f(.)$ and the superscript zeros indicate the substitution of $\mathbf{b}_i$ with zero in these quantities. The value of c is an inflation factor representing the amount by which the within-cluster variability should be increased in determining R^*. In our simulation we used $c = 1$. Note that the inverse of $\frac{1}{k}\sum_{i=1}^{k} \mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i(\boldsymbol{\beta})\mathbf{Z}_i$ exists by the full-rank assumption on $\mathbf{Z}_i$. Thus, calculation of $\tilde{\mathbf{R}}$ is straightforward, requiring only a few matrix operations and knowledge of the form of the weight matrix $\mathbf{W}_i$ for the particular exponential family under consideration McCullagh and Nelder (1989), pp. 30.

We now offer two heuristic justifications for $\tilde{\mathbf{R}}$. The first arises from the approximate shrinkage estimate of $\mathbf{b}_i$ — that is, $\widetilde{\mathbf{b}}_i = \mathbf{D}\mathbf{Z}_i^{\mathrm{T}}\left(\mathbf{W}_i^{-1}(\boldsymbol{\beta}) + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{\mathrm{T}}\right)^{-1}\left(\mathbf{Y}_i^* - h(\mathbf{X}_i\boldsymbol{\beta})\right)$ where $\mathbf{Y}_i^*$ is the working dependent variable Breslow and Clayton (1993). After some

matrix manipulations, it can be shown that $\widetilde{\mathbf{b}}_i$ may be expressed as

$$\widetilde{\mathbf{b}}_i \quad = \quad \mathbf{S}_i\mathbf{0} + (\mathbf{I} - \mathbf{S}_i)\,\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,(\mathbf{Y}_i^* - h\,(\mathbf{X}_i\boldsymbol{\beta}))\,,$$

where $\mathbf{I}$ is the $q \times q$ identity matrix and $\mathbf{S}_i = \mathbf{I} - \left(\mathbf{D}^{-1} + \mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$. The matrix $\mathbf{S}_i$ controls the relative contribution of the prior mean $\mathbf{0}$ and the data to the posterior update of $\mathbf{b}_i$, and thus offers a natural metric for evaluating the informativeness of a particular prior guess for $\mathbf{D}$. It ranges from $\mathbf{I} - \left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$ when $\mathbf{D} = \infty$, which corresponds to a flat prior for $\mathbf{b}_i$, to $\mathbf{I}$ when $\mathbf{D} = \mathbf{0}$, which corresponds to a point mass prior for $\mathbf{b}_i$ at zero. A prior guess of $\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$ for $\mathbf{D}$ would result in a weight of $\mathbf{I} - \frac{1}{2}\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$, which is exactly half-way between the weights accorded by the two extreme choices of $\mathbf{D}$. Thus, this seems like a reasonable guess for $\mathbf{D}$ in the absence of any other prior knowledge. However, since $\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$ varies with i, we suggest replacing it with its harmonic mean over clusters, which leads to our choice of $\tilde{\mathbf{R}}$.

A second justification arises from considering a maximum likelihood-based Normal approximation to the GLMM in which the exponential family specification is replaced with

$$\widehat{\mathbf{b}}_i \quad \sim \quad \mathrm{N}_q\left(\mathbf{b}_i, \mathbf{I}\left(\widehat{\mathbf{b}}_i\right)\right),$$

where $\widehat{\mathbf{b}}_i$ is the ML estimator of $\mathbf{b}_i$ based on the first-stage likelihood $\prod_{j=1}^{n_i} f\,(Y_{ij}|\mathbf{b}_i, \boldsymbol{\beta})$, and $\mathbf{I}\left(\widehat{\mathbf{b}}_i\right)$ is the observed information evaluated at $\widehat{\mathbf{b}}_i$. It can be shown that $\mathbf{I}\left(\widehat{\mathbf{b}}_i\right) = \left(\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{W}}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$, where $\widehat{\mathbf{W}}_i\,(\boldsymbol{\beta})$ is the GLM weight matrix $\mathbf{W}_i$ defined previously but with $\widehat{\mathbf{b}}_i$ in place of zero. However, when $\widehat{\mathbf{W}}_i\,(\boldsymbol{\beta})$ is close to $\mathbf{W}_i\,(\boldsymbol{\beta})$, the within-cluster variance $\mathbf{I}\left(\widehat{\mathbf{b}}_i\right)$ will be approximated well by $\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$. Thus, a prior guess of $\left(\frac{1}{k}\sum_{i=1}^{k}\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$ for $\mathbf{D}$, corresponds roughly to an *a priori* belief that the between-cluster variance is equal to the harmonic mean of the within-cluster variance.

