261

In Latent Variable & Latent Structure Models, Marcoulides G, Moustaki, I (eds). Chapter 11, pp. 225-244 Lawrence Erlbaum (2002) 0-8058-4046-X

11 Multilevel Factor Analysis Modelling Using Markov Chain Monte Carlo Estimation*

Harvey Goldstein and William Browne

Institute of Education, University of London

11.1 Introduction

Traditional applications of structural equation models have, until recently, ignored complex population data structures. Thus, for example, factor analyses of achievement or ability test scores among students have not adjusted for differences between schools or neighbourhoods. In the case where a substantial part of inter-individual differences can be accounted for by such groupings, inferences which ignore this may be seriously misleading. In the extreme case, if *all* the variation was due to a combination of school and neighbourhood effects, a failure to adjust to these would lead to the detection of apparent individual level factors which would in fact be non-existent.

Recognising this problem, McDonald and Goldstein (1989) presented a multilevel factor analysis and structural equation model where individuals are recognised as belonging to groups and explicit random effects for group effects are incorporated. They presented an algorithm for maximum likelihood estimation (this model was further explored by Longford & Muthèn, 1992; McDonald, 1993). Raudenbush (1995) applied the EM algorithm to estimation for a 2-level structural equation model. Rowe and Hill (in press) show how existing multilevel software can be used to provide approximations to maximum likelihood estimates in general multilevel structural equation models.

In this chapter these models are extended in two ways. First, it is shown how a Markov Chain Monte Carlo (MCMC) algorithm can be used to fit such models. An important feature of the MCMC approach is that it decomposes the computational algorithm into separate steps, for each of which there is a relatively straightforward estimation procedure. This provides a chain sampled from the full posterior distribution of the parameters from which one can calculate uncertainty intervals based upon quantiles etc. The second advantage is that the decomposition into separate steps allows one to easily extend the procedure to the estimation of very general models, and an illustration of how this can be done is provided.

* **Acknowledgements** This work was partly carried out with the support of a research grant from the Economic and Social Research Council for the development of multilevel models in the Social Sciences. We are very grateful to Ian Langford, Ken Rowe and Ian Plewis for helpful comments.

A fairly general 2-level factor model can be written as follows, using standard factor and multilevel model notation:

$$\begin{aligned} &Y = \Lambda_2 v_2 + u + \Lambda_1 v_1 + e \\ &Y = \{y_{rij}\}, \\ &r = 1, \ldots, R \;\; i = 1, \ldots, n_j \;\; j = 1, \ldots, J \end{aligned} \tag{11.1}$$

where the "uniquenesses" u (level 2), e (level 1) are mutually independent with covariance matrix Ψ_1, and there are R response measures. The Λ_1,Λ_2 are the loading matrices for the level 1 and level 2 factors and the v_1, v_2 are the independent factor vectors at level 1 and level 2. Note that there can be different numbers of factors at each level. By adopting the convention of regarding the measurements themselves as constituting the lowest level of the hierarchy, Equation 11.1 is regarded as a 3-level model. Extensions to more levels are straightforward.

Model (11.1) allows for a factor structure to exist at each level and one needs to further specify the factor structure, for example that the factors are orthogonal or patterned with corresponding identifiability constraints. Further restrictions can be imposed. For example, one may wish to model the uniquenesses in terms of further explanatory variables. In addition, one can add measured covariates to the model and extend to the general case of a linear structural or path model (see discussion).

11.2 A Simple Illustration

To illustrate these procedures consider a simple single level model which can be written as

$$\begin{aligned} &y_{ri} = \lambda_r \nu_i + e_{ri}, \qquad r = 1, \ldots, R, \quad i = 1, \ldots, N \\ &\nu_i \sim N(0,1), \quad e_{ri} \sim N(0, \sigma_{er}^2) \end{aligned} \tag{11.2}$$

This can be viewed as a 2-level model with a single level 2 random effect (v_i) with variance constrained to 1 and R level 1 units for each level 2 unit, each with their own (unique) variance.

If the values of the "loadings" λ_r were known, then Model (11.2) could be fitted directly as a 2-level model, with the loading vector as the explanatory variable for the level 2 random effects with variance constrained to be equal to 1; if there are any measured covariates in the model their coefficients could also be estimated at the same time. Conversely, if the values of the random effects v_i were known, one could estimate the loadings; this would now be a single level model with each response variate having its own variance. These considerations suggest that an EM algorithm can be used in the estimation where the random effects are regarded as missing data (see Rubin & Thayer, 1982). In this chapter a stochastic MCMC algorithm is proposed.

