Multiple membership multiple classification (MMMC) models

William J Browne, Harvey Goldstein and Jon Rasbash

Statistical Modelling 2001; 1; 103

DOI: 10.1177/1471082X0100100202

Statistical Modelling 2001; **1**: 103–124

Multiple membership multiple classification (MMMC) models

William J Browne, Harvey Goldstein and Jon Rasbash
Institute of Education, University of London, London, UK

Abstract: In the social and other sciences many data are collected with a known but complex underlying structure. Over the past two decades there has been an increase in the use of multilevel modelling techniques that account for nested data structures. Often however the underlying data structures are more complex and cannot be fitted into a nested structure. First, there are cross-classified models where the classifications in the data are not nested. Secondly, we consider multiple membership models where an observation does not belong simply to one member of a classification. These two extensions when combined allow us to fit models to a large array of underlying structures. Existing frequentist modelling approaches to fitting such data have some important computational limitations. In this paper we consider ways of overcoming such limitations using Bayesian methods, since Bayesian model fitting is easily accomplished using Monte Carlo Markov chain (MCMC) techniques. In examples where we have been able to make direct comparisons, Bayesian methods in conjunction with suitable 'diffuse' prior distributions lead to similar inferences to existing frequentist techniques. In this paper we illustrate our techniques with examples in the fields of education, veterinary epidemiology, demography, and public health illustrating the diversity of models that fit into our framework.

Key words: multilevel modelling; hierarchical modelling; Monte Carlo Markov chain (MCMC); cross-classified models; multiple membership models; complex data structures; Bayesian GLMM modelling

Data and software available from: http://stat.uibk.ac.at/SMIJ

Received October 2000; **revised** June and August 2001; **accepted** August 2001

1 Introduction

Over the past two decades or so there has been a great interest in fitting realistically complex statistical models to the large datasets that occur in the social sciences and other application areas. These complex models account for the underlying structure in such datasets through the use of random effects. Historically some of the first random effects models fitted to large datasets were in the field of education and here the structures of interest were generally pupils within classes within schools and other nested or hierarchical structures. The fitting of these multilevel or hierarchical models (e.g., Goldstein, 1995; Bryk and Raudenbush, 1992 and Draper, 2001) is now commonplace in many application areas and several special purpose software packages have been developed to fit such models (Rasbash *et al*., 2000b; Bryk *et al*., 1988).

To fit these models it is necessary to use either iterative procedures, e.g., iterative generalized least squares (IGLS (Goldstein, 1986)) or simulation based methods, e.g., Gibbs

Address for correspondence: WJ Browne, Institute for Education, University of London, 20 Bedford Way, London WC1 0AL, UK. E-mail: w.browne@ioe.ac.uk

 0962-2802(01)ST008RA

sampling (Gelfand and Smith, 1990). Due to the computational intensity of the Monte Carlo Markov chain (MCMC) simulation based methods and the speed of computers at the time, the iterative maximum likelihood (ML) procedures were implemented in software packages (ML2 (Rasbash, 1989), HLM (Bryk *et al*., 1988)) several years prior to the MCMC methods.

Those who used these early statistical modelling packages then discovered datasets whose structures did not fit into the standard multilevel framework. Two such structures are cross-classified models and multiple membership models. Methods like IGLS exploit the nested structure of the data in multilevel (hierarchical) models. As these two types of structures are not strictly nested the initial solution was to convert these structures into nested models with constraints (Rasbash and Goldstein, 1994). This approach along with its problems will be discussed further in a later section.

In fact such models are part of a larger family of models known as generalized linear mixed models (GLMMs). These models are a combination of the linear mixed model (Harville, 1977) and the generalized linear model framework (Nelder and Wedderburn, 1972). The (normal) linear mixed model can be written

$$y = X\beta + Zu + e$$
$$u \sim MVN(0, \Sigma_\theta) \qquad e \sim N(0, \sigma_e^2) \tag{1}$$

Here the formulation of Σ_θ will control the type of mixed model produced. For multilevel modelling u will contain the random effects and Σ_θ is block diagonal, i.e., the u are split into independent subsets, one subset for each level. Clayton and Rasbash (1999) also consider cross-classified models as a special case of the GLMM and use a technique they call the 'alternating imputation posterior (AIP) algorithm' which we will describe later.

From a Bayesian viewpoint, Clayton (1996) shows the flexibility of different specifications of the random effects precision matrix, Σ_θ^{-1}. In this paper the models that we consider will all have block diagonal Σ_θ and we will actually split u into its independent subsets in the equations that follow. Additional complexity, in the form of cross-classified models and multiple membership models will then be achieved by modifications to the Z matrix. We will also later deal with the non-normal GLMM case.

A serious problem with the increase in complexity of the models is to establish a notation, particularly for the indices, that captures the structure of the models (see Rasbash and Browne, 2001 for more details and for a notation that extends the standard multilevel notation). Nevertheless we will see that the MCMC methods that we use in this paper do not require the exact nesting structures in the model for estimation purposes. In this paper we develop some new terminology and notation that hopefully will make the equations and estimation algorithms for these complex models simpler.

In Section 2 we introduce these new definitions and notation and demonstrate how they work for a simple two level model. Then in Sections 3 and 4 we consider the two advances to the basic structure of the multilevel model, namely cross-classified effects and multiple membership models with examples. We then describe our general framework of MMMC models that encompasses these two advances. We finally demonstrate through three actual data examples the kinds of models that can be fitted in this framework.

2 Classifications

Consider a problem that has one response variable (multivariate responses are a simple extension) and assume that there is a unique response in our dataset for each of N lowest level units. Here the lowest level units could be individuals, time points or even areas.

We now define a *classification* as a function, c, that maps from the set Θ of N lowest level units to a set Φ of size M where $M \leq N$, and we define the resulting set Φ of M objects as the *classification units*. So we have $c(n_i) = \Phi_i$, where the lowest level unit $n_i \in \Theta$ and $\Phi_i \subset \Phi$.

We will consider two types of classifications. A *single membership classification* is a function c from Θ to Φ that maps each $n_i \in \Theta$ to a unique $m_j \in \Phi$. A *multiple membership classification* is a map c from Θ to Φ that maps each $n_i \in \Theta$ to a subset (possibly of size 1) Φ_i of Φ. Note that we will still maintain that $M \leq N$ to avoid identifiability problems in the estimation that follows.

A special classification is the identity classification that maps every $n_i \in \Theta$ to $n_i \in \Phi$ where $\Theta = \Phi$. Given these definitions we will now see that all the sets of random effects that feature in multilevel models, cross-classified models and multiple membership models will each have an associated classification. Note that different classifications may share the same set Φ of classification units, for example, the areas and neighbours classifications in the lip cancer example in Section 8.

2.1 The importance of unique identifiers in nested models

One potential problem in fitting multilevel (hierarchical) models in the framework that we are introducing is the problem of unique identifiers. For example, in education, a very common structure is to have pupils within classes within schools. Here we could have class 1 in school 1 and class 1 in school 2. Hierarchical data structures are in fact a special case of the cross-classified data structures that we study next that have no crossings. Therefore we could fit a hierarchical model as a cross-classified model, however in this case we would need to differentiate between the two class 1s as they are not the same classification unit. We shall therefore assume that all classification unit identifiers are unique across a dataset.

2.2 Classification diagrams

Assuming that we have unique identifiers in nested models we do not need to know the nesting structure to fit a model with several groups of random effects. The equations that will follow use the classification notation and consequently also do not show the nestings. It is however useful to display the structure of the classifications involved in a model and for this we advocate the use of a classification diagram. The classification diagrams for the simple two level model, a cross-classified model and a multiple membership model are shown in Figure 1. Here each set of classification units (including the lowest level units themselves) is represented by a box and the classifications themselves are represented by arrows from the lowest level units to the classification units. A single membership classification is represented by a single arrow whilst a multiple membership classification is represented by a pair of arrows. If there is nesting between classifications then this can be represented by the arrow that represents the 'higher level' classification being drawn from the 'lower level' classification rather than from the lowest level units.