Note that our specification for the prior on $\mathbf{D}$ depends on $\boldsymbol{\beta}$ through $\mu_{ij}^{\mathbf{b}}$, which appears in $\mathbf{W}_i\,(\boldsymbol{\beta})$, and is thus a specification of the conditional distribution of $\mathbf{D}$ given $\boldsymbol{\beta}$. A consequence of this appearance of $\boldsymbol{\beta}$ is that the full conditional distribution of $\boldsymbol{\beta}$ given the data and all other parameters will no longer be free of $\mathbf{D}$. Although this presents no substantial difficulties, the simplicity of the standard assumption of independence of $\boldsymbol{\beta}$ and $\mathbf{D}$ (together with a uniform or Normal prior on $\boldsymbol{\beta}$) enables particularly straightforward MCMC implementation via Gibbs sampling (Zeger and Karim, 1991). Thus, we propose a slight modification to the prior given above: we replace the family of conditional distributions of $\mathbf{D}$ given $\boldsymbol{\beta}$ by the single conditional distribution of $\mathbf{D}$ given $\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is an estimate of the regression coefficients from the GLM model obtained by pooling all the data and setting $\mathbf{b}_i = \mathbf{0}$, for all i. That is, we

specify the default inverted Wishart by $\rho = q$, and

$$\mathbf{R}^* = c \cdot \left(\frac{1}{k} \sum_{i=1}^{k} \mathbf{Z}_i^{\mathrm{T}} \mathbf{W}_i \left(\widehat{\boldsymbol{\beta}} \right) \mathbf{Z}_i \right)^{-1}. \tag{2}$$

Note that $\mathbf{W}_i \left(\widehat{\boldsymbol{\beta}} \right)$ is a $O_p \left(k^{-1/2} \right)$ consistent estimator of $\mathbf{W}_i \left(\boldsymbol{\beta} \right)$, and that the estimate $\widehat{\boldsymbol{\beta}}$ may be obtained in a simple pre-calculation.

2.3 Asymptotic irrelevance of the data-dependence in the modified prior

Our modified default prior now depends on the data through the replacement of the conditional prior $\pi \left(\mathbf{D} | \boldsymbol{\beta} \right)$ with $\pi \left(\mathbf{D} | \widehat{\boldsymbol{\beta}} \right)$. It is possible for such a data-dependent substitution to yield very misleading inferences. For example, in the one-sample Normal problem using the conjugate family of prior distributions on the mean μ and variance σ^2: $\pi \left(\mu | \sigma^2 \right) = N \left(\mu_0, \lambda_0 \sigma^2 \right)$, $\pi \left(\sigma^2 \right) = \mathrm{IG} \left(\alpha_0, \beta_0 \right)$, one might take $\widehat{\sigma}$ to be the standard error of the sample mean and substitute it for σ in the Normal prior $\pi \left(\mu | \sigma^2 \right)$. This results in a prior whose informativeness is derived from the data; indeed, it would count the data twice, and is clearly an unreasonable procedure. The substitution we have made, however, is quite different: it does not carry the same amount of information as the full data set, but in fact carries less information than does a single observation (that is, a single cluster).

More formally, let $\boldsymbol{\lambda} = (\boldsymbol{\beta}, \mathbf{D})$, $\pi_{def} \left(\boldsymbol{\lambda} \right)$ and $\pi_{mod} \left(\boldsymbol{\lambda} \right)$ be the original default conjugate prior and its modification, and $q \left(\boldsymbol{\lambda} \right)$ be any alternative non-data-dependent prior. Also, let $G \left(\boldsymbol{\lambda} \right)$ be a function to be estimated and let $E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, \pi_{def} \right)$ be the posterior expectation of $G \left(\boldsymbol{\lambda} \right)$ based on $\pi_{def} \left(\boldsymbol{\lambda} \right)$, and similarly for the other two priors. Then, as $k \to \infty$, we have

$$E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, \pi_{def} \right) = E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, q \right) \left(1 + O_p(k^{-1}) \right), \tag{3}$$

which is one way of saying that, in large samples, the effect of changing the prior is roughly that of changing a single observation. If an informative data-dependent prior were used (analogous to that mentioned for the one-sample Normal) in place of $q \left(\boldsymbol{\lambda} \right)$, Equation (3) would no longer hold. Our modified prior produces

$$E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, \pi_{def} \right) = E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, \pi_{mod} \right) \left(1 + O_p(k^{-1}) \right). \tag{4}$$