MCMC works by simulating new values for each unknown parameter in turn from their respective conditional posterior distributions assuming the other parameters are known. This can be shown to be equivalent (upon convergence) to sampling from the joint posterior distribution. MCMC procedures generally incorporate prior information about parameter values and so are fully Bayesian procedures. In this chapter diffuse prior information is assumed, although algorithms are given that assume generic prior distributions (see below). Inference is based upon the chain values: conventionally the means of the parameter chains are used as point estimates but medians and modes (which will often be close to maximum likelihood estimates) are also available, as illustrated. This procedure has several advantages. In principle it allows one to provide estimates for complex multilevel factor analysis models with exact inferences available. Since the model is an extension of a general multilevel model one can theoretically extend other existing multilevel models in a similar way. Thus, for example, one could consider cross-classified structures and discrete responses as well as conditioning on measured covariates. Another example is the model proposed by Blozis and Cudeck (1999) where second level residuals in a repeated measures model are assumed to have a factor structure. The following section describes the procedure by applying it to the simple example of Equation (11.2) and then applies it to more complex examples.

11.3 A Simple Implementation of the Algorithm

The computations have all been carried out in a development version of the program *MLwiN* (Rasbash et al., 2000). The essentials of the procedure are described next.

Assume that the factor loadings have normal prior distributions, $p(\lambda_r) \sim N(\lambda_r^*, \sigma_{\lambda r}^2)$ and that the level 1 variance parameters have independent inverse Gamma priors, $p(\sigma_{er}^2) \sim \Gamma^{-1}(a_{er}^*, b_{er}^*)$. The * superscript is used to denote the appropriate parameters of the prior distributions.

11.3.1 This model can be updated using a very simple three step Gibbs sampling algorithm

Step 1: Update λ_r (r=1,...,R) from the following distribution: $p(\lambda_r) \sim N(\hat{\lambda}_r, D_r)$ where

$$D_r = \left(\frac{\sum_i \nu_i^2}{\sigma_{er}^2} + \frac{1}{\sigma_{\lambda r}^2}\right)^{-1} \quad and \quad \hat{\lambda}_r = D_r\left(\frac{\sum_i \nu_i y_{ri}}{\sigma_{er}^2} + \frac{\lambda_r^*}{\sigma_{\lambda r}^2}\right)$$

Step 2: Update ν_i (i=1,...,N) from the following distribution: $p(\nu_i) \sim N(\hat{\nu}_i, D_i)$ where

$$D_i = \left(\frac{\sum_r \lambda_r^2}{\sigma_{er}^2} + 1\right)^{-1} \quad and \quad \hat{\nu}_i = D_i\left(\frac{\sum_r \lambda_r y_{ri}}{\sigma_{er}^2}\right)$$

Step 3: Update σ_{er}^2 from the following distribution: $p(\sigma_{er}^2) \sim \Gamma^{-1}(\hat{a}_{er}, \hat{b}_{er})$ where $\hat{a}_{er} = N/2 + a_{er}^*$ and $\hat{b}_{er} = \frac{1}{2}\sum_i e_{ri}^2 + b_{er}^*$.

To study the performance of the procedure, a small data set using the following model and parameters was simulated:

$$\lambda = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \quad \Psi_1 = \begin{pmatrix} 0.2 & & & \\ & 0.3 & & \\ & & 0.4 & \\ & & & 0.5 \end{pmatrix}, \quad \mathrm{N} = 20, \quad \mathrm{R} = 4 \tag{11.3}$$

$$y_{ri} = \lambda_r v_i + e_{ri}, \tag{11.4}$$

The lower triangle of the correlation matrix of the responses is

$$\begin{bmatrix} 1 \\ 0.93\ 1 \\ 0.92\ 0.97\ 1 \\ 0.89\ 0.97\ 0.99\ 1 \end{bmatrix}$$

All the variables have positively skewed distributions, and the chain loading estimates also have highly significant positive skewness and kurtosis. The initial starting value for each loading was 2 and for each level 1 variance (uniqueness) was 0.2. Good starting values will speed up the convergence of the MCMC chains.

Table 11.1 shows the maximum likelihood estimates produced by the AMOS factor analysis package (Arbuckle, 1997) together with the MCMC results. The factor analysis program carries out a prior standardisation so that the response variates have zero means. In terms of the MCMC algorithm, this is equivalent to adding covariates as an "intercept" term to Equation 4, one for each response variable; these could be estimated by adding an extra step to the above algorithm. Prior centring of the observed responses can be carried out to improve convergence.

