ELSEVIER

Computational Statistics & Data Analysis 39 (2002) 203–225

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

# Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation

William J. Browne[a], David Draper[b,*,1], Harvey Goldstein[a], Jon Rasbash[a]

[a] *Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK*
[b] *Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, 1156 High Street, Santa Cruz, CA 95064, USA*

Received 1 July 2000; received in revised form 1 June 2001

**Abstract**

In multilevel modelling it is common practice to assume constant variance at level 1 across individuals. In this paper we consider situations where the level-1 variance depends on predictor variables. We examine two cases using a dataset from educational research; in the first case the variance at level 1 of a test score depends on a continuous "intake score" predictor, and in the second case the variance is assumed to differ according to gender. We contrast two maximum-likelihood methods based on iterative generalised least squares with two Markov chain Monte Carlo (MCMC) methods based on adaptive hybrid versions of the Metropolis-Hastings (MH) algorithm, and we use two simulation experiments to compare these four methods. We find that all four approaches have good repeated-sampling behaviour in the classes of models we simulate. We conclude by contrasting raw- and log-scale formulations of the level-1 variance function, and we find that adaptive MH sampling is considerably more efficient than adaptive rejection sampling when the heteroscedasticity is modelled polynomially on the log scale.


*Keywords:* Adaptive Metropolis-Hastings sampling; Educational data; Heteroscedasticity; Hierarchical modelling; IGLS; Markov chain Monte Carlo (MCMC); Maximum-likelihood methods; MCMC efficiency; Multilevel modelling; RIGLS

* Corresponding author.
*E-mail addresses:* w.browne@ioe.ac.uk (W.J. Browne), draper@ams.ucsc.edu (D. Draper), h.goldstein@ioe.ac.uk (H. Goldstein), j.rasbash@ioe.ac.uk (J. Rasbash).
[1] Web: www.soe.ucsc.edu/∼draper.

0167-9473/02/$ - see front matter 
PII: S 0167-9473(01)00058-5

## 1. Introduction

Over the past 15 years or so, fitting multilevel models to data with a hierarchical or nested structure has become increasingly common for statisticians in many application areas (e.g., Goldstein, 1986, 1995; Bryk and Raudenbush, 1992; Draper, 2002). The main purpose of fitting such models is to partition the variation in a response variable as a function of levels in the hierarchy and relate this variability to descriptions of the data structure. In education, for example, multilevel modelling can be used to calculate the proportion of variation in an observation that is explained by the variability between students, classes, and schools in a 3-level nested structure. Random-effects modelling of this kind is generally combined with fixed-effects modelling, in which predictors are additionally related to the response variable as covariates.

Generally these models assume a constant level-1 variance for the error or residual term for all observations (in our notation students are at level 1 in the 3-level structure above), but there is no reason why this should be true in all applications. An alternative is to allow heteroscedasticity—in other words, to fit models that relate the amount of level-1 variability to predictor variables. We will refer to this here as *complex level-1 variation*. Heteroscedasticity is a common modelling concern in the standard fitting of linear models to data lacking a hierarchical or multilevel structure (e.g., Weisberg, 1985), but far less attention has been paid to this topic with multilevel data.

As our main motivating example we consider a dataset studied in Rasbash et al. (2000), which was originally analysed in Goldstein et al. (1993). This dataset contains exam results for 4059 pupils from 65 schools sampled from six inner London Education Authorities. The response variable of interest is the total score achieved in GCSE examinations (a standardised test taken at age 16 by these pupils). This variable has already been normalised (transformed by replacing each value by its standard normal score) in the dataset we consider.

Table 1 contains mean and variance estimates for the response variable for various partitions of the dataset. One of the main predictors of interest is a score on a reading test (LRT) that all pupils took at age 11. For purposes of partitioning we have divided the pupils into 7 groups of roughly equal sample size based on a standardised version of the LRT score. From the mean column of the table it is clear that girls generally do a bit better than boys and that the LRT score is positively correlated with the exam score. It can also be seen that boys' exam scores are slightly more variable than girls' scores and that the variance of the exam score bears a roughly quadratic relationship to LRT score. Both of these conclusions mean that in fitting a multilevel model to this dataset it will be worth considering the need for complex variation at level 1.

The plan of the paper is as follows. In Section 2 we describe two versions of a maximum-likelihood approach to the fitting of multilevel models with complex level-1 variation and examine several examples of complex variance structures. Sections 3 and 4 present a Markov chain Monte Carlo (MCMC) method for Bayesian fitting of such models based on adaptive Metropolis-Hastings sampling, using two

Table 1
A comparison of means and variances of normalised exam scores for various partitions of the GCSE dataset

| Partition | Sample size | Mean | Variance |
|---|---|---|---|
| Whole dataset | 4,059 | 0.000 | 1.000 |
| Boys | 1,623 | −0.140 | 1.052 |
| Girls | 2,436 | 0.093 | 0.940 |
| Standardised LRT $< -1$ | 612 | −0.887 | 0.731 |
| $-1 <$ standardised LRT $< -0.5$ | 594 | −0.499 | 0.599 |
| $-0.5 <$ standardised LRT $< -0.1$ | 619 | −0.191 | 0.650 |
| $-0.1 <$ standardised LRT $< 0.3$ | 710 | 0.044 | 0.658 |
| $0.3 <$ standardised LRT $< 0.7$ | 547 | 0.279 | 0.659 |
| $0.7 <$ standardised LRT $< 1.1$ | 428 | 0.571 | 0.678 |
| $1.1 <$ standardised LRT | 549 | 0.963 | 0.703 |

different proposal distributions. In Section 5 we give results from two simulation studies investigating the bias and interval coverage properties, in repeated sampling, of the four fitting methods described in the previous three sections. Section 6 examines alternatives (a) to our MCMC methods and (b) to our formulation of complex variance structures, and Section 7 discusses our conclusions and suggests extensions of the work presented here.

## 2. Maximum-likelihood-based methods and complex variance structures

We begin by describing a general 2-level model with complex variation (later sections will examine methods to fit alternatives to this general model with additional constraints added). The basic structure for a general Gaussian multilevel model is

$$y \sim N_n(X\beta, V). \tag{1}$$

Here $y$ is an $(n \times 1)$ vector of responses, not necessarily independently distributed, with $\beta$ a $(p_f \times 1)$ vector of fixed-effect coefficients of the predictors in the $(n \times p_f)$ matrix $X$; $n$ is the total number of level-1 observations in the data set (4059 students, in the example in Section 1), and $p_f$ is the number of fixed effects in the model. The $(n \times n)$ covariance matrix $V$ for the responses contains all the random structure in the model; for the two-level case we can write the variance term $V_{ij,ij} = \Sigma_{e,ij} + \Sigma_{u,ij}$ for observation $i$ in level-2 unit $j$ (in our example $j$ runs from 1 to 65, the number of schools in the GCSE data set). In this expression the variance has been partitioned into separate terms for the two levels, with $e$ and $u$ denoting random effects at levels 1 and 2, respectively. The covariances between the responses have the form $V_{ij,i'j} = f(\Sigma_{u,ij}, \Sigma_{u,i'j})$ if the two observations are in the same level-2 unit, and $V_{ij,i'j'} = 0$ otherwise (also $V_{ij,ij'} = 0$ for $j \neq j'$). This means that if the $y$ vector

is ordered so that all the observations in each level-2 unit are grouped together, $V$ has a block diagonal form.

In this general formulation the level-1 and level-2 variances and covariances are potentially different for each pair of observations, but important special cases exist with simpler structure, e.g., variance-components models where both the level-1 and level-2 variances are constant across observations. The covariates $X$ may make an appearance in the random structure of the model, leading to a further partition of $\Sigma_{u,ij}$. An example is a random-slopes regression model with a single predictor $X_{1,ij}$, in which $\Sigma_{u,ij} = \Omega_{u,00} + 2X_{1,ij}\Omega_{u,01} + X_{1,ij}^2\Omega_{u,11}$. Here $\Omega_u$ consists of the variance and covariance terms at level 2 expressed as a matrix (with structural zeroes where necessary; for example, the $\Sigma_{u,ij}$ expression above is the product of the matrix

$$\begin{pmatrix} \Omega_{u,00} & \Omega_{u,01} & 0 \\ 0 & \Omega_{u,01} & \Omega_{u,11} \end{pmatrix}$$

with the vector $(1, X_{1,ij}, X_{1,ij}^2)^{\mathrm{T}}$). Using this notation the (general) within-block covariance term can be written $f(\Sigma_{u,ij}, \Sigma_{u,i'j}) = X_{ij}^{\mathrm{T}}\Omega_u X_{i'j}$, where $X_{ij}$ is a vector of predictors.