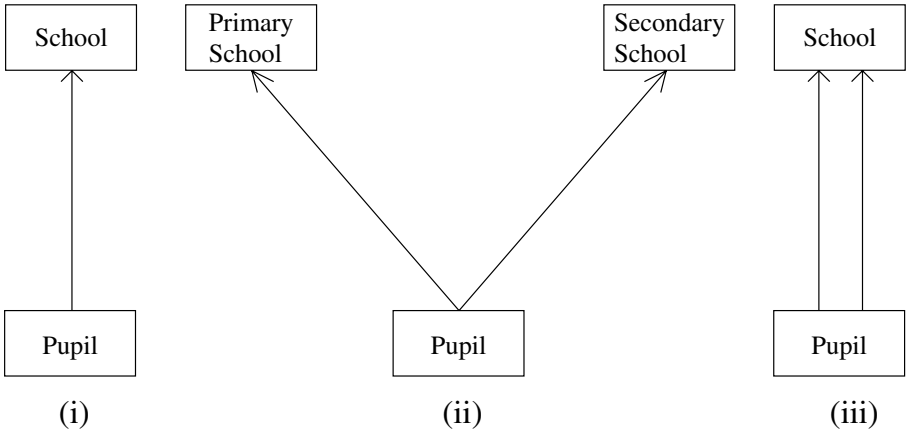

Figure 1 Classification diagrams for (i) simple two level nested model (model 3) (ii) cross-classified model (model 4) and (iii) multiple membership model (model 5)

2.3 Multilevel/hierarchical models

For illustration we will consider a simple two level normal multilevel model from the field of education that has been analysed in Rasbash *et al.* (2000a). Here we have 4059 pupils who are each classified as belonging to one of 65 schools. We will consider a simple two level variance components model with no predictors, except for an intercept, that is analysed in Browne and Draper (2001) where the response of interest is a (normalized) exam score. This model can be written in standard multilevel notation as follows:

$$y_{ij} = \beta_0 + u_j + e_{ij} \quad i = 1, \ldots, n_j \;\; j = 1, \ldots, J$$
$$\sum_{j=1}^{J} n_j = N \qquad u_j \sim N(0, \sigma_u^2) \qquad e_{ij} \sim N(0, \sigma_e^2) \tag{2}$$

The model can then be rewritten in the classification notation as follows:

$$y_i = \beta_0 + u^{(2)}_{\text{School}(i)} + e_i \quad i = 1, \ldots, N \qquad \text{School}(i) \in (1, \ldots, J)$$
$$u^{(2)}_{\text{School}(i)} \sim N(0, \sigma^2_{u(2)}) \qquad e_i \sim N(0, \sigma_e^2) \tag{3}$$

Note that we start numbering classifications from 2 upwards as classification 1 is the 'identity' classification that applies to the lowest level. Note also that in this case we could simplify the notation by writing $u^{(2)}$ as u and $\sigma^2_{u(2)}$ as σ^2_u as we have only one higher classification.

3 Cross-classified models

When the classifications in a model are not completely nested this is known as a cross-classified model. We will now describe the various existing methodologies for fitting cross-classified models before considering an example.

3.1 Alternative methodology

There are several frequentist approaches that have been considered for fitting cross-classified models. In the examples that follow we will only compare the approach of Rasbash and Goldstein (1994) with a corresponding Bayesian model fit accomplished via MCMC, but will list the other approaches for completeness.

Rasbash and Goldstein (1994) describe a likelihood-based approach that involves transforming the cross-classified model into a constrained nested model. Then the standard IGLS algorithm can be used to fit the resulting constrained model. For datasets with large numbers of units in each classification this approach requires large amounts of memory to cope with the constraints. Also for examples that deviate away from a close to nested design there can be numerical instabilities in the method.

Clayton and Rasbash (1999) introduced a technique that uses a data augmentation approach (Tanner and Wong, 1987; Schafer, 1997). Their alternating imputation posterior (AIP) method consists of treating the various nested hierarchies ('wings' in their terminology) in turn whilst including terms from the other 'wings' as offset terms. For each 'wing' a maximum likelihood or quasi-likelihood method is used and then a stochastic draw of the residuals is taken. Although this method works reasonably well, if the response is a binary variable and quasi-likelihood methods need to be used then this method is still affected by the bias that is inherent in quasi-likelihood methods for binary response multilevel models (see Goldstein and Rasbash, 1996).

Raudenbush (1993) considers an empirical Bayes approach to fitting cross-classified models based on the EM algorithm. He considers the specific case of two classifications where one of the classifications has many units whilst the other has far fewer and shows two educational examples to illustrate the method.

Two other recent approaches that can be used for fitting cross-classified models, in particular with non-normal responses are Gauss–Hermite quadrature within PQL estimation (Pan and Thompson, 2000) and the hierarchical generalized linear model (HGLM) framework as described in Lee and Nelder (2001). Neither of these approaches has been designed with speed of estimation in mind and so they are currently not feasible for the size of some of the problems that we will consider in this paper.

The MCMC algorithms for cross-classified models using the classification notation above are essentially identical to the algorithms for a nested model, as the MCMC method treats each classification as a random additive term and does not need to construct the global block-diagonal V matrix used in the IGLS algorithm. We are implementing the MCMC approach as a Bayesian method and consequently in the models that follow we will need to add prior distributions for unknown parameters. In order to compare with the ML based IGLS method (and because we have no additional prior information) we will use 'diffuse' prior distributions in all the examples that follow. Note that unless otherwise stated in all the models that follow we use normal$(0, 10^6)$ priors for all fixed effects and $\Gamma^{-1}(\epsilon, \epsilon)$ priors for all variance components, where $\epsilon = 10^{-3}$.

3.2 An example

The data that we will consider come from Fife in Scotland. As a response variable we have the exam results at age 16 of 3435 schoolchildren who attended 19 secondary schools and 148 primary schools. Here there is a cross-classification of primary schools and secondary

Table 1 Point estimates for the Fife educational dataset

Parameter	IGLS (s.e.)	MCMC (s.e.)
Mean achievement (β_0)	5.50 (0.17)	5.51 (0.18)
Between secondary school variance ($\sigma^2_{u(2)}$)	0.35 (0.16)	0.41 (0.21)
Between primary school variance ($\sigma^2_{u(3)}$)	1.12 (0.20)	1.15 (0.21)
Between individual variance (σ^2_e)	8.10 (0.20)	8.12 (0.20)

schools since not every child who went to a particular primary school then proceeded to the same secondary school. Often in education particular primary schools are feeder schools to a particular secondary school. In our example 89 out of 148 primary schools had children who went to different secondary schools. If we define the main secondary school for primary school i as the secondary school which the largest number of pupils in school i attended, then we find that only 288 out of 3435 children went to a secondary other than their main secondary. So although we have a cross-classified design, the distribution of pupils is close to nested.

We will fit the following simple cross-classified variance components model to the dataset

$$y_i = \beta_0 + u^{(2)}_{\mathrm{SEC}(i)} + u^{(3)}_{\mathrm{PRIM}(i)} + e_i$$
$$u^{(2)}_{\mathrm{SEC}(i)} \sim N(0, \sigma^2_{u(2)}) \qquad u^{(3)}_{\mathrm{PRIM}(i)} \sim N(0, \sigma^2_{u(3)}) \qquad e_i \sim N(0, \sigma^2_e) \tag{4}$$

where y_i is the exam score for the ith pupil in the dataset, SEC(i) is the secondary school they attended and PRIM(i) the primary school they attended. $u^{(2)}_{\mathrm{SEC}(i)}$ is the random effect for secondary school SEC(i), $u^{(3)}_{\mathrm{PRIM}(i)}$ is the random effect for primary school PRIM(i) and e_i is a level 1 residual for the ith pupil in the dataset. This model is illustrated in the second classification diagram in Figure 1. To complete the Bayesian specification of this model for the MCMC method we include 'diffuse' priors as described earlier.

