Statistical Modelling 2006; **6**: 23–42

Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood

Edmond SW Ng[1], James R Carpenter[2], Harvey Goldstein[1] and Jon Rasbash[1]
[1]Centre for Multilevel Modelling, Graduate School of Education, University of Bristol, UK
[2]Medical Statistics Unit, London School of Hygiene and Tropical Medicine, University of London, UK

Abstract: Fitting multilevel models to discrete outcome data is problematic because the discrete distribution of the response variable implies an analytically intractable log-likelihood function. Among a number of approximate methods proposed, second-order penalised quasi-likelihood (PQL) is commonly used and is one of the most accurate. Unfortunately, even the second-order PQL approximation has been shown to produce estimates biased toward zero in certain circumstances. This bias can be marked especially when the data are sparse. One option to reduce this bias is to use Monte-Carlo simulation. A bootstrap bias correction method proposed by Kuk has been implemented in MLwiN. However, a similar technique based on the Robbins–Monro (RM) algorithm is potentially more efficient. An alternative is to use simulated maximum likelihood (SML), either alone or to refine estimates identified by other methods. In this article, we first compare bias correction using the RM algorithm, Kuk's method and SML. We find that SML performs as efficiently as the other two methods and also yields standard errors of the bias-corrected parameter estimates and an estimate of the log-likelihood at the maximum, with which nested models can be compared. Secondly, using simulated and real data examples, we compare SML, second-order Laplace approximation (as implemented in HLM), Markov Chain Monte-Carlo (MCMC) (in MLwiN) and numerical integration using adaptive quadrature methods (in Stata's GLLAMM and in SAS's proc NLMIXED). We find that when the data are sparse, the second-order Laplace approximation produces markedly lower parameter estimates, whereas the MCMC method produces estimates that are noticeably higher than those from the SML and quadrature methods. Although proc NLMIXED is much faster than GLLAMM, it is not designed to fit models of more than two levels. SML produces parameter estimates and log-likelihoods very similar to those from quadrature methods. Further our SML approach extends to handle other link functions, discrete data distributions, non-normal random effects and higher-level models.

Key words: Bias correction; Kuk's method; Monte-Carlo integration; numerical integration; Robbins–Monro algorithm; simulated maximum likelihood

Data and code link available from: http://stat.uibk.ac.at/SMIJ
Received October 2004; **revised** July, November and December 2005; **accepted** January 2005

1 Introduction

Multilevel models for discrete data are extremely useful in a wide range of social science and medical settings, where the principal outcome is discrete and the data are naturally

Address for correspondence: Edmond S.W. Ng, Centre for Multilevel Modelling, Graduate School of Education, University of Bristol, 35 Berkeley Square, Bristol BS8 1JA, UK. E-mail: e.ng@bristol.ac.uk

10.1191/1471082X06st106oa

clustered. For example, consider the data on the investigation of contraceptive use (yes/no) where individuals are clustered in localities, (discussed in Section 2).

However, unlike when the data are continuous, when (restricted) maximum likelihood estimates can be readily obtained using, for example, the iterative techniques developed by Goldstein (1986, 1989), fitting such models is not straightforward. This is because the discrete distribution of the response means that the likelihood is computed by integrating a product of discrete and normal densities, which has no analytical solution.

To tackle this problem, two classes of approximation have been developed; marginal and penalised quasi-likelihood (Breslow and Clayton, 1993; Goldstein and Rasbash, 1996). Roughly speaking, both methods approximate the discrete multilevel likelihood by a Gaussian multilevel likelihood, which can then be fitted in the usual way. However, both approximations generally yield parameter estimates that are biased towards zero (Rodríguez and Goldman, 1995; Lin and Breslow, 1996). This is particularly the case when the response data are binary, and the fitted probabilities are close to 1 or 0. This bias is more marked if there are relatively few observations at the bottom level of the hierarchy for each unit at the second level of the hierarchy (for example few repeated observations on each individual). Further, these approximate methods do not yield an estimate of the log-likelihood for comparing nested models.

The problems with marginal and penalised quasi-likelihood have led to the development of a variety of proposals for reducing the bias. Lin and Breslow (1996) have proposed an analytical bias correction. An alternative is to extend the Taylor expansion in the penalised quasi-likelihood (PQL) to higher order. One such example is by Raudenbush *et al.* (2000) who used a sixth-order Laplace approximation (Laplace6) to the marginal likelihood with maximization by Fisher scoring. This approach is implemented in HLM 5. The Fisher scoring step has subsequently been replaced by EM fitting of a second-order Laplace approximation in HLM 6.01 (henceforth EM-Laplace2) (Congdon, 2005). However, the quasi-likelihood approach does not provide an estimate of the log-likelihood at the maximum and the higher order Taylor expansions may be difficult to fit.

Another option is to use Monte-Carlo simulation. One proposal, due to Kuk (1995), is implemented in MLwiN (Rasbash *et al.* 2000); it is known as bootstrap bias correction. The Robbins–Monro (RM) stochastic approximation method (Wetherill and Glazebrook, 1986) can also be applied to this problem, and is potentially fully efficient. Unfortunately, neither of these two methods readily gives standard errors of the bias-corrected parameters. Another contender is simulated maximum likelihood (SML) (McCulloch, 1997). SML is attractive because it yields standard errors of the bias-corrected parameter estimates and an estimate of the log-likelihood at the maximum. Further, it can be used to fit models with non-normal random effects.

Alternatively, a Bayesian approach can be pursued using Markov Chain Monte-Carlo (MCMC) methods where exact inferences can be made based on their posterior distributions of the parameters. MCMC methods have also been implemented in MLwiN (Browne, 2000).

1.1 Outline

The main purpose of this article is to evaluate three different simulation-based approaches to bias correction in discrete two-level models: Kuk's method, the RM search

and SML. A secondary aim is to use simulated and real data examples to compare and contrast the best simulation-based bias correction method with promising alternative methods, specifically EM-Laplace2, MCMC and numerical integration (using adaptive quadrature).

The plan for the article is as follows. Section 2 presents data from studies that helped to motivate this research: these are typical of discrete, sparse, multilevel data sets. Section 3 describes and compares bias correction using Kuk's method and the RM search. Properties of the methods are highlighted with a simulation study. Section 4 describes the application of PQL (2) followed by SML, and compares it with the other two methods and numerical quadrature as implemented in SAS proc NLMIXED.

The illustrative analysis of two data sets (using second-order PQL, second-order PQL followed by SML, EM-Laplace2, MCMC and numerical quadrature) is given in Section 5 and a comparison of SML and numerical integration approaches using simulated data sets is given in Section 6. We conclude with a discussion in Section 7.