In the language of this section, what was referred to earlier as complex variation at level 1 simply means partitioning the level-1 variance so that it depends in a natural way on predictor variables. Fig. 1 presents several potential variance structures that can be fitted to the GCSE dataset described earlier. The corresponding models are

$$\begin{aligned} &y_{ij} \sim N(\beta_0 + \beta_1 X_{1,ij}, V), \\ &\Sigma_{e,ij} = \Omega_{e,00} + 2X_{1,ij}\Omega_{e,01} + X_{1,ij}^2\Omega_{e,11}; \end{aligned} \tag{2}$$

$$\begin{aligned} &y_{ij} \sim N(\beta_0 + \beta_1 X_{1,ij}, V), \\ &\Sigma_{u,ij} = \Omega_{u,00}, \\ &\Sigma_{e,ij} = \Omega_{e,00} + 2X_{1,ij}\Omega_{e,01} + X_{1,ij}^2\Omega_{e,11}; \end{aligned} \tag{3}$$

$$\begin{aligned} &y_{ij} \sim N(\beta_0 + \beta_1 X_{1,ij}, V), \\ &\Sigma_{u,ij} = \Omega_{u,00} + 2X_{1,ij}\Omega_{u,01} + X_{1,ij}^2\Omega_{u,11}, \\ &\Sigma_{e,ij} = \Omega_{e,00} + 2X_{1,ij}\Omega_{e,01} + X_{1,ij}^2\Omega_{e,11}; \end{aligned} \tag{4}$$

and

$$\begin{aligned} &y_{ij} \sim N(\beta_0 + \beta_1 X_{1,ij} + \beta_2 X_{2,ij}, V), \\ &\Sigma_{u,ij} = \Omega_{u,00} + 2X_{1,ij}\Omega_{u,01} + X_{1,ij}^2\Omega_{u,11}, \\ &\Sigma_{e,ij} = \Omega_{e,00} + 2X_{1,ij}\Omega_{e,01} + 2X_{1,ij}X_{2,ij}\Omega_{e,12} + X_{2,ij}\Omega_{e,22}. \end{aligned} \tag{5}$$

In all these models $X_1$ refers to the standardised LRT score and $X_2$ refers to gender (coded 0 for boys and 1 for girls). In Eq. (2) we have a simple one-level regression model with a quadratic variance relationship with LRT; the other models involve fitting increasingly complex variance structures to the data in a two-level framework.

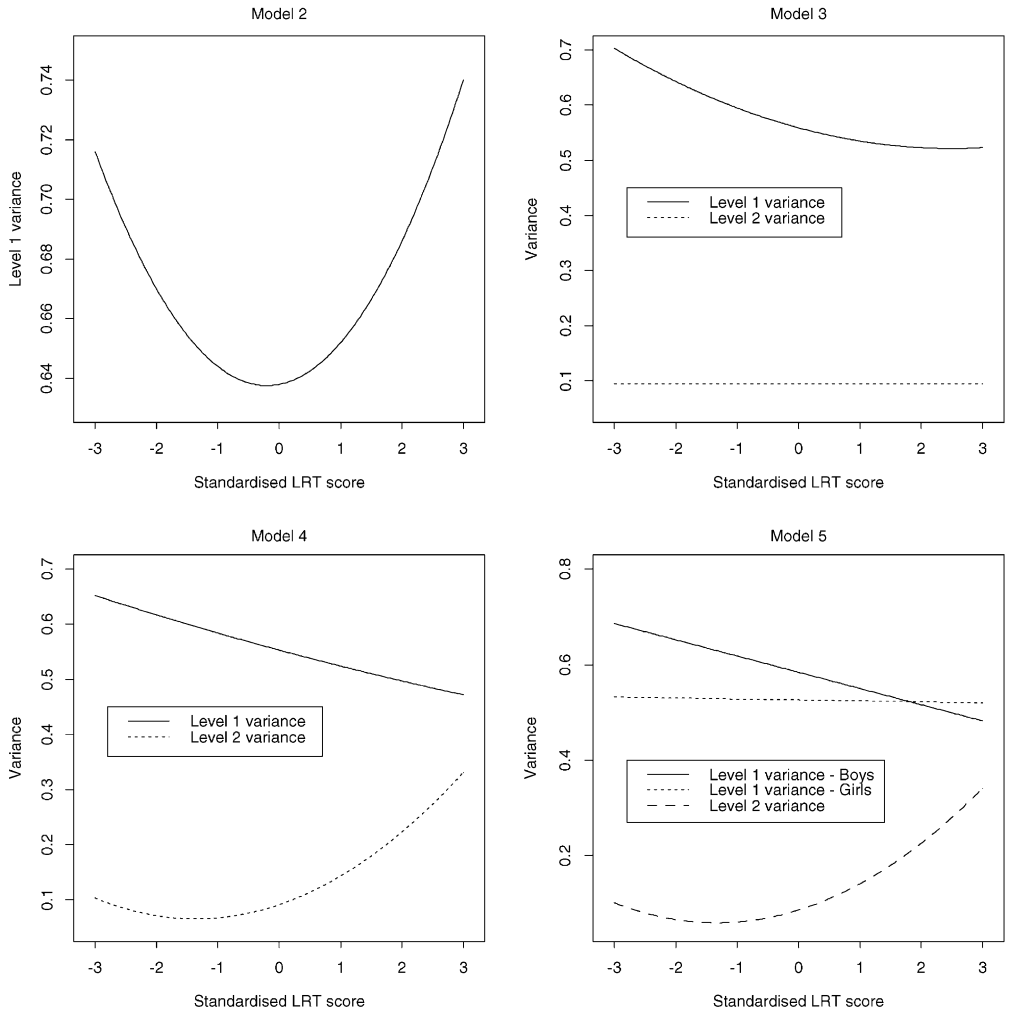


Fig. 1. Four different variance structures fitted to the GCSE dataset.

One approach to fitting models such as (2–5) via maximum likelihood (ML) is based on *iterative generalised least squares* (IGLS), and its restricted variant (RIGLS, also known as REML) which corrects for bias. The basic idea is similar to that of the EM algorithm (Dempster et al., 1977) in that (a) an estimate $\hat{\beta}$ of $\beta$ in (1) is obtained using a current estimate of $V$ and (b) an estimate of $V$ is then obtained using the $\hat{\beta}$ from (a), but in IGLS/RIGLS the estimation of the covariance matrix $V$ is recast as a regression problem and weighted least squares is used in both steps (see Goldstein, 1986, 1989 for details).

Table 2 gives IGLS estimates obtained for models (2–5) applied to the GCSE data. Both gender and LRT score are evidently useful in predicting GCSE score. Model (2), which naively ignores the hierarchical nature of the data, hints at heteroscedasticity (the ML estimate of $\Omega_{e,11}$ is about as big as its standard error (SE)),

Table 2
IGLS estimates for models (2–5) fitted to the GCSE dataset (standard errors (SEs) in parentheses)

| Parameter | Model | | | |
|---|---|---|---|---|
| | (2) | (3) | (4) | (5) |
| $\beta_0$ | −0.002 (0.013) | 0.002 (0.040) | −0.012 (0.040) | −0.112 (0.043) |
| $\beta_1$ | 0.596 (0.013) | 0.565 (0.013) | 0.558 (0.020) | 0.554 (0.020) |
| $\beta_2$ | — | — | — | 0.175 (0.032) |
| $\Omega_{u,00}$ | — | 0.094 (0.018) | 0.091 (0.018) | 0.086 (0.017) |
| $\Omega_{u,01}$ | — | — | 0.019 (0.007) | 0.020 (0.007) |
| $\Omega_{u,11}$ | — | — | 0.014 (0.004) | 0.015 (0.004) |
| $\Omega_{e,00}$ | 0.638 (0.017) | 0.559 (0.015) | 0.553 (0.015) | 0.584 (0.021) |
| $\Omega_{e,01}$ | 0.002 (0.007) | −0.015 (0.007) | −0.015 (0.006) | −0.034 (0.010) |
| $\Omega_{e,11}$ | 0.010 (0.011) | 0.006 (0.009) | 0.001 (0.009) | — |
| $\Omega_{e,12}$ | — | — | — | 0.032 (0.013) |
| $\Omega_{e,22}$ | — | — | — | −0.058 (0.026) |

but (from the estimates of $\Omega_{u,00}$ in Eqs. (3)–(5)) there is a clear need for two-level modelling, and the full complexity of what is required to describe the data only comes into focus with model (5) (in which every estimate is at least 2.2 times as large as its SE).

## 3. An MCMC method for a general 2-level Gaussian model with complex level-1 variation

Browne and Draper (2000, 2001) gave Gibbs-sampling algorithms for Bayesian fitting of 2-level variance-components and random-slopes-regression models, respectively. In this section we consider a general 2-level model with complex variation at level 1; this can easily be generalised to an $N$-level model via an approach similar to the method detailed in Browne (1998). For MCMC fitting of model (1) it is useful to rewrite it as follows, with $y_{ij}$ denoting the (scalar) outcome for (level-1) observation $i$ in level-2 unit $j$:

$$y_{ij} = X_{ij}\beta + Z_{ij}u_j + X_{ij}^{\mathrm{C}}e_{ij},$$
$$u_j \sim N_{p_2}(0, \Omega_u), \quad e_{ij} \sim N_{p_1}(0, \Omega_e). \tag{6}$$

Here $\beta, u_j$, and $e_{ij}$ are $(p_f \times 1), (p_2 \times 1)$, and $(p_1 \times 1)$ vectors of fixed-effects parameters and level-2 and level-1 residuals, respectively; $X_{ij}, Z_{ij}$, and $X_{ij}^{\mathrm{C}}$ are vectors of predictor values (the C stands for composite; see below); and $p_1$ and $p_2$ are the numbers of parameters specifying the random effects at levels 1 and 2, respectively. The IGLS/RIGLS methods do not directly estimate the $u_j$ and $e_{ij}$, but they can be estimated after fitting the model using a method given in Goldstein (1995). In Eq. (6), $\Omega_e$ and $\Omega_u$ are the variance terms at level 1 and level 2 written as $(p_1 \times p_1)$ and $(p_2 \times p_2)$ matrices, respectively.

