Statistical Modelling 2009; **9**(**3**): 173–197

Multilevel models with multivariate mixed response types

Harvey Goldstein[1], James Carpenter[2], Michael G Kenward[2] and Kate A Levin[3]
[1]University of Bristol
[2]London School of Hygiene and Tropical Medicine
[3]University of Edinburgh

Abstract: We build upon the existing literature to formulate a class of models for multivariate mixtures of Gaussian, ordered or unordered categorical responses and continuous distributions that are not Gaussian, each of which can be defined at any level of a multilevel data hierarchy. We describe a Markov chain Monte Carlo algorithm for fitting such models. We show how this unifies a number of disparate problems, including partially observed data and missing data in generalized linear modelling. The two-level model is considered in detail with worked examples of applications to a prediction problem and to multiple imputation for missing data. We conclude with a discussion outlining possible extensions and connections in the literature. Software for estimating the models is freely available.

Key words: Box–Cox transformation; data augmentation; data coarsening; latent Gaussian model; maximum indicant model; MCMC; missing data; mixed response models; multilevel; multiple imputation; multivariate; normalising transformations; partially known values; prediction; prior-informed imputation; probit model

Received November 2007; **first revision**: March 2008; **second revision**: April 2008; **accepted**: April 2008

1 Introduction

Multilevel models are routinely used for the analysis of data that have a hierarchical structure, such as students nested within schools or repeated measures data with measurement occasions nested within individuals. Goldstein (2003) discusses a wide range of such models with examples. These models have been developed for different univariate response types including Gaussian, unordered and ordered responses.

There exists a wide range of extensions of such models that allow multivariate mixtures of Gaussian, ordered or unordered categorical responses, see for example,

Address for correspondence: Harvey Goldstein, Graduate School of Education, University of Bristol, Bristol, BS8 1JA, UK E-mail: h.goldstein@bristol.ac.uk

 10.1177/1471082X0800900301

Rabe-Hesketh *et al.* (2005), Dunson (2000), Goldstein (2003), Browne (2004), Imai and van Dyk (2005), Asparouhov and Muthen (2007). A number of approaches are based upon maximum likelihood (ML) estimation. Thus, Rabe-Hesketh *et al.* (2005) use numerical integration based upon quadrature to develop a very general framework, to obtain ML estimates, that incorporates many of the features that are discussed in the present paper. Similarly, Asparouhov and Muthen (2007) use the expectation–maximization (EM) algorithm to provide an overlapping framework of models. MCMC procedures are used by Dunson (2000) and Imai and van Dyk (2007) for particular cases of such models.

The aim of this paper is to provide a unifying framework that incorporates all these existing models and to show how this extends to new models and procedures, especially for handling coarsened data (e.g., measurement error, probabilistic linkage and missing data). Our model allows multivariate response variables to exist at all levels of a data hierarchy; these are linked across levels via their associated random effects. We also show how Box–Cox type normalizing transformations for continuous non-Gaussian responses can be incorporated.

Drawing on the existing literature we develop a general MCMC algorithm for fitting our models which has two principal attractions. First, it conveniently allows the inclusion of informative prior information, which is particularly important in certain problems involving missing data, e.g., in probabilistic data linkage. Second, MCMC estimation is computationally efficient for complex models since, unlike numerical integration (and also the EM algorithm where numerical integration is required for the expectation step), computational load increases linearly with the number of parameters. A procedure such as ours also provides a model fitting algorithm that readily allows extensions via the insertion of samplers for additional components or distributions. The properties of our procedures derive from the general properties of MCMC algorithms. Thus, in principle, they will provide full posterior distributions that accommodate all sources of random variation, but they are also subject to problems of 'convergence' and stability to which we return in the discussion section.

A number of special cases of our models can be fitted within some existing software packages such as WINBUGS (Spiegelhalter *et al.*, 1999), MPLUS (Muthen and Muthen 2004) and GLLAMM (Rabe-Hesketh *et al.*, 2001). Nevertheless, it appears that none of these can accommodate the full generality that we propose. Where the set of responses is entirely at the lowest level of the data hierarchy, our model is similar to that of Dunson (2000), although his model is formulated from a more factor analytic viewpoint. We have developed our own software that will handle the core models discussed in this paper. This is freely available for download as a set of compiled Matlab programs (Mathworks, 2004), together with training materials, from http://www.cmm.bristol.ac.uk/.

There are many examples of multivariate data where responses are at more than one level of a data hierarchy. Thus, in longitudinal studies, we may have repeated measures on individuals constituting the lowest level of the hierarchy together with

measures that are constant for each individual at level 2 of the hierarchy (Goldstein, 2003, Chapter 5). An example, which we discuss in a later section, is growth data where there are repeated measures of a variable during a growth period (level 1) and a single measure at adulthood (level 2) when growth has stopped. Fitting all these as responses allows us to estimate the correlations between the adult measure and the parameters that describe growth, such as the mean height and the rate of change of height with age. This in turn provides a flexible procedure for making growth predictions.

Another important case that we discuss in detail is where we have missing data in a multilevel model. Full multiple imputation procedures consider all the variables with missing data as a set of multivariate responses, and if some of these are at different levels of the data hierarchy, this requires the procedures we are considering. We also extend the multiple imputation model to consider data where the values are unknown but we have a probability distribution available for these values so that they become 'partially known'. In that case we treat the probability distribution as a prior within the Bayesian framework.

In all these cases, we may wish to consider non-Gaussian responses that are either discrete or continuous and our procedures include such cases. This relates to work by Yucel (2008). Thus, in the repeated measures case, we may have responses that are on ordered scales such as examination grades or responses to graded attitude scales, and in the missing data case we will often have discrete or other non-Gaussian variables with missing values.

We formulate our model as follows. We first consider creating an underlying set of latent multivariate Gaussian responses; one Gaussian response for each binary, ordered or non-Gaussian continuous variable and a set of Gaussian responses for each multicategory response variable. This reduces the analysis to a multivariate Gaussian model that allows us to apply standard MCMC samplers.

The paper is structured as follows. We start in Section 2 with responses occurring at level 1 and introduce our notation and describe the multivariate Gaussian model. Section 3 describes the procedures for sampling the latent Gaussian variables for non-Gaussian responses, the extension to two levels and missing data. Section 4 gives examples of applications and Section 5 contains the discussion.

2 Model and notation

2.1 Multivariate Gaussian data

Let $j = 1, \ldots, J$ index level 2 units and $i = 1, \ldots, I_j$ index level 1 units, nested within the level 2 units. For example, the level 1 units might be students and the level 2 units schools. The underlying multivariate Gaussian model structure we consider,

for a two-level model, is as follows:

$$\begin{aligned} y_{ij}^{(1)} &= X_{1ij}\beta^{(1)} + Z_{1ij}u_j^{(1)} + e_{ij}^{(1)} \\ y_j^{(2)} &= X_{2j}\beta^{(2)} + Z_{2j}u_j^{(2)} \\ e_{ij}^{(1)} &\sim MVN(0, \Omega_1), \quad u_j = (u_j^{(1)}, u_j^{(2)})^T, \quad u_j \sim MVN(0, \Omega_2). \end{aligned} \tag{2.1}$$

The superscripts denote the level at which a variable is measured or defined. Thus, $y_{ij}^{(1)}$ $(p_1 \times 1)$ contains the (latent or actual) Gaussian responses that are defined at level 1 and $y_j^{(2)}(p_2 \times 1)$ contains the responses that are defined at level 2. Without loss of generality, we assume the same set of predictors for each response at level 1 and likewise at level 2. Let X_{1ij} $(1 \times f_1)$ contain the level 1 predictor variables and $\beta^{(1)}$ $(f_1 \times p_1)$ be the matrix containing the fixed coefficients for these predictors. Similarly, $Z_{1ij}(1 \times q_1)$ is the matrix that contains the predictor variables for the q_1 level 2 random effects denoted by $u_j^{(1)}$ $(q_1 \times p_1)$ for the level 1 responses. The level 1 residuals $e_{ij}^{(1)}$ are calculated by subtracting the current estimate of the linear component of the model from each of the level 1 responses.