The loading estimates are summarised by taking both the mean and medians of the chain. The mode can also be computed, but in this data set for the variances it is very poorly estimated and is given only for the loadings. In fact, the likelihood surface with respect to the variances is very flat. The MCMC chains can be summarised using a Normal kernel density smoothing method (Silverman 1986).

The estimates and standard errors from the MCMC chain are larger than the maximum likelihood estimates. The standard errors for the latter will generally be underestimates, especially for such a small data set since they use the estimated (plug in) parameter values. The distributions for the variances in particular are skew so that median rather than mean estimates seem preferable. Since sampling is from the likelihood, the maximum likelihood estimate will be located at the joint parameter mode. Although this has not been computed, as can be seen from the loading estimates the univariate modes are closer to the maximum likelihood estimates than the means or medians. Table 11.2 shows good agreement between the variable means and the fitted intercept terms.

Table 11.1. Maximum likelihood estimates for simulated data set together with MCMC estimates using chain length 50,000 burn in 20.

Parameter	ML estimate (s.e.)	MCMC mean estimates (s.d.)	MCMC median estimates	MCMC modal estimates
λ_1	0.92 (0.17)	1.03 (0.22)	1.00	0.98
λ_2	2.41 (0.41)	2.71 (0.52)	2.65	2.59
λ_3	3.86 (0.57)	3.91 (0.72)	3.82	3.71
λ_4	4.30 (0.71)	4.82 (0.90)	4.71	4.58
σ^2_{e1}	0.15 (0.05)	0.17 (0.07)	0.16	
σ^2_{e2}	0.25 (0.09)	0.31 (0.14)	0.28	
σ^2_{e3}	0.09 (0.10)	0.10 (0.17)	0.06	
σ^2_{e4}	0.43 (0.20)	0.55 (0.31)	0.50	

Table 11.2. Variable means and fitted intercepts

Variable	Mean	Intercept
1	0.54	0.57
2	0.64	0.71
3	1.12	1.21
4	1.28	1.36

The structure described by Equations 11.3 and 11.4 was also fitted with a simulated data set of 200 cases rather than 20. The results are given in Table 11.3 for the maximum likelihood estimates and the means and medians of the MCMC procedure. A closer agreement can be seen in this table. The MCMC estimates are slightly higher (by up to 2%) than the maximum likelihood ones, with the modal estimates being closest. In more complex examples one may need to run the chain longer with a longer burn in and also try more than one chain with different starting values. For example, a conventional single level factor model could be fitted using standard software to obtain approximations to the level 1 loadings and unique variances.

11.4 Other Procedures

Geweke and Zhou (1996) consider the single level factor model with uncorrelated factors. They use Gibbs sampling and consider identifiability constraints. Zhu and Lee (1999) also consider single level structures including non-linear models that involve factor products and powers of factors. They use Gibbs steps for the parameters and a Metropolis Hastings algorithm for simulating from the conditional distribution of the factors. They also provide

Table 11.3. Model (11.3) & (11.4) with 200 simulated individuals. 5000 cycles.

Parameter	ML estimate (s.e.)	MCMC mean estimates (s.d.)	MCMC median estimates	MCMC mode estimates
λ_1	0.95 (0.06)	0.97 (0.06)	0.96	0.96
λ_2	1.86 (0.10)	1.89 (0.10)	1.89	1.88
λ_3	2.92 (0.15)	2.98 (0.16)	2.97	2.97
λ_4	3.86 (0.20)	3.94 (0.20)	3.93	3.92
σ^2_{e1}	0.22 (0.023)	0.23 (0.024)	0.22	0.22
σ^2_{e2}	0.27 (0.033)	0.27 (0.033)	0.27	0.27
σ^2_{e3}	0.38 (0.058)	0.38 (0.060)	0.38	0.38
σ^2_{e4}	0.39 (0.085)	0.39 (0.087)	0.38	0.38

a goodness-of-fit criterion (see discussion). It appears, however, that their algorithm requires individuals to have complete data vectors with no missing responses, whereas the procedure described in this chapter has no such restriction.

Scheines, Hoijtink, and Boomsma, (1999) also use MCMC and take as data the sample covariance matrix, for a single level structure, where covariates are assumed to have been incorporated into the means. They assume a multivariate normal prior with truncation at zero for the variances. Rejection sampling is used to produce the posterior distribution. They discuss the problem of identification, and point out that identification issues may be resolved by specifying an informative prior.

