MCMC Sampling for a Multilevel Model With Nonindependent Residuals Within and Between Cluster Units

William Browne and Harvey Goldstein

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2010 35: 453
DOI: 10.3102/1076998609359788

Journal of Educational and Behavioral Statistics
August 2010, Vol. 35, No. 4, pp. 453–473
DOI: 10.3102/1076998609359788

MCMC Sampling for a Multilevel Model With Nonindependent Residuals Within and Between Cluster Units

William Browne
Harvey Goldstein
University of Bristol

In this article, we discuss the effect of removing the independence assumptions between the residuals in two-level random effect models. We first consider removing the independence between the Level 2 residuals and instead assume that the vector of all residuals at the cluster level follows a general multivariate normal distribution. We demonstrate how this assumption can allow us to fit higher levels of clustering and school competition effects via an example from education. We then consider removing the assumption of independence between Level 1 residuals within clusters. We show how this extension can allow time series type models. Both normal and binary responses are considered.

Keywords: *correlated residuals; education; MCMC; multilevel models; time series*

1. Introduction

Multilevel models have been in use now for several decades and allow statistical modeling of response variables that involve some dependence due to clustering. Here cluster membership influences the response and hence the responses of two units in the same cluster are generally more alike than responses from two randomly chosen units. For example, in education, pupils from the same school might be expected to have more correlated responses than a randomly selected sample of pupils from across all schools.

In this article, we will consider only two-level random intercept models, that is models that exhibit one level of clustering and where the correlation induced by this clustering can be expressed by a single term. We shall show, however, that certain three-level models can also be described within our general framework and we will describe extensions to further models in the discussion. To be precise, all our models can be written in the Form 1:

$$\begin{aligned} &y_{ij} = \mathbf{X_{ij}}\beta + u_j + e_{ij} \\ &E(u_j) = 0, \ \ \mathrm{var}(u_j) = \sigma_u^2, \ \ E(e_{ij}) = 0, \ \ \mathrm{var}(e_{ij}) = \sigma_e^2 \\ &i = 1, ..., n_j, j = 1, ..., J. \end{aligned} \tag{1}$$

Here y_{ij} is the response of the ith Level 1 unit within the jth cluster, $\mathbf{X}_{ij}$ is a vector of p predictor variables with associated fixed effects β, u_j are cluster-specific random effects, and e_{ij} are unit-specific residual terms. In the standard modeling framework, we would also assume that the u_j and e_{ij} are independent and identically (normally) distributed, that is,

$$u_j \sim N(0, \sigma_u^2), \quad e_{ij} \sim N(0, \sigma_e^2).$$

In this article, we consider the effect of changing these assumptions by first removing the assumption of independence of the Level 2 residuals u_j and second by removing the assumption of independence between the Level 1 residuals e_{ij} within the same cluster. We will consider these two changes independently in the next two sections and then show some illustrative examples and finish with some discussion and extensions to nonnormal responses.

2. Nonindependence of the Level 2 Residuals

In this section, we will consider removing the assumption of independence between the Level 2 random effects associated with different clusters. We may for example believe that some pairs of clusters are more similar to each other than to other clusters.

Such a belief is what often drives spatial modeling where the relative locations of clusters of data are expected to influence the correlation between them. Typically, this spatial correlation would be represented either as a function of distance between cluster centroids or by creating a lattice structure with neighbouring clusters linked. This second form is often used in conditional autoregressive (CAR) modeling (Besag & Kooperberg, 1995; Besag, York, & Mollie, 1991), where individual cluster effects are dependent on their neighboring clusters effect, and the approach is different from what we describe below. CAR models are used extensively in epidemiology for disease mapping (Clayton & Kaldor, 1987). In fact, the normal CAR model can generally be represented in terms of a (nonconditional) multivariate normal distribution with a corresponding structure for the cluster correlations. These correlations, however, as in the first form, will depend on the single set of weights used to specify the spatial proximity matrix in the CAR model, and for our proposed model, we show how a correlation structure can be formulated that is allowed to depend on more than one set of distance metrics. Our approach also models the correlation as an explicit, rather than implicit, function of distance. In addition, we extend our approach to modeling correlations among the Level 1 random effects, with particular applications to time series. We will therefore not deal with these earlier models further.

Another way in which nonindependence occurs is when clusters themselves can be clustered into further higher level clusters; for example, in education, pupils may be clustered into classes that themselves are clustered into schools.

This will often be fitted as an additional level of random effects but, as we show later, the same clustering can be accounted for in a two-level modeling framework by removing the independence assumptions.

We begin by describing how the independence assumption is removed from the standard two-level model. Consider the Model 1 described in the introduction and then let $\mathbf{u} = (u_1, u_2, ..., u_J)^T$, that is all the (independent) residuals at Level 2 stacked as a vector of length J. A more general model that allows dependence among these residuals will then have $\mathbf{u} \sim \text{MVN}(0, \mathbf{\Omega_u})$ with the earlier independence assumption as a special case where $\mathbf{\Omega_u}$ is diagonal with σ_u^2 on the diagonal. We now have flexibility in how we parameterize the covariance matrix $\mathbf{\Omega_u}$. We could consider an unconstrained representation and estimate all parameters in the covariance matrix $\mathbf{\Omega_u}$ but this will result in $J \times (J+1)/2$ parameters.

There may also be identifiability issues, for example, if we assume we have just a random intercept for each cluster and a different variance for each cluster intercept, then we do not have the data to identify these different variances without making some further assumptions about the variances, for example, by expressing an informative prior for the variances or by modeling the individual variances as functions of cluster level covariates. In practice, however, analysts will often assume a common variance for each cluster intercept, as we do in this article, and then focus on modeling the correlations between the cluster intercepts. We shall return to this issue in the Discussion.

We thus write $\mathbf{\Omega_u} = \sigma_u^2 \mathbf{D_u}$ where $\mathbf{D_u}$ is the correlation matrix of the u's and σ_u^2 is the common variance term. Let us also write $\rho_{j_1 j_2}$ to represent the correlation between the intercepts for clusters j_1 and j_2. We will then model the corresponding correlations using a functional form $f^{-1}(j_1, j_2, \alpha)$, which involves a set of distance measures for the Level 2 units j_1 and j_2 and a set of parameters α. We shall assume here a generalized linear function of the form:

$$f(\rho_{j_1 j_2}) = \alpha_1 g_1(j_1, j_2) + \alpha_2 g_2(j_1, j_2) + \cdots \tag{2}$$

We can choose the inverse hyperbolic function

$$\begin{aligned} f_{j_1 j_2} &= f(\rho_{j_1 j_2}) = 2 \tan \text{h}^{-1}(\rho_{j_1 j_2}) = \log\left(\tfrac{1+\rho_{j_1 j_2}}{1-\rho_{j_1 j_2}}\right), \\ \rho_{j_1 j_2} &= (e^{f_{j_1 j_2}} - 1)/(e^{f_{j_1 j_2}} + 1) \end{aligned}$$

which is effectively the Fisher z transformation for a correlation coefficient and where the g_h, $h = 1 \ldots p$ are known. This function ensures that each correlation lies in the interval $(-1, 1)$ (although of itself this does not guarantee that the covariance matrix is positive definite).