Correspondingly, X_{2j} $(1 \times f_2)$ is the vector that contains predictor variables for higher level unit j and $\beta^{(2)}$ $(f_2 \times p_2)$ contains the fixed coefficients. The matrix Z_{2j} $(q_2 \times p_2)$ contains the level 2 random effects for the level 2 responses. Finally, $u_j^{(2)}$ $(q_2 \times p_2)$ is the matrix of level 2 residuals for the level 2 responses and these are correlated with the level 2 residuals for the level 1 responses $u_j^{(1)}$. In the examples of this paper we shall assume that $q_2 = 1$.

Our estimation strategy is to use Gibbs sampling, drawing from the appropriate known posterior conditional distribution, and where, for particular parameters, this does not correspond to a known distribution, we use Metropolis–Hastings (MH) sampling. We will assume and describe appropriate default priors.

2.2 Fitting a multivariate Gaussian model with level 1 responses

We now consider model (2.1) but without the second line, that is a model with only level 1 responses:

$$\begin{aligned} y_{ij} &= X_{1ij}\beta + z_{ij}u_j + e_{ij} \\ e_{ij} &\sim MVN(0, \Omega_1), \quad u_j \sim MVN(0, \Omega_2) \end{aligned}$$

where, for simplicity, we now omit superscripts. The Gibbs sampling steps are standard (Browne and Draper, 2006) and we summarize them for completeness. To sample β, we assume a uniform prior and sample from the posterior distribution

which is multivariate Gaussian with mean

$$\left[\sum_{ij}(I_{(p_1\times p_1)}\otimes X_{ij})^T\Omega_1^{-1}(I_{(p_1\times p_1)}\otimes X_{ij})\right]^{-1}\sum_{ij}(I_{(p_1\times p_1)}\otimes X_{ij})^T\Omega_1^{-1}\tilde{y}_{ij}^T,\quad \tilde{y}_{ij}=y_{ij}-z_{ij}u_j$$

and covariance matrix $\left[\sum_{ij}(I_{(p_1\times p_1)}\otimes X_{ij})^T\Omega_1^{-1}(I_{(p_1\times p_1)}\otimes X_{ij})\right]^{-1}$.

We sample u_j from the multivariate Gaussian distribution

$$MVN\left(\left[\sum_i z_{ij}^T\Omega_1^{-1}z_{ij}+\Omega_2^{-1}\right]^{-1}\left[\sum_i z_{ij}^T\Omega_1^{-1}(y_{ij}-X_{ij}\beta)\right],\ \left[\sum_i z_{ij}^T\Omega_1^{-1}z_{ij}+\Omega_2^{-1}\right]^{-1}\right).$$

We sample a new level 2 covariance matrix from its posterior distribution

$$\Omega_2^{-1}\sim Wishart\,(v_u, S_u),\quad v_u=m+v_p,\quad S_u=\left(\sum_{j=1}^{m}u_ju_j^T+S_p\right)^{-1},$$

where m is the number of level 2 units, u_j is the row vector of residuals for the jth level 2 unit and the prior $p(\Omega_2^{-1})\sim Wishart(v_p, S_p)$, where v_u are the degrees of freedom—the sum of the number of level 2 units and degrees of freedom associated with the prior. One choice, which we will use when we consider non-Gaussian responses, is $v_p=-3,\ S_p=0$, which is equivalent to choosing a uniform prior for Ω_2.

The level 1 covariance matrix is sampled in the same way (for our present model this has only one random effect for each response). Finally, the level 1 residuals are obtained by subtraction.

In some cases, we may wish to impose a linear constraint on a subset of the fixed parameters. For example, if we wish to have a different set of predictors for each response, we can fit a maximal model constraining appropriate subsets to zero. Suppose we wish to impose a set of q independent linear constraints on some or all of the elements of β. These constraints will involve $q^*\geq q$ distinct elements of β, and we re-order β so that q of these q^* elements appear first.

We can write the set of constraints as

$$C\beta=k,\quad C \text{ is } (q\times p),\quad k \text{ is } (q\times 1).$$

Write the QR decomposition of C as $C=QR$, $R=(T\ W)$, where Q is orthogonal $(q\times q)$ and T is upper triangular $[(q\times(p-q))]$ and write $\beta^T=(\beta_1^T\ \beta_2^T)$, where β_1 contains the first q elements of (the re-ordered) β and β_2 contains the last $p-q$ elements.

We now have $Q^Tk=Q^TQR\beta=(T\ W)\beta$ and we construct

$$\begin{pmatrix}T & W\\ 0 & I\end{pmatrix}\beta=\begin{pmatrix}(T\ W)\beta\\ \beta_2\end{pmatrix}=\begin{pmatrix}Q^Tk\\ \beta_2\end{pmatrix},$$

which gives

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \beta = \begin{pmatrix} T & W \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} Q^T k \\ \beta_2 \end{pmatrix} = \begin{pmatrix} T^{-1} & -T^{-1}W \\ 0 & I \end{pmatrix} \begin{pmatrix} Q^T k \\ \beta_2 \end{pmatrix}$$
$$= \begin{pmatrix} T^{-1}Q^T k - T^{-1}W\beta_2 \\ \beta_2 \end{pmatrix},$$

so that we can sample the p-q elements of β_2 freely and compute the q elements of β_1 as the function of β_2 given by the expression $T^{-1}Q^T k - T^{-1}W\beta_2$.

3 Fitting non-Gaussian responses and missing data

3.1 Multicategory (unordered) responses

We assume a 'maximum indicant' model (Aitchison and Bennett 1970). Suppose that we have just a single multinomial response vector with p categories, where the response, y, is (0,1) in each category. That is, we expand the actual response for level 1 unit i (a categorical variable with values 1 to p) into p (0,1) variables where only one of these is 1. Thus, $y_{hi} = 1$ if response is in category $h (h = 1, \ldots, p)$ for individual i, 0 otherwise. For each y_{hi} we assume that an underlying latent Gaussian variable v_{hi} exists and that we have the following multivariate model for these, where for now we omit the level 2 random effects:

$$v_{hi} = X_{1hi}\beta_{1h} + e_{hi}, \quad e_i \sim MVN(0, \Sigma),$$

Σ is a $p \times p$ correlation matrix, e_i mutually independent vectors

$$X_{1hi} \text{ is } (1 \times s), \quad \beta_{1h} \text{ is } (s \times 1), \quad e_i \text{ is } (p \times 1), \quad \beta_1 = \{\beta_{11}^T, \ldots, \beta_{1p}^T\}^{\mathrm{T}} \text{ is } (ps \times 1), \tag{3.1}$$

where β_1 is the subset of β, corresponding to these latent Gaussian responses. The maximum indicant model states that we observe category h for individual i iff $v_{hi} > v_{h^*i} \quad \forall\, h^* \neq h$.

For identifiability purposes, and since we have no substantive interest in the structure of Σ, we will model directly only the first $p - 1$ categories and assume that Σ is diagonal with variances equal to 1. The final category is derived as shown below. This is just the model considered by Aitchison and Bennett (1970).

Now, let Y_{1hi}^* be the set of remaining responses other than the multicategory response being sampled, adjusted for X_1 predictors (common to all responses) and (possible) random effects at higher levels. When sampling the v_{hi} we condition on this set so that (3.1) becomes

$$v_{hi} = X_{1hi}\beta_{1h} + Y_{1hi}^*\beta_{2h} + e_{hi}. \tag{3.2}$$

Thus, if Ω_1 is the current residual covariance matrix for the full set of model responses, we write $\Omega_1 = \begin{pmatrix} \Sigma_1 & \\ \Sigma_{12} & \Sigma_2 \end{pmatrix}$, where Σ_1 is the residual covariance matrix for the Y_1^* and $\Sigma_2 = I_{p-1}$.

We therefore have $\beta_2 = \Sigma_{12}\Sigma_1^{-1}$.