We see in Table 1 that in this example there is more variation between primary schools than between secondary schools. The MCMC (posterior mean) estimates (based on a main run of 50 000 iterations after a burn-in of 500 iterations from a simple special case of the algorithm in Appendix A) replicate the IGLS estimates from the Rasbash and Goldstein (1994) method with slightly greater higher level variances due to the skewness of the posterior distribution. A further discussion of this dataset is given in Goldstein (1995).

4 Multiple membership models

Our second extension to the standard multilevel framework considers the case when a lowest level unit is a member of more than one higher classification unit. These models are commonly known as multiple membership models (Hill and Goldstein, 1998; Rasbash and Browne, 2001). For example, in medical studies a hospital patient may be treated by several nurses and each nurse will then have an effect on the patient's progress. Of course different nurses will spend different amounts of time with each patient and so we would also like to incorporate this information in our model. To do this we use a weighting scheme so that for

the nurse classification each patient will have weights for all the nurses that treated them that typically sum to 1. One obvious way of choosing weights would be to make them proportional to the length of time each nurse spends with a patient.

Other examples where we may have a multiple membership model are in education with children being taught by several teachers in the process of their schooling, and in demography where individuals will belong to several different households over a period of time.

4.1 A simulated example

Here we will consider a simulation of a realistic educational example based on the educational hierarchical dataset (Rasbash *et al.*, 2000a) described earlier. We will assume that 90% of children stayed in the same school throughout their schooling and that the other 10% changed school (to another school chosen at random) at some point during this period. For the purposes of this simulation we will assume that a child only changes school at most once and that both schools they are members of are given equal weighting (0.5 each). Neither of these restrictions are necessary as will become clear in the real data examples in a later section.

The model is then as follows:

$$\begin{aligned} y_i &= \beta_0 + \sum_{j \in \text{School}(i)} w_{i,j}^{(2)} u_j^{(2)} + e_i \\ i &= 1, \ldots, N \qquad \text{School}(i) \subset (1, \ldots, J) \\ u_j^{(2)} &\sim N(0, \sigma^2_{u(2)}) \quad e_i \sim N(0, \sigma^2_e) \end{aligned} \tag{5}$$

and is shown in the third classification diagram in Figure 1. Again, for the MCMC method, we will use diffuse prior distributions. As the multiple membership model is a special case of the family of models introduced in the next section, the MCMC algorithm here is a Gibbs sampler that is a special case of the algorithm given in Appendix A.

One thousand sets of response variables were generated with known parameters and the results obtained from the Rasbash and Goldstein IGLS method and MCMC with the above priors are shown in Table 2. Here the estimates given by the two methods are the average values over the 1000 simulated datasets. The 90% interval estimates for the MCMC method were constructed from the 5th and 95th percentiles of the chains, whilst for IGLS we used

Table 2 Summary of simulations for a simple multiple membership model

Parameter	True	IGLS est. (MCSE)	MCMC est. (MCSE)
Mean achievement (β_0)	0	−0.0019 (0.0014)	−0.0014 (0.0013)
School variance ($\sigma^2_{u(2)}$)	0.1	0.097 (0.0006)	0.102 (0.0006)
Individual variance (σ^2_e)	0.6	0.600 (0.004)	0.600 (0.0004)
Actual coverage of nominal 90%/95% intervals			
Mean achievement (β_0)	–	83.9% /89.6%	89.9%/94.7%
School variance ($\sigma^2_{u(2)}$)	–	88.0%/92.0%	90.3%/94.4%
Individual variance (σ^2_e)	–	93.9%/96.4%	90.0%/94.3%

symmetric point estimate ± 1.645 estimated standard deviation intervals (see Browne and Draper, 2001 for similar simulations on a variance components model). In Table 2 we see that our MCMC method gives both very little bias and far better coverage properties than the IGLS method for this model.

5 Multiple membership multiple classification (MMMC) models

5.1 A general three classification normal model with one multiple membership classification

As mentioned earlier the MCMC algorithm, unlike the IGLS algorithm, does not require details of the nestings in the classification structure (assuming unique identifiers) when fitting complex random effects models. Consequently, there is no unique ordering for the sets of random effects which are additive terms in the model. This means that we only need consider a three classification model that includes the 'identity' classification for the lowest level, a single member classification and a multiple member classification since further classifications will involve similar steps. In this section we will define this general model for a normal response. Two of our three later examples actually have binomial and Poisson responses so the extension to these responses is also described. The general normal response model can be written as

$$\begin{aligned} y_i &= X_i\beta + Z_i^{(2)}u_{C_2(i)}^{(2)} + \sum_{j\in C_3(i)} w_{i,j}^{(3)} Z_i^{(3)} u_j^{(3)} + e_i \\ u_{C_2(i)}^{(2)} &\sim N(0, \Sigma_{u(2)}) \qquad u_j^{(3)} \sim N(0, \Sigma_{u(3)}) \qquad e_i \sim N(0, \sigma_e^2) \end{aligned} \tag{6}$$

Here y is an N vector, β is a vector of p_f fixed effect parameters, and $u_i^{(2)}, u_i^{(3)}$ are the vectors of residuals for the p_2 and p_3 random effects for classifications 2 and 3 respectively. The e_i are scalar lowest level unit residuals. $X_i, Z_i^{(2)}$ and $Z_i^{(3)}$ are vectors of predictor values and $w_{i,j}^{(3)}$ is a scalar weight for the classification 3 unit j for lowest level unit i.

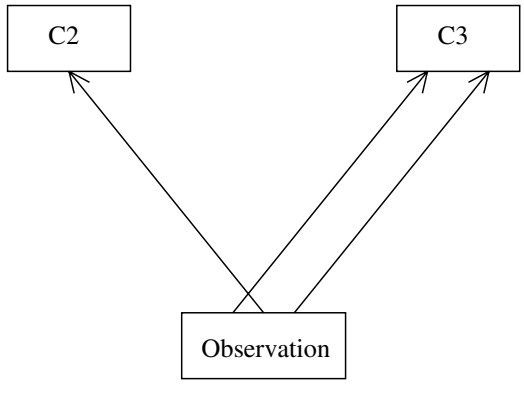

Figure 2 Classification diagram for the general 3 classification model (model 6)

For prior distributions we use a multivariate normal prior for the fixed effect parameters, $\beta \sim N_{p_f}(\mu_p, S_p)$, for the classification 3 variance matrix an inverse Wishart prior $\Sigma_{u(3)} \sim W^{-1}_{p_3}(\nu_3, S_3)$, for the classification 2 variance matrix an inverse Wishart prior $\Sigma_{u(2)} \sim W^{-1}_{p_2}(\nu_2, S_2)$ and for the lowest level unit variance a scaled inverse χ^2 prior $\sigma^2_e \sim SI\chi^2(\nu_e, s^2_e)$. Note that a $\Gamma^{-1}(\epsilon, \epsilon)$ prior as used in the examples is a special case of this prior where $\nu_e = 2\epsilon$ and $s^2_e = 1$.

This model can then be fitted using six Gibbs sampling steps as shown in Appendix A.

5.2 Extensions to other response types

An MCMC algorithm is given in Appendix B for fitting the corresponding binomial and Poisson MMMC models

$$\begin{gathered} y_i \sim \text{Binomial}(m_i, \pi_i) \\ \text{logit}(\pi_i) = X_i\beta + Z^{(2)}_i u^{(2)}_{C_2(i)} + \sum_{j \in C_3(i)} w^{(3)}_{i,j} Z^{(3)}_i u^{(3)}_j \\ u^{(2)}_{C_2(i)} \sim N(0, \Sigma_{u(2)}) \qquad u^{(3)}_j \sim N(0, \Sigma_{u(3)}) \end{gathered} \tag{7}$$

$$\begin{gathered} y_i \sim \text{Poisson}(\lambda_i) \\ \log(\lambda_i) = X_i\beta + Z^{(2)}_i u^{(2)}_{C_2(i)} + \sum_{j \in C_3(i)} w^{(3)}_{i,j} Z^{(3)}_i u^{(3)}_j \\ u^{(2)}_{C_2(i)} \sim N(0, \Sigma_{u(2)}) \qquad u^{(3)}_j \sim N(0, \Sigma_{u(3)}) \end{gathered} \tag{8}$$

The same generic prior distributions used for model (6) are used for both these models. This algorithm is based on a combination of univariate Metropolis Hastings (MH) steps and Gibbs steps and has been implemented in a development version of the MLwiN software package (Rasbash *et al.*, 2000b) that will be available publicly in late 2001. Note that the purely multilevel (nested) MCMC algorithm for binary and Poisson response models is implemented in the current version of MLwiN. These models can also be fitted using the Adaptive Rejection (AR) algorithm (Gilks and Wild, 1992) in the software package WinBUGS (Spiegelhalter *et al.*, 2000). Choosing between these two approaches will be discussed in Example 1.