2 The data

Here, we briefly describe two data sets that we analyse in Section 5. *USAsmall* is a simulated data set based on a hierarchical structure similar to that of a multistage clustered survey of prenatal health care in a developing country described in Rodríguez and Goldman (1995). The response is 'whether or not mothers received modern prenatal care during pregnancies' with a set of covariates that are 'measured' at the individual, family and community levels. In this data set, 737 children (level 1 units) are nested within 479 families (level 2 units). The interesting feature of this data set is that a high proportion (56%) of the families have only one child, as this sparseness may cause marked biases in the quasi-likelihood estimates.

The second data set, *BANG*, is a sub-sample of the 1988 Bangladesh fertility survey of contraceptive use (Huq and Cleland, 1990). In this survey, respondents were selected using a two-stage cluster sample design. Clusters, corresponding roughly to villages in rural areas and *mohalla* (neighbourhoods) in urban areas, were taken as the primary sampling units (PSUs) and selected with probability proportional to size. They were stratified by district and urban/rural area. Amin *et al.* (1997) analysed the data using a three-level logistic regression model with women (level 1) nested within PSUs (level 2) and the PSUs nested within districts (level 3).

Approximately, 20% of the women (level 1 units) in the original sample were chosen randomly to be included in this sub-sample. The second level (PSUs) is ignored in our illustrative analysis in Section 5, as SML has not yet been implemented for higher-level models. Thus, the sub-sample analysed consists of 1934 women (level 1 units) from 60 districts (level 2 units). To be included in the sample, a woman had to have been married at some time (ie, be an 'ever married' woman). On average, there are 32 women for each district, but this ranges from 2 to 118. The researchers were interested in exploring the relationship between the use of contraceptive at the time of the survey and a number of covariates including age, the number of existing children and the type of region of residence.

3 Comparison of Kuk's method and RM search

In this section, we briefly outline and discuss two computer intensive methods for bias correction: Kuk's method and the RM algorithm, as it applies to this problem.

Background: Consider, two-level data consisting, for example, of repeated observations on individuals. Let y_i denote the ith response from a subject ($i = 1, \ldots, I$), and assume that responses from different subjects are independent. For the sake of simplicity, the subscript, j, for subject is omitted throughout. Let u denote a p-vector of subject-specific random effects, assumed to follow a multivariate normal distribution $N(0, \Sigma)$, with density $f(u, \Sigma)$. Conditional on subject-specific random effects and covariates, the two-level multilevel model assumes individual repeated observations are independent. They follow a discrete distribution, denoted by $g(y_i|u, \beta, x_i, z_i)$, $(i = 1, \ldots, I)$, where

$$E[y_i|u, x_i] = \lambda^{-1}(x_i^T \beta + z_i^T u)$$

x_i is a $(q \times 1)$ vector of covariates at the ith response with $(q \times 1)$ coefficient vector β, z_i is a $(p \times 1)$ vector of covariates for the random effects and λ is the link function (eg, logistic, probit, log).

The likelihood for the I observations on this individual is then

$$L(y; \beta, \Sigma) = \int \prod_{i=1}^{I} \{g(y_i|u, \beta, x_i, z_i)\} f(u, \Sigma)\, du \tag{3.1}$$

where the integral is over the p-dimensional normal distribution of the random effects. As already discussed, when the response model, g, is normal (and λ is the identity link), the integral is tractable and (restricted) maximum likelihood estimates can be found using, for example, (R)IGLS (Goldstein, 1989). However, when the response is discrete, there is no closed form expression for the integral. There are two approaches to obtain approximate maximum likelihood estimates. One is to replace the likelihood above by a quasi-likelihood (marginal or penalised), which can be maximised using the algorithm for normal models (Breslow and Clayton, 1993; Goldstein and Rasbash, 1996). The other is numerical or Monte-Carlo integration, which we return to below.

While quasi-likelihood estimates are computationally relatively quick to calculate, the resulting fixed and random parameters tend to be biased towards zero. This bias is particularly noticeable for the variance of the random effects, Σ, and persists even if a second-order PQL approximation is used, particularly if there are few level 1 units for each level 2 unit and/or the probabilities are close to 0 or 1. Accordingly, it is often desirable to carry out a second stage of bias correction.

Below, we compare the method proposed by Kuk (1995) with an application of the RM search algorithm to this problem. As the RM algorithm has certain optimality properties, its use may be preferable in practice. First, we briefly outline the method for bias correction proposed by Kuk (1995) and then describe our application of the RM search algorithm to the problem.

3.1 Kuk's method for bias correction

Let θ be the vector of parameters in the model (fixed and random) and suppose that quasi-likelihood estimation on the observed data gives the biased estimate $\tilde{\theta}$. Start the algorithm with $\hat{\theta}^0_{\mathrm{BC}} = \tilde{\theta}$. At step i:

1) Use a parametric bootstrap to simulate H data sets from the model with parameters $\hat{\theta}^{i-1}_{\mathrm{BC}}$.
2) Fit the model (using quasi-likelihood) to each of these H data sets, obtaining H bootstrap estimates $\tilde{\theta}^*_h, h = 1, \ldots, H$. Calculate the average of these, $\bar{\theta}^* = \sum_{h=1}^{H} \tilde{\theta}^*_h / H$.
3) Set $\hat{\theta}^i_{\mathrm{BC}} = \hat{\theta}^{i-1}_{\mathrm{BC}} + (\tilde{\theta} - \bar{\theta}^*)$.
4) Repeat steps 1 to 3 until $|\hat{\theta}^i_{\mathrm{BC}} - \hat{\theta}^{i-1}_{\mathrm{BC}}|$ is acceptably small.

Typically, H is 300 to 1000, and at least five iterations are needed.

Full details are given by Rasbash *et al.* (2000). Kuk's method has been implemented in the MLwiN software; it is fairly reliable but computationally intensive. A drawback is that there is no ready estimate of the standard error of the bias-corrected estimates.

To see why, recall that θ is defined to be the vector of parameter estimates, and note that the bias-corrected estimate can be thought of as some function $h(\tilde{\theta})$, with approximate variance (from a Taylor expansion) $\nabla h \, \mathrm{Var}(\tilde{\theta}) \nabla h^T$. While quasi-likelihood estimation gives an estimate of $\mathrm{Var}(\tilde{\theta})$, bias correction generates little information about ∇h in the neighbourhood of $h(\tilde{\theta})$ and additional simulation is needed to estimate this.

3.2 Application of the RM method to bias correction

Kuk (1995) does not discuss the efficiency of his method, in terms of obtaining the smallest MSE for the estimated bias of the PQL parameter estimates given a fixed number of simulated data sets. In contrast, the RM search process is known to be fully efficient under certain circumstances (Wetherill and Glazebrook, 1986, chapter 9). Further, it has been usefully applied to the problem of finding confidence intervals by test inversion (eg, Carpenter, 1999).