While the same set of model predictors, X_1, applies to each category, the coefficients in general are specific to each category. We therefore have

$$\begin{aligned} &X_{1hi} = X_{1i}, \quad v_i = (X_{1i}^*\beta_1) + e_i, \quad v_i \text{ is } ((p-1)\times 1), \\ &X_{1i}^* = I_{p-1} \otimes X_{1i} \text{ is } ((p-1)\times(p-1)s). \end{aligned} \tag{3.3}$$

Since we observe category h for individual i iff $v_{hi} > v_{h^*i} \quad \forall\, h^* \neq h$, the category probabilities are given by

$$\pi_{hi} = pr[X_{1hi}\beta_h + e_{hi} > X_{1hi}\beta_{h^*} + e_{h^*i}] \;\forall\, h^* \neq h. \tag{3.4}$$

If we now add level 2 random effects (j indexes level 2), (3.1) becomes $v_{hij} = X_{1hij}\beta_{1i} + z_{ij}u_{hj} + e_{hij}$ where u_{hj} is $(q\times 1)$ and we write $u_j = \{u_{hj}\}^T$ which is a $(q(p-1)\times 1)$ vector with $\Omega_u = \mathrm{cov}(u_j)$. We also now write $z_{ij}^* = I_{p-1} \otimes z_{ij}$ which is $((p-1)\times q(p-1))$ and z_{ij} is $(1\times q)$.

To sample the latent Gaussian responses $v_{ij} = \{v_{hij}\}$, we select a sample of $p-1$ values from $N(X_{1i}^*\beta_1 + Y_{1i}^*\beta_2 + z_{ij}^*u_j, \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12}^T)$ and accept this draw to replace the current set of $p-1$ values if and only if the maximum of these $p-1$ values actually occurs in the category where a response variable value of 1 is observed and if this maximum is greater than 0 'or' is less than or equal to 0 and a value of 1 is observed in the final category. If not, we select another sample.

3.2 Ordered responses

Suppose we have an ordered p-category response, ordered categories numbered $1, \ldots, p$. We adopt the probit link cumulative probability model

$$\gamma_h = \int_{-\infty}^{\alpha_h - (X_1\beta_1 + Y_1^*\beta_2 + ZU)} \varphi(t)dt,$$

$$\gamma_h = \sum_{g=1}^{h} \pi_g \quad \text{categories}, \quad h = 1, \ldots p-1,$$

where π_g is the probability that the observation occurs in category $g[g = 1, \ldots, p]$. The Y_1^*, β_2 are as before, and the underlying latent Gaussian variable is given by

$$Y^* = e^* + (X_1\beta_1 + Y_1^*\beta_2 + ZU), \quad e^* \sim N(0,\ 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12}).$$

We assume that the intercept term is incorporated in the fixed part predictor so that $\alpha_1 = 0$. The MCMC steps required to sample the Gaussian variable and the threshold parameters are described in Goldstein *et al.* (2007).

3.3 Missing responses

Where level 1 responses are missing we sample new responses, omitting detailed subscripts, by drawing from $MVN(X_2^*\beta_2^* + e_1^*\beta_1^* + z_2^*u_2^*, \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$, where Σ_2 is the current covariance matrix of residuals for the missing responses, Σ_1 is the covariance matrix of residuals for the observed responses and Σ_{12} is the matrix of covariances between the observed and missing residuals. The $X_2^*\beta_2^*$ and $z_2^*u_2^*$ are the fixed predictor and level 2 residual contribution for the missing responses, $\beta_1^* = \Sigma_1^{-1}\Sigma_{12}$ and e_1^* are the level 1 residuals for the observed responses.

Following all these steps for categorical responses, we obtain a set of Gaussian responses which are combined with any observed Gaussian responses to give a complete set of responses which have a multivariate Gaussian distribution (Geweke, 1991) which can then be modelled as in Section 2.

3.4 Transforming non-Gaussian continuous distributions

For a wide class of distributions, we can apply a normalizing transformation that is a function of one or more parameters, and then incorporate this in a similar way to that described for discrete responses. For example, the Box–Cox transformation (Box and Cox, 1964) for $y \geq 0$ is

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)\lambda^{-1}, & \lambda \neq 0, \\ \log(y), & \lambda = 0, \end{cases}$$

and requires a step for sampling the parameter λ and using this to transform the responses. This step can be carried out using MH with a suitable proposal, where the relevant component of the log likelihood for the untransformed y is

$$-\sum_i \frac{(y_i^{(\lambda)} - \mu_i)^2}{2\sigma^2} + (\lambda - 1)\sum_i \log(y_i),$$

where μ_i comprises the fixed predictor, including level 2 random effects for level 1 responses plus conditioning on the remaining random effects at the same level and σ^2 is the conditional variance on the transformed scale. The second term in this expression is derived from the Jacobian of the transformation. Starting values for the parameter and a Gaussian proposal distribution variance can be obtained from an initial ML estimation for the relevant variable.

3.5 General continuous distributions and survival data

When we have continuous response distributions where no suitable normalizing transformation can be found, we can proceed as follows. We categorize the distribution scale and treat the categories as an ordered classification. We can view this

as a semiparametric procedure for any distribution that can be described by a set of ordered categories to a given approximation. One important example is survival or event history data. If the survival time interval is divided by cut-off points into a set of sub intervals, the resulting set of categories can be treated as an ordered categorization to which the procedures of this paper can be applied. An advantage of this formulation is that multivariate survival models as well as models where survival times are at different levels of a data hierarchy can be incorporated. Any kind of censoring is readily handled using the missing data procedures. Where a large number of thresholds are involved, the threshold parameters themselves can be modelled as a smooth function of time or interval number, e.g., using a spline or fractional polynomial, so reducing the number of parameters (Goldstein, 2003, Chapter 11). Work on these models is currently being pursued (Goldstein and Kounali, 2009, Goldstein, 2009).

3.6 Sampling the level 1 (multivariate) covariance matrix

For all the categorical responses, the level 1 variances are fixed to be equal to 1.0, with zero correlations among the categories of each unordered categorical variable, but non-zero correlations between these categories and other categorical and continuous variables. Thus, for this set of correlations and for the unconstrained variances, we use an MH sampling procedure as follows. We assume uniform priors.

Let $\Omega_{1,lm}$ denote the (l,m)th element of the covariance matrix. We update these covariance parameters using a Metropolis step and a Gaussian random walk proposal as follows.

At iteration t generat $\Omega^*_{1,lm} \sim \mathrm{N}\left(\Omega^{(t-1)}_{1,lm}, \sigma^2_{plm}\right)$, where σ^2_{plm} is a proposal distribution variance that has to be set for each covariance and variance. Then form a proposed new matrix Ω^*_1 by replacing the (l,m)th element of $\Omega^{(t-1)}_1$by this proposed value unless Ω^*_1 is not positive definite in which case set $\Omega^{(t)}_{1,lm} = \Omega^{(t-1)}_{1,lm}$. That is set $\Omega^{(t)}_{1,lm} = \Omega^*_{1,lm}$ with probability min $\left[1, p(\Omega^*_1|e_{ij})/p(\Omega^{(t-1)}_1|e_{ij})\right]$ and $\Omega^{(t)}_{1,lm} = \Omega^{(t-1)}_{1,lm}$ otherwise.

The components of the likelihood ratio are

$$p(\Omega^*_1|e_{ij}) = \prod_{ij} |\Omega^*_1|^{-1/2} \exp(-(e_{ij})^T(\Omega^*_1)^{-1}e_{ij}/2) \quad \text{and}$$
$$p(\Omega^{(t-1)}_1|e_{ij}) = \prod_{ij} |\Omega^{(t-1)}_1|^{-1/2} \exp(-(e_{ij})^T(\Omega^{(t-1)}_1)^{-1}e_{ij}/2).$$

We can use an adaptive procedure (Browne, 2004) to select the proposal distribution parameters.

3.7 Responses at both level 1 and level 2

We now return to model (2.1) containing multivariate Gaussian responses at both levels where these are sampled from either Gaussian responses or, following the

appropriate steps, non-Gaussian discrete or continuous responses. For non-Gaussian level 1 responses we sample as described above. We note that in sampling the level 2 random effects for the level 1 responses, we sample as in the model with responses only at level 1 since the likelihood for these random effects has no component deriving from the level 2 responses. For the level 2 responses, we have the following summary steps.

Step 1: For non-Gaussian level 2 responses, we sample as for level 1 conditioning on all the remaining level 2 random effects.