Gibbs sampling procedures for fitting multilevel models such as (6) proceed most smoothly by treating the level-2 residuals as latent variables when forming the full conditional posterior distributions. In a multilevel model with simple (homoscedastic) variation at level 1, the level-1 residuals may be calculated at each iteration by subtraction. In the above model we cannot explicitly compute the individual level-1 residuals; instead we deal with the "composite" residuals $X_{ij}^{C} e_{ij}$ as these can be calculated by subtraction. The important part of the algorithm that follows is to store the composite level-1 variance function for each individual

$$\Sigma_{e,ij} = (X_{ij}^{C})^{T} \Omega_e X_{ij}^{C}. \tag{7}$$

All the other parameters then depend on the level-1 covariance matrix $\Omega_e$ through these individual variances. This means that the algorithm that follows, apart from the updating step for $\Omega_e$, is almost identical to the algorithm for the same model without complex variation (Browne, 1998).

### 3.1. Inverse-Wishart proposals for the level-1 covariance matrix

In the first MCMC method examined in this paper, we collect together the terms in the variance equation at level 1, $\Sigma_{e,ij}$, into the covariance matrix $\Omega_e$. Updating $\Omega_e$ using a Metropolis-Hastings (MH) algorithm therefore requires a proposal distribution that generates positive-definite matrices (later we will relax this restriction). We use an inverse-Wishart proposal distribution with expectation the current estimate $\Omega_e^{(t)}$ at iteration $t$ to generate $\Omega_e^{(t+1)}$. In the parameterisation used, for example, by Gelman et al. (1995a), the inverse-Wishart distribution $W_k^{-1}(v,S)$ for a $(k \times k)$ matrix has expectation $(v-k-1)^{-1}S$. So if we let $v = w + k + 1$ and $S = w\Omega_e^{(t)}$, where $w$ is a positive integer degrees of freedom parameter, this will produce a distribution with expectation $\Omega_e^{(t)}$. The parameter $w$ is a tuning constant which may be set to an integer value that gives the desired MH acceptance rate.

For prior distributions on the parameters in model (6) we make the following choices in this algorithm: a generic prior $p(\Omega_e)$ (to be specified in Section 4.2; we use the same prior for both MCMC methods for comparability) for the level-1 covariance matrix, an inverse-Wishart prior $\Omega_u \sim W_{p_2}^{-1}(v_2, S_2)$ for the level-2 covariance matrix, and a multivariate normal prior $\beta \sim N_{p_f}(\mu_p, S_p)$ for the fixed effects parameter vector. The algorithm, which is detailed in Appendix $A$, is a hybrid of Gibbs and MH steps; it divides the parameters and latent variables in (6) into four blocks and uses multivariate normal Gibbs updates for $\beta$ and the $u_j$, inverse-Wishart Gibbs updates for $\Omega_u$, and inverse-Wishart MH proposals for $\Omega_e$.

### 3.2. An adaptive method for choosing the tuning constant w

Browne and Draper (2001) describe an adaptive hybrid Metropolis-Gibbs sampler for fitting random-effects logistic regression models. Gibbs sampling may be used in such models for variance parameters, but Metropolis updates are needed for fixed effects and latent residuals. Browne and Draper employ a series of univariate normal proposal distributions (PDs) for these quantities, and give a procedure for adaptive

Table 3
An illustration of the adaptive MH procedure of Section 3.2 with model (4) applied to the GCSE data

| Iterations | Acceptance rate (%) | $w$ | Within tolerance? |
|---|---|---|---|
| 0 | — | 100 | 0 |
| 100 | 20 | 138 | 0 |
| 200 | 19 | 195 | 0 |
| 300 | 30 | 208 | 1 |
| 400 | 31 | 215 | 2 |
| 500 | 30 | 229 | 3 |

choice of appropriate values for the variances of these PDs to achieve efficient MH acceptance rates. Here we provide a modification of this procedure for the case of inverse-Wishart proposals.

We set the tuning parameter $w$ described above to an arbitrary starting value (100 in the example that follows) and run the algorithm in batches of 100 iterations. The goal is to achieve an acceptance rate for the level-1 covariance matrix that lies within a specified tolerance interval $(r-\Delta, r+\Delta)$. We compare the empirical acceptance rate $r^*$ for the current batch of 100 iterations with the tolerance interval, and modify the proposal distribution appropriately before proceeding with the next batch of 100. The modification performed at the end of each batch is as follows:

$$\text{If } r^* \geqslant r, \quad \text{then } w \to \frac{w}{[2-(1-r^*)/(1-r)]} - 1,$$

$$\text{else } w \to w\left(2 - \frac{r^*}{r}\right) + 1, \tag{8}$$

where only the integer part of $w$ is used in (8). The amount by which $w$ is altered in each iteration of this procedure is an increasing function of the distance between $r$ and $r^*$ (the use of the multiplicative factors $[2-(1-r^*)/(1-r)]^{-1}$ and $(2-r^*/r)$ approximates a binary search in which the distance between the current and estimated optimal $w$ is halved at each step). The adaptive procedure ends when three successive $r^*$ values lie within the tolerance interval; the value of $w$ is then fixed and we proceed with the usual burn-in and monitoring periods. In the rest of the paper we refer to the procedure described here as the *adaptive MH method.*

### 3.3. An example

We consider the model in Section 2 which has a quadratic relationship between the variance and the LRT predictor (model (4)). The adaptive procedure was run for this model with a target acceptance rate of $r = 32\%$ (based on a recommendation in Gelman et al., 1995b) and a tolerance of $\Delta = 5\%$. Table 3 summarises the progress of the adaptive method in this example; here only 500 iterations are required to adjust the proposal distribution to give the desired acceptance rate (500–2000 iterations are typically needed in the applications we have examined).

Table 4
Parameter estimates for four methods of fitting model (4) to the London schools dataset (SEs/posterior standard deviations in parentheses). The MCMC methods were monitored for 50,000 iterations after the adaptive procedure and a burn-in of 500 iterations from IGLS starting values

| Parameter | | | MCMC method | |
|---|---|---|---|---|
| | IGLS | RIGLS | 1 | 2 |
| $\beta_0$ | −0.012 (0.040) | −0.012 (0.040) | −0.011 (0.040) | −0.012 (0.040) |
| $\beta_1$ | 0.558 (0.020) | 0.558 (0.020) | 0.560 (0.020) | 0.559 (0.020) |
| $\Omega_{u,00}$ | 0.091 (0.018) | 0.093 (0.018) | 0.095 (0.020) | 0.095 (0.020) |
| $\Omega_{u,01}$ | 0.019 (0.007) | 0.019 (0.007) | 0.020 (0.007) | 0.020 (0.007) |
| $\Omega_{u,11}$ | 0.014 (0.004) | 0.015 (0.004) | 0.014 (0.005) | 0.014 (0.005) |
| $\Omega_{e,00}$ | 0.553 (0.015) | 0.553 (0.015) | 0.547 (0.014) | 0.553 (0.015) |
| $\Omega_{e,01}$ | −0.015 (0.006) | −0.015 (0.006) | −0.015 (0.007) | −0.015 (0.007) |
| $\Omega_{e,11}$ | 0.001 (0.009) | 0.001 (0.009) | 0.009 (0.007) | 0.003 (0.009) |

Table 4 compares the estimates produced by this MCMC method for model (4) to those (a) from the IGLS and RIGLS procedures and (b) from another MCMC method to be described in the next section (here and throughout the paper, MCMC point estimates are posterior means, and we have used the MCMC diagnostics in the software package CODA (Best et al., 1995) to develop a monitoring strategy that ensures good mixing and accurate posterior summaries). We used a slightly informative inverse-Wishart prior $\Omega_u \sim W_{p_2}^{-1}(v_2, S_2)$ for the level-2 covariance matrix for the MCMC methods based on the RIGLS estimate $S_2$ (taking a value of $v_2$ small enough to create an essentially flat prior), and a uniform prior for the level-1 covariance matrix. We have found that short burn-in periods from maximum-likelihood (ML) starting values are sufficient to yield good MCMC results; in all cases in this paper we use burn-ins of 500 iterations from ML initial values. In this example the results for all the methods are fairly similar, with one exception: the estimate of $\Omega_{e,11}$ is noticeably larger using MCMC method 1. This difference highlights the fact that the first MCMC approach actually fits a model with an extra positive-definite constraint: we are forcing $\Omega_{e,11}$ to be positive, which inflates the point estimate. The second MCMC method, which we consider below, is based on different constraints; when we examined the chain of values it produced for $\Omega_{e,11}$ we found that nearly 40% of the values were negative. There is no inconsistency in this result: in the model to be examined in the next section, $\Omega_{e,11}$ is not a variance.