Where we have independence between two random effects, these functions can be given values of zero. This can be achieved by introducing an indicator vector $\delta_{j_1 j_2}$ to produce a final correlation structure defined by $\delta_{j_1 j_2} \rho_{j_1 j_2}$.

An alternative link function, if we wish to restrict the correlations to be positive is the logit given by

$$\rho_{j_1 j_2} = e^{f_{j_1 j_2}} / (e^{f_{j_1 j_2}} + 1)$$

or the log link function given by

$$\rho_{j_1 j_2} = e^{f_{j_1 j_2}}.$$

In this case, the correlations are positive and f is restricted to be negative to ensure the resulting correlations are less than 1.

These three functions are illustrated in Figure 1. We now will describe a Markov chain Monte Carlo (MCMC) algorithm for these models.

3. MCMC Algorithm for a Two-Level Variance Components Model With Correlated Level 2 Residuals

We are interested in fitting the following model:

$$\begin{array}{l} y_{ij} = \mathbf{X_{ij}}\beta + u_j + e_{ij} \\ \mathbf{u} \sim \text{MVN}(0, \mathbf{\Omega_u}) \text{ where } \mathbf{u} = (u_1, u_2, ..., u_J)^T \\ E(e_{ij}) = 0, \;\; \text{var}(e_{ij}) = \sigma_e^2 \\ i = 1, ..., n_j, j = 1, ..., J. \end{array} \tag{3}$$

To fit this model in a Bayesian framework, we need to include prior distributions for the unknown parameters, in this case β, $\mathbf{\Omega_u}$, and σ_e^2. In situations where we do not have additional information about these parameters, we would like to use prior distributions that correspond to our lack of information and therefore we use "diffuse" priors. Here we use (improper) uniform priors for β and the commonly used $\Gamma^{-1}(10^{-3}, 10^{-3})$ prior for σ_e^2 (approximately equivalent to a uniform prior for the log of the variance). $\mathbf{\Omega_u}$ will be parameterized by a restricted set of parameters and we will describe the priors used for these parameters in the subsection that follows.

Apart from the different prior for $\mathbf{\Omega_u}$, this model is identical to a standard variance components model. This change of prior has no effect on the full conditional distributions for β and σ_e^2, which are multivariate normal and inverse gamma, respectively, and so these parameters will be updated using Gibbs sampling in the usual way.

We next consider sampling the Level 2 covariance matrix and the Level 2 random effects.

3.1. Sampling the Level 2 Covariance Matrix

The Level 2 covariance matrix $\mathbf{\Omega_u} = \sigma_u^2 \mathbf{D_u}$ and we are then parameterising $\mathbf{D_u}$ in terms of parameters α and functions $\mathbf{g}$ of the pairs of Level 2 units as detailed earlier. Thus, to specify a prior for $\mathbf{\Omega_u}$, we need simply specify priors for σ_u^2 and for α, and we will use the following shorthand $p(\mathbf{\Omega_u}) = p(\sigma_u^2)p(\alpha)$.

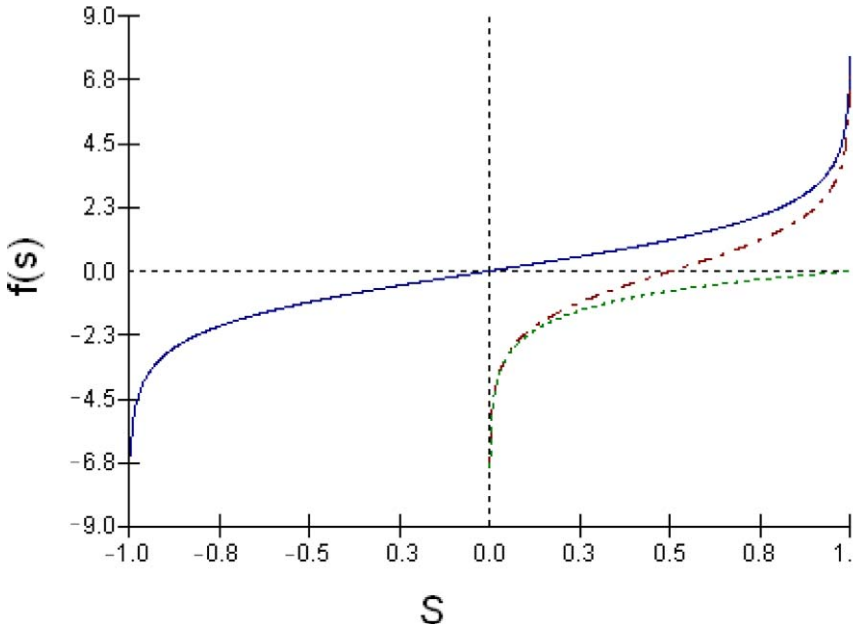

FIGURE 1. *Link function f(s). From left to right; hyperbolic, logit, log.*

In practice, we will use either an improper uniform or $\Gamma^{-1}(10^{-3}, 10^{-3})$ for σ_u^2 and uniform priors for α.

Updating the matrix $\mathbf{\Omega_u}$ then consists of two steps, first updating σ_u^2 and second α.

At iteration t, we generate $\sigma_u^2(*) \sim N(\sigma_u^2(t-1), \sigma_p^2)$ where σ_p^2 is a proposal distribution variance that has to be set. If we generate a negative $\sigma_u^2(*)$, then we set $\sigma_u^2(t) = \sigma_u^2(t-1)$, otherwise we form a proposed new matrix $\mathbf{\Omega_u^*}$ by calculating $\mathbf{\Omega_u^*} = \sigma_u^2(*)\mathbf{D_u(t-1)}$.