Step 2: For the full level 2 covariance matrix, we use Gibbs sampling with an inverse Wishart as described earlier if all the level 2 responses are Gaussian, since the components of this matrix derived from the level 1 responses are assumed to be Gaussian by the model. If any of the level 2 responses are categorical then, because of constraints on variances and covariances, as in sampling the level 1 covariance matrix, we need to use MH sampling element by element. The procedure is similar to that for the level 1 covariance matrix but now the components of the likelihood ratio for a particular level 2 covariance matrix Ω_2 are as follows:

$$
\begin{aligned}
p(\Omega_2^*|u_j^{(2)}) &= \prod_{ij} |\Omega_2^*|^{-1/2} \exp(-(u_j^{(2)})^T (\Omega_2^*)^{-1} u_j^{(2)}/2), \\
p(\Omega_2^{(t-1)}|u_j^{(2)}) &= \prod_{ij} |\Omega_2^{(t-1)}|^{-1/2} \exp(-(u_j^{(2)})^T (\Omega_2^{(t-1)})^{-1} u_j^{(2)}/2).
\end{aligned}
\tag{3.5}
$$

Step 3: The level 2 response fixed effects are estimated using the multivariate (regression) model specified by the second line of (2.1).

Step 4: The level 2 random effects for the level 2 responses are obtained by subtraction. We note that where level 2 responses are missing we draw a sample from $MVN(0, \Omega_2)$, where Ω_2 now incorporates level 2 random effects from responses at both levels, and we need to condition on the non-missing level 2 random effects (see Section 3.8).

3.8 Missing responses

At any cycle of the MCMC algorithm, we can sample a set of Gaussian responses for any missing response values. To impute the corresponding category responses given these values we proceed as follows. For an ordered variable, we use the current threshold parameter values to assign the Gaussian sample value to the corresponding category. For an unordered variable, we sample into the category indicated by the maximum from a draw from the associated multivariate Gaussian distribution. For transformed non-Gaussian continuous variables, we apply the appropriate back transformation using the parameter values at the current iteration of the algorithm.

3.9 Multiple imputation for multilevel missing data

The model and fitting procedures we describe have important and direct application to multilevel missing data problems where some or all of the response or predictor variables (covariates) with missing data are non-Gaussian. Suppose we have a multilevel dataset, comprising a mix of continuous and discrete data, with a non-monotone pattern of missing observations. We have some model of interest (MOI) we wish to fit to these data. We assume that the missingness mechanism is missing at random (MAR), i.e., for each unit, given the observed data, the missingness mechanism does not depend on the values of the unseen data. However, this is not usually sufficient to enable us to obtain valid inference for our MOI by simply fitting it to the observed data. This is because some of the covariates in our MOI will typically be missing, and we do not obtain valid inference under MAR by fitting the model to the observed data unless (i) the only missing values are among the responses and (ii) we have included the fully observed variables required for MAR to hold as covariates in our model.

Instead, we can use multiple imputation (Rubin, 1987, Kenward and Carpenter, 2007), with model (2.1) as our imputation model, to obtain valid inference for our MOI under the MAR assumption. From all the variables in our dataset, we form two groups. Group 1 is the variables we need to fit our MOI. If there were no missing data among these variables, we would use them to estimate the parameters in the MOI directly. However, they are only partially observed, so we cannot do this. Group 2 is the variables that are needed so that, for each unit, the assumption that missing data are MAR holds. Groups 1 and 2 need not exhaust the dataset. We then proceed as follows (see Schafer, 1997):

1. Using the variables in Groups 1 and 2, apply our MCMC algorithm to estimate model (2.1) from the observed data. This gives valid estimates of the parameters because all the data are responses in this model, and integrating the likelihood over the missing responses leaves the likelihood for the observed data.
2. Once the sampler has converged continue running it to create K imputations of the missing data—in other words K 'completed' datasets.
3. Fit the MOI to each of the K 'completed' datasets and then combine the parameter estimates using the usual rules for multiple imputation (Rubin 1987) in order to obtain final estimates and standard errors.

Assuming MAR, this gives us valid inference for our MOI. Of course, we can never be sure that MAR holds; data may not be missing at random (MNAR). For example, see Rubin (1987).

In essence, our approach generalizes the models for imputing data under MAR proposed by Schafer (1997). However, our approach is significantly more general. First, we can allow a mix of Gaussian, continuous non-Gaussian, binary, ordinal and unordered categorical variables. Thus, we can handle most of the common types of data. Second, our model is multilevel. We have shown elsewhere (Carpenter and

Goldstein, 2004) that ignoring multilevel structure can lead to bias if the dataset is unbalanced; it will also typically mean that the variance of the imputation distribution is underestimated (so the resulting inferences will be too precise). Third, our models can incorporate survival data, hitherto relatively awkward to handle via multiple imputation.

Carrying out multiple imputation for missing data can never be fully automatic, as care must be taken (i) to include the variables necessary for MAR and (ii) that the imputation model has a structure consistent (congenial) with the MOI. Nevertheless, we believe our model provides a flexible tool for reaching these goals.

An alternative approach to multiple imputation using a joint model is the multiple imputation chained equation approach, in which a set of conditional univariate models are used for imputation without specifying a joint model (Van Buuren, 2007). By contrast to the joint model approach, neither does it have a well-established theoretical grounding nor does it naturally extend to multilevel structures, for example longitudinal data measured in continuous time.

3.10 Response variables with partially known values

In certain cases, we may have a response where some values from a continuous distribution are known accurately but others are only known to lie within a given range. One example is retrospective data where the time since an event is measured and where some individuals can only provide an interval estimate. An illustration is in the measurement of pregnancy gestation length where only some individuals can supply accurate values of the timing of their last menstrual period. Since we typically do not know which are the accurate values an additional step is required to provide a probability distribution which will assign a probability for the observed value close to 1 where this value is in fact accurate, and a more variable distribution of values where it is not. Mixture modelling (see for example Qin *et al.*, 2007) provides one approach to this. Another extreme example is that of truncation where all measurements below a given value are known but for the remainder we only have the information that they lie above the given value. An extension of this is where we have several possible intervals and associated information about which interval an observation lies in. In other cases, we may have probabilistic information associated with different possible ranges of values for a variable. For categorical responses we may be able to specify a set of probabilities across a set of categories.

All these are examples of data coarsening (Heitjan and Rubin, 1991), where what is observed is intermediate between fully missing and fully known. We would propose to treat the above probabilities as constituting a prior distribution for the unknown (missing) value and we refer to this as 'prior-informed imputation'. In the Gaussian case, we may use rejection sampling as follows. First, select a sample from the conditional distribution as for fully missing values. If the sampled value is admissible, e.g. lies in a valid interval, then accept it according to the probability associated with the interval it lies in. If the value is not admissible or not accepted, then select another sample. In the categorical case, we sample a value on the underlying Gaussian scale

as for Gaussian responses. It is admissible if it corresponds to a category with a non-zero prior probability. It is then accepted using the corresponding category prior probability. If not valid or accepted then select another sample. We note that if more than one response for a unit has a partially known value, all responses have to be selected at one (multivariate) draw. As an example, for ordered categories, when computing the log-likelihood contribution in MH sampling, we form a weighted log-likelihood over the valid categories with the weights being the 'prior' probabilities.

An example for categorical variables is where individuals are asked to choose a response category but some cannot make as fine a distinction as other individuals, so that for these we can assign a prior over several categories. Another case is where data are to be coded, e.g., into social groupings, but where for some individuals we only have an assignment to a group of several categories. We assume that the occurrence of such assignments is independent of the model parameter values.

In probabilistic data linkage procedures (see for example Scheuren and Winkler, 1993), estimated linkage probabilities for individuals can be used in a model to correct standard estimators and to produce unbiased estimates. Such procedures, however, operate at the individual record level, whereas it is often the case that we have different probabilities for different variables. In such a situation we can use these variable-specific probabilities as priors in order to provide efficient estimates.

In some cases, where the number of admissible categories is small, a good approximation may sometimes be obtained by sampling only with respect to the prior distributions and further research on this would be useful.

In all these cases, we impute a value or category, so that prior-informed imputation can be regarded as an extension to the missing data procedure described above, to provide a complete dataset.

4 Applications

4.1 A simulation study in multiple imputation

Here, we report a simulation study to assess the use of model (2.1) for multilevel multiple imputation with a mix of data types. Our dataset is the 'tutorial' dataset, distributed as an example with the *MLwiN* software package (Rasbash *et al.*, 2004). Briefly, this consists of a normalized measure of educational achievement at 16 years (the response) and a number of covariates at both level 1 (student) and level 2 (school), as detailed in Table 1.