Accordingly, we investigated applying this method to bias-correcting quasi-likelihood parameter estimates. Recalling θ is the vector of parameter estimates (fixed and random) the RM algorithm is as follows:

Denote by $\tilde{\theta}$ the possibly biased quasi-likelihood estimate of θ obtained by fitting the model to the observed data. Now let $\hat{\theta}^i_{\mathrm{BC}}$ be the ith bias-corrected estimate from the RM algorithm and set $\hat{\theta}^0_{\mathrm{BC}} = \tilde{\theta}$. At step i:

1) Simulate a data set with parameter values $\hat{\theta}^{i-1}_{\mathrm{BC}}$, and find the quasi-likelihood parameter estimates, denoted $\tilde{\theta}^*$.
2) Set $\hat{\theta}^i_{\mathrm{BC}} = \hat{\theta}^{i-1}_{\mathrm{BC}} + a_i(\tilde{\theta} - \tilde{\theta}^*)$, where $a_i = c/i$, and the choice of c is discussed below.
3) Repeat steps 1 and 2 until convergence.

In contrast to Kuk's method, the RM algorithm only requires 1, not H, data sets to be simulated at each step; however, the number of steps required till convergence is typically much greater. The choice of the constant c is important; if it is chosen optimally, then the RM algorithm makes efficient use of the simulations; more formally, it achieves the Cramér-Rao lower bound for estimation of the bias of the PQL estimates in the limit as $i \to \infty$ (Wetherill and Glazebrook, 1986, chapter 9).

Both Kuk's method and the RM algorithm can be thought of as stochastic versions of a Newton-Raphson search. To see this, consider the problem of finding bias-corrected estimates as the problem of finding the zero of the equation $k(\theta) = \tilde{\theta} - b(\theta)$, where b gives the expected quasi-likelihood estimate from a data set with true parameter θ. If k were known analytically, given the last estimate $\hat{\theta}_{\mathrm{BC}}^{i-1}$, the new estimate is

$$\hat{\theta}_{\mathrm{BC}}^{i} = \hat{\theta}_{\mathrm{BC}}^{i-1} + k(\hat{\theta}_{\mathrm{BC}}^{i-1})\{\nabla k(\hat{\theta}_{\mathrm{BC}}^{i-1})\}^{-1}$$

where, in the general case of more than one parameter, k is a vector function with derivative ∇k.

In Kuk's method, we assume that $\nabla k = 1$, and use a Monte-Carlo estimate of $k(\theta)$, based on H bootstrap data sets,

$$\tilde{\theta} - \bar{\theta}^{*} = \tilde{\theta} - \left\{\sum_{h=1}^{H} \tilde{\theta}_{h}^{*}/H\right\}.$$

In contrast, the RM approach uses $H = 1$, and convergence occurs because the update steps, c/i, go to 0 as $i \to \infty$. Also the RM method asymptotically (as $i \to \infty$) achieves the Cramér-Rao lower bound for unbiased non-parametric estimators if the constant c is chosen as $c = \{\nabla k(\theta)\}^{-1}$, where θ is the true parameter value.

The attraction of the RM algorithm is that the efficiency declines slowly from the optimum provided c is over-estimated. However, if it is too small, the efficiency can decline quite fast (Wetherill and Glazebrook, 1986).

However, as in practice the true value of θ is unknown, and ∇k is usually set equal to 1, it is unclear which method will be more efficient. We, therefore, carried out a simulation study to investigate this, which we briefly report in the next subsection.

3.3 Simulation study to compare Kuk's method and RM algorithm for bias correction

Following Kuk (1995), 100 data sets were simulated from a two-level logistic model, $\mathrm{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j$, with two level 1 units per level 2 unit (ie, $i = 1 \cdots 2$, $j = 1 \cdots 15$), x_{ij} from -14 to 15 in steps of 1, $\beta_0 = 0.2$, $\beta_1 = 0.1$, $u_j \sim N(0, \sigma_u^2)$ and $\sigma_u^2 = 1$. The response is binomial with $n = 6$.

All experiments used 500 simulations in total. Various choices for the two bias correction methods were explored. For Kuk's method, we performed 10 iterations of bias correction with $H = 50$ simulations each and 50 iterations with $H = 10$ simulations each. For the RM algorithm, the constant, c, is chosen to be 1 and 1.5 with 500 iterations

Table 1 Comparison of bias-correction by Kuk's, Robbins-Monro, SML and numerical quadrature (SAS NLMIXED, see Section (4.3)) methods

			Kuk's		Robbins-Monro		Simulated maximum likelihood	Numerical quadrature (SAS NLMIXED)	
	True value	Average of glmmPQL estimates	10[a]	50[a]	$c = 1$[a]	$c = 1.5$[a]	Not applicable	Not applicable	
			50[b]	10[b]	500[b]	500[b]	500[b] (draws)	(a)	(b)
β_0	0.2	0.181	0.188	0.189	0.185	0.187	0.203	0.190	0.193
β_1	0.1	0.093	0.097	0.095	0.095	0.096	0.099	0.099	0.099
σ_u^2	1.0	0.796	1.025	1.018	0.986	0.980	0.939	0.927	0.968
Mean Squared Error									
β_0			0.126	0.140	0.133	0.127	0.136	0.135	0.133
β_1			0.00133	0.00144	0.00134	0.00135	0.00134	0.00134	0.00133
σ_u^2			0.678	0.712	0.549	0.526	0.561	0.478	0.478

(For SAS NLMIXED, column (a) reports the average over the 99 simulated datasets for which NLMIXED converged with positive Hessian eigenvalues; column (b) reports the average over the 84 simulations for which NLMIXED reported all projected gradients <0.001 at the final step.)
[a]Number of steps.
[b]Number of simulations.

in both cases. First-order PQL estimation using the glmmPQL function in the MASS library (Venables and Ripley, 2002) of *R* (version 1.6.2) was used to obtain the initial estimates, which were then bias-corrected. Note the algorithm implemented in glmmPQL differs from that of Goldstein and Rasbash (1996). In brief, the variance in glmmPQL is estimated as a proportion of $V(y_i|u)$; hence, it allows under- or over-dispersion by default. The results are in Table 1.

All three glmmPQL estimates are biased downwards towards zero and the bias is largest for the variance. In line with theory suggesting the RM algorithm, it is fairly efficient if c is close to its optimum value, mean squared errors from the RM algorithm with $c = 1.5$ are close to or smaller than those from Kuk's method. This suggests that the RM algorithm with $c = 1.5$ is preferable. However, in other examples c may be different; moreover neither method gives ready estimates of the variance of the bias-corrected estimates.

4 Simulated maximum likelihood

An alternative approach to finding maximum likelihood estimates for multilevel models with discrete responses is to use Monte-Carlo integration to repeatedly evaluate the likelihood given by Equation (3.1) as a search is carried out over plausible values of β and Σ to find those which give the maximum.