We then perform a Metropolis update step by setting $\sigma_u^2(t) = \sigma_u^2(*)$ with probability $\min[1, p(\mathbf{\Omega_u^*}|u)/\mathbf{p}(\mathbf{\Omega_u^{(t-1)}}|u)]$ and setting $\sigma_u^2(t) = \sigma_u^2(t-1)$ otherwise.

The components of the likelihood ratio are

$$p(\mathbf{\Omega_u^*}|\mathbf{u}) = |\mathbf{\Omega_u^*}|^{-1/2}\exp - (\mathbf{u^T}(\mathbf{\Omega_u^*})^{-1}\mathbf{u}/2)\mathbf{p}(\mathbf{\Omega_u^*})$$

$$\text{and } p(\mathbf{\Omega_u^{(t-1)}}|\mathbf{u}) = |\mathbf{\Omega_u^{(t-1)}}|^{-1/2}\exp - (\mathbf{u^T}(\mathbf{\Omega_u^{(t-1)}})^{-1}\mathbf{u}/2)p(\mathbf{\Omega_u^{(t-1)}})$$

with current estimates substituted.

For each element l of $\{\alpha\}$ at iteration t we then generate $\alpha_l^* \sim N(\alpha_l^{(t-1)}, \sigma_{\alpha,l}^2)$ where $\sigma_{\alpha,l}^2$ is a proposal distribution variance that has to be set for each element.

We then form a new correlation matrix $\mathbf{D}_{\mathbf{u}}^{*}$ by substituting the value α_l^{*} in place of $\alpha_l^{(t-1)}$ and check that the matrix formed is positive definite. If the matrix is not positive definite, we reject the proposal and set $\alpha_l^{(t)} = \alpha_l^{(t-1)}$ and proceed to the next element of the α vector.

Assuming $\mathbf{D}_{\mathbf{u}}^{*}$ is positive definite, we then form $\mathbf{\Omega}_{\mathbf{u}}^{*} = \sigma_u^2(t)\mathbf{D}_{\mathbf{u}}^{*}$ and again perform a Metropolis step by setting $\alpha_l^{(t)} = \alpha_l^{*}$ with probability $\min[1, p(\mathbf{\Omega}_{\mathbf{u}}^{*}|\mathbf{u})/\mathbf{p}(\mathbf{\Omega}_{\mathbf{u}}^{(t-1)}|\mathbf{u})]$ and setting $\alpha_l^{(t)} = \alpha_l^{(t-1)}$ otherwise, where the components of the likelihood ratio are as for the step updating σ_u^2.

This procedure is repeated for each of the elements of α in turn. The proposal distribution variances can be chosen by an adaptive sampling procedure (see below).

3.2. Sampling the Level 2 Residuals

The conditional posterior distribution for the Level 2 residuals $\mathbf{u}$ for a general two-level model is as follows:

$$p(\mathbf{u}|\mathbf{y}, \mathbf{\Omega}_{\mathbf{u}}, \sigma_e^2) \propto \left(\frac{1}{\sigma_e^2}\right)^{N/2} \exp[-\tfrac{1}{2\sigma_e^2}(\mathbf{y} - (\mathbf{X}\beta) - (\mathbf{Z}\mathbf{u}))^T(\mathbf{y} - (\mathbf{X}\beta) - (\mathbf{Z}\mathbf{u})) + \sigma_e^2\mathbf{u}^T\mathbf{\Omega}_{\mathbf{u}}^{-1}\mathbf{u}] \quad , \tag{4}$$

so that we now sample from

$$\mathbf{u} \sim N(\widehat{\mathbf{u}}, \widehat{\mathbf{D}})$$

where specifically for the variance components case, we have

$$\begin{aligned} \widehat{\mathbf{D}} &= \sigma_e^2\left[\sum Z_{ij}^T Z_{ij} + \sigma_e^2\mathbf{\Omega}_{\mathbf{u}}^{-\mathbf{1}}\right]^{-1} = \sigma_e^2\left[\mathrm{diag}(n_j) + \sigma_e^2\mathbf{\Omega}_{\mathbf{u}}^{-\mathbf{1}}\right]^{-1} \\ \widehat{\mathbf{u}} &= \left[\sum Z_{ij}^T Z_{ij} + \sigma_e^2\mathbf{\Omega}_{\mathbf{u}}^{-\mathbf{1}}\right]^{-1}\left[\sum Z_{ij}^T(y_{ij} - (X\beta)_{ij})\right] = \widehat{\mathbf{D}}\sigma_e^{-2}\widetilde{\mathbf{y}}. \\ \widetilde{\mathbf{y}} &= \{\widetilde{\mathbf{y}}_j\}, \ \ \widetilde{\mathbf{y}}_j = \sum_{i=1}^{n_j}(y_{ij} - (X\beta)_{ij}). \end{aligned} \tag{5}$$

3.3. Adaptive Proposal Distributions

The proposal distributions are determined adaptively (Browne & Draper, 2000). We choose a desirable acceptance rate r, in the present case .5, and we choose a batch size $B = 100$, as used by Browne and Draper (2000). For each batch of iterations during the burn-in period, we compute the acceptance rate r^{*} for each parameter. We update the proposal distribution, for each parameter, according to the following rule, from suitable starting values supplied by the user.

$$\text{If } r^* \geq r,\ \ \theta_t = \theta_{t-1}[2 - \left(\frac{1 - r^*}{1 - r}\right)],\ \text{otherwise } \theta_t = \theta_{t-1}(2 - \frac{r^*}{r})^{-1} \tag{6}$$

where r is the desired acceptance rate and θ_t is the normal proposal distribution standard deviation for the parameter under consideration at iteration t. Unlike in Browne and Draper (2000), here we simply adapt for the whole of the burn-in period.

3.4. The Deviance Information Criterion (DIC) Diagnostic

To compare the models in later sections, we use the DIC diagnostic (Spiegelhalter, Best, Carlin, & van der Linde, 2002). While the use of the DIC has been criticized for certain classes of model, it has found widespread use in random effects models. It allows us to assess any improvements in model fit for different assumed distributions for the u vector.