To evaluate using model (2.1) for multiple imputation, we started by fitting the following substantive model (our MOI) to the full dataset:

$$y_{ij} = X_{ij}\beta + u_j + e_{ij},$$
$$e_{ij} \sim N\left(0, \sigma_e^2\right), \quad u_j \sim N\left(0, \sigma_u^2\right).$$

This gave the parameter estimates in the second column of Table 1. We then did the following.

Table 1 Simulation study model. Parameter estimates and standard errors in brackets. One hundred simulated datasets. MCMC estimation used a burn-in of 2000 with five imputed datasets at iterations 1, 500, 1000, 1500, 2000. Estimates are computed using restricted ML

Parameter	Complete dataset	Imputation	Relative bias (%)	Imputation standard error*
Intercept	0.260 (0.056)	0.263	1.2	0.0021
Reading test score	0.391 (0.017)	0.391	0.0	0.0007
Verbal reasoning band 2**	−0.417 (0.032)	−0.414	−0.7	0.0014
Verbal reasoning band 3**	−0.765 (0.054)	−0.768	0.4	0.0024
School gender category 2***	0.099 (0.108)	0.091	−8.1	0.0040
School gender category 3***	0.241 (0.084)	0.230	−4.6	0.0038
Level 2 variance	0.079 (0.016)	0.080	1.3	0.0005
Level 1 variance	0.536 (0.012)	0.536	0.0	0.0004

Notes: *The imputation standard error is the standard error for each parameter over the 100 simulations.
**Verbal reasoning band has three categories: category 1 (the reference category) is the top 25% of original verbal reasoning scores, category 2 is the middle 50% of verbal reasoning scores and category 3 is the bottom 25% of verbal reasoning scores.
***School gender has three categories: mixed schools (the reference category), category 2 is boys' school and category 3 is girls' school.

Of the 65 schools, 10 were randomly sampled for each simulated dataset where each school had the same probability of inclusion. For these, the school gender was set to be missing. We note that other sampling schemes, for example selecting schools with probability proportional to size, will not satisfy the MCAR condition. In addition, 10% of the response values were randomly set to be missing and 5% of the verbal reasoning band categories were set to be missing. This yielded between 25% and 30% of records with any missing data. One hundred datasets were simulated. The sampling procedure was checked by fitting the MOI to the simulated datasets using listwise deletion; this gave essentially unbiased estimates.

With the exception of the school gender category 3, all the full data values lie within a 95% Gaussian confidence interval. The results show negligible biases for all the level 1 parameters and the variance estimates. For the level 2 categorical variable, school gender, there appears to be a small downward bias. We suggest two possible reasons for this. First, it may be due to the rather small number of schools in the study combined with the assumption of a uniform prior for the level 2 covariance matrix in the imputation model. Second, and more subtly, the MOI (a conditional model of normalized exam score given covariates) is not exactly compatible (congenial) with the joint distribution implied by the imputation model, although we believe it is likely to be close because the imputation model relies on an underlying joint multivariate Gaussian distribution with unstructured covariance matrices at levels 1 and 2. We return to this point in the discussion. This would be sufficient to account for the small biases we observe.

Table 2 Two level growth model

Coefficient	Estimate	Standard error
Level 1 model intercept	153.05	0.69
Age (about age 13.0)	7.07	0.16
Age-squared	0.294	0.054
Age-cubed	−0.208	0.029
Level 2 model intercept	174.70	0.80
Level 2 covariance matrix		
$\begin{pmatrix} 55.77 & 1.29 & 50.01 \\ 1.30 & 0.53 & 1.24 \\ 50.01 & 1.24 & 69.42 \end{pmatrix}$		
Level 1 variance	3.21	

4.2 A growth data example

Our first example uses the growth data analysed by Goldstein (1989). This dataset consists of 108 children with height measured on up to six occasions around the age of 13 together with their adult heights, altogether 436 growth measurements and 108 adult height measurements. We shall fit a cubic growth curve to the level 1 (within child) measures and a single intercept for the adult height measurement. We will also allow the age slope to vary at level 2 so that each child is allowed to grow at different rates. In general, we may wish to allow the higher order polynomial coefficients to vary across individuals and to introduce further covariates such as gender, but for purposes of illustration we shall fit a simple explanatory model.

This model can be written as follows:

$$
\begin{aligned}
y_j^{(2)} &= \gamma_0 + u_{0j}^{(2)}, \\
y_{ij}^{(1)} &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + u_{0j}^{(1)} + u_{1j}^{(1)} t_{ij} + e_{ij}, \qquad (4.1) \\
\begin{pmatrix} u_{0j}^{(1)} \\ u_{1j}^{(1)} \\ u_{0j}^{(2)} \end{pmatrix} &\sim MVN(0, \Omega_2), \quad \Omega_2 = \begin{pmatrix} \sigma_{u0}^{(1)^2} & & \\ \sigma_{u01}^{(1,1)} & \sigma_{u1}^{(1)^2} & \\ \sigma_{u00}^{(1,2)} & \sigma_{u10}^{(1,2)} & \sigma_{u0}^{(2)^2} \end{pmatrix}, \quad e_{ij} \sim N(0, \sigma_e^2),
\end{aligned}
$$

where $y_{ij}^{(1)}$ is the ith measurement around the age of 13 for the jth child, $y_j^{(2)}$ is the adult height of the jth child and t_{ij} is age. We are assuming a multivariate Gaussian distribution for all the responses, one of which (adult height) is at level 2.

The results from fitting this model with a burn-in of 500 and 5000 iterations are shown in Table 2.

The chains are well behaved and that for the slope variance (at level 2) is as shown in Figure 1.

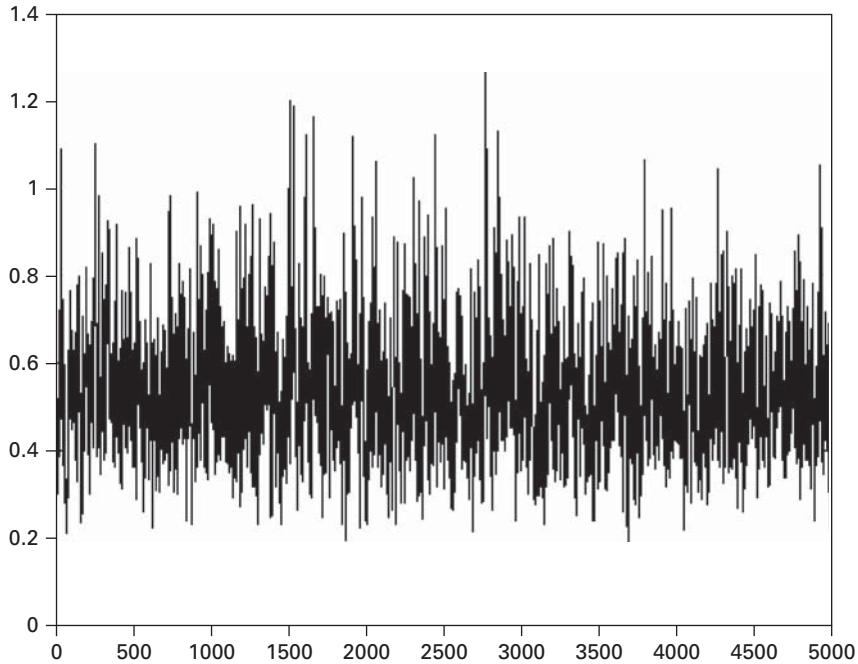

Figure 1 Chain for 5000 iterations for slope variance parameter

The 95% interval for the slope variance is from 0.30 to 0.86. The average height at age 13.0 years is 153.1 (standard error 0.69) and the average adult height is 174.7 (standard error 0.80).