We now describe how this can be done using importance sampling (see eg, Ripley, 1987, p. 122). Briefly, if a random variable X has a probability density function $p(x)$, and $q(x)$ is a probability density function defined on the same support, then, for any

function k,

$$\int k(x)p(x)\,dx \cong \frac{1}{H}\sum_{h=1}^{H} k(x_h)p(x_h)/q(x_h) \tag{4.1}$$

where $(x_1, \ldots, x_H)$ is an independent identically distributed sample from the density $q(x)$, and the approximation becomes exact as $H \to \infty$. The probability density $q(x)$ is known as the *importance density*.

We can apply this idea to Equation (3.1) as follows. Let $\tilde{\Sigma}$ be the quasi-likelihood estimate of Σ. Choose as the importance density, the multivariate normal density with mean 0 and covariance matrix $\tilde{\Sigma}$, which we write as $f(u, \tilde{\Sigma})$. Simulate a large number, H, of draws from this distribution, $(u_1, \ldots, u_H)$. Then applying the importance principle (4.1),

$$\begin{aligned} L(y; \beta, \Sigma) &= \int \prod_{i=1}^{I} \{g(y_i|u, \beta, x_i, z_i)\} f(u, \Sigma)\,du \\ &\cong \frac{1}{H}\sum_{h=1}^{H}\prod_{i=1}^{I} \frac{\{g(y_i|u_h, \beta, x_i, z_i)\} f(u_h, \Sigma)}{f(u_h, \tilde{\Sigma})}. \end{aligned} \tag{4.2}$$

This approximation can be used in two ways. First, we can obtain an estimate of the log-likelihood at the quasi-likelihood parameter estimates by setting (β, Σ) equal to the quasi-likelihood estimates $(\tilde{\beta}, \tilde{\Sigma})$. In this case, the importance ratio will be equal to 1.

Second, by repeatedly calculating the right hand side of this expression at various values of (β, Σ), we can search for the parameter values that maximise the simulated likelihood. Note that we do not simulate a new sample, $(u_1, \ldots, u_H)$, as the search progresses. Rather, we simulate a sample at the beginning, and, conditional on this, find (β, Σ) to maximise the Monte-Carlo estimate. If a new sample is drawn, then the estimate of the log-likelihood will be slightly different at the same parameter values, which is sufficient for optimising software to report an error.

In order to search for the maximum, optimising software (which uses versions of the Newton–Raphson search) requires the derivative of Equation (4.2) with respect to (β, Σ). Although this can be worked out numerically, this is prohibitively slow for practical use. We have, therefore, calculated the derivatives of the right hand side of Equation (4.2), with respect to (β, Σ), for a sample of size H. Expressions for these are given in Appendix A.

Once the maximum has been found, the analytic expressions for the derivative can be used to compute an estimate of the second derivative of the log-likelihood at the maximum, whose inverse is the asymptotic variance matrix of the parameter estimates. Thus, unlike bias correction of quasi-likelihood estimates using Kuk's method or a RM search, SML gives standard errors of the bias-corrected parameter estimates with relatively little additional computational effort.

4.1 Two-stage algorithm approach to obtain maximum likelihood estimates

McCulloch (1997) reports that SML only works well if the importance distribution is fairly accurate. If this is not the case, then one option is to use the iterative procedure described in Appendix B. Although, given time, this should lead to the maximum likelihood estimates, it is too slow for practical use; further it ignores the fact that good starting values are available from the quasi-likelihood estimates. The most accurate of these are the second-order PQL estimates (Goldstein and Rasbash, 1996). Accordingly, we propose the following algorithm:
Two-stage algorithm:

1) Fit the model using second-order PQL to give estimates denoted $(\tilde{\beta}, \tilde{\Sigma})$.
2) Set the importance density $\Sigma_I = \tilde{\Sigma}$ and simulate $(u_1, \ldots, u_H)$ from the multivariate normal distribution with covariance matrix Σ_I.
3) Fixing $(u_1, \ldots, u_H)$, find the values of (β, Σ) that maximise the Monte-Carlo estimate of the likelihood in Equation (4.2), using the derivatives in Appendix A. Report these as the maximum likelihood estimates.
4) Use the analytical form of the first derivatives to obtain a numerical estimate of the second derivative of the log-likelihood function at the maximum. Invert this for an estimate of the variance matrix of the parameter estimates.

In the next section, we describe a simulation study comparing this approach with Kuk's method for bias correction.

4.2 Comparison of Kuk's method and 'two-stage algorithm' approach for bias correction

The 100 simulated data sets used in the comparison between Kuk's and the RM methods are used again in this simulation study. For the 'two-stage' SML approach, H (the number of draws from the importance density) is set at 500. The results are shown in Table 1.

An unconstrained search routine ('fminunc' in Matlab) based on the interior-reflective Newton method is used in the SML approach. This method is described in Coleman and Li (1996, 1994). Three simulated data sets resulted in non-convergence. In these cases, we performed a constrained search where the variance is constrained to be positive.

The results suggest SML gives comparable mean squared errors, but with the advantage of giving standard errors of bias-corrected parameter estimates and an estimate of the log-likelihood at the maximum.

4.3 Comparison of 'two-stage algorithm' and numerical quadrature

To complete this section, we report the results of applying SAS proc NLMIXED to the 100 simulated data sets. Proc NLMIXED uses numerical integration using quadrature. It selects the number of quadrature points adaptively. It does this by evaluating the log-likelihood function at the starting values with increasing numbers of quadrature points until the difference between two successive evaluations is less than the value set by the option, QTOL. We use the default value of 1E-4. Proc NLMIXED would not converge

unless starting values were provided. We chose these to be 0 for both fixed parameters and 1 for the random intercept variance. Proc NLMIXED reports the projected gradient at the final parameter estimates, which should be close to 0 if they are close to the maximum likelihood estimates. With these starting values, NLMIXED did not converge properly for one data set (final Hessian had a negative eigenvalue, and a gradient of 35 for the variance); for 15 others NLMIXED issued a note that at least one projected (absoluted) gradient exceeded 0.001; the maximum of these was 0.012. For 10 others, NLMIXED reported a negative eigenvalue while iterating, but converged satisfactorily with all projected gradients <0.001.

Table 1, rightmost columns, reports the results for both (a) the 99 simulations for which NLMIXED converged and, for reference, (b) the 84 for which NLMIXED converged with all projected gradients <0.001. Unsurprisingly, NLMIXED performed better in (b), but we believe (a) is the fairer comparison with our two-stage procedure (which fitted all 100 datasets). Comparing the NLMIXED results (a) with our two stage procedure, we see that the bias is greater with proc NLMIXED. However, the mean square error is slightly smaller. This is because the variability of the two-stage bias-corrected estimates with only 500 simulations exceeds that of those from NLMIXED. The difference is greatest for the estimated variance. For this parameter, the sample variance of the estimates from NLMIXED is 0.688^2; that from the two-stage approach is 0.747^2. Assuming this variability is proportional to 1/(no. of simulations), we would need $500 \times 0.747^2/0.688^2 = 589$ simulations for similar MSEs.