6 Example 1: Danish poultry salmonella outbreaks

Rasbash and Browne (2001) consider an example from veterinary epidemiology concerning the outbreaks of *Salmonella typhimurium* in flocks of chickens in poultry farms in Denmark between 1995 and 1997. The response here is whether *S. typhimurium* is present in a flock, and in the data collected 6.3% of flocks had the disease. At the lowest level, each unit represents a flock of chickens. The basic data have a simple hierarchical structure as each flock is kept in a house on a farm until slaughter. As flocks live for a short time before they are slaughtered several flocks will stay in the same house each year. The hierarchy is as follows 10 127 child flocks within 725 houses on 304 farms.

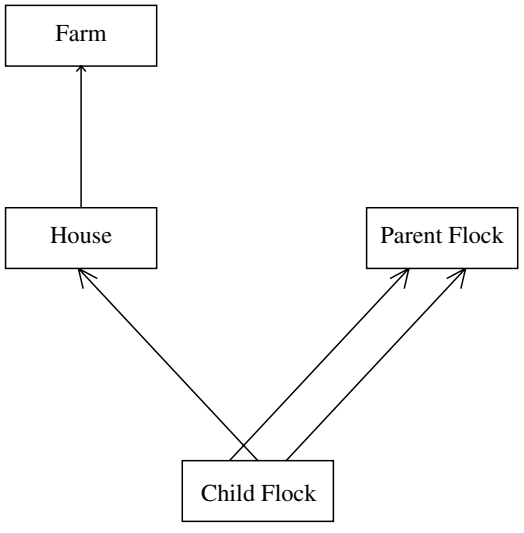

Figure 3 Classification diagram for the Danish poultry model

Each flock is created from a mixture of parent flocks (up to 6) of which there are 200 in Denmark and so we have a crossing between the child flock hierarchy and the multiple membership parent flock classification. The classification diagram is shown in Figure 3. We also know the exact makeup of each child flock (in terms of parent flocks) and so can use these as weights for each of the parent flocks. We are interested in assessing how much of the variability in salmonella incidence can be attributed to houses, farms and parent flocks.

There are also four hatcheries in which all the eggs from the parent flocks are hatched. We will therefore fit a variance components model that allows for different average rates of Salmonella for each year with hatchery included in the fixed part as follows:

$$\begin{aligned} \text{salmonella}_i &\sim \text{Bernouilli}(\pi_i) \\ \text{logit}(\pi_i) &= \beta_0 + Y96_i * \beta_1 + Y97_i * \beta_2 + \text{hatch2}_i * \beta_3 + \text{hatch3}_i * \beta_4 \\ &\quad + \text{hatch4}_i * \beta_5 + u^{(2)}_{\text{House}(i)} + u^{(3)}_{\text{Farm}(i)} + \sum_{j \in P.\text{flock}(i)} w^{(4)}_{i,j} u^{(4)}_j \qquad (9) \\ u^{(2)}_{\text{House}(i)} &\sim N(0, \sigma^2_{u(2)}) \qquad u^{(3)}_{\text{Farm}(i)} \sim N(0, \sigma^2_{u(3)}) \qquad u^{(4)}_j \sim N(0, \sigma^2_{u(4)}) \end{aligned}$$

Rasbash and Browne (2001) considered a frequentist analysis for this problem and used quasi-likelihood methods. They found them to be numerically very unstable for this problem. Here we will instead concentrate on a Bayesian analysis and compare the MH-Gibbs hybrid algorithm from this paper (programmed in MLwiN) with the adaptive rejection (AR) method (Gilks and Wild, 1992) used in the WinBUGS package. To fit this model in a Bayesian framework we need to include priors for the three variance parameters and the fixed effects. As we have no prior information we will use 'diffuse' priors as defined earlier.

The results of fitting model 9 using both these MCMC methods can be seen in Table 3. The MCMC results for both methods were based on a run of 50 000 iterations after a

Table 3 Results for the Danish poultry example

Parameter	MH estimates	AR estimates
Intercept (β_0)	−2.329 (0.216)	−2.331 (0.208)
1996 effect (β_1)	−1.238 (0.165)	−1.242 (0.164)
1997 effect (β_2)	−1.159 (0.194)	−1.163 (0.193)
Hatchery 2 effect (β_3)	−1.730 (0.259)	−1.733 (0.255)
Hatchery 3 effect (β_4)	−0.201 (0.247)	−0.200 (0.252)
Hatchery 4 effect (β_5)	−1.056 (0.381)	−1.054 (0.380)
Parent flock variance ($\sigma^2_{u(4)}$)	0.884 (0.182)	0.890 (0.181)
Farm variance ($\sigma^2_{u(3)}$)	0.922 (0.203)	0.924 (0.193)
House variance ($\sigma^2_{u(2)}$)	0.199 (0.112)	0.202 (0.113)

burn-in of 20 000, as we used arbitrary starting values and so the chain took some time to converge.

From Table 3 we can see close agreement between the two methods, which is to be expected as they are fitting exactly the same model. As reported in Browne and Draper (2000) for other logistic regression problems, which method is preferable is a balance between the speed of the MH–Gibbs method and the reduced Markov chain autocorrelation of the AR method. Here the MH method took 2 h 2 min whilst WinBUGS took 8 h 48 min. Although for the fixed effects the expected required run lengths based on the Raftery–Lewis diagnostic (Raftery–Lewis, 1992) were 2–3 times longer for the MH method, the worst mixing parameter was the between house variance. This parameter is updated via a Gibbs sampling step in each method and therefore has similar expected run lengths.

Examining the model estimates we can see here that there are large effects for the year the chickens were born and for hatchery. There is also a large variability for both the parent flock effects and the farm effects, which are of similar magnitude. There is less variability between houses within farms.

6.1 Co-linearity of random effects

In this example we have fitted two nested classifications (houses within farms) that are crossed with a multiple membership classification (parent flocks). When we consider two fixed effects that are highly correlated then we generally discard one of the two from the model as there is confounding of the effects. The same can be true of sets of random effects particularly in nested models. An extreme case is the often complete confounding between teachers and classes in education. If each teacher teaches one class then we cannot differentiate the variation in the dataset that is due to the teacher from the variation that is due to the class. Even when a few teachers teach two classes or some classes have two teachers it would be overly ambitious to try and fit both sets of random effects. Even when situations are less extreme, for example two classes in every school, then there may still be problems, particularly if the lower (nested) level does not have significant variation. This is often true in binary response models where the response has a limited (two) number of values.

MCMC methods will quickly identify this as the variance chains will show poor mixing properties and high negative cross-chain correlations. In our example 39.4% of farms have only one house (and a further 27% of farms have only two houses) but there appears to be enough information in the dataset to separate the effects of the houses and the farms (the

cross chain correlation between $\sigma^2_{u(2)}$ and $\sigma^2_{u(3)}$ is only -0.19) although the house variance does show worse mixing than the other parameters.