4. An Example

We have simulated data to mimic the educational "tutorial" data set (Rasbash, Steele, Browne, & Prosser, 2004). The data set consists of examination scores measured at the age of 16 years on 4,059 students in 65 schools in London. The examination score is the response with a reading test score taken at the age of 11 years as a covariate. In the original data set, both the response and covariate were transformed to normality using the normal equivalent deviates computed from the sample data. We use the maximum likelihood estimates for the model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij},\ u_j \sim N(0, \sigma_u^2),\ e_{ij} \sim N(0, \sigma_e^2)$$

where we use the values

$$\beta_0 = .002,\ \beta_1 = .563,\ \sigma_u^2 = .092, \text{and } \sigma_e^2 = .566.$$

The Level 2 residuals are then generated as follows:

$$g_1(u_j, u_k) = |j - k|^{-1},\ \ \alpha_1 = 1 \tag{7}$$

The Equation 7 says that the correlation is a monotonically decreasing function of the difference between the two Level 2 identifier numbers. Note that here the correlations are all positive and where we wish to constrain correlations to be positive we might wish to choose a different link function such as the logit.

Generating data from the above model and then using the MCMC algorithm in the last section and a standard algorithm for a model with assumed independence between the u_j, we obtain the results in Table 1.

Here, we see that the posterior mean estimates from the fit of the nonindependence model (which is the generating model for the data) for the variances are positively biased. This is in part due to the right-skew of the distribution and

TABLE 1
Simulated Data Estimates for the Model 7

	Nonindependence Model Estimates	Independence Model Estimates	Generating Model Values
Intercept	−0.006 (0.078)	−0.010 (0.077)	0.002
Slope	0.567 (0.013)	0.562 (0.012)	0.563
Level 2 variance	0.098 (0.036)	0.089 (0.036)	0.092
Level 1 variance	0.573 (0.022)	0.566 (0.014)	0.566
α_1	1.01 (0.39)	—	1.0

Note: Markov chain Monte Carlo (MCMC) was run for 2,500 iterations following a 500 iteration burn-in. The results are based on 50 simulations with standard errors over simulations given in brackets.

hence the mean being larger than the mode for this parameter. The alternative model that assumes fitting independent Level 2 residuals results in an underestimate of the Level 2 variance. We next move on to looking at incorporating a correlated residual structure within Level 2 units.

5. Level 1 Nonindependent Residuals

In the previous sections, we have investigated models where the independence assumption between the cluster level random effects is relaxed. By introducing cluster level random effects in a multilevel model, we have removed the dependence among the observation level residuals by effectively splitting them into a shared cluster level residual and an (adjusted) observation level residual. These latter residuals are then assumed independent and this section relaxes that assumption. Our focus is on the structure of residuals within clusters rather than all residuals in the model, and we add some (structured) correlations between these residuals.

The types of models that such a generalization is useful for include repeated measures at Level 1. In the traditional approach to modeling repeated measures on individuals over time, we capture the nonindependence due to repeated measures coming from the same individual via a cluster (individual) level random effect, while still assuming that the Level 1 residuals are independent, conditional on the Level 2 random effects. In some applications, however, especially where measures are close together in time, this independence assumption may be violated and we will therefore need to model the correlation between residuals, typically as a function of the time differences. A detailed discussion of this model and its applications is given by Goldstein, Healy, and Rasbash (1994).

Our model now becomes

$$y_{ij} = \mathbf{X_{ij}}\beta + u_j + e_{ij},\ \ \mathbf{e_j} \sim \text{MVN}(0, \mathbf{\Omega_{ej}}),\ \ \mathbf{u} \sim \text{MVN}(0, \mathbf{\Omega_u}),\ \ \mathbf{e_j} = \{e_{ij}\} \tag{8}$$

where we assume that there is just a single residual term per observation at Level 1.

We again assume the following form for the correlations although this time we are considering correlations at Level 1:

$$\begin{array}{l} f(\rho_{jk}^{(1)}) = \alpha_1^{(1)} g_1^{(1)}(t_j, t_k) + \alpha_2^{(1)} g_2^{(1)}(t_j, t_k) \ldots \\ \text{and} \\ f(s) = 2\tan h^{-1}(s) = \log\left(\frac{1+s}{1-s}\right) \end{array} \tag{9}$$

where the superscript (1) denotes Level 1.

The sampling for the Level 1 covariance matrix involves essentially the same steps as for the Level 2 matrix described in the earlier algorithm and we omit the details. For sampling the variance and correlation parameters, the components of the likelihood ratio become

$$p(\mathbf{\Omega_e^*}|\mathbf{e}) = \prod_j |\mathbf{\Omega_{ej}^*}|^{-1/2} \exp - (\mathbf{e_j^T}(\mathbf{\Omega_{ej}^*})^{-1}\mathbf{e_j}/2)$$

and

$$p(\mathbf{\Omega_e^{(t-1)}}|\mathbf{e}) = \prod_j |\mathbf{\Omega_{ej}^{(t-1)}}|^{-1/2} \exp - (\mathbf{e_j^T}(\mathbf{\Omega_{ej}^{(t-1)}})^{-1}\mathbf{e_j}/2).$$

In this case, the explanatory variables $g_k^{(1)}\ \ (k = 1, \ldots)$ must be specified for each Level 2 unit. The Level 1 residuals are obtained by subtraction given the current fixed effects and Level 2 residual estimates.

When sampling the fixed effects, because the Level 1 residuals are no longer independent, the standard MCMC step is modified as follows. We assume a "diffuse" prior $p(\beta) \propto 1$ so that

$$p(\beta|\mathbf{y}, \mathbf{\Omega_e}, \mathbf{u}) \propto \prod_j |\mathbf{\Omega_{ej}}|^{-1/2} \exp[-\tilde{\mathbf{y}}_j^T \mathbf{\Omega_{ej}^{-1}} \tilde{\mathbf{y}}_j/2]$$

where

$$\tilde{\mathbf{y}}_\mathbf{j} = \{\tilde{y}_{ij}\},\ \ \tilde{y}_{ij} = y_{ij} - (\mathbf{X}\beta)_{ij} - u_j$$

so that we sample from

$$\begin{array}{c} \beta \sim \text{MVN}(\widehat{\beta}, \widehat{\mathbf{D}}_\beta) \\ \widehat{\mathbf{D}}_\beta = \left[\sum_j \mathbf{X_j^T} \Omega_{\mathbf{ej}}^{-1} \mathbf{X_j}\right]^{-1} \\ \widehat{\beta} = \widehat{\mathbf{D}}_\beta \left[\sum_j \mathbf{X_j^T} \Omega_{\mathbf{ej}}^{-1} (\mathbf{y_j} - (\mathbf{Zu})_\mathbf{j})\right]. \end{array} \tag{10}$$

Note here that for the variance components case $\mathbf{Z_j}$ simply links the correct Level 2 unit to each Level 1 unit. Likewise when sampling the Level 2 random effects, Equation 5 becomes

$$\begin{aligned} \widehat{D}_u &= \left[\text{diag}(Z_j^T \Omega_{ej}^{-1} Z_j) + \Omega_u^{-1}\right]^{-1} \\ \widehat{u} &= \widehat{D}_u \left[\{Z_j^T \Omega_{ej}^{-1} (y_j - (X\beta)_j)\}\right]. \end{aligned} \tag{11}$$

We note that the form of these sampling steps is the same as those in Section 4.2.1 of Browne (2006) for a multivariate multilevel model.