This result may be obtained from asymptotic expansions, as in Kass and Steffey (1989), Equation (3.14), using the MLE-based version mentioned just after that equation). The essential observation is that for any $\boldsymbol{\lambda}$ within an order $O(k^{-1/2})$ neighborhood of the true value (toward which a $\sqrt{k}$-consistent estimator will converge) the ratio of the original to modified priors satifies $\pi_{def} \left(\boldsymbol{\lambda} \right) / \pi_{mod} \left(\boldsymbol{\lambda} \right) = 1 + O_p \left(k^{-1/2} \right)$.

Kass and Steffey (1989) pointed out that when the empirical Bayes substitution of an MLE of $\boldsymbol{\lambda}$ is made, the resulting posterior variance of a random effect is too small,

and no longer approximates to order $O_p(k^{-1})$ the correct posterior variance. This is another example of the use of data-dependent priors that may have strong, undesirable effects on inference. It is worth noting, again by way of contrast, that an expression analogous to (4) holds for posterior variances:

$$\text{var}\left(G\left(\boldsymbol{\lambda}\right)|\mathbf{Y},\pi_{def}\right)=\text{var}\left(G\left(\boldsymbol{\lambda}\right)|\mathbf{Y},\pi_{mod}\right)\left(1+O_p(k^{-1})\right).$$

2.4 Simulation study

We ran three simulations, with generally similar results, and report the most dramatic of them here. Unfortunately, while this illustrates the potential value of the default conjugate prior, it is yet again a scalar example.

We compared the performance of the default conjugate prior with three other priors: an inverted Wishart with $\rho = q$ and $\mathbf{R}$ given by the MLE of $\mathbf{D}$, an "ideal" inverted Wishart with $\rho = q$ and $\mathbf{R}$ given by the true value of $\mathbf{D}$, and the approximate uniform shrinkage prior π_{us} (Natarajan and Kass, 2000). The ideal prior provides an unattainable target for the other Wishart priors.

All priors were used in conjunction with a uniform prior for $\boldsymbol{\beta}$. The conditions under which this gives a proper posterior for GLMMs has been derived by Natarajan and Kass (2000), and were verified for the data here. Inferences for the four priors were based on 2,000 samples generated from their posterior distributions for each data set. Posterior sampling was performed using the Gibbs sampler and followed the implementation described by Zeger and Karim (1991) for the inverted Wishart priors, and Natarajan and Kass for π_{us}.

Breslow (1984) presented mutagenicity assay data on the number of revertant colonies of TA98 Salmonella (Y) at six doses of quinoline ($x = 0, 10, 33, 100, 333, 1000$). Three plates were processed at each of the six dose levels resulting in a total of 18 observations. He considered the following Poisson GLMM for these data:

$$\begin{aligned} Y_i|b_i &\sim \text{Poisson}\left(\mu_i^b\right), \quad i=1,\ldots,18, \\ b_i &\sim \text{N}\left(0,\theta\right), \end{aligned} \tag{5}$$

with $\mu_i^b = exp\left(\beta_0+\beta_1\ln\left(x_i+10\right)+\beta_2 x_i+b_i\right)$. The single variance component θ captures the overdispersion due to plate-to-plate variability. The default conjugate prior is $\text{IG}\left(1,R^*\right)$ where $R^* = 18/\sum_{i=1}^{18} w_i$, $w_i = exp\left(\widehat{\beta}_0+\widehat{\beta}_1\ln(x_i+10)+\widehat{\beta}_2 x_i\right)$ and we estimated $\widehat{\boldsymbol{\beta}}$ from the first-stage Poisson likelihood function with $b_i = 0$. The approximate uniform shrinkage prior is $\pi_{us}\left(\theta\right) \propto 1/\left(1+\theta/R^*\right)^2$.