6.2 Complex random effects

The model described by (9) is essentially a variance components model but we could fit a model that has complex variation at one of the higher classifications. To illustrate this we will modify the farm classification variance to account for different variability between years at the farm classification, that is we replace the simple farm classification random effects, $u^{(3)}_{\text{Farm}(i)}$ with 3 sets of effects one for each year. Our expanded model is then as follows:

$$\begin{aligned}
\text{salmonella}_i &\sim \text{Bernouilli}(\pi_i) \\
\text{logit}(\pi_i) &= \beta_0 + Y96_i * \beta_1 + Y97_i * \beta_2 + \text{hatch2}_i * \beta_3 \\
&\quad + \text{hatch3}_i * \beta_4 + \text{hatch4}_i * \beta_5 + u^{(2)}_{\text{House}(i)} + Y95_i * u^{(3)}_{\text{Farm}(i),0} \\
&\quad + Y96_i * u^{(3)}_{\text{Farm}(i),1} + Y97_i * u^{(3)}_{\text{Farm}(i),2} + \sum_{j \in P.\text{flock}(i)} w^{(4)}_{i,j} u^{(4)}_j \\
u^{(2)}_{\text{House}(i)} &\sim N(0, \sigma^2_{u(2)}), \quad u^{(3)}_{\text{Farm}(i)} \sim N_3(0, \Sigma_{u(3)}), \quad u^{(4)}_j \sim N(0, \sigma^2_{u(4)})
\end{aligned} \tag{10}$$

The farm classification variance is now a matrix and so in a Bayesian formulation we need to set a prior for this matrix. Following the example of Spiegelhalter *et al.* (2000) in their birats example we use a vaguely informative Wishart prior with parameters, $S = I_3$ (the 3×3 identity matrix) and $\nu = 3$. For the fixed effects and other variances we use the same priors as in model (9).

The parameter estimates for this extended model for both the MH and AR methods are given in Table 4. Again both methods give similar estimates as would be expected and this time the MH method takes 2 h 16 min as opposed to 10 h 54 min for the AR method. The mixing properties of the Markov chain were similar to the last model; the MH method giving expected run lengths generally 2–3 times greater for fixed effects but again the worst mixing

Table 4 Estimates for the parameters in model 10

Parameter	MH estimates (s.e.)	AR estimates (s.e.)
Intercept (β_0)	−2.560 (0.234)	−2.573 (0.237)
1996 effect (β_1)	−1.188 (0.272)	−1.158 (0.270)
1997 effect (β_2)	−1.122 (0.292)	−1.120 (0.286)
Hatchery 2 effect (β_3)	−1.806 (0.270)	−1.801 (0.270)
Hatchery 3 effect (β_4)	−0.146 (0.248)	−0.145 (0.254)
Hatchery 4 effect (β_5)	−1.053 (0.393)	−1.059 (0.394)
Parent flock variance ($\sigma^2_{u(4)}$)	0.890 (0.184)	0.892 (0.188)
Farm year 95 variance ($\Sigma_{u(3)}[0,0]$)	1.447 (0.329)	1.476 (0.327)
Farm 95/96 covariance ($\Sigma_{u(3)}[0,1]$)	0.439 (0.276)	0.435 (0.265)
Farm 95/97 covariance ($\Sigma_{u(3)}[0,2]$)	0.479 (0.270)	0.478 (0.274)
Farm year 96 variance ($\Sigma_{u(3)}[1,1]$)	1.427 (0.535)	1.368 (0.516)
Farm 96/97 covariance ($\Sigma_{u(3)}[1,2]$)	0.664 (0.368)	0.661 (0.356)
Farm year 97 variance ($\Sigma_{u(3)}[2,2]$)	1.353 (0.498)	1.370 (0.490)
House variance ($\sigma^2_{u(2)}$)	0.290 (0.124)	0.281 (0.129)

parameter was the house classification variance with longer expected run length for the AR method.

It can be seen that the fixed effects estimates for this model are fairly similar to model 9. It is interesting to see that all the covariances in the farm level variance matrix are positive. This suggests that after adjusting for other factors, if a farm has an incidence of salmonella in 1995 then it is more likely to have an incidence again in 1996 and in 1997. In fact the corresponding correlation estimates are 0.30, 0.34 and 0.48 respectively, showing that, in particular, there is a fairly strong correlation between salmonella infection in farms in 1996 and 1997.

7 Example 2: Belgium household migration

Goldstein *et al.* (2000) consider the problem of assessing the propensity of individuals to move household by considering a longitudinal dataset. This dataset contains the addresses of all inhabitants, over a 5 year period, in the town of Charleroi, recorded every 6 months. The response is the average duration that an individual has stayed in a household based on all the households they have been members of up to and including the current household. We consider at each occasion the individual to be a multiple member of all the previous households including the current household, as their membership of previous households could influence their current household. Each of the multiple membership units are weighted equally, as in Goldstein *et al.* (2000), although other weightings may be valid, for example the current household may be given larger weight.

A household is here defined as a group of people sharing a dwelling for a period of time. Anybody leaving or entering a household at a particular time constitutes a change of household and the current household ceases to exist, being replaced by one or more new households (see Goldstein *et al.* (2000) for more details). We do not include in the dataset the household that the individuals belong to at the end of the 5 year period as the full length of stay in this household is not known. Further research could consider fitting the dataset with these terms included as censored observations. The classification diagram for this dataset can be seen in Figure 4.

The model we will fit is a variance components model with several individual level covariates as follows:

$$\begin{aligned}\text{duration}_i = {} & \beta_0 + \text{gender}_i * \beta_1 + \text{householdsize}_i * \beta_2 + (\text{age}_i - 30) * \beta_3 \\ & + \text{spouse}_i * \beta_4 + \text{child}_i * \beta_5 + \text{married}_i * \beta_6 + \text{Belgian}_i * \beta_7 \\ & + u^{(2)}_{\text{individual}(i)} + \sum_{j \in \text{household}(i)} w^{(3)}_{i,j} u^{(3)}_j + e_i \qquad (11)\end{aligned}$$

$$u^{(2)}_{\text{individual}(i)} \sim N(0, \sigma^2_{u(2)}), \quad u^{(3)}_j \sim N(0, \sigma^2_{u(3)}), \quad e_i \sim N(0, \sigma^2_e)$$

There are 66 624 occasions measured within 37 759 individuals and in total 26 852 households with each individual being a member of up to 10 households.

To fit this model in a Bayesian framework we need to include priors for the three variance parameters and the fixed effects. As we have no prior information we will use 'diffuse' priors as defined earlier.

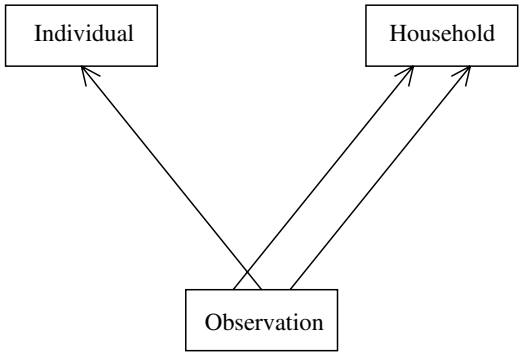

Figure 4 Classification diagram for the Belgium household model

Table 5 Results for the Charleroi population dataset

Parameter	IGLS estimate (s.e.)	MCMC estimate (s.e.)
Intercept (β_0)	1.4780 (0.0101)	1.4790 (0.0093)
Gender (β_1)	−0.0055 (0.0050)	−0.0059 (0.0050)
Size of household (β_2)	−0.0575 (0.0010)	−0.0577 (0.0010)
Age–30 years (β_3)	0.0055 (0.0002)	0.0055 (0.0002)
Spouse (β_4)	0.0418 (0.0031)	0.0417 (0.0032)
Child (β_5)	0.0467 (0.0036)	0.0463 (0.0035)
Is Married (β_6)	0.0472 (0.0039)	0.0468 (0.0038)
Is Belgian national (β_7)	0.0148 (0.0054)	0.0147 (0.0052)
Between household variance ($\sigma^2_{u(3)}$)	1.3340 (0.0122)	1.3340 (0.0126)
Between individual variance ($\sigma^2_{u(2)}$)	0.1520 (0.0015)	0.1520 (0.0015)
Residual variance (σ^2_e)	0.0028 (0.00003)	0.0027 (0.00003)

The results for fitting this model using a special case of the MCMC algorithm in Appendix A can be seen in Table 5. Goldstein *et al*. (2000) used the method of Rasbash and Goldstein (1994) with the IGLS algorithm to fit this dataset. This was possible because although the dataset is very large, the data can be split into disjoint non-intersecting subsets and so the memory required is substantially reduced. We reproduce these estimates for comparison with the MCMC method in Table 5.