One of the main uses of this analysis is to provide a procedure for predicting the adult height of a child for whom we have growth measurements. Thus, we require a prediction formula that we can derive from the parameters of our model. Thus, for example, if we have two growth measurements we will have a linear prediction of the form

$$\hat{y}_j = \gamma_0 + \alpha_1 \tilde{y}_{1j} + \alpha_2 \tilde{y}_{2j}, \tag{4.2}$$

where, from (4.1)

$$\tilde{y}_{ij} = y_{ij}^{(1)} - (\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2), \quad i = 1, 2,$$

is the 'raw' residual for each measurement. The parameters of the prediction equation (4.2) can be derived from the covariance matrix in Table 1, namely

$$\begin{pmatrix} \tilde{y}_{1j} \\ \tilde{y}_{2j} \\ \tilde{y}_j^{(2)} \end{pmatrix} = MVN(0, \Omega),$$

$$\Omega = \begin{pmatrix} \sigma_{u0}^{(1)^2} + \sigma_{u1}^{(1)^2} t_{1j}^2 + 2\sigma_{u01}^{(0,1)} t_{1j} + \sigma_e^2 & & \\ \sigma_{u0}^{(1)^2} + \sigma_{u01}^{(1,1)}(t_{1j} + t_{2j}) + \sigma_{u1}^{(1)^2} t_{1j} t_{2j} & \sigma_{u0}^{(1)^2} + \sigma_{u1}^{(1)^2} t_{2j}^2 + 2\sigma_{u01}^{(0,1)} t_{2j} + \sigma_e^2 & \\ \sigma_{u00}^{(1,2)} + \sigma_{u10}^{(1,2)} t_{1j} & \sigma_{u00}^{(1,2)} + \sigma_{u10}^{(1,2)} t_{2j} & \sigma_{u0}^{(2)^2} \end{pmatrix}$$

$$\tilde{y}_j^{(2)} = y_j^{(2)} - \gamma_0.$$

So that we have

$$\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} \sigma_{u0}^{(1)^2} + \sigma_{u1}^{(1)^2} t_{1j}^2 + 2\sigma_{u01}^{(0,1)} t_{1j} + \sigma_e^2 & \\ \sigma_{u0}^{(1)^2} + \sigma_{u01}^{(1,1)}(t_{1j} + t_{2j}) + \sigma_{u1}^{(1)^2} t_{1j} t_{2j} & \sigma_{u0}^{(1)^2} + \sigma_{u1}^{(1)^2} t_{2j}^2 + 2\sigma_{u01}^{(0,1)} t_{2j} + \sigma_e^2 \end{pmatrix}^{-1}$$
$$\begin{pmatrix} \sigma_{u00}^{(1,2)} \times \sigma_{u10}^{(1,2)} t_{1j} \\ \sigma_{u00}^{(1,2)} \times \sigma_{u10}^{(1,2)} t_{2j} \end{pmatrix},$$

so that a prediction, and confidence interval, can be computed for any set of growth measurements and this will provide the basis for a flexible height prediction system that can be incorporated readily into software.

4.3 A multilevel mixed response multiple imputation example

The data for this example are taken from the Scottish component of the 2005/06 Health Behaviour in School-aged Children (HBSC): A WHO Collaborative Cross-national study (Currie *et al.*, 2008) where 1644 pupils in 75 primary schools were surveyed and asked questions relating to their health behaviour, including their fruit and vegetable eating habits. All schools taking part in the HBSC survey in Scotland received a school survey for a senior member of staff to complete. All 75 primary schools surveyed completed and returned the questionnaires. The response variable chosen for our present analysis is the frequency of fruit intake of the pupil in six categories as described in Figure 2. It is treated as an ordered categorical variable. A histogram of this variable is given in Figure 2.

The predictor variables at the pupil level are pupil gender and father's social class coded in five categories from high (0) to low (4) then economically active (5) and economically inactive (6). At the school level, the predictor variables are the 2001 Carstairs index of social deprivation (a continuous variable) with a mean of 0.35 and standard deviation of 3.4 (Carstairs and Morris, 1991) assigned to each school by its postcode sector, whether the school was actively involved in the National Health Promoting School initiative (Scottish Health Promoting Schools Unit, 2004), whether

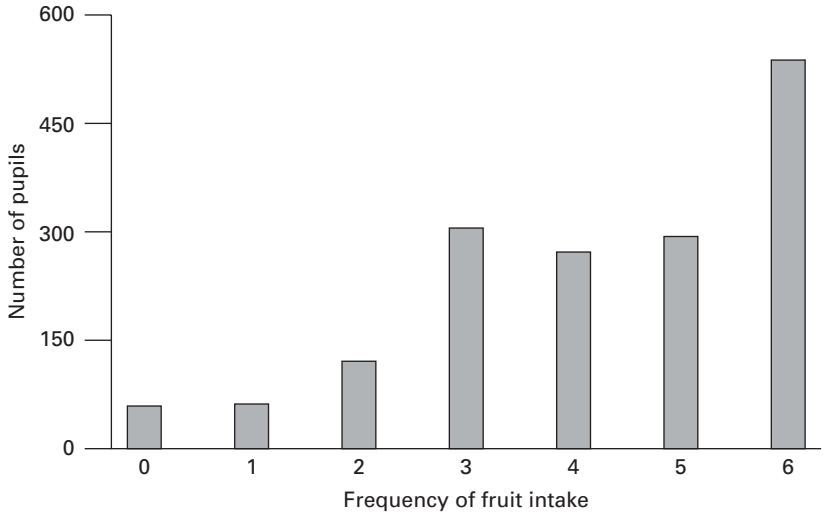

Figure 2 Frequency of fruit intake. Never (0), less than once a week (1), once a week (2), 2–4 days a week (3), 5–6 days a week (4), once a day (5) and more than once a day (6)

it was involved in a national 'Hungry for Success' initiative (Scottish Executive, 2003), and whether pupils can buy fruit at the school (every day, some days, never). For these six variables the first is binary, the second is treated as an unordered categorical variable, the third is treated as Gaussian, the fourth is binary, the fifth is binary and the sixth is ordered.

The MOI that we propose to fit to the data is a two-level cumulative proportions model described above where there are seven categories.

At level 2, we fit a single variance term with the default prior distribution described in section 2.

There are missing data in all variables, except the Carstairs index, ranging from 1.2% in the response to 13.6% for the health promotion initiative question.

The first step is to carry out the multiple imputation for all the missing data using the described algorithms. We have chosen to sample 5000 MCMC iterations, with a burn-in of 1000 and with imputed values computed every 1000 iterations. This yields five completed datasets to each of which the MOI is fitted, using MCMC again with 1000 burn-in and 5000 iterations. This process is then repeated with 10000 MCMC iterations yielding 10 imputed datasets and finally with 20000 iterations to yield 20 imputed datasets. The results are set out in Table 3. This table also shows the results of fitting a model where records with any missing data are deleted; this yields 1138 pupils in 58 schools.

For the listwise deletion model, as expected, we see that there are slightly increased standard errors associated with the estimates, compared to the estimates from the imputation models. There is evidence for a gender difference, with girls more likely to eat fruit and those pupils from highest social group are also more likely to do so.

Table 3 Fitted model for fruit intake for different numbers of imputed datasets, and for analysis based on listwise deletion of all missing data (standard errors in brackets)

Fixed coefficient	5 datasets	10 datasets	20 datasets	Listwise delete
Intercept	2.88 (0.54)	2.86 (0.53)	2.87 (0.54)	2.92 (0.56)
Threshold 1	0.38 (0.05)	0.39 (0.05)	0.39 (0.05)	0.42 (0.07)
Threshold 2	0.81 (0.06)	0.82 (0.06)	0.82 (0.06)	0.85 (0.08)
Threshold 3	1.46 (0.06)	1.47 (0.07)	1.47 (0.06)	1.54 (0.08)
Threshold 4	1.91 (0.07)	1.91 (0.07)	1.91 (0.07)	1.97 (0.08)
Threshold 5	2.38 (0.07)	2.39 (0.07)	2.39 (0.07)	2.48 (0.08)
Gender (girl–boy)	0.24 (0.05)	0.24 (0.05)	0.24 (0.05)	0.21 (0.06)
Father SES cat 2	−0.27 (0.16)	−0.25 (0.18)	−0.25 (0.17)	−0.31 (0.19)
Father SES cat 3	−0.55 (0.17)	−0.51 (0.19)	−0.51 (0.18)	−0.56 (0.20)
Father SES cat 4	−0.47 (0.15)	−0.45 (0.16)	−0.44 (0.16)	−0.50 (0.18)
Father SES cat 5	−0.32 (0.15)	−0.30 (0.17)	−0.31 (0.17)	−0.40 (0.19)
Father SES cat 6	−0.41 (0.15)	−0.39 (0.16)	−0.38 (0.17)	−0.43 (0.18)
Father SES cat 7	−0.58 (0.19)	−0.57 (0.21)	−0.57 (0.20)	−0.65 (0.23)
Carstairs index	−0.01 (0.01)	−0.01 (0.01)	−0.01 (0.01)	−0.01 (0.01)
Health promoting (yes–no)	−0.56 (0.51)	−0.55 (0.50)	−0.56 (0.50)	−0.59 (0.52)
Hungry for success (yes–no)	0.20 (0.18)	0.19 (0.18)	0.20 (0.18)	0.14 (0.21)
Can pupils buy fruit at school (some days–every day)	−0.05 (0.11)	−0.06 (0.11)	−0.06 (0.12)	−0.08 (0.13)
Can pupils buy fruit at school (never–every day)	0.10 (0.10)	0.09 (0.11)	0.08 (0.11)	0.14 (0.13)
Level 1 variance	1.0	1.0	1.0	1.0
Level 2 variance	0.034 (0.017)	0.035 (0.016)	0.035 (0.016)	0.044 (0.020)

Notes: The categories are referred to here as 1, . . . , p. Father's SES categories contrasted with category 1 (highest).