We generated 1,000 data sets from (5) with $\beta_0 = 2.203$, $\beta_1 = .311$, $\beta_2 = -.001$ and $\theta = .040$. These values were chosen because they are close to the estimates obtained for the salmonella data. The estimators of $\boldsymbol{\beta}$ and θ from the four priors were evaluated according to posterior risk and noncoverage probabilities for 95% posterior intervals (the noncoverage probabilities would, ideally, equal .05). The posterior risk was calculated under the squared-error loss function $L\left(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}\right) = \left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ for $\boldsymbol{\beta}$, and

the entropy loss function $L\left(\widehat{\theta},\theta\right) = \left(\widehat{\theta}/\theta - \ln|\widehat{\theta}/\theta| - 1\right)$ for θ. The Bayes estimators corresponding to these loss functions are the posterior mean and harmonic mean respectively. Note that the entropy loss function penalizes underestimation more severely than overestimation in cases when the true value of θ is close to zero. Thus, we would expect the prior π_{us} to have a slightly worse risk than the other priors since it places non-zero mass at zero.

*O*perating Characteristics	IW $(1,\theta)$	IW $\left(1,\widehat{\theta}\right)$	IW $(1,R^*)$	π_{us}
*R*isk				
$\boldsymbol{\beta}$	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
θ	.09 ± .00	.89 ± .05	.12 ± .00	.62 ± .02
*N*oncoverage				
β_0	.046 ± .007	.076 ± .008	.056 ± .007	.070 ± .008
β_1	.054 ± .007	.086 ± .009	.059 ± .007	.067 ± .008
β_2	.051 ± .007	.081 ± .009	.060 ± .007	.075 ± .008
θ	.011 ± .003	.194 ± .012	.007 ± .003	.037 ± .006

Table 1: *S*imulation results: risk and noncoverage probability. $IW(1,\theta)$ denotes the ideal diffuse conjugate prior based on the unknown true value $\theta = .04$. Note that in this one-dimensional case the inverse-Wishart becomes an inverse-gamma. $IW(1,\hat{\theta})$ denotes the diffuse conjugate prior based on the MLE $\hat{\theta}$. $IW(1,R^*)$ denotes the diffuse conjugate prior based on Equation (2), with $c = 1$. The average value of MLE, across the 1,000 data sets, was $\hat{\theta} = .027$ while the average value of R^* was $R^* = .033$.

Table 1 displays these results for $\boldsymbol{\beta}$ and θ under the four priors. An examination of the results shows that the inverted Wishart prior (here, an inverted gamma) centered at the MLE $\widehat{\theta}$ is dominated by the other priors, both in terms of risk and coverage probabilities. The poor risk of this prior is a consequence of the tendency of the MLE to underestimate the true value, while the worse coverage probabilities are due to its failure to account for the extra variability induced by plugging in $\widehat{\theta}$. The default conjugate prior is fairly competitive with the ideal prior and offers slightly better inferences than π_{us} for the regression coefficients.

2.5 Conclusions

There is not much knowledge about the performance of posteriors based on alternative priors for the matrix $\mathbf{D}$ in models of the form (1). The very limited results we have managed to present here are intended to offer the default conjugate prior in (2) as a plausible choice, and we would expect good results for this prior when c is chosen well. Possibly c could be estimated from the data. We hope the future will bring practical

guidance as to when posteriors based on priors for $\mathbf{D}$, including the uniform prior, the default conjugate prior, or other interesting choices such as that recommended by Gelman (2005), are likely to have good frequentist operating characteristics.

References

Breslow, N. E. (1984). "Extra-Poisson variation in log-linear models." *Applied Statistics*, 33: 38–44. 540

Breslow, N. E. and Clayton, D. G. (1993). "Approximate Inference in Generalized Linear Mixed Models." *JASA*, 88: 9–25. 537

Browne, W. J. and Draper, D. (2005). "A comparison of Bayesian and likelihood-based methods for fitting multilevel models." *Bayesian Analysis*. To appear. 535

Diggle, P. J., Haegerty, P., Liang, K. Y., and Zeger, S. L. (2002). *The analysis of Longitudinal Data*. Oxford University Press, 2nd edition.