## 4. Truncated normal proposals for the level-1 variance function

The inverse-Wishart updating method assumes that the variance function at level 1 arises from a positive-definite covariance matrix. We now consider an alternative method that, in a manner similar to IGLS and RIGLS, only requires the variance at level 1 to be a linear function of the parameters. This MCMC solution will still have more constraints than the IGLS solution, because we are still considering

the level-1 and level-2 variances separately and both of these quantities must be positive.

The constraint used in MCMC method 1 that the covariance matrix at level 1 is positive-definite is actually stronger than necessary. Positive-definite matrices will guarantee that any vector $X_{ij}^{C}$ will produce a positive variance in Eq. (6); a milder but still scientifically reasonable constraint is to allow all values of $\Omega_e$ such that $(X_{ij}^{C})^{T}\Omega_e X_{ij}^{C} > 0$ for all $i$ and $j$. This restriction appears complicated to work with, but if we consider each of the parameters in $\Omega_e$ separately and assume the other variables are fixed the constraint becomes manageable. It is once again useful to rewrite model (1), this time as follows:

$$y_{ij} = X_{ij}\beta + Z_{ij}u_j + e_{ij}^{*},$$
$$u_j \sim N_{p_2}(0, \Omega_u), \quad e_{ij}^{*} \sim N(0, \Sigma_{e,ij}), \tag{9}$$

where $e_{ij}^{*} = X_{ij}^{C}e_{ij}$ and $\Sigma_{e,ij}$ is given by Eq. (7). Here the composite level-1 residuals $e_{ij}^{*}$ are normally distributed with variances that depend on the predictors; consequently the constraint that the level-1 variance is always positive is still satisfied but $\Omega_e$ need not be positive-definite.

### 4.1. MH updating: method 2

Our second method is identical to the first for $\beta$, the $u_j$, and $\Omega_u$ (steps 1, 2, and 4 in Appendix $A$) but involves a Hastings update with a different proposal distribution for $\Omega_e$. We update each parameter in the level-1 variance equation in turn, always requiring for all $i$ and $j$ at every iteration $t$ in the Markov chain that

$$\Sigma_{e,ij} = (X_{ij}^{C})^{T}\Omega_e^{(t)}X_{ij}^{C} > 0. \tag{10}$$

Considering first the diagonal terms, $\Omega_{e,kk}$, for each $k = 1, \ldots, p_1$ constraint (10) can be written

$$\Sigma_{e,ij} = (X_{ij(k)}^{C})^{2}\Omega_{e,kk}^{(t)} - d_{ij(kk)}^{C} > 0,$$

where

$$d_{ij(kk)}^{C} = (X_{ij(k)}^{C})^{2}\Omega_{e,kk}^{(t)} - (X_{ij}^{C})^{T}\Omega_e^{(t)}X_{ij}^{C}; \tag{11}$$

here $X_{ij(k)}^{C}$ is the $k$th element of the vector $X_{ij}^{C}$. This is equivalent to requiring that

$$\Omega_{e,kk}^{(t)} > \max_{e,kk} \equiv \max_{i,j} \frac{d_{ij(kk)}^{C}}{(X_{ij(k)}^{C})^{2}}. \tag{12}$$

We use a normal proposal distribution with variance $s_{kk}^{2}$ but reject generated values that fail to satisfy (12). This amounts to using a truncated normal proposal, as shown in Fig. 2(i). The Hastings ratio $R$ can then be calculated as the ratio of the two truncated normal distributions shown in Fig. 2(i) and (ii). Letting the value for

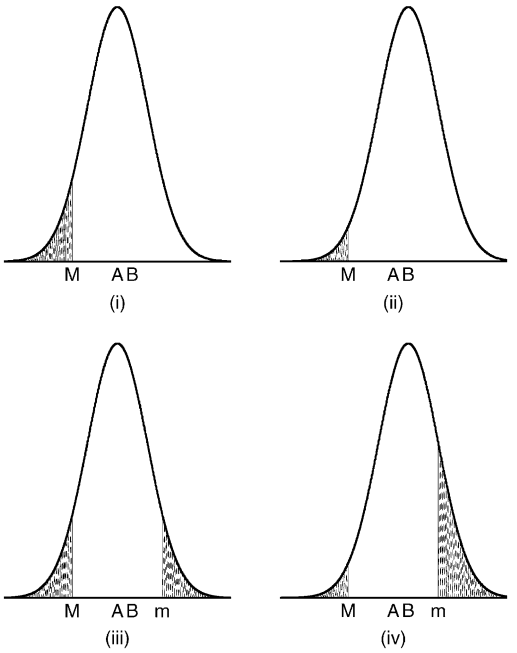


Fig. 2. Plots of truncated univariate normal proposal distributions for a parameter $\theta$. $A$ is the current value $\theta^c$ and $B$ is the proposed new value $\theta^*$. $M$ is $\max_\theta$ and $m$ is $\min_\theta$, the truncation points. The distributions in (i) and (iii) have mean $\theta^c$, while the distributions in (ii) and (iv) have mean $\theta^*$.

$\Omega_{e,kk}$ at time $t$ be $A$ and the proposed value for time $(t+1)$ be $B$,

$$R = \frac{1 - \Phi[(\max_{e,kk} - B)/s_{kk}]}{1 - \Phi[(\max_{e,kk} - A)/s_{kk}]}. \tag{13}$$

The update step is then as follows:

$$\Omega_{e,kk}^{(t+1)} = \left\{ \begin{array}{ll} \Omega_{e,kk}^{*} & \text{with probability } \min[1, R\,\frac{p(\Omega_{e,kk}^{*}|y,\beta,u,\Omega_u)}{p(\Omega_{e,kk}^{(t)}|y,\beta,u,\Omega_u)}] \\ \Omega_{e,kk}^{(t)} & \text{otherwise} \end{array} \right\}, \tag{14}$$

where $p(\Omega_{e,kk}^{*}|y,\beta,u,\Omega_u)$ and the corresponding density in the denominator of (14) are given by (28).

The diagonal terms are a special case as they are always multiplied by a positive quantity in the variance equation, so that the proposal distribution needs only one truncation point. More generally for the non-diagonal terms $\Omega_{e,kl}$ we get the following. As before, at time $t$ for all $i$ and $j$ constraint (10) must be satisfied; for each $1 \leqslant k < l \leqslant p_1$ this can be rewritten

$$\Sigma_{e,ij} = 2X_{ij(k)}^{\mathrm{C}} X_{ij(l)}^{\mathrm{C}} \Omega_{e,kl}^{(t)} - d_{ij(kl)}^{\mathrm{C}} > 0,$$

where

$$d_{ij(kl)}^{\mathrm{C}} = 2X_{ij(k)}^{\mathrm{C}} X_{ij(l)}^{\mathrm{C}} \Omega_{e,kl}^{(t)} - (X_{ij}^{\mathrm{C}})^{\mathrm{T}} \Omega_e^{(t)} X_{ij}^{\mathrm{C}}. \tag{15}$$

This is equivalent to the two constraints

$$\Omega_{e,kl}^{(t)} > \max_{e,kl^+} \equiv \max_{ij} \left\{ \frac{d^{\mathrm{C}}_{ij(kl)}}{2X^{\mathrm{C}}_{ij(k)}X^{\mathrm{C}}_{ij(l)}} \text{ over all } (i,j) \text{ such that } X^{\mathrm{C}}_{ij(k)}X^{\mathrm{C}}_{ij(l)} > 0 \right\},$$

$$\Omega_{e,kl}^{(t)} < \min_{e,kl^-} \equiv \min_{ij} \left\{ \frac{d^{\mathrm{C}}_{ij(kl)}}{2X^{\mathrm{C}}_{ij(k)}X^{\mathrm{C}}_{ij(l)}} \text{ over all } (i,j) \text{ such that } X^{\mathrm{C}}_{ij(k)}X^{\mathrm{C}}_{ij(l)} < 0 \right\}. \tag{16}$$

We again use a normal proposal distribution, this time with variance $s^2_{kl}$, and again values failing to satisfy (16) are rejected. This leads to the truncated normal proposal shown in Fig. 2(iii). The Hastings ratio $R$ is then simply the ratio of the two truncated normal distributions shown in Fig. 2(iii) and (iv). Letting the value for $\Omega_{e,kl}$ at time $t$ be $A$ and the proposed value for time $(t+1)$ be $B$,

$$R = \frac{\Phi[(\min_{e,kl^-} - B)/s_{kl}] - \Phi[(\max_{e,kl^+} - B)/s_{kl}]}{\Phi[(\min_{e,kl^-} - A)/s_{kl}] - \Phi[(\max_{e,kl^+} - A)/s_{kl}]}. \tag{17}$$

The update step is then similar to (14) with subscripts $kl$ in place of $kk$ in the $\Omega_e$ terms.