Consider the fitted tutorial data model with estimates for the parameters given

$$\begin{aligned} &y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \\ &\beta_0 = .035(0.040) \\ &\beta_1 = .567(0.012) \\ &\sigma_u^2 = .088(0.018) \\ &\sigma_e^2 = .566(0.013) \\ &\text{DIC} = 9265.7 \\ &\text{PD} = 59.4 \end{aligned} \quad , \tag{12}$$

where these are the MCMC estimates using inverse gamma priors for the variances. We now fit the model where we assume an equal (non zero) correlation structure at Level 1, that is $g_1^{(1)} = 1$ with an inverse tanh link. We obtain estimates for the following parameters, with a burn-in of 500 and 1,000 samples.

$$\begin{aligned} &\beta_0 = 0.032\ (.040) \\ &\beta_1 = 0.567\ (.013) \\ &\sigma_u^2 = .0000005\ (.00000001) \\ &\sigma_e^2 = .657\ (.023) \\ &\alpha_1^{(1)} = .277\ (.050) \\ &\rho = .138 \\ &\text{DIC} = 9356.5 \\ &\text{PD} = 3.8. \end{aligned} \tag{13}$$

For a two-level variance components model, the full covariance matrix among the Level 1 units in a Level 2 unit can be written in the form

$$\begin{pmatrix} \sigma_e^2 + \sigma_u^2 & & & \\ \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 \end{pmatrix}, \tag{14}$$

where in this case, there are four Level 1 units (Goldstein, 2003, chap. 2). For the model with an equal correlation structure at Level 1 and no Level 2 variation, the corresponding covariance matrix is

$$\begin{pmatrix} \sigma^2_{e*} & & & \\ \rho\sigma^2_{e*} & \sigma^2_{e*} & & \\ \rho\sigma^2_{e*} & \rho\sigma^2_{e*} & \sigma^2_{e*} & \\ \rho\sigma^2_{e*} & \rho\sigma^2_{e*} & \rho\sigma^2_{e*} & \sigma^2_{e*} \end{pmatrix}, \tag{15}$$

which has an equivalent structure with $\rho\sigma^2_{e*} = \sigma^2_u$, $\sigma^2_{e*} = \sigma^2_u + \sigma^2_e$, and ρ can be interpreted in the usual way as the intraunit correlation.

Thus, the parameters of Model 13 are not separately identified but we do see that this provides essentially the same parameter estimates as Equation 12 because $\rho\sigma^2_e$ in Equation 13 is .090 which is close to σ^2_u in Equation 12, and $\sigma^2_e + \sigma^2_u$ in Equation 12 is .654, which is close to the value .657 for σ^2_e in Equation 13. The same results are obtained using the logit and the log link functions. If we constrain $\sigma^2_u = 0$ and fit the model with equal correlations among Level 1 units within each Level 2 unit, we obtain a value of .091 compared to .088 in Equation 12 for the Level 2 variance and .567 for the Level 1 variance compared to .566 in Equation 12. We also notice that the DIC statistics are different because we are actually fitting different models. Model 12 includes random effects whereas Model 13 does not, and this can be seen in the values for the estimated effective number of parameters (PD) which is just under 4 in Equation 13, reflecting the number of actual parameters in that model as opposed to just over 59 in Equation 1, reflecting the inclusion of the random effects as parameters.

Special cases of models with correlated Level 1 residuals have been studied elsewhere. For example, Goldstein et al. (1994) present a maximum likelihood estimation procedure for Equation 7 with a log link function. Several software packages such as SAS (www.sas.com), MPLUS (www.statmodel.com), and GLLAMM (www.gllamm.org) implement maximum likelihood estimation for the case of discrete time models with a finite number of time points and allow a variety of patterns including autoregressive and unstructured correlation matrices.

6. Applications

6.1. An Educational Data Set

We illustrate our procedure with an example of educational examination results where we have a three-level model consisting of schools at Level 3, cohorts or year groups at Level 2, and students at Level 1. We shall show how such a three-level structure, where typically independent residuals are assumed, can be modeled as a two-level structure with a particular correlation pattern among the Level 2 residuals. The data are taken from the Pupil Level Annual Schools Census (PLASC) and the National Pupil Database (NPD), which have been set up for all students in the English state education system (Goldstein, Burgess, & McConnell, 2007). These contain longitudinal records of test results for individuals, together with a limited amount of contextual data. For our

purposes, we use 16-year-old General Cerificate of Secondary Examinartion (GCSE) data that constitutes the annual end of compulsory school leaving examination. It consists of papers taken in different curriculum subjects and students can take as many papers as they wish, subject to timetabling constraints. In the current analysis, we have used the best eight results for each student and this is converted into a total score for each student (Goldstein et al., 2007). There is also a composite test score at the age of 11 (Year 6) for each student prior to entering secondary school and we use the data from three cohorts of students, where each student has both age 11 and GCSE scores, where the latter are for the years 2004–2006. The data are restricted geographically to one local authority containing 54 secondary schools with a total of 29,506 students. Students are allocated to the school they are attending at the time of the GCSE examination. This is not strictly appropriate because it ignores the contributions from previous schools that may have been attended and this issue is discussed by Goldstein et al. (2007) who show how multiple membership models can be used. For current purposes, and in order to illustrate our methodology without undue complications, we shall ignore this issue.

A standard procedure for fitting such data where there are repeated measurements for cohorts attending the same set of schools is to specify a three-level model with students (i) at Level 1, year group/cohort (j) at Level 2, and school (k) at Level 3. We can write this model as

$$\begin{aligned} y_{ijk} &= \beta_0 + \beta_1 x_{ijk} + \sum_{h=2}^{3} \beta_h c_{h,jk} + v_{0k} + u_{0jk} + e_{0ijk} \\ v_{0k} &\sim N(0, \sigma^2_{v0}), \; , \; u_{0jk} \sim N(0, \sigma^2_{u0}), \; e_{0ijk} \sim N(0, \sigma^2_{e0}) \end{aligned} \tag{16}$$

where x is the prior test score and c_h is a dummy variable for the year and h denotes the cohorts coded 1, ..., 3. Table 2 shows the results of fitting the three-level variance components Model 16. We use MLwiN (Rasbash, Browne, Healy, Cameron, & Charlton, 2000) using MCMC with standard default prior distributions.