In the light of this, rather than identifying one or other method as the 'gold standard', we take the view that when both methods give similar results, we can be confident of being close to the maximum likelihood estimates.

5 Data analysis

In this section, we give the results of applying our method to the analysis of two data sets and compare with the results from using EM-Laplace2 (implemented in HLM version 6.02), MCMC (implemented in MLwiN 2.02) and numerical (adaptive) quadrature methods (implemented in (Generalised Linear Latent And Mixed Models (GLLAMM)) (Rabe-Hesketh *et al.*, 2005) in Stata 8.2 and proc NLMIXED in SAS release 8.

5.1 USAsmall data

In these data there is a binary response on 737 children (level 1) from 479 families (level 2) and three covariates. Fifty six percent of families have only one child, making this a 'sparse' data set for which quasi-likelihood estimates are most likely to suffer noticeable bias.

We fit the following model, where the binary response from child i in family j is denoted as Y_{ij}:

$$\text{logit}\{\Pr(Y_{ij} = 1)\} = (\beta_0 + u_j) + \beta_1 X1_{ij} + \beta_2 X2_{ij} + \beta_3 X3_{ij}$$

Table 2 Parameter estimates for the model for the *USAsmall* data from PQL(2), SML, EM-Laplace2, MCMC, proc NLMIXED and GLLAMM

			Estimation method			
	second-order PQL	second-order PQL followed by SML ($H = 3000$)	EM-Laplace2	MCMC	Proc NLMIXED	GLLAMM
Fixed parameters						
β_0	0.905 (0.244)	0.954 (0.275)	0.862 (0.209)	1.020 (0.299)	0.962 (0.276)	0.960 (0.274)
β_1	0.573 (0.445)	0.581 (0.466)	0.487 (0.418)	0.644 (0.491)	0.596 (0.471)	0.599 (0.470)
β_2	1.431 (0.218)	1.501 (0.258)	1.344 (0.208)	1.594 (0.282)	1.515 (0.259)	1.514 (0.259)
β_3	1.798 (0.351)	1.883 (0.408)	1.712 (0.316)	2.014 (0.449)	1.905 (0.410)	1.901 (0.408)
Random parameters						
σ_u^2	2.402 (0.473)	2.800 (1.110)	1.546 (NA)	3.569 (1.399)	2.894 (1.087)	2.874 (1.073)
$-2 \times$ log-likelihood	NA	792.0	2146.4	(DIC) 754.7	791.5	791.5

NA, not available; DIC, Deviance Information Criterion; NLMIXED used adaptive quadrature (Section 4.3); GLLAMM: 8 quadrature points were used.

where u_j is the family-specific random effect with variance σ_u^2. Using i) second-order PQL and ii) second-order PQL followed by SML, as in our two-stage algorithm earlier, gives the results in Table 2.

The second-order PQL estimates of random and fixed parameters are consistent with the downward biases found in Table 1. This is most marked for the variance, σ_u^2, which is 14% smaller than the SML estimate. Further, the standard errors appear to be underestimated by second-order PQL. Again, this is most marked for the standard error of the estimate of σ_u^2.

In addition, SML, like other methods based on numerical integration, gives an estimate of the log-likelihood at the maximum, shown in the bottom row of the table above. This is useful for comparing models. For example, refitting the model without X1 gives $-2 \times$ log-likelihood $= 793.0$ (1 dp). The likelihood ratio test for the importance of X1 is thus $793.0 - 791.4 = 1.6$, which comparing with χ_1^2 is not significant at the 5% level. As expected, this agrees, to 1 dp, with the Wald test $(0.581/0.466)^2 = 1.6$. Also the SML estimates for the fixed and the random parameters are very close to those from the two numerical integration methods, proc NLMIXED and GLLAMM.

Relative to SML and NLMIXED, EM-Laplace2 parameter estimates appear too small; in particular the random parameter, σ_u^2, is 45% less than the SML estimate. The reported deviance from EM-Laplace2 is more than twice those from SML and the other two numerical methods, presumably because it includes an additional constant term from the likelihood. Changes in the deviance are similar, though.

It is interesting to note that the sixth-order Laplace approximation with Fisher scoring (Laplace6) implemented in HLM 5.04 failed to converge in this example. We presume that such convergence difficulties lie behind the replacement of i) sixth-order Laplace approximation with a second-order approximation and ii) the Fisher scoring algorithm with the EM algorithm, in HLM 6.01 and 6.02 (Raudenbush *et al.*, 2004). However, standard errors are currently unavailable for the variance components estimated using

EM-Laplace2 in HLM 6.01 or 6.02. The latter is the latest version available at the time of writing.

The MCMC estimate for σ_u^2 is about 27% higher than the SML estimate. The Gibbs sampling with Gamma diffuse priors ($\alpha = 0.001$ and $\beta = 0.001$) used in the MCMC run in MLwiN is based on a burn-in of 5000 iterations and a chain-length of 100000. It was run for another 400000 iterations to confirm the relatively large level 2 variance. The mean (SD), median and effective sample size (assuming successive values in chain independent) of the resulting chain are 3.511 (1.339), 3.126 and 2014. Our results are probably because of the fact that, although the prior is diffuse, it gives zero weight to a zero variance, so that when there is relatively little information in the data, both the posterior mean and median are pulled away from zero relative to the maximum likelihood estimate. Interestingly, the Bayesian Deviance Information Criterion (DIC) is fairly close to those from SML and the numerical methods.

5.2 BANG data

This data set comes from a study of the fertility of women in Bangladesh and is typical of social science data to which our method can usefully be applied.

In the sub-sample analysed here, the response is the contraceptive use of women (1 = yes, 0 = no) at the time of the survey. Three covariates are considered in our illustrative analysis: number of living children (four categories: 0, 1, 2 and ≥3), age (centred at the sample mean) and region of residence (1 = urban, 0 = rural).

Let Y_{ij} denote contraceptive use of women i in district j, and 1[] an indicator function for the event in brackets. We fit the following model:

$$\begin{aligned}\text{logit}\{Pr(Y_{ij} = 1)\} = (\beta_0 + u_j) + (\beta_1 + v_j) \times 1[\text{urban resident}] + \beta_2 \times (\text{centred age})_{ij} \\ + \beta_3 \times 1[1 \text{ child}] + \beta_4 \times 1[2 \text{ children}] + \beta_5 \times 1[\geq 3 \text{ children}]\end{aligned}$$

$$\begin{pmatrix} u_j \\ v_j \end{pmatrix} \sim \text{N}(0, \Omega); \quad \Omega = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}.$$

In this example, region of residence is not taken as a level in the data hierarchy but is treated as a level 1 covariate in the model. The dummy variable for urban residential effect on the use of contraceptive is included in the random part of the model thus allowing variance between districts to differ between rural and urban regions.