### 4.2. Proposal distribution variances and prior distributions

In the method outlined above we consider each parameter in $\Omega_e$ separately. This means that we use a separate truncated univariate normal proposal distribution for each parameter subject to the constraints that the value generated will produce a positive level-1 variance $\Sigma_{e,ij}$ for all $i$ and $j$. We therefore need to choose a proposal distribution variance for each parameter. Two possible solutions are to use the variance of the parameter estimate from the RIGLS procedure multiplied by a suitable positive scale factor, or to use an adaptive approach before the burn-in and monitoring run of the simulation. See Browne and Draper (2001) for a description of both of these methods in the case of random effects logistic regression models.

Prior distributions using this method must take account of the constraints imposed on the parameters. In all the analyses we perform in this paper with this method, we use a series of marginal uniform priors for the level-1 variance terms subject to the constraints; in other words, all valid combinations of parameter estimates for $\Omega_e$ are a priori equally likely. Other attempts at specifying prior distributions may well fail to respect the constraints.

### 4.3. Examples

Model (4) was fitted to the GCSE data in Section 3.3, and the estimates produced by both MCMC methods are shown in Table 4. For the truncated normal method we used the adaptive MH procedure, in this case with a desired acceptance rate of 50% as the parameters are updated separately (following the univariate recommendations of Gelman et al. (1995b)). The advantage of the truncated normal method is that

Table 5
Parameter estimates for three methods fitted to model (18) for the GCSE dataset. MCMC method 2, using truncated normal proposals, was monitored for 50,000 iterations following the adapting period and a burn-in of 500 iterations from IGLS starting values

| Parameter | IGLS | RIGLS | MCMC method 2 |
|---|---|---|---|
| $\beta_0$ | −0.161 (0.058) | −0.161 (0.058) | −0.160 (0.060) |
| $\beta_1$ | 0.261 (0.041) | 0.261 (0.041) | 0.260 (0.040) |
| $\Omega_{u,00}$ | 0.162 (0.031) | 0.165 (0.032) | 0.171 (0.035) |
| $\Omega_{e,00}$ | 0.913 (0.032) | 0.914 (0.032) | 0.916 (0.032) |
| $\Omega_{e,01}$ | −0.062 (0.020) | −0.062 (0.020) | −0.062 (0.020) |

it can handle variance functions that would not necessarily have a positive-definite matrix form. For illustration we consider a simple case which the inverse-Wishart method cannot fit. Our model is as follows:

$$y_{ij} \sim N(\beta_0 + \text{girl}_{ij}\beta_1, V),$$

where

$$V = \Omega_{u,00} + \Omega_{e,00} + 2\,\text{girl}_{ij}\Omega_{e,01}. \tag{18}$$

This model includes a variance for boys and a term that represents the difference in variance between boys and girls. The results from fitting this model are given in Table 5; all methods give roughly the same estimates for the level-1 variance terms. The total variances produced by the model for boys ($\Omega_{u,00} + \Omega_{e,00}$) and girls ($\Omega_{u,00} + \Omega_{e,00} + 2\,\Omega_{e,01}$) are similar to the values given in the part of Table 1 where the variance in the response is calculated for boys and girls separately.

## 5. Simulation studies

In this section we examine the bias and interval-coverage properties, in repeated sampling, of the four methods described above, in two sets of simulated models with complex level-1 variation based on the GCSE example. We first consider model (4), which features a quadratic variance relationship with the input reading test (LRT) predictor. As true (population) parameters for our simulation we used values close to the estimates obtained in the actual data, with one exception: we increased $\Omega_{e,11}$ so that the correlation of the random effects at level 1 was reduced. This is because sample datasets drawn from multilevel models with high correlation cause convergence problems with the IGLS and RIGLS methods (Browne and Draper, 2001).

One thousand datasets were generated randomly according to model (4)—with the same numbers of level-1 and level-2 units (4059 and 65, respectively) as in the original GCSE data set, and the same distribution of level-1 observations within level-2 units—and fitted using the four methods, with the results presented in Table 6. For the MCMC methods (in both of the simulation studies) the posterior distribution with each dataset was monitored for 10,000 iterations after the adapting period

Table 6
Summary of results for the first simulation study, with LRT score random at levels 1 and 2. Bias results in (a) are relative except those in brackets, which are absolute (the true value in those cases is zero). Monte Carlo standard errors (SEs) in (a) are given in parentheses. The Monte Carlo SEs for the estimated interval coverages in (b) range from 0.7% to 1.0%

(a) *Relative bias of point estimates* (%)

| Parameter {True Value} | IGLS | RIGLS | MCMC method 1 | MCMC method 2 |
|---|---|---|---|---|
| $\beta_0$ {0.0} | [ − 0.00 (0.001)] | [ − 0.00 (0.001)] | [ − 0.00 (0.001)] | [ − 0.00 (0.001)] |
| $\beta_1$ {0.5} | − 0.30 (0.14) | − 0.30 (0.14) | − 0.30 (0.14) | − 0.30 (0.14) |
| $\Omega_{u,00}$ {0.1} | − 1.81 (0.59) | − 0.08 (0.60) | 2.79 (0.62) | 2.79 (0.62) |
| $\Omega_{u,01}$ {0.02} | − 2.27 (1.23) | − 0.73 (1.25) | 3.01 (1.29) | 2.98 (1.30) |
| $\Omega_{u,11}$ {0.02} | − 3.00 (0.96) | − 0.47 (0.92) | − 1.75 (0.97) | − 1.79 (0.97) |
| $\Omega_{e,00}$ {0.5} | − 0.10 (0.09) | − 0.10 (0.09) | 0.01 (0.09) | − 0.01 (0.09) |
| $\Omega_{e,01}$ {−0.02} | − 0.61 (1.15) | − 0.61 (1.15) | − 1.53 (1.16) | − 1.53 (1.16) |
| $\Omega_{e,11}$ {0.05} | 0.64 (0.70) | 0.65 (0.70) | 4.63 (0.71) | 4.77 (0.70) |

(b) *Interval coverage probabilities at nominal levels* 90%/95%

| Parameter | IGLS | RIGLS | MCMC method 1 | MCMC method 2 |
|---|---|---|---|---|
| $\beta_0$ | 89.3/94.4 | 89.5/94.5 | 89.7/95.3 | 89.7/95.2 |
| $\beta_1$ | 87.4/94.4 | 87.6/94.7 | 87.6/94.6 | 87.7/94.5 |
| $\Omega_{u,00}$ | 89.4/93.1 | 90.7/93.6 | 91.1/96.0 | 91.1/96.0 |
| $\Omega_{u,01}$ | 90.0/94.4 | 90.3/94.6 | 88.7/94.1 | 88.8/94.1 |
| $\Omega_{u,11}$ | 86.9/91.0 | 87.6/92.4 | 85.7/91.6 | 85.8/92.1 |
| $\Omega_{e,00}$ | 90.7/94.1 | 90.7/94.1 | 90.2/94.1 | 90.9/94.8 |
| $\Omega_{e,01}$ | 90.2/95.0 | 90.2/95.0 | 89.8/95.1 | 90.5/94.9 |
| $\Omega_{e,11}$ | 90.6/95.1 | 90.7/95.1 | 90.0/95.0 | 90.9/95.4 |

(c) *Mean interval widths at nominal levels* 90%/95%

| Parameter | IGLS | RIGLS | MCMC method 1 | MCMC method 2 |
|---|---|---|---|---|
| $\beta_0$ | 0.135/0.161 | 0.136/0.162 | 0.138/0.165 | 0.138/0.165 |
| $\beta_1$ | 0.073/0.087 | 0.074/0.088 | 0.073/0.088 | 0.073/0.088 |
| $\Omega_{u,00}$ | 0.063/0.075 | 0.064/0.077 | 0.067/0.081 | 0.067/0.081 |
| $\Omega_{u,01}$ | 0.025/0.030 | 0.026/0.031 | 0.026/0.032 | 0.026/0.032 |
| $\Omega_{u,11}$ | 0.018/0.022 | 0.019/0.022 | 0.018/0.022 | 0.018/0.022 |
| $\Omega_{e,00}$ | 0.048/0.057 | 0.048/0.057 | 0.048/0.057 | 0.048/0.057 |
| $\Omega_{e,01}$ | 0.024/0.028 | 0.024/0.028 | 0.024/0.029 | 0.024/0.029 |
| $\Omega_{e,11}$ | 0.037/0.044 | 0.037/0.044 | 0.037/0.044 | 0.038/0.045 |

and a burn-in of 500 from IGLS starting values. Uniform priors were used for the level-1 variances and fixed effects. A (slightly) informative inverse-Wishart prior was used for the level-2 covariance matrix in line with the results in Browne and Draper (2000). Interval estimates at nominal level $100(1-\alpha)\%$ with the IGLS and RIGLS approaches were of the form $\hat{\theta}\pm\Phi^{-1}(1-\alpha/2)\widehat{SE}(\hat{\theta})$ based on the large-sample normal