The MCMC algorithm (in MLwiN) takes a long time initially calculating the indexing arrays (13 min on a 733 MHz PC), due to the huge numbers of random effects, but this then enables the sampler to run faster (70 iterations a minute). The IGLS algorithm in this example takes about 3 minutes per iteration and needs five iterations to converge.

Table 5 shows that the results for the two methods are almost identical. This is to be expected as the variance estimates are based on large numbers of higher level units and so their distributions are fairly symmetric. From the estimates we can see that people stay longer in the same household if they are older, if they are children or spouses rather than heads of household, if they are married or if they are Belgian nationals.

8 Example 3: Scottish lip cancer data

The Scottish lip cancer dataset (Clayton and Kaldor, 1987) has been analysed many times using many different models that attempt to account for spatial random variation. The response variable is the observed count of male lip cancer in the period 1975–80, by region, for the 56 regions of Scotland. Research has focused on the effect of sun exposure using the surrogate measure percentage of the workforce working in outdoor occupations. We can fit a spatial model into the MMMC framework by considering the areas as one classification and the neighbours as another multiple membership classification. The model is then as follows:

$$\begin{gathered}\text{obs}_i \sim \text{Poisson}(\lambda_i) \\ \log_e(\lambda_i) = \log_e(\exp_i) + \beta_0 + X_i\beta_1 + u^{(2)}_{\text{Area}(i)} + \sum_{j \in \text{Neighbour}(i)} w^{(3)}_{i,j} u^{(3)}_j \\ u^{(2)}_{\text{Area}(i)} \sim N(0, \sigma^2_{u(2)}) \qquad u^{(3)}_j \sim N(0, \sigma^2_{u(3)})\end{gathered} \tag{12}$$

Here the weights $w^{(3)}_{i,j} = 1/r_i$ where r_i is the number of neighbouring regions for region i. The one predictor variable X_i is the percentage of the workforce involved in agriculture, fishing or forestry (divided by 10). This model can be represented in a classification diagram as shown below.

The model as it stands can be fitted using either quasi-likelihood methods in a frequentist setting (which we do not consider here) or MCMC in a Bayesian framework. Again to complete a Bayesian formulation of this model we require the addition of prior distributions. In the comparison experiment that follows in the next section we will again use 'diffuse' priors as described earlier.

Langford *et al.* (1999) use quasi-likelihood methods and extend this model by incorporating a covariance between the two sets of random effects, $u^{(2)}$ and $u^{(3)}$. This is possible for certain random effects models where the two classifications use the same set of classification units, so the concept of a correlation between the random effects has a meaning. This model is however not part of the general MMMC framework which assumes conditional

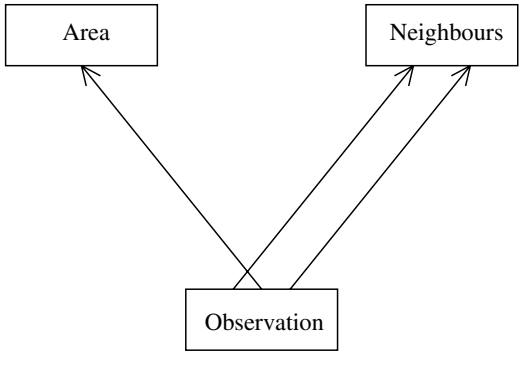

Figure 5 Classification diagram for the lip cancer model

independence between the random effects in different classifications and so will not be considered here.

8.1 Alternative spatial models to MMMC

The standard Bayesian spatial Poisson models are based on the conditional autoregressive (CAR) prior (Besag *et al.*, 1991) that was originally used in image analysis. These priors were used on the Scottish lip cancer dataset (Breslow and Clayton, 1993) and the model can be written as follows:

$$\begin{aligned} \text{obs}_i &\sim \text{Poisson}(\lambda_i) \\ \log_e(\lambda_i) &= \log_e(\exp_i) + \beta_0 + X_i\beta_1 + u_i + v_i \\ u_i &\sim N(0, \sigma_u^2) \qquad v_i \sim N(\bar{v}_i, \sigma_v^2/r_i) \\ \text{where } \bar{v}_i &= \sum_{j \in \text{Neighbour}(i)} v_j/r_i \end{aligned} \tag{13}$$

Here r_i is the number of neighbouring regions for region i. To fit a CAR model using MCMC methods, again prior distributions are required and we use the same 'diffuse priors' as in the MMMC model. This model is in fact similar to the MMMC model with two sets of random effects, except that spatial correlation is achieved through the variance structure rather than through the multiple membership relationship and so the neighbourhood random effects are not independent. It is also possible given additional data, such as the distances between neighbours, to incorporate this information either in the weight matrix, as in the MMMC model, or in the CAR model framework, although this is not considered in this example.

8.2 Comparison of models

To compare the MMMC and CAR models we performed a cross validation study of the lip cancer dataset. For each region in turn we set its actual deaths to be a missing value and fitted both the MMMC and CAR models to the data. At each iteration the missing value was imputed and consequently we obtained an estimate of the posterior distribution of the (unknown) observed number of cases for the region. As an MCMC method for the CAR model is not available currently in the MLwiN package we instead used the AR algorithm in the WinBUGS package for both models.

Both methods appeared to give reasonable interval estimates and Table 6 gives the results for regions where the interval estimates from one of the methods did not contain the actual number of deaths. The MMMC method gave 95% intervals that contained the true value 53 out of 56 times and the CAR method 54 out of 56 times, but further work involving cross validation needs to be done here.

In terms of point estimates, Table 7 contains some additional information about the two methods. Both the mean and median estimates were (on average) closer using the CAR model and this method also had smaller average intervals. However, the MMMC models mean squared difference (MSD) estimates were inflated by one or two poor estimates. Considering the 56 regions individually we find that the MMMC mean estimate is closer nearly 50% of the time. We will consider a more extensive cross-validation study and

Table 6 A list of regions whose intervals do not contain the actual number of deaths for the lip cancer dataset

Area	Actual	Expected	MMMC interval	CAR interval
Moray	26	8.11	(3,24)	(6,40)
Kirkcaldy	19	15.47	(1,18)	(5,41)
Dundee	6	19.62	(6,57)	(8,59)
Annandale	0	4.16	(1,19)	(1,15)

Table 7 Some further comparisons between the two models fit to the lip cancer dataset

	MMMC model	CAR model
Mean squared difference (means)	155.86	57.21
Mean squared difference (medians)	114.07	45.17
Mean interval width	26.00	22.50
Closer estimate (mean)	27/56	29/56

compare results with the Langford model in a further paper, along with comparisons using the DIC (Spiegelhalter *et al*., 2001) criterion.

9 Conclusions

In this paper we have extended the now standard multilevel modelling framework to encompass both crossed random effects and multiple membership random effects. We have developed notation based on mappings that allows these models to be easily specified, and have given several examples that show the power and scope of this extended family of models. We have shown how a Bayesian analysis can be easily implemented using MCMC based algorithms, and how these algorithms do not have the disadvantage of frequentist maximum likelihood based methods that need large amounts of memory when the model structure becomes complex, and the datasets become large. However it is also important to note that in some examples (like the Belgian population example studied here) that are almost nested, through clever partitioning of the data the maximum likelihood based methods can still be applied and may be faster. The models were all programmed in a development version of the MLwiN package (Rasbash *et al*., 2000b) which will be available to the user community in 2002. More information on MLwiN is available at the multilevel modelling project website (http://multilevel.ioe.ac.uk/).

10 Extensions

In the models in this paper we consider a simple variance at the lowest level of the model. Browne *et al*. (2001) show how to incorporate complex level 1 variation within the MCMC algorithm to allow for heteroscedasticity in normal response models. This extension to the

simple multilevel modelling framework can easily be combined with the cross-classified and multiple membership models that are covered in this paper.