Gelman, A. (2005). "Prior distributions for variance parameters in hierarchical models (Comment on an article by Browne and Draper)." *Bayesian Analysis*. To appear. 542

Kass, R. E. and Steffey, D. (1989). "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)." *JASA*, 84: 717–726. 539

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd edition. 537

Natarajan, R. and Kass, R. E. (2000). "Reference Bayesian methods for generalized linear mixed models." *JASA*, 95: 227–237. 536, 540

Zeger, S. L. and Karim, M. R. (1991). "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach." *JASA*, 86: 79–86. 540

Bayesian Analysis (2006) **1**, Number 3, pp. 543–546

Comment on article by Browne and Draper

Paul C Lambert*
University of Leicester, UK

I would like to congratulate the authors on a clearly written and detailed paper. Large scale simulation studies are important to understand the properties of complex models which we are increasingly able to fit. The amount of computing time needed for the simulation studies performed by Browne and Draper (stated in the Appendix) demonstrates that this can be a time consuming task.

As stated by the authors, the use of multilevel models has grown substantially over the last few years. However, as listed in the first paragraph of section 1, there are a number of competing methods proposed for their estimation, both Bayesian and likelihood based. Within the Bayesian framework there is of course the added issue of the choice of prior distributions for the various model parameters. It is worth noting here that the increased use of Bayesian methods over the last decade or so has not necessarily been due to a philosophical shift, but rather a desire to fit complex models, with software such as WinBUGS enabling users to do this. Many of these users want their 'data to dominate' and therefore want all prior distributions to be non-informative. However, this is rarely straightforward and in hierarchical models it is the choice of prior distribution for the hierarchical variance parameters that has been shown to be most crucial, particularly in small samples. In earlier work we conducted a simulation study on the choice of prior distribution for the variance component (between study variance) in a meta-analysis of aggregated data (Lambert et al. 2005). One of the advantages of using aggregated data is that models are quicker to fit and we were able to compare 13 different prior distributions for 9 different scenarios. When the number of level 2 units is large the choice of prior distribution becomes less important. However, for many real applications in medicine one would expect the number of level 2 units to be small, for example meta-analysis (Sutton and Abrams 2001) and cluster randomised trials (Turner et al. 2001). It is to the situations where there are only a small number of level 2 units that I wish to address most of my comments.

- The inverse-gamma (ϵ, ϵ) distribution is by far the most common prior distribution used for variance components. One reason for this is that in the set of BUGS examples (Spiegelhalter et al. 1996a,b) it is the only prior distribution used for variance components, with $\epsilon = 0.001$. As Browne and Draper point out, the inverse-gamma (ϵ, ϵ) distribution has a spike near zero and that this can create problems for low values of σ_u^2 or when the number of level 2 units is small. These problems have recently been demonstrated by Gelman (Gelman 2006). My view is that there is a need to educate users to move away from tradition and avoid using this prior distribution for hierarchical variance parameters, particularly when the number of level 2 units is small.

*University of Leicester, UK, http://www.hs.le.ac.uk/personal/pl4/

- One of the problems with both the prior distributions investigated by Browne and Draper is that with a small number of level 2 units, the posterior distribution may include implausibility large values for σ_u^2. The use of weakly informative prior distributions that will give low (or zero) probability to values that are clearly implausible are likely to produce more realistic estimates (Gelman 2006; Lambert et al. 2005).

- A disadvantage of the two prior distributions chosen by Browne and Draper is that interpretation on the variance or precision scales is less obvious and for this reason I prefer prior distributions on the standard deviation scale, particularly if using informative or weakly informative prior distributions, as these will be on the same scale as the model and thus provides greater transparency. Two such prior distributions are the uniform or half-normal distributions. In addition the half-Cauchy distribution used by Gelman looks particularly promising for situations with a small number of level 2 units (Gelman 2006)

- Another important point illustrated in the paper is that the choice of summary statistic (mean, median or mode) can lead to very different point estimates, particularly in small samples. This is of course to be expected when the posterior distribution is skewed, but does illustrate the importance in reporting which summary measure has been used. It is also worthwhile noting that the majority of WinBUGS users rarely report the mode for the simple reason that the standard output does not report it.

- The results of the simulation for the random effects logistic regression (RELR) are particularly interesting with the quasi-likelihood methods performing poorly even with a large number of level 2 and level 3 units. It is for these types of models that the Bayesian approach is particularly advantageous. This is demonstated by their use in genetic epidemiology where complex random effects models are used to model genetic and enviromental associations in pedigree data (Burton et al. 1999). The RELR simulation study has a large number of units in comparison to the variance components simulations and one would expect similar problems to occur regarding the choice of prior distributions when the number of level 2 (or level 3) units are small. I agree with Browne and Draper that other likelihood based approaches need further investigation, in particular the use of adaptive quadrature based methods (Pinheiro and Bates 1995) and hierarchical generalized linear models (Lee and Nelder 1996). However, due to flexibility and potential to extend the models I think it is likely that a Bayesian approach is the most sensible in these situations.