Table 3 gives the parameter estimates for the various approaches considered. For SML, we see the adjusted odds of using contraception is 2.3 times greater for women in urban areas than those in rural areas, reduced by about 0.8 times for each decade-increase in age, roughly three times greater with the first child and four times greater with two or more children (compared with having no children). There appears to be significant variability between districts, but this is reduced by $2 \times -0.41 + 0.69 = -0.13$ (log-odds scale) in urban regions.

To carry out a likelihood ratio test for the latter effect, we refitted the model without the random term for urban residence and its covariance with the random intercept. This gave a $-2 \times$ log-likelihood = 2413.0. Thus, the likelihood ratio test for this effect is

Table 3 Parameter estimates for the model for the *BANG* data from PQL(2), SML, EM-Laplace2, MCMC, proc NLMIXED and GLLAMM

				Estimation method			
	Second-order PQL	Second-order PQL followed by SML ($H = 3000$)		EM-Laplace2	MCMC	Proc NLMIXED	GLLAMM
Fixed parameters			Odds ratio				
β_0	−1.713 (0.159)	−1.713 (0.162)	0.18	−1.710 (0.140)	−1.727 (0.162)	−1.711 (0.161)	−1.712 (0.161)
β_1	0.816 (0.170)	0.813 (0.173)	2.25	0.815 (0.173)	0.827 (0.178)	0.815 (0.171)	0.816 (0.172)
β_2	−0.026 (0.008)	−0.026 (0.008)	0.97	−0.026 (0.010)	−0.027 (0.008)	−0.026 (0.008)	−0.026 (0.008)
β_3	1.135 (0.159)	1.133 (0.160)	3.10	1.131 (0.138)	1.137 (0.160)	1.133 (0.160)	1.133 (0.160)
β_4	1.360 (0.176)	1.359 (0.177)	3.90	1.357 (0.202)	1.366 (0.177)	1.358 (0.177)	1.358 (0.177)
β_5	1.357 (0.181)	1.355 (0.183)	3.90	1.352 (0.194)	1.364 (0.180)	1.354 (0.183)	1.354 (0.183)
Random parameters							
σ_u^2	0.396 (0.118)	0.398 (0.132)		0.379 (NA)	0.435 (0.144)	0.388 (0.129)	0.390 (0.129)
σ_{uv}	−0.414 (0.160)	−0.407 (0.177)		−0.391 (NA)	−0.455 (0.191)	−0.404 (0.175)	−0.406 (0.176)
σ_v^2	0.686 (0.284)	0.661 (0.339)		0.627 (NA)	0.770 (0.340)	0.664 (0.322)	0.666 (0.322)
−2× log-likelihood							
Current model	NA	2398.9		5953.1	(DIC) 2386.4	2398.7	2398.6
No random effect for urban residence	NA	2413.0		5968.1	(DIC) 2407.7	2413.7	2413.7
Change in deviance (to 1 dp)	NApp	14.1		15.0	NApp	15.0	15.0

NA, Not available; NApp, Not applicable; GLLAMM, Eight-quadrature points were used; NLMIXED used adaptive quadrature (Section 4.3).

$2413.0 - 2398.9 = 14.1$. Comparing this with the χ^2_2 distribution, we conclude that there is strong evidence to support a reduction in the variability of contraceptive use in urban regions. (Note this is an approximate test, because of the constraints on the variance terms. Corrections to allow for this are discussed by Self and Liang (1987). However, the conclusions are unchanged here.)

The fixed and random parameters obtained from using EM-Laplace2 are close to those from SML and the two numerical integration methods. This is different from the previous *USAsmall* example where parameter estimates were markedly lower when compared with SML and the other methods considered. The reported deviance from EM-Laplace2 is 5953.1. Refitting the same model without allowing random effect for urban residence gave a deviance of 5968.1. This yields a change in deviance of 15.0, to 1 dp, on 2 d.f. It is slightly higher than that from SML but the same as those from the two numerical methods as shown in Table 3. Note the larger value of the deviance (relative to the other software) just reflects that the EM-Laplace2 adds a different (arbitrary) additive constant to the deviance. This does not affect inferences, which are based on changes in the deviance.

The estimates for the three random parameters from MCMC with Gamma diffuse priors ($\alpha = 0.001$ and $\beta = 0.001$) are again somewhat higher (on average 13%) though to a lesser extent than those from SML and the two numerical methods.

It is interesting to note that proc NLMIXED failed to converge for the model considered in this example when user-supplied starting values were not provided (default starting values for all parameters are 1). For this and the earlier *USAsmall* example PQL(2) estimates were provided as starting values for proc NLMIXED. User-supplied starting values were not needed for GLLAMM in either of these examples. It uses parameter estimates obtained from ordinary logistic model as starting values.

Finally, note that, in this example, there is little difference between the second-order PQL estimates and those obtained by following second-order PQL with SML. This reflects the fact that there are a reasonable number of women (level 1 units) in each district (level 2 unit).

6 Comparison with numerical integration methods in SAS

To conclude, we present a further comparison of PQL followed by SML with the numerical integration method implemented in SAS. We found the latter considerably faster than Stata's GLLAMM (see below for some timings) especially when there are many random effects in the model.

We simulated data from a two-level logistic model,

$$\text{logit}\{\Pr(Y_{ij} = 1)\} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})X1_{ij} + (\beta_2 + u_{2j})X2_{ij} + (\beta_3 + u_{3j})X3_{ij}$$

where

$$\beta = [1111]', \quad X1 \text{ to } X3 \sim \text{N}(0, 1) \text{ and } \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \end{bmatrix} \sim N\left(0, \begin{bmatrix} 0.5 & & & \\ 0.05 & 0.5 & & \\ 0.05 & 0.05 & 0.5 & \\ 0.05 & 0.05 & 0.05 & 0.5 \end{bmatrix}\right).$$

Table 4 Results from SML ($H = 5000$) and SAS proc NLMIXED

Random parameters in model	SML			SAS proc NLMIXED		
	Time (min)	Max. |gradient|	Log-likelihood	Time (min)	Max. |gradient|	Log-likelihood
1	3.4	0.004121	−1033.097	0.05	1.723e-6	−1033.021
3	4.9	0.017214	−1019.218	0.4	0.000719	−1019.143
6	30.0	0.006752	−1006.041	4.65	0.000437	−1005.950
10	86.1	0.004275	−995.822	32.35	0.00323	−997.042

There are 200 level 2 units with 1 to 19 level 1 units each.