In the Belgian example we have omitted the censored data of the last household that individuals belong to. The data are also all left censored as it is assumed that every individual starts in their first household 6 months before the first data collection point. It would be useful to reanalyse this example while also modelling the censored data using a survival type model. This could easily be implemented in an MCMC framework by imputing the true data based on the censored data and a known prior distribution.

In the lip cancer example we performed some initial comparisons between an MMMC model that can be applied to spatial data and the standard CAR spatial models. We intend to extend this comparison work while considering Bayesian formulations for the MMMC model with correlated random effects as described in Langford *et al*. (1999).

Acknowledgements

We are grateful to (a) the ESRC for financial support, (b) Mariann Chriel for supplying the Danish poultry data (c) Michel Poulain for supplying the Belgian household data and (d) to Leo Knorr-Held and Min Yang for references and comments on early drafts of this paper. 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.

References

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.

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

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

Browne WJ, Draper D (2001) A comparison of Bayesian and likelihood methods for fitting multilevel models (*Submitted*).

Browne WJ, Draper D, Goldstein H, Rasbash J (2001) Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computational Statistics and Data Analysis* (to appear).

Bryk AS, Raudenbush SW (1992) *Hierarchical linear models: applications and data analysis methods*. London: Sage.

Bryk AS, Raudenbush SW, Seltzer M, Congdon R (1988) *An introduction to HLM: computer program and user's guide*. Chicago: University of Chicago Department of Education.

Clayton DG (1996) Generalized linear mixed models. In: Gilks WR, Richardson S, Spiegelhalter DJ, eds. *Markov Chain Monte Carlo in practice*. London: Chapman and Hall.

Clayton DG, Kaldor J (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671–81.

Clayton DG, Rasbash J (1999) Estimation in large crossed random-effects models by data augmentation. *Journal of the Royal Statistics Society, Series A* **162**, 425–36.

Draper D (2001) *Bayesian hierarchical modeling*. New York: Springer-Verlag (forthcoming).

Gelfand AE, Smith AFM (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gilks WR, Wild P (1992) Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society, Series C* **41**, 337–48.

Goldstein H (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43–56.

Goldstein H (1995) *Multilevel statistical models*. 2nd edn. London: Edward Arnold.

Goldstein H, Rasbash J (1996) Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A* **159**, 505–13.

Goldstein H, Rasbash J, Browne WJ, Woodhouse G, Poulain M (2000) Multilevel models in the study of dynamic household structures. *European Journal of population* **16**, 373–87.

Harville D (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–40.

Hill PW, Goldstein H (1998) Multilevel modelling of educational data with cross-classification and missing identification of units. *Journal of Educational and Behavioral Statistics* **23**, 117–28.

Langford IH, Leyland AH, Rasbash J, Goldstein H, Day RJ, McDonald A-L (1999) Multilevel modelling of area-based health data. In: Lawson A, Biggeri A, Bohning D, Lesaffre E, Viel J-F, Bertollini R, eds. *Disease mapping and risk assessment for public health*. Chichester: Wiley.

Lee Y, Nelder JA (2001) Hierarchical generalised linear models: a synthesis of generalised linear models, random-effects models, and structured dispersions. *Biometrika* (to appear).

Nelder J, Wedderburn R (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–84.

Pan J-X, Thompson R (2000) Generalized linear mixed models: An improved estimating procedure. In: Bethlehem JG, van der Heijden PGM, eds. *COMPSTAT: Proceedings in computational statistics*, 2000. Heidelberg: Physica-Verlag, 373–78.

Raferty AE, Lewis SM (1992) How many iterations in the Gibbs sampler? In Bernardo JM, Berger JO, David AP, Smith AFM eds. *Bayesian Statistics 4*. Oxford: Oxford University Press.

Rasbash J (1989) *The ML2 software package*. London: Institute of Education, University of London.

Rasbash J, Goldstein H (1994) Efficient analysis of mixed hierarchical and crossed random structures using a multilevel model. *Journal of Behavioural Statistics* **19**, 337–50.

Rasbash J, Browne WJ (2001) Non-hierarchical multilevel models. In: Leyland A, Goldstein H, eds. *Multilevel modelling of health statistics*. Chichester: Wiley.

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

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

Raudenbush SW (1993) A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Education Statistics* **18**, 321–50.

Schafer JL (1997) *Analysis of incomplete multivariate data*. London: Chapman and Hall.

Spiegelhalter DJ, Thomas A, Best NG (2000) *WINBUGS version 1.3: user manual*. Cambridge: Medical Research Council Biostatistics Unit.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2001) Bayesian measures of model complexity and fit. Submitted. Available from the author.

Tanner M, Wong W (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–50.

Appendix A: details of MCMC algorithm for normal model in Section 5.1

For the Normal model in Section 5.1 we use a Gibbs sampling algorithm which involves random draws from the following six full-conditional distributions.

Step 1: Update the fixed effects parameter vector β from its full conditional distribution which is multivariate normal with dimension p_f

$$p(\beta \mid y, u^{(2)}, u^{(3)}, \Sigma_{u(2)}, \Sigma_{u(3)}, \sigma_e^2) \sim N_{p_f}(\widehat{\beta}, \widehat{D})$$

$$\text{where } \widehat{D} = \left[\sum_{i=1}^{N} \frac{(X_i)^T X_i}{\sigma_e^2} + S_p^{-1}\right]^{-1} \text{ and } \widehat{\beta} = \widehat{D}\left[\sum_i \frac{(X_i)^T d_i}{\sigma_e^2} + S_p^{-1}\mu_p\right]$$

$$\text{where } d_i = y_i - Z_i^{(2)} u_{C_2(i)}^{(2)} - \sum_{j \in C_3(i)} w_{i,j}^{(3)} Z_i^{(3)} u_j^{(3)} \tag{14}$$

Step 2: Update the 'simple' classification 2 units, $u_k^{(2)}$, from their multivariate normal full conditional distribution with dimension p_2

$$p(u_k^{(2)} \mid y, \beta, u^{(3)}, \Sigma_{u(2)}, \Sigma_{u(3)}, \sigma_e^2) \sim N_{p_2}(\widehat{u}_k^{(2)}, \widehat{D}_k^{(2)})$$

$$\text{where } \widehat{D}_k^{(2)} = \left[\sum_{i, C_2(i)=k} \frac{(Z_i^{(2)})^T Z_i^{(2)}}{\sigma_e^2} + \Sigma_{u(2)}^{-1}\right]^{-1} \text{ and } \widehat{u}_k^{(2)} = \widehat{D}_k^{(2)}\left[\sum_{i, C_2(i)=k} \frac{(Z_i^{(2)})^T d_i^{(2)}}{\sigma_e^2}\right]$$

$$\text{where } d_i^{(2)} = y_i - X_i\beta - \sum_{j \in C_3(i)} w_{i,j}^{(3)} Z_i^{(3)} u_j^{(3)} \tag{15}$$

Step 3: Update the 'multiple membership' classification 3 units, $u_k^{(3)}$, from their multivariate normal full conditional distribution with dimension p_3

$$p(u_k^{(3)} \mid y, \beta, u^{(2)}, \Sigma_{u(2)}, \Sigma_{u(3)}, \sigma_e^2) \sim N_{p_2}(\widehat{u}_k^{(3)}, \widehat{D}_k^{(3)})$$

$$\text{where } \widehat{D}_k^{(3)} = \left[\sum_{i,k \in C_3(i)} \frac{(w_{i,k}^{(3)})^2 (Z_i^{(3)})^T Z_i^{(3)}}{\sigma_e^2} + \Sigma_{u(3)}^{-1}\right]^{-1} \text{ and } \widehat{u}_k^{(3)} = \widehat{D}_k^{(3)}\left[\sum_{i,k \in C_3(i)} \frac{w_{i,k}^{(3)} (Z_i^{(3)})^T d_{i,k}^{(3)}}{\sigma_e^2}\right]$$

$$\text{where } d_{i,k}^{(3)} = y_i - X_i\beta - Z_i^{(2)} u_{C_2(i)}^{(2)} \tag{16}$$

Step 4: Update the lowest level variance σ_e^2 by drawing from the Gamma full conditional distribution for $1/\sigma_e^2$

$$p(1/\sigma_e^2 \mid y, \beta, u^{(2)}, u_k^{(3)}, \Sigma_{u(2)}, \Sigma_{u(3)}) \sim \text{Gamma}\left[\frac{N + \nu_e}{2}, \frac{1}{2}\sum_i e_i^2 + \nu_e s_e^2\right] \tag{17}$$

Step 5: Update the classification 2 variance matrix, $\Sigma_{u(2)}$. Expressed as a Wishart draw of $\Sigma_{u(2)}^{-1}$ the full conditional is

$$p(\Sigma_{u(2)}^{-1} \mid y, \beta, u^{(2)}, u_k^{(3)}, \Sigma_{u(3)}, \sigma_e^2) \sim W_{p_2}\left[n_2 + \nu_2, \left(\sum_{j=1}^{n_2} u_j^{(2)} (u_j^{(2)})^T + S_2\right)^{-1}\right] \tag{18}$$

where n_2 is the number of classification 2 units and p_2 is the number of rows or columns in $\Sigma_{u(2)}$. An improper uniform prior on $\Sigma_{u(2)}$ corresponds to the choice $(\nu_2, S_2) = (-p_2 - 1, 0)$.