We note that the standard three-level model in effect makes some simplifying assumptions, and the saturated model is given by Equation 17, which fits a 3×3 full covariance matrix at the school level, that is the three cohorts in each school are given separate random effects that are correlated. The results are given in Table 3.

$$\begin{aligned} y_{ik} &= \beta_0 + \beta_1 x_{ik} + \sum_{h=2}^{3} \beta_h c_{h,k} + \sum_{h=1}^{3} u_{hk} c_{h,k} + e_{0ik}, \\ \mathbf{u}_k &\sim \mathrm{MVN}(0, \boldsymbol{\Omega}_{\mathbf{u}}), \; e_{0ij} \sim N(0, \sigma^2_{e0}) \end{aligned} \tag{17}$$

where as before i indexes pupils, k indexes school, and h indexes the year the pupil takes the exam.

TABLE 2
A Three-Level Variance Components Model for Examination Data

Parameter	Estimate	Standard Error
Intercept	0.018	0.025
Year 2	−0.041	0.017
Year 3	−0.003	0.017
Pretest	0.719	0.004
Level 1 variance	0.467	0.004
Level 2 variance	0.005	0.001
Level 3 variance	0.029	0.006
DIC (PD)	61,398.0 (127.0)	

Note: Burn-in = 500, sample = 5,000. Year Group 1 (2004) chosen as base category. Uniform priors for variances.

TABLE 3
A Two-Level Random Coefficient Model for Examination Data

Parameter	Estimate	Standard Error
Intercept	0.015	0.027
Year 2	−0.043	0.020
Year 3	−0.004	0.019
Pretest	0.719	0.004
Pupil level variance	0.467	0.004
School level covariance matrix	$\begin{bmatrix} 0.036 & & \\ 0.032 & 0.043 & \\ 0.028 & 0.034 & 0.033 \end{bmatrix}$	$\begin{bmatrix} 0.008 & & \\ 0.008 & 0.010 & \\ 0.007 & 0.008 & 0.008 \end{bmatrix}$
DIC (PD)	61,399.8 (128.2)	

Note: Burn-in = 500, sample = 5,000. Year 1 (2004) chosen as base category. Uniform priors for variances.

We note that the variance components structure in Table 2 corresponds to a covariance matrix in the two-level formulation, as in Table 3, with a common diagonal variance of .034 and a common covariance of .029.

We now illustrate our methodology by fitting a further series of two-level models with the covariances across years defined in different ways. Our first model is specified as in Equation 17 but with a common variance for the three cohorts and the correlation structure at Level 2 defined as follows:

$$f_{h_1h_2} = \alpha_1 g_1(u_{h_1}u_{h_2}) = \alpha_1|h_1 - h_2|^{-1}, \quad \rho_{h_1h_2} = (e^{f_{h_1h_2}} - 1)/(e^{f_{h_1h_2}} + 1)\,. \tag{18}$$

TABLE 4
A Two-Level Model for Examination Data With Correlation Structure at Level 2 Specified by Equation 18

Parameter	Estimate	Standard Error
Intercept	0.017	0.024
Year 2	−0.040	0.018
Year 3	−0.003	0.025
Pretest	0.719	0.004
Alpha	2.063	0.183
Level 1 variance	0.467	0.004
Level 2 covariance matrix	$\begin{bmatrix} 0.028 & & \\ 0.021 & 0.028 & \\ 0.013 & 0.021 & 0.028 \end{bmatrix}$	
DIC (PD)	61,402.3 (129.6)	

Note: Inverse tanh link function. Burn-in = 500, sample = 5,000. Year 1 (2004) chosen as base category. Uniform priors for variances.

This is similar to a first-order autoregressive model and the results are given in Table 4. The structure of the explanatory variable vector for the Level 2 correlations for three schools, written in matrix form for the off diagonal elements of the 9×9 covariance matrix for the three cohorts in the three schools is given by Equation 19. The variance term is the sum of the between-school variance and the between-cohort variance after fitting the correlation structure.

$$\begin{pmatrix} 1 & & & & & & & & \\ 0.5 & 1 & & & & & & & \\ 0 & 0 & 0 & & & & & & \\ 0 & 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & 0.5 & 1 & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{19}$$

Because there are 54 schools, this vector has length $(162 \times 161)/2 = 13{,}041$. Choosing the inverse tanh link allows us to model a covariance structure with decreasing correlations as we move away from the diagonal. We note that this model may be more conveniently fitted using a three-level formulation with no variation at Level 3 and with the correlation structure specified at Level 2. The current analysis serves to illustrate the connection between a two-level and a three-level formulation.

We can see that this model tends to underestimate the variance and covariances. If, however, we fit the same model using a logit link function we obtain the following results in Table 5.

TABLE 5
A Two-Level Model for Examination Data With Correlation Structure at Level 2 Specified by Equation 18

Parameter	Estimate	Standard Error
Intercept	0.015	0.024
Year 2	−0.041	0.017
Year 3	−0.004	0.017
Pretest	0.719	0.004
Alpha	1.693	0.342
Level 1 variance	0.467	0.004
Level 2 covariance matrix	$\begin{bmatrix} 0.034 & & \\ 0.029 & 0.034 & \\ 0.024 & 0.029 & 0.034 \end{bmatrix}$	
DIC (PD)	61,398.5 (125.8)	

Note: Logit link function. Burn-in = 500, sample = 5,000. Year 1 (2004) chosen as base category. Uniform priors for variances.

We see now that this is a rather better fit to the data and effectively makes use of the fact that the correlations are positive, which is a constraint imposed by the logit link function. Finally, we fit a model where the correlation structure is modeled by two α parameters as follows:

$$\begin{aligned} f_{h_1 h_2} &= \alpha_1 g_1(u_{h_1} u_{h_2}) + \alpha_2 g_2(u_{h_1} u_{h_2}) = \alpha_1 |h_1 - h_2|^{-1} + \alpha_2 .1, \\ \rho_{h_1 h_2} &= (e^{f_{h_1 h_2}} - 1)/(e^{f_{h_1 h_2}} + 1) \end{aligned} \tag{20}$$

Here, we have added a constant term α_2 to the correlation function. The results are given in Table 6.