The MCMC chains for the fixed coefficients and the level 2 variance show good mixing, but those for the threshold parameters do not, although the chain appears stationary. Figure 3 shows the chain for the first threshold parameter and Figure 4 for the level 2 variance.

For the completed datasets we obtain similar pictures for the chains.

From Table 3 we see some small changes in parameter values moving from 5 to 10 completed datasets and a smaller change is seen moving from 10 to 20 sets, which suggests that 10 completed datasets are adequate. We also note that, as we would expect, the standard errors for all the parameters are smaller than those for the listwise deleted model, in some cases substantially so.

5 Discussion

5.1 Model fit

We propose the use of the deviance information criterion (DIC, Spiegelhalter *et al.*, 2002) to assess model fit. This requires the calculation of the likelihood at each cycle

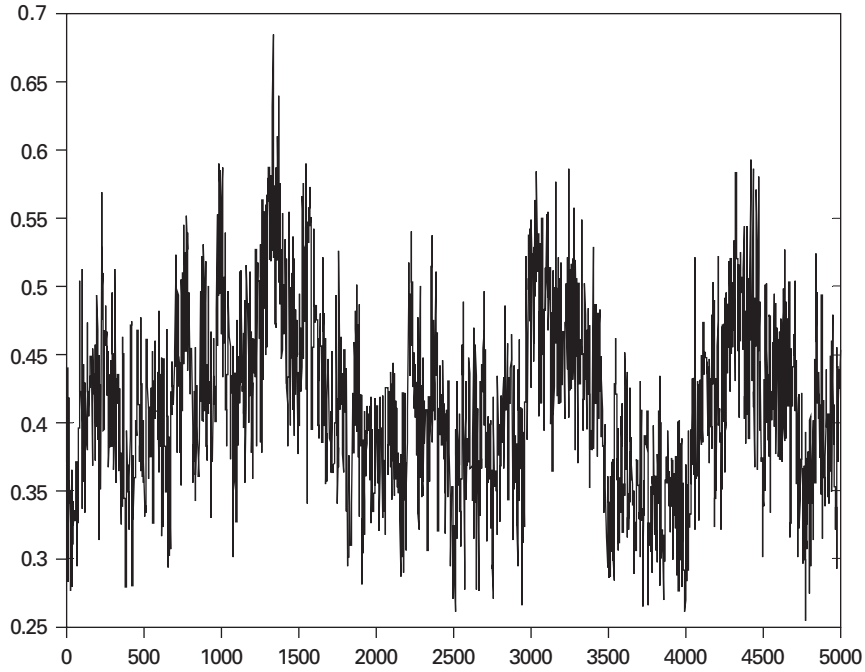

Figure 3 MCMC chain of 5000 samples for threshold parameter 1

of the MCMC algorithm. To do this we can write the full set of responses as

$$Y = \left\{ \begin{array}{c} Y_1 \\ Y_2 \\ Y_3 \end{array} \right\},$$

where Y_1, Y_2, Y_3 refer, respectively, to the Gaussian, ordered and unordered categorical sets of variables. Consider just the set of level 1 responses first.
We write the likelihood as

$$P(Y) = P(Y_1|Y_2, Y_3)P(Y_2|Y_3)P(Y_3).$$

The first term on the right-hand side is the multivariate Gaussian likelihood adjusted for the remaining responses. Where we have a non-Gaussian continuous response we can include this using the expression given in Section 3.4. The computations for the second term can be carried out by writing $P(Y_2|Y_3)$ as a product of individual conditional variables so that we only require the evaluation of univariate density intervals. Evaluation of the third term involves evaluating the joint distribution of relevant-order statistics, and this again can be expressed as a set of conditional distributions, one for each categorical variable.

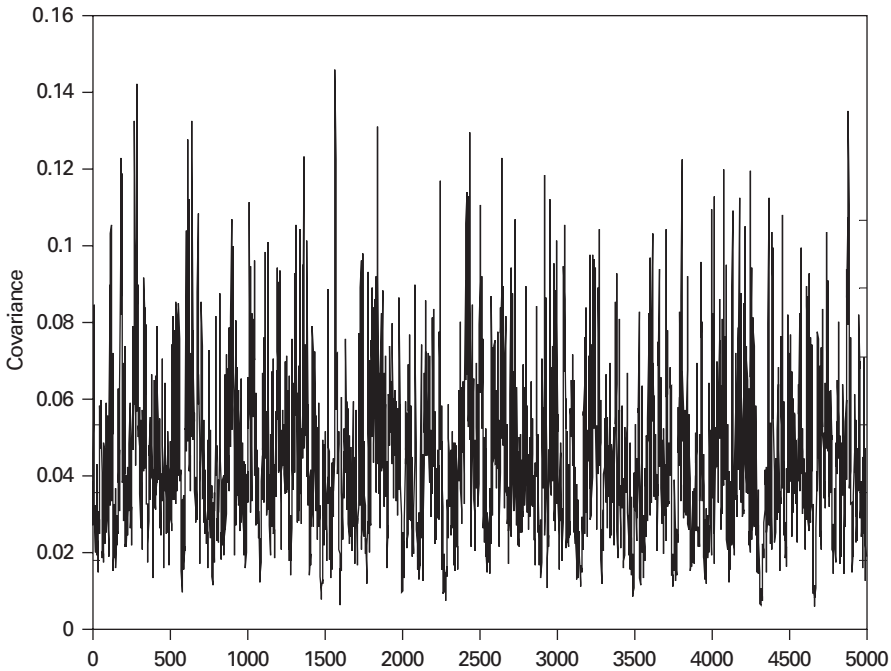

Figure 4 MCMC chain of 5000 samples for level 2 variance

A practical procedure for carrying out these computations is by simulating from the model defined by the current parameter estimates at each iteration to approximate the likelihood.

Given the likelihood at each cycle, and similarly for the final parameter estimates, the DIC statistic is calculated as

$$DIC = \bar{D} + p_D, \quad p_D = \bar{D} - D(\theta),$$

Where D is the deviance at each cycle (i.e., -2 *log-likelihood)*, $\bar{D}$ is the average value over the chain, θ is the set of final parameter estimates and P_D is the effective number of parameters (Spiegelhalter *et al*., 2002).
We have not yet implemented this procedure and it is the subject of further research.

5.2 General discrete distributions

Further extensions can be made for general discrete distributions, where the probability density function for the discrete values provides the equivalent of the probability of category membership and is a function of one or more distribution parameters. Thus, e.g., for a Poisson distribution, there is just a single such parameter that

requires updating. At any given cycle of the MCMC algorithm, given current parameter values, we can then sample a latent Gaussian value as for the general ordered case described above. Further work on such models is in progress (Goldstein and Kounali, 2009).

5.3 Copula models

Our procedures have similarities with the Gaussian copula model (see e.g., Pitt *et al*., 2006). In this model, a set of p latent variables following a $(p \times p)$ multivariate Gaussian distribution have a one-to-one mapping on to p marginal discrete or continuous distributions which are modelled as functions of covariates. Our procedures for deriving the latent Gaussian variables can be viewed as a multilevel generalization of this copula model, and they also extend the standard copula model to allow the incorporation of multinomial marginal distributions.