Step 6: Update the classification 3 variance matrix, $\Sigma_{u(3)}$. Expressed as a Wishart draw of $\Sigma_{u(3)}^{-1}$ the full conditional is

$$p(\Sigma_{u(3)}^{-1} \mid y, \beta, u^{(2)}, u_k^{(3)}, \Sigma_{u(2)}, \sigma_e^2) \sim W_{p_3}\left[n_3 + \nu_3, \left(\sum_{j=1}^{n_3} u_j^{(3)}(u_j^{(3)})^T + S_3\right)^{-1}\right] \tag{19}$$

where n_3 is the number of classification 3 units and p_3 is the number of rows or columns in $\Sigma_{u(3)}$. An improper uniform prior on $\Sigma_{u(3)}$ corresponds to the choice $(\nu_3, S_3) = (-p_3 - 1, 0)$.

Appendix B: details of MCMC algorithm for binomial and Poisson models in Section 5.2

The above algorithm for the general three classification normal response model can be easily adapted to both binomial and Poisson response models. Here we will use univariate Metropolis updates for the fixed effects and sets of classification 2 and 3 residuals as the full conditionals do not have standard forms. For ease of notation in the full conditionals that follow we define $\pi_i = X_i\beta + Z_i^{(2)}u_{C_2(i)}^{(2)} + \sum_{j\in C_3(i)} w_{i,j}^{(3)} Z_i^{(3)} u_j^{(3)}$.

Step 1: Update β using univariate random walk Metropolis at time t as follows: for $l = 1, \ldots, p_f$ and with $\beta_{(-l)}$ signifying the beta vector without component l

$$\begin{aligned}\beta_l(t) &= \beta_l^* \text{ with probability } \min\left[1, \frac{p(\beta_l^* \mid y, u^{(2)}, u^{(3)}, \beta_{(-l)})}{p(\beta_l(t-1) \mid y, u^{(2)}, u^{(3)}, \beta_{(-l)})}\right] \\ &= \beta_l(t-1) \text{ otherwise}\end{aligned} \tag{20}$$

where $\beta_l^* \sim N(\beta_l(t-1), \sigma_{1l}^2)$ and

$$p(\beta_l \mid y, u^{(2)}, u^{(3)}, \beta_{(-l)}) \propto \prod_i [1 + e^{-\pi_i}]^{-y_i}[1 + e^{\pi_i}]^{y_i - m_i} \tag{21}$$

for the logistic binomial and

$$p(\beta_l \mid y, u^{(2)}, u^{(3)}, \beta_{(-l)}) \propto \prod_i [e^{\pi_i}]^{y_i} e^{-e^{\pi_i}} \tag{22}$$

for the Poisson regression.

Step 2: Update the 'simple' classification 2 units, $u_k^{(2)}$, using univariate random walk Metropolis at time t as follows: for $k = 1, \ldots, n_2$ and $l = 1, \ldots, p_2$ and with $u_{k(-l)}^{(2)}$ signifying the $u_k^{(2)}$ vector without component l

$$\begin{aligned}u_{k(l)}^{(2)}(t) &= u_{k(l)}^{(2)*} \text{ with probability } \min\left[1, \frac{p(u_{k(l)}^{(2)*} \mid y, u_{k(-l)}^{(2*)}, u^{(3)}, \beta, \Sigma_{u(2)})}{p(u_{k(l)}^{(2)}(t-1) \mid y, u_{k(-l)}^{(2)}(t-1), u^{(3)}, \beta, \Sigma_{u(2)})}\right] \\ &= u_{k(l)}^{(2)}(t-1) \text{ otherwise}\end{aligned} \tag{23}$$

where $u^{(2)*}_{k(l)} \sim N(u^{(2)}_{k(l)}(t-1), \sigma^2_{2kl})$ and

$$p(u^{(2)}_{k(l)} \mid y, u^{(2)}_{k(-l)}, u^{(3)}, \beta, \Sigma_{u(2)}) \propto \exp(-\frac{1}{2}(u^{(2)}_k)^T \Sigma^{-1}_{u(2)} u^{(2)}_k) \prod_{i,k=C_2(i)} [1+e^{-\pi_i}]^{-y_i}[1+e^{\pi_i}]^{y_i-m_i} \tag{24}$$

for the logistic binomial and

$$p(u^{(2)}_{k(l)} \mid y, u^{(2)}_{k(-l)}, u^{(3)}, \beta, \Sigma_{u(2)}) \propto \exp(-\frac{1}{2}(u^{(2)}_k)^T \Sigma^{-1}_{u(2)} u^{(2)}_k) \prod_{i,k=C_2(i)} [e^{\pi_i}]^{y_i} e^{-e^{\pi_i}} \tag{25}$$

for the Poisson regression.

Step 3: Update the 'multiple membership' classification 3 units, $u^{(3)}_k$, using univariate random walk Metropolis at time t as follows: for $k = 1, \ldots, n_3$ and $l = 1, \ldots, p_3$ and with $u^{(3)}_{k(-l)}$ signifying the $u^{(3)}_k$ vector without component l

$$u^{(3)}_{k(l)}(t) = u^{(3)*}_{k(l)} \text{ with probability } \min\left[1, \frac{p(u^{(3)*}_{k(l)} \mid y, u^{(3*)}_{k(-l)}, u^{(2)}, \beta, \Sigma_{u(3)})}{p(u^{(3)}_{k(l)}(t-1) \mid y, u^{(3)}_{k(-l)}(t-1), u^{(2)}, \beta, \Sigma_{u(3)})}\right] \tag{26}$$

$$= u^{(3)}_{k(l)}(t-1) \text{ otherwise}$$

where $u^{(3)*}_{k(l)} \sim N(u^{(3)}_{k(l)}(t-1), \sigma^2_{3kl})$ and

$$p(u^{(3)}_{k(l)} \mid y, u^{(3)}_{k(-l)}, u^{(2)}, \beta, \Sigma_{u(3)}) \propto \exp\left(-\frac{1}{2}(u^{(3)}_k)^T \Sigma^{-1}_{u(3)} u^{(3)}_k\right) \prod_{i,k\in C_3(i)} [1+e^{-\pi_i}]^{-y_i}[1+e^{\pi_i}]^{y_i-m_i} \tag{27}$$

for the logistic binomial and

$$p(u^{(3)}_{k(l)} \mid y, u^{(3)}_{k(-l)}, u^{(2)}, \beta, \Sigma_{u(3)}) \propto \exp\left(-\frac{1}{2}(u^{(3)}_k)^T \Sigma^{-1}_{u(3)} u^{(3)}_k\right) \prod_{i,k\in C_3(i)} [e^{\pi_i}]^{y_i} e^{-e^{\pi_i}} \tag{28}$$

for the Poisson regression.

Step 4 of the previous algorithm is redundant as the lowest level variance is defined by the response type and is not estimated. Steps 5 and 6 of the algorithm are the same as in the algorithm in Appendix A.