Table 7
Summary of results for the second simulation study, with separate variances at level 1 for boys and girls. Monte Carlo standard errors (SEs) in (a) are given in parentheses. The Monte Carlo SEs for the estimated interval coverages in (b) range from 0.7% to 1.0%

(a) *Relative bias of point estimates* (%)

| Parameter {True Value} | IGLS | RIGLS | MCMC method 2 |
|---|---|---|---|
| $\beta_0$ {−0.15} | −0.45 (1.31) | −0.45 (1.31) | −0.46 (1.31) |
| $\beta_1$ {0.25} | −0.83 (0.53) | −0.83 (0.53) | −0.83 (0.53) |
| $\Omega_{u,00}$ {0.2} | −1.38 (0.59) | 0.41 (0.60) | 3.71 (0.62) |
| $\Omega_{e,00}$ {0.9} | −0.19 (0.11) | −0.16 (0.11) | 0.09 (0.11) |
| $\Omega_{e,01}$ {−0.05} | 1.42 (1.24) | 1.29 (1.24) | 0.32 (1.24) |

(b) *Interval coverage probabilities at nominal levels* 90%/95%

| Parameter | IGLS | RIGLS | MCMC method 2 |
|---|---|---|---|
| $\beta_0$ | 90.5/94.7 | 90.7/94.9 | 90.8/95.1 |
| $\beta_1$ | 89.1/94.3 | 89.4/94.3 | 89.7/94.3 |
| $\Omega_{u,00}$ | 88.3/92.5 | 89.2/93.4 | 90.0/95.3 |
| $\Omega_{e,00}$ | 89.2/94.8 | 89.3/94.8 | 89.6/95.0 |
| $\Omega_{e,01}$ | 90.3/95.7 | 90.4/95.7 | 90.4/95.1 |

(c) *Mean interval widths at nominal levels* 90%/95%

| Parameter | IGLS | RIGLS | MCMC method 2 |
|---|---|---|---|
| $\beta_0$ | 0.204/0.243 | 0.206/0.245 | 0.208/0.249 |
| $\beta_1$ | 0.134/0.159 | 0.134/0.159 | 0.134/0.160 |
| $\Omega_{u,00}$ | 0.124/0.147 | 0.126/0.150 | 0.133/0.160 |
| $\Omega_{e,00}$ | 0.105/0.125 | 0.105/0.125 | 0.105/0.125 |
| $\Omega_{e,01}$ | 0.065/0.077 | 0.065/0.077 | 0.065/0.078 |

approximation; this is what users of most multilevel packages such as `MLwiN` (Rasbash et al., 2000) and `HLM` (Bryk et al., 1988) would report, if they provide interval estimates at all (such packages routinely report only point estimates and estimated asymptotic standard errors with maximum-likelihood methods). With the Bayesian MCMC methods we give results based on posterior means as point estimates and 90%/95% central posterior intervals.

Our second simulation study (Table 7) was based on model (18) from Section 4.3, in which the male and female subsamples had different level-1 variances (MCMC method 1 is not available for this model). We again created 1000 simulation datasets with population values similar to the estimates obtained with the GCSE dataset. With the Bayesian approach to fitting, uniform priors were used for the level-1 variances and fixed effects, and a $\Gamma(\varepsilon,\varepsilon)$ prior (with $\varepsilon = 0.001$; this distribution has mean 1 and variance $\varepsilon^{-1}$) was used for the reciprocal of the level-2 variance parameter $\Omega_{u,00}$, in line with the results in Browne and Draper (2001).

It is evident from Tables 6 and 7 that all four methods performed reasonably well in both models. RIGLS succeeded in reducing the (already small) biases arising from IGLS estimation in most cases, and the relative biases of the MCMC methods are also small (ranging from 0% to 4.8%, with a median absolute value of 1.5%). Interval coverages for all four methods were all close to nominal, with actual coverages ranging from 86–91% and 91–96% at nominal 90% and 95%, respectively; all four methods achieved this level of coverage with intervals of comparable length; and the ratios of 95% and 90% interval lengths for each method were all close to the value $(\Phi^{-1}(0.975)/\Phi^{-1}(0.95))$ to be expected under normality. The ML methods have the clear advantage of speed (on the original GCSE data set IGLS/RIGLS and MCMC methods 1 and 2 took 2, 168, and 248 s on a 500 MHz Pentium PC, respectively, with the MCMC methods based on 10,000 monitoring iterations), but the ML approach has two potential disadvantages: on data sets with small numbers of level-1 and level-2 units, it requires more sophisticated methods for constructing interval estimates for variance parameters (to achieve good coverage properties) than the large-sample normal approximation used here (Browne and Draper, 2001), and it may fail to converge when the $\Omega_e$ and/or $\Omega_u$ matrices exhibit a high degree of correlation between the parameters quantifying the random effects. The Bayesian methods are considerably slower but have the additional advantage that inferences about arbitrary functions of the model parameters are automatic once the model parameters themselves have been monitored.

## 6. Other MCMC methods

### 6.1. Gibbs sampling

There are special cases of the problem of complex level-1 variation that can be fitted using a standard Gibbs sampler. The model (Eq. (18)) used in the second simulation, where we use a different level-1 variance term for each gender, is one such example. Here we could reparameterise the model with two variances, one for boys ($\sigma_b^2$) and one for girls ($\sigma_g^2$), rather than a boys' variance plus a difference. Scaled-inverse-$\chi^2$ priors (see, e.g., Gelman et al., 1995a) can be used for these two variances, with parameters $(\nu_b, s_b^2)$ and $(\nu_g, s_g^2)$, respectively. If we divide the children into boys' and girls' subgroups $B$ and $G$, of size $n_b$ and $n_g$, then step 3 of the algorithm given in Appendix $A$ can be rewritten as two Gibbs sampling steps as follows: the full conditional for $\sigma_b^2$ is

$$p(\sigma_b^2 \mid y, \beta, u, \Omega_u) \sim \Gamma^{-1}(a_b, b_b),$$

where

$$a_b = \frac{n_b + \nu_b}{2} \quad \text{and} \quad b_b = \frac{1}{2}\left(\nu_b s_b^2 + \sum_{i,j \in B} e_{ij}^2\right) \tag{19}$$

and the full conditional for $\sigma_g^2$ is exactly analogous. Now the level-1 variance is $\Sigma_{e,ij} = \sigma_b^2 I(i,j \in B) + \sigma_g^2 I(i,j \in G)$ and the other steps of the algorithm are as before.

### 6.2. Modelling the variance on the log scale

The developers of the software package BUGS (Spiegelhalter et al., 1997) use a different approach to fitting complex level-1 variation in one of their examples, the Schools data set (example 9 in Vol. 2 of Spiegelhalter et al., 1996). They model the logarithm of the level-1 precision as a function of predictors and other parameters:

$$\begin{aligned} &y_{ij} = X_{ij}\beta + Z_{ij}u_j + e_{ij}, \\ &e_{ij} \sim N(0, \tau_{ij}^{-1}), \quad u_j \sim N_{p_2}(0, \Omega_u), \\ &\log(\tau_{ij}) = (X_{ij}^{C})^{T}\Omega_e X_{ij}^{C}. \end{aligned} \tag{20}$$

This results in a multiplicative, rather than an additive, variance function:

$$\begin{aligned} \Sigma_{e,ij} = \tau_{ij}^{-1} &= \exp[-(X_{ij}^{C})^{T}\Omega_e X_{ij}^{C}] \\ &= \exp[-(X_{ij(1)}^{C})^{T}\Omega_{e,11} X_{ij(1)}^{C}] \cdots \exp[-(X_{ij(n)}^{C})^{T}\Omega_{e,nn} X_{ij(n)}^{C}]. \end{aligned} \tag{21}$$

The advantages of this approach are that the parameters are now unconstrained, the level-1 variance will never be negative, and it is easier to specify a prior with this method. The disadvantages are that the interpretation of the individual coefficients is not as easy and computation for these models is slower. The interpretation difficulty will be apparent mainly when the $X$ variables are categorical.

Model (20) can be fitted in BUGS using adaptive rejection (AR) sampling (Gilks and Wild, 1992). Alternatively the adaptive MH method used in the truncated normal algorithm in Section 4.1 can be used, this time with no parameter constraints and hence no truncation in the normal proposal distributions. Goldstein (1995, Appendix 5.1) shows how to obtain ML estimates for this model; see Yang et al. (2000) for a set of MLwiN macros to do this.