5.4 Conclusions

In this paper, we have sought to provide a general structure that unifies a wide range of applications. Many further developments are possible. For example, it is relatively straightforward to add further modelling steps to accommodate, e.g., measurement errors (Goldstein, 2003, Chapter 13) or latent variables (Goldstein and Browne, 2005).

The practicality of using the procedures described in this paper needs to be investigated for various data configurations. Thus, e.g., ordered data with a large number of categories may lead to convergence problems due to the large number of threshold parameters. In our ordered response example, we have seen how the chains for the threshold parameters can have a high level of autocorrelation. Similar problems may be experienced with unordered categorical variables having a large number of categories. Furthermore, our algorithm provides estimates that may be sensitive to assumptions about prior distributions, especially for covariance matrices, and we recognize that there is further work to be carried out in that area. We are also aware that there will often be more efficient samplers available that can speed up computation, especially for the threshold parameters (see e.g., VAN Dyk and Meng, 2001). Using multiple chains is also important in practice to make suitable judgements about convergence. Exploration of these issues is currently being carried out in conjunction with extensions to the procedures in the present paper.

For data where there is missing information, the ability to deal with non-Gaussian data greatly extends the usefulness of existing multiple imputation procedures, especially within a multilevel framework. The further ability to handle partially known data is also important for a number of applications, such as linkage studies and retrospective surveys, where such partial information is common. Nevertheless, with a flexible model such as the one developed here, multiple imputation is not a totally automatic process. First, we need to think about whether the imputation model is

compatible with the MOI. This is relatively easy to do if the data are Gaussian, but harder if we have a mix of response types. Nevertheless, we believe our model is flexible enough to ensure parameter estimates from multiple imputation have small bias in a range of models. Second, the class of imputation models we propose assumes joint multivariate normality and if this is only approximately true, as may occur, e.g., if we are using a Box–Cox transformation, some bias may result. We note, however, that our sampling procedures for generating Gaussian variables, do so by conditioning on linear functions of the remaining variables, which if these are Gaussian will ensure a multivariate Gaussian distribution, or yield a good approximation for certain cases involving ordered responses.

A further potential application is to mixture models where individuals are assumed to belong to each of a number of groups with a set of probabilities associated with group membership and where each group will have its own parameter values for a given model. If a subset of individuals can be assigned prior group membership probabilities, some of which may equal one, then this can be formulated as a missing data problem where, for a set of indicators denoting group membership, the values are (probabilistically) known for the sample subset. The remaining values are imputed, as are the values for the subset conditioned additionally on the assigned priors. Research into these models is also currently under way.

Acknowledgements

This work was partly funded by a grant from the Economic and Social Research Council, UK, RES-000-23-0140, a grant from the Medical Research Council, UK, G0600599 and a further grant from the Economic and Social Research Council, UK, PTA 035-250027. The Health Behaviour in School-aged Children (HBSC) study is an international survey conducted in collaboration with the World Health Organization Regional Office for Europe. The authors would like to thank Candace Currie, the HBSC International Coordinator and Principal Investigator for Scotland for allowing us to use the Scottish 2005/06 HBSC data. The paper also acknowledges the HBSC International Research Network in 41 countries which developed the study's research protocol. The authors would also like to thank the following for helpful comments: William Browne, Paul Clarke, Daphne Kounali, Anthony Robinson, Jon Rasbash and Fiona Steele. The final version has benefited considerably from comments by the referees and an associate editor of the journal.

References

Aitchison J and Bennett JA (1970) Polychotomous quantal response by maximum indicant. *Biometrika*, **57**, 253–62.

Asparouhov T and Muthen B (2007) Multilevel mixture models. In Hancock GR and Samuelson KM (eds). *Advances in latent mixture models* Charlotte, NC:

Information Age Publishing, Inc., 27–51.
Box GEP and Cox DR (1964) An analysis of transformations (with discussion). *Journal of the Royal Statistical Society*, *B*, **26**, 211–52.
Browne WJ (2009) *MCMC estimation in MLwiN*. Bristol: University of Bristol.
Browne WJ and Draper D (2006) A comparison of Bayesian and likelihood based methods for fitting multilevel models. *Bayesian Analysis*, **1**, 473–514.
Carpenter J and Goldstein H (2004) Multiple imputation using MLwiN. *Multilevel Modelling Newsletter*, **16**, 9–18.
Carstairs V and Morris R (1991) *Deprivation and health in Scotland*. Aberdeen, Scotland: Aberdeen University Press.
Currie C, Levin K, and Todd J (2008) Health behaviour in school-aged children: findings from the 2006 HBSC survey in Scotland. Child and Adolescent Health Research Unit, University of Edinburgh.
Dunson DB (2000) Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, **62**, 355–66.
Geweke J (1991) Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. *Computing science and statistics: Proceedings of the 23rd symposium on the interface*. Fairfa Station, VA: Interface foundation of North America.
Goldstein H (1989) Models for multilevel response variables with an application to Growth Curves. In Bock RD (ed). *Multilevel analysis of educational data*. New York: Academic Press, 107–25.
Goldstein H (2003) *Multilevel statistical models. Third edition*. London: Edward Arnold.
Goldstein H (2009) The analysis of survival and event history data using a latent normal model. (*in press*).
Goldstein H, Bonnet G, and Rocher T (2007) Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and behavioural Statistics*, **32**, 252–86.
Goldstein H and Browne W (2005) Multilevel factor analysis models for continuous and discrete data. In Olivares A and McArdle JJ (eds). *Contemporary psychometrics. A Festschrift to Roderick P. McDonald*. Mahwah, NJ: Lawrence Erlbaum.
Goldstein H and Kounali D (2009) Multivariate multilevel modelling of childhood growth, members of growth measurements and adult characteristics. *Journal of the Royal Statistical Society*, A **172**, 599–613.
Heitjan DF and Rubin DB (1991) Ignorability and coarse data. *Annals of Statistics*, **19**, 2244–53.
Imai K and van Dyk DA (2005) A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, **124**, 311–34.
Kenward M and Carpenter J (2007) Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, **16**, 199–218.
Mathworks (2004) *Matlab*. Available at http://www.mathworks.co.uk.
Muthen LK and Muthen BO (2004) *MPLUS users guide version 5*. Los Angeles: University of California, Graduate School of Education.
Pitt M, Chan D, and Kohn R (2006) Efficient Bayesian inference for Guassian copula regression models. *Biometrika*, **93**, 537–54.
Qin C, Dietz PM, England LJ, Martin JA, *et al.* (2007) Effects of different data-editing methods on trends in race-specific delivery rates, United States, 1990–2002. *Pediatric and Perinatal Epidemiology*, **21**, 41–49.
Rabe-Hesketh S, Pickles A, and Skrondal A (2001) GLLAMM: a general class of multilevel models and a STATA program.

Multilevel Modelling Newsletter, **13**, 17–23.

Rabe-Hesketh S, Skrondal A, and Pickles A (2005) Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, **128**, 301–23.

Rasbash J, Steele F, Browne W and Goldstein H (2009). *A user's guide to MLwiN version 2.10*. Bristol, Centre for Multilevel Modelling, University of Bristol.

Rubin DB (1987) *Multiple imputation for non response in surveys*. Chichester: Wiley.

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

Scheuren F and Winkler WE (1993) Regression analysis of data files that are computer matched. *Survey Methodology*, **19**, 35–38.

Scottish Executive (2003) *Hungry for success: a whole school approach to school meals in Scotland*. Edinburgh: The Stationary Office.

Scottish Health Promoting Unit (2004) *Being well–doing well: a framework for health promoting schools in Scotland 2004*. Dundee: SHPSU.

Spiegelhalter D, Best N, Carlin BP, and Van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, B, **64**, 583–640.

Spiegelhalter DJ, Thomas A, and Best NG (1999) *WINBUGS version 1.2, user manual*. Cambridge: MRC Biostatistics Unit.

Van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16**, 219–42.

Van Dyk D and Meng X (2001) The art of data augmentation. *Journal of Computational and Graphical Statistics*, **10**, 1–30.

Yucel R (2008) Multiple imputation for multilevel continuous data. Philosophical transactions of the Royal Society, A, **2**, 2389–403.