- It is clear is that for any Bayesian hierarchical model involving a small number of units, the role of the prior distribution for the hierarchical variance parameters is crucial and that there is unlikely to be an 'off-the-shelf' vague prior distribution suitable for all scenarios. Therefore a sensitivity analysis should routinely be performed. Finally, it is worth reiterating the importance of reporting all prior distributions used, in both the main and sensitivity analyses, and their impact on results.

References

Burton, P. R., Tiller, K. J., Gurrin, L. C., Cookson, W. O. C. M., Musk, A. W., and Palmer, L. J. (1999). "Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMS) and Gibbs sampling." *Genetic Epidemiology*, 17: 118–140. 544

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, ??: ???–??? 543, 544

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and D. R. Jones, D. (2005). "How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS." *Statistics in Medicine*, 24: 2401–2428. 543, 544

Lee, Y. and Nelder, J. A. (1996). "Hierarchical generalized linear models (with discussion)." *Journal of the Royal Statistical Society Series B*, 58: 619–656. 544

Pinheiro, J. C. and Bates, D. M. (1995). "Approximations to the log-likelihood function in the nonlinear mixed-effects model." *Journal of Computational and Graphical Statistics*, 4: 12–35. 544

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996a). *BUGS Examples Volume 1, Version 0.5*. Cambridge: MRC Biostatistics Unit. 543

— (1996b). *BUGS Examples Volume 2, Version 0.5*. Cambridge: MRC Biostatistics Unit. 543

Sutton, A. J. and Abrams, K. R. (2001). "Bayesian methods in meta-analysis and evidence synthesis." *Statistical Methods in Medical Research*, 10: 277–303. 543

Turner, R. M., Omar, R. Z., and Thompson, S. G. (2001). "Bayesian methods of analysis for cluster randomized trials with binary outcome data." *Statistics in Medicine*, 20: 453–472. 543

Bayesian Analysis (2006) **1**, Number 3, pp. 547–550

Rejoinder

William J. Browne*, and David Draper†

We are grateful to Gelman, Kass and Natarajan, and Lambert for their thoughtful comments (and indeed for the original research that they summarize in their papers), and we offer the following remarks by way of rejoinder.

- Many of the results presented in our article were obtained more than a few years ago (based, as they were, on part of the work in Browne (1998)) and are only now seeing the light of publication largely due to, shall we say, the vagaries of non-Bayesian refereeing. We focused on the $\Gamma^{-1}(\epsilon, \epsilon)$ prior for random-effects variances in some of our work because—under the influence of the `WinBUGS` package and the examples distributed with it—this was very much the most common prior in use in hierarchical/multilevel modeling in the mid to late 1990s. Lambert expresses the opinion that this is still true today, although it appears to us that the pendulum is shifting away from this prior, for reasons like those mentioned by Gelman. (To be fair to the `WinBUGS` development group, in many of the examples distributed with release 1.4.1 they currently offer analyses with both $\Gamma(0.001, 0.001)$ priors on random-effects precision parameters τ and Uniform priors on the corresponding standard deviation parameters $\sigma = \tau^{-1/2}$, although they send a distinctly mixed message by building in default values of 0.001 for each of the shape and scale parameters whenever a parameter is given a Gamma distribution in the `DoodleBUGS` part of the package.)

  It is interesting to see that in 2006 there is still no consensus on a general-purpose choice of diffuse prior for this situation, although the work summarized in both the Gelman and Kass-Natarajan contributions to this discussion may go some distance toward achieving this goal. We have found ourselves recently gravitating toward Uniform priors on random-effects standard deviations, which accord with one of Gelman's suggestions, although instead of using Uniform$(0, \infty)$ (or Uniform$(0, A)$ for huge A) we prefer Uniform$(0, c)$ where c is chosen just large enough not to truncate the marginal likelihood for σ (and, in an interesting resurrection of the sometimes appropriately maligned Gamma prior, c can often be chosen well by making a preliminary fitting with a $\Gamma^{-1}(0.001, 0.001)$ prior on σ^2 and looking at the marginal posterior for σ). It is also interesting that $\Gamma^{-1}(\epsilon, \epsilon)$ priors were originally chosen for computational convenience (through their conditional conjugacy), and the half t family mentioned by Gelman again has surfaced due to computational benefits, this time arising from model expansion. One of us (Browne (2004)) has also seen these benefits in a more complex random effects model, reinforcing Gelman's comments on efficiency of MCMC chains.