To explore the differences between log-variance modelling and our earlier approach, we fitted four different level-1 variance functions to the GCSE dataset to model the effect of LRT score ($X_1$) on the level-1 variance. We considered the quadratic relationship examined earlier (model (4)), and the simpler linear relationship

$$\Sigma_{e,ij} = \Omega_{e,00} + 2X_{1,ij}\Omega_{e,01}; \tag{22}$$

we also considered two exponential relationships

$$\begin{aligned} &\Sigma_{e,ij} = \exp(-\Omega_{e,00} - 2X_{1,ij}\Omega_{e,01}), \\ &\Sigma_{e,ij} = \exp(-\Omega_{e,00} - 2X_{1,ij}\Omega_{e,01} - X_{1,ij}^2\Omega_{e,11}). \end{aligned} \tag{23}$$

In each of the four models the level-2 variance structure and fixed effects were as in Eq. (4). Fig. 3 plots the resulting level-1 estimated variances as a function of LRT

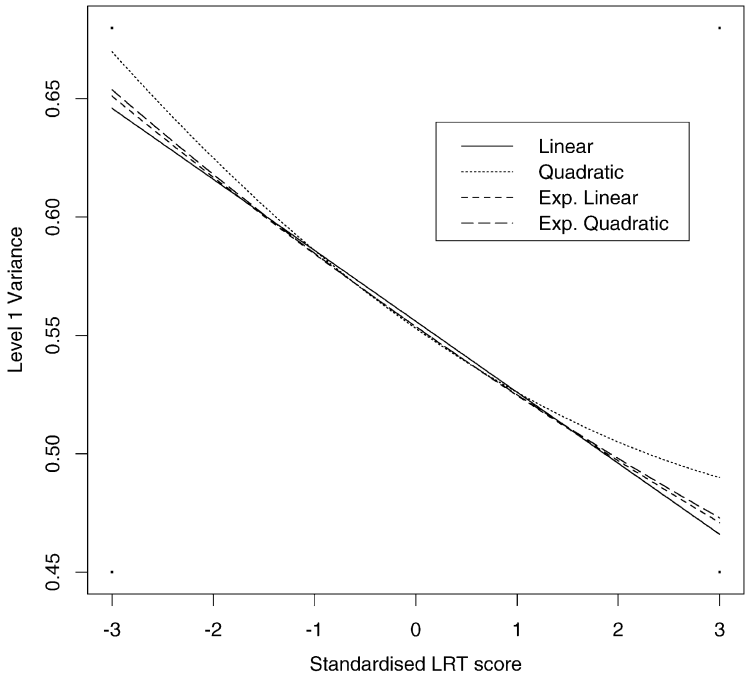


Fig. 3. Four ways to model the effect of standardised LRT score on the level-1 variance in the GCSE dataset.

score. In this case for the majority of the data the variance estimates produced by all four models are fairly similar, with any discrepancies between the models occurring at the extremes of the LRT range where there are relatively few observations.

Table 8 presents estimates of the four models (using MH method 2), together with Raftery and Lewis (1992) default diagnostics (for both MH and AR sampling in the exponential models (23)) and comparative timings. From part (b) of the table it is evident that the parameter with the worst MCMC mixing for both methods is the intercept $\beta_0$. This means that, although the MH method requires longer monitoring runs than the AR approach for the level-1 variance parameters, the run lengths required to ensure that all parameter estimates have a specified accuracy (with respect to 95% interval estimation) will be roughly equal (since the length of monitoring run chosen will be determined by the largest of the Raftery–Lewis estimates). From part (c) of the table it can be seen that the MH approach is 4–9 times faster in real-time execution speed in this example. Results in Table 8 are based on a single data set but are typical of findings we have obtained with other similar models.

## 7. Conclusions and extensions

In this paper we have presented several methods for modelling non-constant level-1 variance functions with multilevel data. We have introduced two new adaptive

Table 8
Parameter estimates for four different level-1 variance functions applied to the GCSE dataset and fit by MCMC. All methods used a monitoring period of 50,000 iterations after a burn-in of 500 iterations from maximum-likelihood starting values. Methods with an adaptive MH step at level 1 were run using a development version of `MLwiN`; those with an adaptive rejection (AR) step at level 1 were run in `WinBUGS`. Posterior standard deviations are given in parentheses in (a)[a]

(a) *Parameter estimates*

| Parameter | Linear | Quadratic | Exponential Linear | Exponential Quadratic |
|---|---|---|---|---|
| $\beta_0$ | −0.011 (0.041) | −0.011 (0.041) | −0.011 (0.041) | −0.012 (0.041) |
| $\beta_1$ | 0.558 (0.020) | 0.559 (0.020) | 0.558 (0.020) | 0.559 (0.020) |
| $\Omega_{u,00}$ | 0.095 (0.020) | 0.095 (0.020) | 0.095 (0.020) | 0.095 (0.020) |
| $\Omega_{u,01}$ | 0.020 (0.007) | 0.020 (0.007) | 0.020 (0.007) | 0.020 (0.007) |
| $\Omega_{u,11}$ | 0.014 (0.005) | 0.014 (0.005) | 0.014 (0.005) | 0.014 (0.005) |
| $\Omega_{e,00}$ | 0.556 (0.012) | 0.553 (0.015) | 0.591 (0.023) | 0.591 (0.027) |
| $\Omega_{e,01}$ | −0.015 (0.006) | −0.015 (0.007) | 0.027 (0.012) | 0.027 (0.012) |
| $\Omega_{e,11}$ | — | 0.003 (0.009) | — | −0.0005 (0.016) |

(b) *Raftery–Lewis values* (*in thousands of iterations*); *main entries apply to MH method 2, with the corresponding values for AR in parentheses*

| $\hat{N}$ | Linear | Quadratic | Exponential Linear | Exponential Quadratic |
|---|---|---|---|---|
| $\beta_0$ | 16.3 | 16.0 | 16.8 (17.1) | 17.5 (17.7) |
| $\beta_1$ | 7.4 | 6.9 | 7.4 (7.5) | 7.4 (7.0) |
| $\Omega_{u,00}$ | 4.3 | 4.3 | 4.4 (4.3) | 4.3 (4.1) |
| $\Omega_{u,01}$ | 5.9 | 5.4 | 5.4 (5.6) | 5.7 (5.7) |
| $\Omega_{u,11}$ | 9.6 | 10.3 | 9.8 (9.2) | 9.4 (9.5) |
| $\Omega_{e,00}$ | 14.4 | 16.0 | 14.8 (3.7) | 16.2 (4.7) |
| $\Omega_{e,01}$ | 14.4 | 14.8 | 13.7 (3.8) | 14.8 (3.9) |
| $\Omega_{e,11}$ | — | 16.9 | — | 15.6 (4.6) |

(c) *Timings* (*in minutes at 500 Pentium MHz*)

| Method | Linear | Quadratic | Exponential Linear | Exponential Quadratic |
|---|---|---|---|---|
| Metropolis-Hastings | 19 | 20 | 23 | 25 |
| Adaptive Rejection | — | — | 95 | 227 |

[a]Note: The Raftery–Lewis values in (b) estimate the lengths of monitoring runs necessary to ensure that the actual coverages of the nominal 95% central posterior intervals for the given parameters are in the range 94–96% with Monte Carlo probability at least 95%.

Metropolis-Hastings sampling methods for fitting such functions subject to different constraints. The two methods give similar estimates for models where the true parameter values are not affected by the constraints, but if the true values do not satisfy the additional positive-definite matrix constraint of the inverse-Wishart proposal method then the estimates from the two methods will differ.

The main advantage of the inverse-Wishart method is that it models the level-1 variance function as a matrix in a manner analogous to the usual treatment of the level-2 variance function, meaning (among other things) that informative inverse-Wishart priors at level 1 can be used with this approach. The main advantage of the truncated normal proposal method is that it is more general and can deal with any variance function at level 1. Both methods have bias and interval coverage properties that are similar to those from the maximum-likelihood (IGLS and RIGLS) approaches, and all four methods perform satisfactorily in repeated sampling in this regard.

In Section 6.2 we considered an alternative formulation of the level-1 variance function in terms of the log of the precision at level 1. This method has two advantages: there is no need to impose constraints on the terms in the resulting variance function, and it is therefore easier to contemplate a variety of prior distributions for the resulting variance parameters. The main disadvantage of this approach is that the individual terms in the variance function may not be as easily interpreted, making it potentially difficult to construct sensible informative priors. Table 8 shows clearly, however, that adaptive-rejection sampling is much less efficient than adaptive Metropolis-Hastings sampling to achieve default MCMC accuracy standards with variance (or precision) functions that are exponential in the parameters.

Our examination of multiple models (e.g., models (2–5) for the GCSE data in Section 2, and the alternative formulations of the variance function in Section 6) brings up the topic of model choice with heteroscedastic multilevel data. We have found two approaches useful: (i) gradually expanding simple models such as (2) in directions suggested by deficiencies uncovered from residual plots, monitoring the significance of the new parameters which index the model expansions (for instance, reading the columns of Table 2 from left to right) and stopping when new model expansion parameters are no longer significant; and (ii) employing predictive diagnostics, in which (a) the data are divided into non-overlapping modeling and validation subsamples $M$ and $V$ in a way that respects the multilevel structure, (b) each of the models under scrutiny is fit to $M$, yielding predictive distributions for all observations in $V$ from each model, and (c) the actual data values in $V$ are compared with their predictive distributions under each model using, e.g., a log scoring rule as in Gelfand and Ghosh (1998) and Draper (2002).