This provides a slightly better fit than the model in Table 4 but the second parameter α_2 is not significant.

6.2. A Growth Data Set

We reanalyze a data set of longitudinal measurements at nine occasions on a sample of 21 boys, discussed by Goldstein (2003, Chapter 5). There, a two-level polynomial growth model was fitted with terms up to the fourth order and with Level 2 random coefficients for the intercept, linear, and quadratic terms. In Goldstein (2003), the following log link function was used at Level 1 to describe the correlation structure:

$$f_{t_1 t_2} = \alpha |t_1 - t_2|, \;\; \rho_{t_1 t_2} = e^{f_{t_1 t_2}}. \tag{21}$$

The nine target occasions were nominally 3 months apart, but there was variation around this interval and we therefore fit our model in continuous time. In discrete time, it becomes a first-order autoregressive model.

TABLE 6
A Two-Level Model for Examination Data With Correlation Structure at Level 2 Specified by Equation 19

Parameter	Estimate	Standard Error
Intercept	0.017	0.026
Year 2	−0.041	0.017
Year 3	−0.002	0.018
Pretest	0.719	0.004
α_1	2.144	0.612
α_2	0.389	0.680
Level 1 variance	0.467	0.004
Level 2 covariance matrix	$\begin{bmatrix} 0.033 & & \\ 0.028 & 0.033 & \\ 0.021 & 0.028 & 0.033 \end{bmatrix}$	
DIC (PD)	61,401.0 (126.4)	

Note: Inverse tanh link function. Burn-in = 500, sample = 5,000. Year 1 (2004) chosen as base category. Uniform priors for variances.

Table 7 shows the results from fitting the model from Goldstein (2003) but here using the inverse tanh link.

The estimate of α is close to zero suggesting little evidence for any autocorrelation structure. Figure 2 shows the chain for α using the inverse tanh link. The chain shows reasonably good mixing with no discernible trend.

The analysis was also carried out using the log and logit links (details omitted). For these, α was estimated with large negative values and poorly mixing chains, and this compares with the MLE in Goldstein (2003) of −6.9 with *SE* of 2.0.

7. Discrete Responses

Our procedures can be extended to handle discrete responses. Here, we consider the binary response case in detail, using a probit link function.

We consider first the case where the Level 1 residuals are independent. We write

$$z \sim N(\mu, 1), \quad \mu = \mathbf{X}\beta + Zu,$$

where we observe a positive (=1) response for our binary response y if z is positive, that is

$$z_{ij} = \mu + e_{ij} > 0 \text{ or}$$
$$e_{ij} > -\mu.$$

TABLE 7
Height as a Fourth-Degree Polynomial on Age, Measured About 13.0 Years

Fixed			
Intercept	148.9 (1.3)		
Age	6.16 (0.35)		
Age^2	2.16 (0.47)		
Age^3	0.39 (0.16)		
Age^4	−1.55 (0.46)		
Cos (time)	−0.24 (0.07)		
Random			
Level 2 covariance matrix			
	Intercept	Age	Age squared
Intercept	65.9 (19.7)		
Age	8.5 (3.5)	3.0 (0.9)	
Age squared	1.5 (1.6)	0.9 (0.4)	0.64 (0.25)
Level 1 variance			
σ_e^2	0.21 (0.03)		
α (mean)	−0.020 (0.20)		
α (median + 95% interval)	−0.064 (−0.28 0.42)		
DIC (PD)	344.0 (58.1)		

Note: Standard errors in brackets. Markov chain Monte Carlo (MCMC) estimates. Burn-in = 5,000, sample = 50,000. After adapting, proposal distribution $SD = 0.08$. Inverse tanh link.

We have

$$\text{Prob}(y = 1) = \text{Prob}(e_{ij} > -\mu) = \int_{-\mu}^{\infty} \phi(t)\mathrm{d}t = \int_{-\infty}^{\mu} \phi(t)\mathrm{d}t \tag{22}$$

where $\phi(t)$ is the density function of the standard normal distribution. This is the standard probit model. This leads to the MCMC sampling step whereby when a 1 is observed, we sample z from $\int_{-\infty}^{\mu} \phi(t)\mathrm{d}t$ and when a 0 is observed we sample z from $\int_{\mu}^{\infty} \phi(t)\mathrm{d}t = \int_{-\infty}^{-\mu} \phi(t)\mathrm{d}t$.

Using the above, we can sample the z latent variables, and our model and hence our algorithm is as before but with the inclusion of the extra sampling step for z. For the case of nonindependence at Level 1, we modify this step as follows.

For Level 2 unit j, suppose that the values at the start of the current iteration of the latent normal Level 1 random effects (residuals) are given by $e_j = \{e_{ij}\}$. We sample each random effect, e_{ij}, conditioning on the remaining random effects in the Level 2 unit. That is, for the conditional sampling, the distribution $N(\mu, 1)$ is replaced by

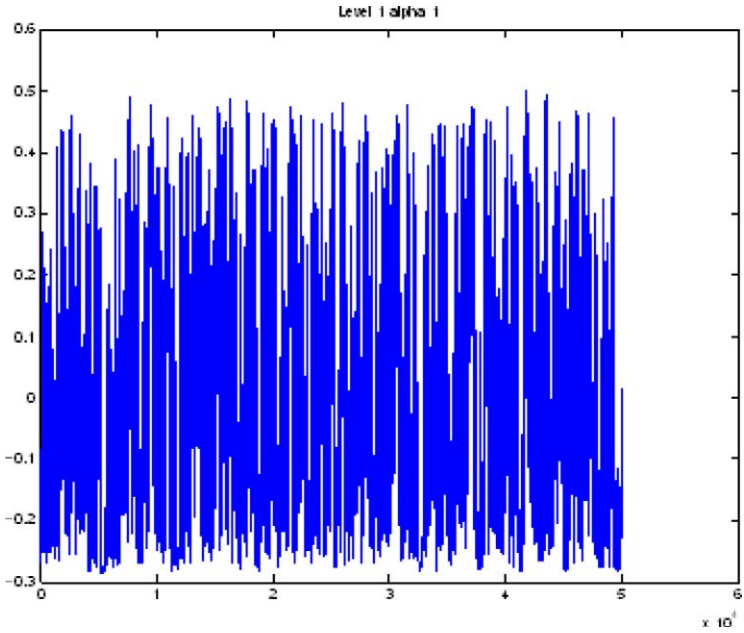

FIGURE 2. *Chain for alpha in Table 7.*

$$N[\mu_{kj} + e_{k,j}^T \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_1^{-1}, \sigma_k^2], \;\; \sigma_k^2 = 1 - \boldsymbol{\Sigma}_{12}^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_{12}, \;\; e_{k,j} = e_{j(j \neq k)}\,, \tag{23}$$

where we partition $\Omega_{ej} = \begin{pmatrix} \boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_{12} \quad 1 \end{pmatrix}$. This sequence will produce a multivariate normal distribution for the latent variables within each Level 2 unit.

