

Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models

William J. Browne¹ and David Draper²

¹Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, England

²Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, England

Summary

We use simulation studies (a) to compare Bayesian and likelihood fitting methods, in terms of validity of conclusions, in two-level random-slopes regression (RSR) models, and (b) to compare several Bayesian estimation methods based on Markov chain Monte Carlo, in terms of computational efficiency, in random-effects logistic regression (RELR) models. We find (a) that the Bayesian approach with a particular choice of diffuse inverse Wishart prior distribution for the (co)variance parameters performs at least as well—in terms of bias of estimates and actual coverage of nominal 95% intervals—as maximum likelihood methods in RSR models with medium sample sizes (expressed in terms of the number J of level-2 units), but neither approach performs as well as might be hoped with small J ; and (b) that an adaptive hybrid Metropolis-Gibbs sampling method we have developed for use in the multilevel modeling package *MLwiN* outperforms adaptive rejection Gibbs sampling in the RELR models we have considered, sometimes by a wide margin.

Keywords: Adaptive Metropolis Sampling, Diffuse Prior Distributions, Educational Data, Gibbs Sampling, Hierarchical Modeling, IGLS, Markov Chain Monte Carlo (MCMC), MCMC Efficiency, Maximum Likelihood Methods, Random-Effects Logistic Regression, Random-Slopes Regression, RIGLS, Variance Components.

1 Introduction

Multilevel models, for data having a nested or hierarchical structure, have become an important component of the applied statistician's tool-chest in the past 15 years (e.g., Bryk and Raudenbush 1992, Goldstein 1995, Draper 2000). Examples include variance-components (VC), random-slopes regression (RSR), and random effects logistic regression (RELR) models, all of which we will visit in what follows. In the early days of multilevel modeling the only available fitting methods were based on maximum likelihood: iterative generalized least squares (IGLS) and restricted IGLS (RIGLS)—or related methods such as Fisher scoring (Longford 1987), restricted maximum likelihood (REML), and empirical Bayes estimation (Bryk et al. 1988)—for models with Gaussian outcomes (Goldstein 1986, 1989); and marginal quasi-likelihood (MQL) and penalized (or predictive) quasi-likelihood (PQL) for data sets with dichotomous outcomes (e.g., Breslow and Clayton 1993). More recently fully Bayesian analyses based on Markov chain Monte Carlo (MCMC) methods have become possible in packages such as BUGS (Spiegelhalter et al. 1997) and MLwiN (Rasbash et al. 1999). Recent alternatives for fitting multilevel models, which we do not pursue here, include integrated-likelihood approaches based on Gaussian quadrature (e.g., Pinheiro and Bates 1995) and Laplace approximations (e.g., Raudenbush et al. 2000).

We (the authors of this article) are the co-developers of the Bayesian MCMC capabilities in MLwiN. Below we examine (a) the relative performance, in the sense of point and interval estimation accuracy, of likelihood and Bayesian fitting methods in RSR models, and (b) some performance comparisons in RELR models—in the sense of required CPU time to achieve a given accuracy of posterior summary—between several MCMC fitting methods, including adaptive rejection sampling (Gilks and Wild 1992) and an approach we have developed specifically for MLwiN based on adaptive hybrid Metropolis-Gibbs sampling. In a companion article to this one (Browne and Draper 1999, hereafter BD99) we compare likelihood and Bayesian fitting methods in VC and RELR models (also see Hoijsink [this issue] for an MCMC investigation of a random-intercept model).

2 Random-slopes regression (RSR) models

A multilevel modeling data set which we have found useful in fixing ideas was collected by Mortimore et al. (1988) in a study called the Junior School Project (JSP). This was a longitudinal investigation of roughly 2,000 pupils from 50 primary schools chosen randomly from the 636 Inner London Education Authority (ILEA) schools in 1980. Woodhouse et al. (1995) examined a random subsample of $N = 887$ students at $J = 48$ of these schools; here we will refer to this subsample as the JSP data. One focus of principal interest was the relationship between mathematics test scores at year 3 (math3) and

Table
label
in par
(for tl

9

year 5
not fa
appro
of the
school
5 to 6
so it v
nature
the ch
unstak

$u_j =$

where
math3
over al
of the
schools
having