There are two obvious extensions of this work, to arbitrary variance structures at higher levels and to multivariate normal responses. Two approaches to fitting random effects at level 2 and above appear common in current applied work: modelling all random effects independently, or fitting fully dependent random effects with a complete covariance matrix at each level (see the Birats example in Spiegelhalter et al. (1996) for an illustration of both formulations). It is fairly easy to fit any block-diagonal covariance structure at a higher level using Gibbs sampling, in a straightforward extension of the approach given in Section 6.1. The adaptive MH sampler with a truncated normal proposal (method 2, Section 4.1) can be used to fit any dependence structure among the random effects at the higher levels, including non block-diagonal covariance matrices.

With multivariate normal response models the variance function at the lowest level includes variances for each response plus covariances between responses. This variance function could also be extended to include predictors that may influence the variance of individual responses in an analogous way to the univariate model. We intend to report on MCMC sampling algorithms for general multivariate-response multilevel models elsewhere.

## Acknowledgements

We are grateful to the EPSRC, ESRC, and European Commission for financial support, and to David Spiegelhalter, Nicky Best, and other participants in the BUGS project for references and comments on the set of multilevel modelling papers based on the first author's Ph.D. dissertation. 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 A. Details of MCMC method 1

- In step 1 of the algorithm described in Section 3.1, the full conditional distribution in the Gibbs update for the fixed effects parameter vector $\beta$ is multivariate normal: with $p_f$ as the number of fixed effects

$$p(\beta \mid y, u, \Omega_u, \Omega_e) \sim N_{p_f}(\hat{\beta}, \hat{D}),$$

where

$$\hat{\beta} = \hat{D}\left[\sum_{ij} \frac{X_{ij}^{\mathrm{T}}(y_{ij} - Z_{ij}u_j)}{\Sigma_{e,ij}} + S_p^{-1}\mu_p\right]$$

and

$$\hat{D} = \left[\sum_{ij} \frac{X_{ij}^{\mathrm{T}}X_{ij}}{\Sigma_{e,ij}} + S_p^{-1}\right]^{-1}. \tag{24}$$

- Step 2 involves a Gibbs update of the level-2 residuals, $u_j$, also with a multivariate normal full conditional distribution: with $p_2$ the number of parameters describing the random effects at level-2 and $n_j^*$ the number of level-1 units in level-2 unit $j$

$$p(u_j \mid y, \beta, \Omega_u, \Omega_e) \sim N_{p_2}(\hat{u}_j, \hat{D}_j),$$

where

$$\hat{u}_j = \hat{D}_j \sum_{i=1}^{n_j^*} \frac{Z_{ij}^{\mathrm{T}}(y_{ij} - X_{ij}\beta)}{\Sigma_{e,ij}}$$

and

$$\hat{D}_j = \left[ \sum_{i=1}^{n_j^*} \frac{Z_{ij}^{\mathrm{T}} Z_{ij}}{\Sigma_{e,ij}} + \Omega_u^{-1} \right]^{-1}. \tag{25}$$

- Step 3 employs a Hastings update using an inverse-Wishart proposal distribution for the level-1 covariance matrix $\Omega_e$. Specifically, the Markov chain moves from $\Omega_e^{(t-1)}$ at time $(t-1)$ to $\Omega_e^{(t)}$ as follows:

$$\Omega_e^{(t)} = \begin{cases} \Omega_e^* & \text{with probability min} \left[ 1, R \frac{p(\Omega_e^* \mid y,\beta,u,\Omega_u)}{p(\Omega_e^{(t-1)} \mid y,\beta,u,\Omega_u)} \right] \\ \Omega_e^{(t-1)} & \text{otherwise} \end{cases}. \tag{26}$$

Here (a) $\Omega_e^* \sim W_{p_1}^{-1}(w + p_1 + 1, w\,\Omega_e^{(t)})$, where $w$ is chosen as in Section 3.2 and $p_1$ is the number of rows or columns in $\Omega_e$; (b) the Hastings ratio $R$ in (26) is

$$R = \left( \frac{|\Omega_e^*|}{|\Omega_e^{(t-1)}|} \right)^{\alpha} \exp\left( \frac{w}{2} \{ \mathrm{tr}[\Omega_e^{(t-1)} (\Omega_e^*)^{-1}] - \mathrm{tr}[\Omega_e^* (\Omega_e^{(t-1)})^{-1}] \} \right), \tag{27}$$

where $\alpha = (2w + 3p_1 + 3)/2$; and (c) the full conditional distribution for $\Omega_e$ in (26) is

$$p(\Omega_e \mid y, \beta, u, \Omega_u) \propto \prod_{ij} \left\{ \Sigma_{e,ij}^{-1/2} \exp\left[ -\frac{1}{2\Sigma_{e,ij}} (y_{ij} - X_{ij}\beta - Z_{ij}u_j)^2 \right] \right\}, \tag{28}$$

where we have expressed the right-hand side of (28) for convenience in terms of $\Sigma_{e,ij}$ as in Eq. (7).

Finally, step 4 involves a Gibbs update of the level-2 covariance matrix $\Omega_u$: expressed as a Wishart draw of $\Omega_u^{-1}$ the full conditional is

$$p(\Omega_u^{-1} \mid y, u, \beta, \Omega_e) \sim W_{p_2} \left[ J + \nu_2, \left( \sum_{j=1}^{J} u_j (u_j)^{\mathrm{T}} + S_2 \right)^{-1} \right], \tag{29}$$

where $p_2$ is the number of rows or columns in $\Omega_u$ and $J$ is the number of level-2 units in the data set. An improper uniform prior on $\Omega_u$ corresponds to the choice $(\nu_2, S_2) = (-p_2 - 1, 0)$.

## Appendix B. Computing details

`MLwiN` is a package for performing maximum-likelihood and Bayesian calculations in a wide variety of hierarchical and other multilevel models, developed by the *Multilevel Models Project* at the University of London. It is available for a nominal charge from the developers' web site, `http://multilevel.ioe.ac.uk/`. Sample code for the models in this paper is not easy to distribute, because `MLwiN` uses a point-and-click graphical user interface for constructing its models, but examples like the ones examined here may be found in the `MLwiN` user's guide (Rasbash et al., 2000). `HLM5` is a package which performs maximum-likelihood and

generalised-estimating-equations calculations in a large variety of hierarchical models. It is available for a nominal fee at www.ssicentral.com/hlm/hlm.htm.

## References

Best, N.G., Cowles, M.K., Vines, S.K., 1995. CODA Manual version 0.30. MRC Biostatistics Unit, Cambridge, UK.

Browne, W.J., 1998. Applying MCMC Methods to Multilevel Models. Ph.D. dissertation, Department of Mathematical Sciences, University of Bath, UK.

Browne, W.J., Draper, D., 2000. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Comput. Statist. 15, 391–420.

Browne, W.J., Draper, D., 2001. A comparison of Bayesian and likelihood methods for fitting multilevel models. Under review.

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

Bryk, A.S., Raudenbush, S.W., Seltzer, M., Congdon, R., 1988. An Introduction to HLM: Computer Program and User's Guide, 2nd Edition. University of Chicago, Department of Education, Chicago.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Royal Statist. Soc. Ser. B 39, 1–88.

Draper, D., 2002. Bayesian Hierarchical Modeling. Springer, New York, forthcoming.

Gelfand, A.E., Ghosh, S.K., 1998. Model choice: a minimum posterior predictive loss approach. Biometrika 85, 1–11.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995a. Bayesian Data Analysis. Chapman & Hall, London.

Gelman, A., Roberts, G.O., Gilks, W.R., 1995b. Efficient Metropolis jumping rules. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 5. Oxford University Press, Oxford, pp. 599–607.

Gilks, W.R., Wild, P., 1992. Adaptive rejection sampling for Gibbs sampling. J. Royal Statist. Soc. Ser. C 41, 337–348.

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

Goldstein, H., 1989. Restricted unbiased iterative generalised least squares estimation. Biometrika 76, 622–623.

Goldstein, H., 1995. Multilevel Statistical Models, 2nd Edition. Edward Arnold, London.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., 1993. A multilevel analysis of school examination results. Oxford Rev. Education 19, 425–433.

Raftery, A.E., Lewis, S.M., 1992. How many iterations in the Gibbs sampler?. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4. Oxford University Press, Oxford, pp. 763–773.

Rasbash, J., Browne, W.J., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., Lewis, T., 2000. A User's Guide to MLwiN (Version 2.1). Institute of Education, University of London, London.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Gilks, W.R., 1996. BUGS 0.5 Examples (Version ii). Medical Research Council Biostatistics Unit, Cambridge.

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

Weisberg, S., 1985. Applied Linear Regression, 2nd Edition. Wiley, New York.

Yang, M., Rasbash, J., Goldstein, H., Barbosa, M., 2000. MLwiN Macros for Advanced Multilevel Modelling (Version 2.0). Institute of Education, University of London, London.