In the comparison, models with increasing numbers of random effects are fitted. Results from our 'two-stage algorithm' are compared with those from proc NLMIXED. The optimisation technique used in proc NLMIXED is quasi-Newton and adaptive Gaussian quadrature is used to integrate over the random effects numerically. At sensible precision, parameter estimates and standard errors were very similar for both methods (results not shown). Table 4 shows the relative time taken to converge, the log-likelihood value and absolute value of the largest gradient at convergence.

The log-likelihood values are comparable between the two methods. The gradients at convergence are lower in NLMIXED. However, imposing a more stringent convergence criterion on SML would only be sensible if H were increased. Elapsed time goes up at an exponential rate in NLMIXED, while we were able to achieve comparable accuracy with a more linear increase in time using SML.

The similar parameter estimates, standard errors and log-likelihoods are encouraging. As the relative increase in computational time for SAS with increasing number of random effects is much faster SML would appear to have the edge for models with a large number of random effects. We have not optimised our code in a low-level language, so it is plausible that further optimisation could bring the baseline times for the two methods closer. In addition, as discussed below, SML has the potential to extend to non-normal random effects and more than two levels.

Finally, note that GLLAMM took 22 s, 4.5 min and 93 min for the models with one, three and six random parameters. The model with 10 random parameters had not converged after 15 h.

7 Discussion and conclusions

We have investigated methods for bias correction of parameter estimates in two-level models for binary data. Initially, we used a simulation study to compare three simulation-based methods: i) Kuk's; ii) the RM search; and iii) SML, as a means to bias-correct quasi-likelihood estimates. Our results show that SML performs comparably with the other methods, but has the advantage that it yields estimates of the variances of the parameter estimates together with an estimate of the log-likelihood at the maximum that can be used for comparing models.

Further, SML has some computational advantages over Kuk's method and the RM search. First, SML does not require refitting the model to each bootstrap data set, so it is not prone to problems when unusual bootstrap data sets are simulated, to which either the model cannot be fitted or which gives outlying parameter estimates, such as negative variances. Even when second-order PQL estimates converge readily on the original data, these problems mean Kuk's and the RM approach require ad-hoc 'fudges' to the code to work automatically. In contrast, our experience is that, given second-order PQL estimates have converged, SML is far more robust.

Our experience confirms that of McCulloch (1997), that SML needs good starting values to converge to the maximum likelihood estimates. Using SML after second-order PQL therefore seems sensible. Note that if bias-correction is not required, our code can be used to provide an estimate of the log-likelihood at the maximum.

There are several areas in which refinements could be made to our algorithm. First, convergence could be assessed more formally. For example, a running estimate of the standard error of the estimated likelihood function could be computed, and simulation stopped when it is smaller than a user-specified threshold. Alternatively, Booth *et al.* (2001) note that the variability of SML tends to increase as the random effects get larger, so it might make more efficient use of simulation to concentrate more heavily on level 2 units with larger random effects. Another approach would be to use a level-2 unit-specific importance distribution, based on the second-order PQL estimate of the mean and variance of the subject-specific random effect. Clearly, our code can be generalised to include other distributions (Poisson, negative binomial) and link functions. In addition, the averaging step (the slowest part) of SML is a natural candidate for parallel processing.

It may be possible to further improve the accuracy of SML using carefully chosen so-called 'good lattice points', instead of a random sample from the importance distribution. In empirical studies, the results with such points have been promising (Pan and Thompson, 1998).

In our illustrative analysis of the *USAsmall* data set with its sparse data structure, we found that EM-Laplace2 produced estimates that are markedly lower in both fixed and random parameters when compared with estimates from the other methods considered, including PQL(2). As this marked reduction in magnitude of parameter estimates was not replicated in the analysis of the *BANG* data set where a reasonable number of level 1 units were nested within most of the level 2 units, we are led to believe that EM-Laplace2 in HLM may produce downwardly biased estimates when the data are sparse. However, this would require a more extensive simulation study to verify.

Turning to the comparison with numerical integration, we used SAS NLMIXED to fit the same simulated data sets used in the comparison between Kuk's method, RM search and SML. We found the bias was fractionally larger with NLMIXED, which also had some convergence problems. While mean square error was slightly smaller with NLMIXED, this is not surprising as we only used 500 simulations for SML. For our data sets SML and NLMIXED gave similar parameter estimates. We also found that when increasing number of random effects were included in the model, we were able to obtain comparable results to numerical integration without increasing the power of H by 1 for each new random effect. Thus, the computing time of SAS's numerical integration increased faster than for SML. We are hopeful that if more effort were put into

optimising code it would also be more comparable in terms of speed. The rapid increase in computing power available to researchers will also help here. The advantages are: i) the code we have written would be relatively simple to extend to at least single non-normal random effects; and ii) the method extends to models with more than two levels which SAS's proc NLMIXED is not designed for (although, with the use of certain tricks, savvy users may be able to fit three-level models). Although, using SML with three or more levels requires a modification of our code and a considerable increase in H, nevertheless Carpenter *et al.* (2003) obtained useful results in a practical example.

Finally, note that our approach is complementary to fitting similar models using a Bayesian approach, using MCMC methods in MLwiN, WinBUGS (Spiegelhalter *et al.* 1999) or other Bayesian software. Advantages of our approach are that it is unnecessary to specify priors for variance terms; which even if notionally 'uninformative' may turn out to have unforeseen effects (see Tables 2 and 3); that a likelihood is available for testing, and that Monte-Carlo error can be relatively straightforwardly addressed (eg, Booth and Hobert, 1999).

In summary, we conclude the following: SML (starting from PQL estimates) performed better than bias-corrected versions of PQL by Kuk's method or the RM algorithm. It makes as efficient use of simulations, yet avoids their computational problems and in addition provides standard errors and an estimate of the log-likelihood at the maximum. Our comparison of SML, EM-Laplace2, a Bayesian approach with uninformative priors and numerical integration, came out in favour of SML and numerical integration. The former is more flexible, and in our simulation study slightly less biased. However, provided good starting values are available and taking into account computational speed, SAS proc NLMIXED appears the 'best of the rest' for standard two-level analyses.

Matlab code implementing SML is available from the Journal website.

Acknowledgement

The authors would like to thank Prof. Anthony Davison for his useful comments when this work was presented at the Ecole Polytechnique Fédéral de Lausanne, and the reviewers for their helpful criticisms. This work was funded by the Economic Social Research Council, UK (award reference R000223547).

References

Amin S, Diamond I and Steele F (1997) Contraception and religiosity in Bangladesh. In Jones GW, Caldwell JC, Douglas RM and D'Souza, RM editors, *The Continuing Demographic Transition*. Oxford University Press, 268–89.