When sampling conditionally on correlated random effects, it is possible for the conditional mean to become relatively large and the corresponding residual variance to become relatively small so that for some data points we may be sampling from the extreme tail of the normal distribution. Given machine accuracy, this may lead to the associated tail probability being returned as 1.0 leading to a latent variable value that is coded as infinite. To avoid this problem, a cutoff should be chosen, for example, a value equivalent to 5 on the standard normal scale.

8. Discussion

We have shown how a wide class of nonindependence structures for random effects at different levels can be specified and fitted. There are several extensions to our models including higher levels of the data hierarchy and

cross-classifications where independence across classifications is assumed and also to multiple membership models. In addition, we can consider a fully multivariate version with a set joint responses, and this is under investigation. The above estimation steps could be applied to each relevant classification conditional on current estimates. We can also extend the binary response model to ordered classifications using a probit link function together with a set of "threshold" parameters defining the category boundaries. For further details, see Goldstein, Carpenter, Kenward, and Levin (2009).

As the number of Level 2 units becomes large or the number of Level 1 units within a Level 2 unit becomes large for nonindependent Level 1 models, we will need to handle very large covariance matrices and efficient procedures for this will need to be developed so that we can sample from these efficiently. We also need to take care that our parameters are identified. For example choosing $g_1(j_1,j_2) = |j_1 - j_2|^{-1}$, $g_2(j_1,j_2) = |j_1 - j_2|^{-2}$ leads to nonidentifiability if the inverse tanh, logit, or log link functions are used. The choice of prior distribution for the elements of the covariance matrix could be further studied. Thus, for example, we could choose an inverse gamma prior for the variance parameters, although in our examination scores example this makes little difference to the estimates.

A number of models can be viewed as special cases. Time series models such as autoregressive structures (Goldstein et al., 1994) are one example and within-family sibling relationship models are another. In the latter case, the correlation between sibling characteristics will typically depend on whether they are twins or singletons or on whether they are half or full siblings. In the former case, an important feature of our model is that we can model complex time series structures where the repeated measures do not occur at the same set of regular time intervals for each individual. This distinguishes it from existing approaches that treat the set of common occasions as a special kind of multivariate structure. Another important application of our models is in the modeling of educational and other data where institutions do not behave independently. Thus, for example, in the case of schooling effects, actions taken by one school when competing for limited resources can be expected to affect the actions of nearby schools and partnerships and collaborations will also invalidate assumptions about the independence of school effects on pupil progress or performance. In our first example, we assumed a simple model using an inverse distance function for the correlation between schools. In practice, the choice of one or more such functions will need to be guided by both theory and data.

In our exposition, we have assumed a common variance for the Level 2 residuals and for the Level 1 residuals. A natural generalization is to allow these variances to depend on Level 2 or Level 1 covariates. For example, in a time series model, the variance may be a function of time, for example a seasonal function. To incorporate this at either level, we would simply insert another Metropolis

step into the algorithm that sampled the parameters of such a function to provide a current value for the variance. Browne (2006) describes a procedure for this.

A set of MATLAB (Mathworks, 2004) macros was written to implement these models.

References

Besag, J., & Kooperberg, C. L. (1995). On conditional and intrinsic autoregressions. *Biometrika*, *82*, 733–746.

Besag, J., York, J., & Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, *43*, 1–59.

Browne, W. J. (2006). MCMC algorithms for constrained variance matrices. *Computational Statistics and Data Analysis*, *50*, 1655–1677.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391–420.

Clayton, D. G., & Kaldor, J. (1987). Empirical Bayes estimates of age standardised relative risks for use in disease mapping. *Biometrics*, *43*, 671–681.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.

Goldstein, H., Burgess, S., & McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society, Series A*, *170*, 941–954.

Goldstein, H., Carpenter, J., Kenward, M., & Levin, K. (2009). Multilevel Models with multivariate mixed response types. *Statistical Modeling*, *9*, 173–197.

Goldstein, H., Healy, M. J. R., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, *13*, 1643–1655.

Mathworks. (2004). Matlab. Retrieved from www.mathworks.co.uk

Rasbash, J., Browne, W. J., Healy, M., Cameron, B., & Charlton, C. (2000). *The MLwiN software package version 1.10.* London: Institute of Education, University of London.

Rasbash, J., Steele, F., Browne, W. J., & Prosser, B. (2004). *A user's guide to MLwiN.* London: Institute of Education, University of London.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*, 583–640.

Authors

WILLIAM BROWNE is currently Professor of Biostatistics at the School of Veterinary Science, University of Bristol, Lower Langford, Bristol, BS40 5DU, UK; william.browne @bristol.ac.uk. He has previously worked at the Institute of Education and at the School of Mathematical Sciences, University of Nottingham. He is a currently a member of the Council of the Royal Statistical Society and his recent research has been funded by several UK Research Councils, the Department for Environment, Food and Rural affairs, and by the Wellcome Foundation. He has recently become co-Director of the Centre for Multilevel Modelling at Bristol. His research interests are in the use of statistical modelling

techniques (in particular MCMC methods) for analysing complex datasets in many fields including veterinary epidemiology, ecology and education.

HARVEY GOLDSTEIN is Professor of Social Statistics at the Graduate School of Education, University of Bristol, Bristol, BS8 1JA, UK; e-mail h.goldstein@bristol.ac.uk. He has been a member of the Council of the Royal Statistical Society (RSS), and chair of its Educational Strategy Group. He was awarded the RSS Guy medal in silver in 1998 and was elected a fellow of the British Academy in 1996. He has been the principal applicant on several major UK research council funded projects since 1981. He has two main research interests. The first is the use of statistical modelling techniques in the construction and analysis of educational tests with a particular interest in institutional and international comparisons. The second is in the methodology of multilevel modelling. His major recent book, *Multilevel Statistical Models* (Wiley, 2010, 4th edition) is the standard reference text in this area of statistical data analysis.

Manuscript received February 12, 2009
Revision revised July 16, 2009
Accepted October 6, 2009