McDonald and Goldstein (1989) show how maximum likelihood estimates can be obtained for a 2-level structural equation model. They derive the covariance structure for such a model and show how an efficient algorithm can be constructed to obtain maximum likelihood estimates for the multivariate normal case. Longford and Muthèn (1992) develop this approach. The latter authors, together with Goldstein (1995, Chapter 11) and Rowe and Hill (1997) also point out that consistent estimators can be obtained from a 2-stage process as follows. A 2-level multivariate response linear model is fitted using an efficient procedure such as maximum likelihood. This can be accomplished, for example as pointed out earlier by defining a 3-level model where the lowest level is that of the response variables (see Goldstein, 1995, Chapter 8 and Model (11.5) below). This analysis will produce estimates for the (residual) covariance matrices at each level and each of these can then be structured according to an underlying latent variable model in the usual way. By considering the two matrices as two "populations" one can also impose constraints on, say, the loadings using an algorithm for simultaneously fitting structural equations across several populations.

Rabe-hesketh, Pickles, and Taylor (2000) consider a general formulation, similar to model 7 below, but allowing general link functions, to specify multilevel structural equation generalised linear models (GLLAMM). They consider maximum likelihood estimation using general maximisation algorithms and a set of macros has been written to implement the model in the program STATA.

In the MCMC formulation presented in this chapter, it is possible to deal with incomplete data vectors and also to use informative prior distributions, as described below. The algorithm can also be extended to the non-linear factor case using a Metropolis Hastings step when sampling the factor values, as in Zhu and Lee (1999).

11.5 General Multilevel Bayesian Factor Models

Extensions to models with further factors, patterned loading matrices and higher levels in the data structure are straightforward. Consider the 2-level factor model

$$y_{rij} = \beta_r + \sum_{f=1}^{F} \lambda_{fr}^{(2)} \nu_{fj}^{(2)} + \sum_{g=1}^{G} \lambda_{gr}^{(1)} \nu_{gij}^{(1)} + u_{rj} + e_{rij}$$

$$u_{rj} \sim N(0, \sigma^2_{ur}), \quad e_{rij} \sim N(0, \sigma^2_{er}),$$

$$\nu_{fj}^{(2)} \sim MVN_F(0, \Omega_2), \quad \nu_{gij}^{(1)} \sim MVN_G(0, \Omega_1)$$

$$r = 1, \ldots, R, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J, \quad \sum_{j=1}^{J} n_j = N$$

Here R responses for N individuals are split between J level 2 units. F sets of factors, with $\nu_{fj}^{(2)}$ defined at level 2 and G sets of factors, with $\nu_{gij}^{(1)}$ defined at level 1 are given. For the fixed part of the model, the algorithm is restricted to a single intercept term β_r for each response although it is easy to extend the algorithm to arbitrary fixed terms. The residuals at levels 1 and 2, e_{rij} and $u_{rj}$are assumed to be independent.

Although this allows a very flexible set of factor models it should be noted that in order for such models to be identifiable suitable constraints must be put on the parameters. See Everitt (1984) for further discussion of identifiability. These will consist of fixing the values of some of the elements of the factor variance matrices, Ω_1 and Ω_2 and/or some of the factor loadings, $\lambda_{fr}^{(2)}$ and $\lambda_{gr}^{(1)}$, or specifying informative priors.

The algorithms presented will give steps for all parameters and so any parameter that is constrained will simply maintain its chosen value and will not be updated. It is initially assumed that the factor variance matrices, Ω_1 and Ω_2 are known (completely constrained) and a discussion is provided about how the algorithm can be extended to encompass partially constrained variance matrices. The parameters in the following steps are those available at the current iteration of the algorithm.

11.6 Prior Distributions

For the algorithm assume the following general priors:

$$p(\beta_r) \sim N(\beta_r^*, \sigma_{br}^2)$$

$$p(\lambda_{fr}^{(2)}) \sim N(\lambda_{fr}^{(2)*}, \sigma_{2fr}^2), p(\lambda_{gr}^{(1)}) \sim N(\lambda_{gr}^{(1)*}, \sigma_{1gr}^2)$$

$$p(\sigma_{ur}^2) \sim \Gamma^{-1}(a_{ur}^*, b_{ur}^*), p(\sigma_{er}^2) \sim \Gamma^{-1}(a_{er}^*, b_{er}^*)$$

With the assumption that the factor variance matrices are known, one can use a Gibbs sampling algorithm which will involve updating parameters in turn by generating new values from the following 8 sets of conditional posterior distributions.