Booth JG, Hobert JP and Jank W (2001) A survey of Monte Carlo algorithms for maximising the likelihood of a two-stage hierarchical model. Technical report, University of Florida, Department of Statistics.

Booth JG and Hobert JP (1999) Maximising generalised linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B* **61**, 265–85.

Breslow NE and Clayton DG (1993) Approximate inference in generalised linear

mixed models. *Journal of the American Statistical Association* **88**, 9–25.
Browne WJ (2002) *MCMC estimation in MLwiN*. Institute of Education.
Carpenter JR, Ho T and Pocock S (2003) Modelling a repeated ordered categorical response in clinical trials with smoothing splines: angina grade in the RITA-2 trial. Technical report, Medical Statistics Unit, London School of Hygiene and Tropical Medicine.
Carpenter JR (1999) Test inversion bootstrap confidence intervals. *Journal of the Royal Statistical Society, Series B* **61**, 159–72.
Coleman TF and Li Y (1996) An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimisation* **6**, 418–45.
Coleman TF and Li Y (1994) On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds. *Mathematical Programming* **67**, 189–224.
Congdon R (2005) *Personal Communication.*
Goldstein H (1986) Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika* 73, 43–56.
Goldstein H (1989) Restricted unbiased iterative generalised least squares estimation. *Biometrika* **76**, 622–23.
Goldstein H (2003) *Multilevel statistical models* (3rd edition). Arnold.
Goldstein H and Rasbash J (1996) Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A* **159**, 505–13.
Huq NM and Cleland J (1990) Bangladesh Fertility Survey 1989 (Main Report). National Institute of Population Research and Training.
Kuk AYC (1995) Asymptotically unbiased estimation in generalised linear models with random effects, *Journal of the Royal Statistical Society, Series B* **57**, 395–407.
Lin X and Breslow NE (1996) Bias correction in generalised linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91**, 1007–1016.
Matlab (2003) Release 13, The Mathworks Inc. (www.mathworks.com).
McCulloch CE (1997) Maximum likelihood algorithms for generalised linear mixed models. *Journal of the American Statistical Association* **92**, 162–70.
Pan J-X and Thompson R (1988) Quasi-Monte Carlo EM algorithm for MLEs in generalised linear mixed models. *Compstat 1998: Proceedings on the 13th Symposium in Computational Statistics*, 419–24.
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, Browne W, Goldstein H, Yang M, Plewis I, Healy M, Woodhouse G, Draper D, Langford I and Lewis T (2000) *A user's guide to MLwiN* (version 2.1). Institute of Education.
Raudenbush SW, Yang ML, Yosef M (2000) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics* **9**, 141–57.
Raudenbush SW, Bryk AS, Cheong YF, Congdon R and du Toit M (2004) *HLM 6: Hierarchical Linear and Nonlinear Modeling (Second printing with revisions).* Scientific Software International, Inc., 109.
Ripley BD (1987) *Stochastic simulation.* Wiley.
Rodríguez G and Goldman N (1995) An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A* **158**, 73–89.
Self SG and Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association* **82**, 605–10.
Spiegelhalter DJ, Thomas A and Best, NG (1999) *WinBUGS version 1.2 user manual.* MRC Biostatistics Unit.
Venables WN and Ripley BD (2002) *Modern applied statistics with S.* Springer (4th edition).
Wetherill GB and Glazebrook KD (1986) *Sequential methods in statistics. (3rd edition).* Chapman & Hall, 161–77.

Appendix A

Derivative expressions for SML

For every level 2, in other words jth unit, the first derivative of the right hand side of Equation (4.2) can be written as

$$\frac{\partial l}{\partial \theta} = \frac{\partial \log L}{\partial \theta} = \frac{\partial}{\partial \theta}\left(-\log(H) + \log \sum_{h=1}^{H} \prod_{i=1}^{I} g_i w_h\right)$$

where

$$\theta \in (\beta, \Sigma), \quad g_i = g\,(y_i|u_h, \beta, x_i, z_i) \quad \text{and } w_h = \frac{f(u_h, \Sigma)}{f(u_h, \tilde{\Sigma})}.$$

$$\frac{\partial l}{\partial \beta} = C^{-1} \sum_{h=1}^{H} w_h \sum_{i=1}^{I} \left(\frac{\partial g_i}{\partial \beta} \prod_{i' \neq i}^{I} g_{i'}\right)$$

where

$$C = \sum_{h=1}^{H} \prod_{i=1}^{I} g_i w_h.$$

For the case where the importance density follows a multivariate normal distribution with mean 0 and covariance matrix $\tilde{\Sigma}$,

$$\begin{aligned} w_h &= \frac{(2\pi|\Sigma|)^{-1/2} \exp(-\frac{1}{2} u_h^T \Sigma^{-1} u_h)}{(2\pi|\tilde{\Sigma}|)^{-1/2} \exp(-\frac{1}{2} u_h^T \tilde{\Sigma}^{-1} u_h)} \\ &= \exp\left[\frac{1}{2}\log\left(|\tilde{\Sigma}|/|\Sigma|\right) + \frac{1}{2} u_h^T \left(\tilde{\Sigma}^{-1} - \Sigma^{-1}\right) u_h\right]. \end{aligned}$$

$$\frac{\partial l}{\partial \gamma} = C^{-1} \sum_{h=1}^{H} \frac{\partial w_h}{\partial \gamma} \left(\prod_{i=1}^{I} g_i\right)$$

where γ is an element of the variance–covariance matrix, Σ, and

$$\frac{\partial w_h}{\partial \gamma} = w_h \left[-\frac{1}{2}\text{trace}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \gamma}\right) - u_h^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \gamma} \Sigma^{-1} u_h\right].$$

For example, for a bivariate distribution, $\frac{\partial \Sigma}{\partial \sigma_1^2} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\frac{\partial \Sigma}{\partial \sigma_{12}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Finally, the full first derivative with respect to θ is simply the derivatives summed over j, that is

$$\sum_{j=1}^{J} \frac{\partial l_j}{\partial \theta}.$$

For three or higher level, cross-classifications and multiple membership models, see Appendix 4.2 in Goldstein (2003).

Appendix B

SML algorithm when starting far from maximum

Let $\hat{\Sigma}^i$ denote the estimate of the covariance matrix at step i. Set the covariance of the importance density $\Sigma_I = \hat{\Sigma}^i$. At step i:

1) Simulate $(u_1, \ldots, u_H)$ from the multivariate normal distribution with covariance matrix Σ_I.
2) Fixing $(u_1, \ldots, u_H)$, find the values of (β, Σ) that maximise the Monte-Carlo estimate of the likelihood in Equation (4.2). Denote these $(\hat{\beta}^{i+1}, \hat{\Sigma}^{i+1})$.
3) Set $\Sigma_I = \hat{\Sigma}^{i+1}$.

Iterate steps 1–3 till convergence.