*Division of Statistics, School of Mathematical Sciences, University of Nottingham, UK http://www.maths.nottingham.ac.uk/personal/pmzwjb/

†Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA, http://www.ams.ucsc.edu/~draper

- The IGLS estimation method we use to get maximum likelihood estimates in the paper (see Section 2.1) has other features that are of interest to explore for their potential payoff with MCMC methods. Given a particular random effects model, the IGLS method does not in fact directly fit this model, but rather fits a structured multivariate normal model with the whole set of responses treated as one vector-valued outcome, and with constraints (e.g., positive between-groups variance) included in the covariance matrix of the response; these constraints create equivalence between the multivariate model and the original random effects model. We are currently investigating MCMC algorithms for such structured multivariate normal models; here we have the option of allowing the parameter in this model that corresponds to the between-groups variance in the random effects model to have positive prior probability of taking negative values. This has advantages in performing Bayesian model selection and may help in choosing a reference prior for this family of structured multivariate normal models (although the equivalence with the random effects model is lost by such a prior choice).

- In three places in Gelman's paper (Sections 5.1, 5.2, and 6.2) he refers to what he characterizes as the good performance of a particular choice of prior ("the simplest approach ... seems to perform well"; "this prior distribution appears to perform well in this example"; "the half-Cauchy prior distribution does slightly better than the uniform") without saying what standard of merit he is using to come to these conclusions. We believe that the best way to settle issues of this type is through simulation studies (of the type illustrated in our paper, in Kass and Natarajan's contribution to this discussion, and in Lambert et al. (2005)), in which an environment embodying a particular known truth is created and then a variety of Bayesian inferential methods are compared on their ability to reproduce the known truth. This is a form *calibration* inquiry—how often does my method get the right answer?—that it would seem all statisticians, whether they are using Bayesian methods or not, would be interested in undertaking. (How exactly can Gelman know that the half-Cauchy prior distribution does slightly better than the uniform in his ANOVA example without performing such a simulation? See, e.g., Draper (2006) for some recent thoughts on the importance of combining the notions of coherence (internal consistency) and calibration (external consistency) in contemporary Bayesian inference.) In fact, this simulation approach has by now become so easy to perform—e.g., by embedding calls to `WinBUGS` in a random-data-set-generating environment in `R` (in part thanks to the useful `R` functions Gelman has made available at `www.stat.columbia.edu/~gelman/bugsR`)—and inexpensive computers have become so fast that most questions one might have about the calibration properties of a particular choice of diffuse prior can be answered in a completely problem-specific manner with just an hour or two of programming and a few hours or days of computer time.

  At about the time of Browne (1998), we were the co-developers of the MCMC capabilities in the multilevel modeling package `MLwiN` (Rasbash et al. (2005)), and—since we wanted to give users a default choice of diffuse priors for that package—it was natural to ask calibration questions of the type addressed in our paper. We

believe that similar questions are routinely worth asking, not just by software developers but by essentially all Bayesian analysts, and we hope that the implementation and publication of Bayesian calibration studies of the type discussed here will become considerably more frequent in the not-too-distant future.

References

Browne, W. J. 1998. *Applying MCMC Methods to Multilevel Models*. Ph.D. dissertation: University of Bath, U.K. 547, 548

—. 2004. An illustration of the use of reparameterisation methods for improving MCMC efficiency in crossed random effect models. *Multilevel Modelling Newsletter* **16**: 13–25. 547

Draper, D. 2006. Coherence and calibration: comments on subjectivity and "objectivity" in Bayesian analysis. Discussion of "The case for objective Bayesian analysis" by J. Berger and "Subjective Bayesian analysis: principles and practice" by M. Goldstein. *Bayesian Analysis* (this issue). 548

Lambert, P. C., A. J. Sutton, P. R. Burton, K. R. Abrams, and D. R. Jones. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using `WinBUGS`. *Statistics in Medicine* **24**: 2401–2428. 548

Rasbash, J., F. Steele, W. Browne, and B. Prosser. 2005. *A User's Guide to `MLwiN` Version 2.0*. University of Bristol, U.K. (`www.mlwin.com`). 548