Step 1: Update current value of β_r(r=$1,\ldots,R$) from the following distribution
$p(\beta_r) \sim N(\hat{\beta}_r, D_{br})$ where

$$D_{br} = \left(\frac{N}{\sigma_{er}^2} + \frac{1}{\sigma_{br}^2}\right)^{-1}$$

and

$$\hat{\beta}_r = D_{br}\left(\frac{\sum_{ij} d_{rij}^{\beta}}{\sigma_{er}^2} + \frac{\beta_r^*}{\sigma_{br}^2}\right)$$

where

$$d_{rij}^{\beta} = e_{rij} - \beta_r$$

Step 2: Update $\lambda_{fr}^{(2)}$ (r=$1,\ldots,R$, f=$1,\ldots,F$ where not constrained) from the following distribution: $p(\lambda_{fr}^{(2)}) \sim N(\hat{\lambda}_{fr}^{(2)}, D_{fr}^{(2)})$ where

$$D_{fr}^{(2)} = \left(\frac{\sum_j n_j(\nu_{fj}^{(2)})^2}{\sigma_{er}^2} + \frac{1}{\sigma_{2fr}^2}\right)^{-1}$$

and

$$\hat{\lambda}_{fr}^{(2)} = D_{fr}^{(2)}\left(\frac{\sum_{ij} \nu_{fj}^{(2)} d_{rijf}^{(2)}}{\sigma_{er}^2} + \frac{\lambda_{fr}^{(2)*}}{\sigma_{2fr}^2}\right)$$

where

$$d_{rijf}^{(2)} = e_{rij} - \lambda_{fr}^{(2)}\nu_{fj}^{(2)}$$

Step 3: Update $\lambda_{gr}^{(1)}$ (r=$1,\ldots,R$, g=$1,\ldots,G$ where not constrained) from the following distribution: $p(\lambda_{gr}^{(1)}) \sim N(\hat{\lambda}_{gr}^{(1)}, D_{gr}^{(1)})$ where

$$D_{gr}^{(1)} = \left(\frac{\sum_{ij}(\nu_{gij}^{(1)})^2}{\sigma_{er}^2} + \frac{1}{\sigma_{1gr}^2}\right)^{-1}$$

and

$$\hat{\lambda}_{gr}^{(1)} = D_{gr}^{(1)}\left(\frac{\sum_{ij} \nu_{gij}^{(1)} d_{rijg}^{(1)}}{\sigma_{er}^2} + \frac{\lambda_{gr}^{(1)*}}{\sigma_{1gr}^2}\right)$$

where

$$d_{rijg}^{(1)} = e_{rij} - \lambda_{gr}^{(1)}\nu_{gij}^{(1)}$$

Step 4: Update $\nu_j^{(2)}$(j=$1,\ldots,J$) from the following distribution:
$p(\nu_j^{(2)}) \sim MVN_F(\hat{\nu}_j^{(2)}, D_j^{(2)})$ where

$$D_j^{(2)} = \left(\sum_r \frac{n_j \lambda_r^{(2)}(\lambda_r^{(2)})^T}{\sigma_{er}^2} + \Omega_2^{-1}\right)^{-1}$$

and

$$\hat{\nu}_j^{(2)} = D_j^{(2)}\left(\sum_r \sum_{i=1}^{n_j} \frac{\lambda_r^{(2)} d_{rij}^{(2)}}{\sigma_{er}^2}\right)$$

where

$$d_{rij}^{(2)} = e_{rij} - \sum_{f=1}^{F} \lambda_{fr}^{(2)}\nu_{fj}^{(2)},$$

$$\lambda_r^{(2)} = (\lambda_{1r}^{(2)}, \ldots, \lambda_{Fr}^{(2)})^T,$$

$$\nu_j^{(2)} = (\nu_{1j}^{(2)}, \ldots, \nu_{Fj}^{(2)})^T$$

Step 5: Update $\nu_{ij}^{(1)}$($i = 1, \ldots, n_j, j = 1, \ldots, J$) from the following distribution: $p(\nu_{ij}^{(1)}) \sim MVN_G(\hat{\nu}_{ij}^{(1)}, D_{ij}^{(1)})$ where

$$D_{ij}^{(1)} = \left(\sum_r \frac{\lambda_r^{(1)}(\lambda_r^{(1)})^T}{\sigma_{er}^2} + \Omega_1^{-1}\right)^{-1}$$

and

$$\hat{\nu}_j^{(1)} = D_{ij}^{(1)} \left(\sum_r \frac{\lambda_r^{(1)} d_{rij}^{(1)}}{\sigma_{er}^2} \right)$$

where

$$d_{rij}^{(1)} = e_{rij} - \sum_{g=1}^{G} \lambda_{gr}^{(1)} \nu_{gj}^{(1)},$$

$$\lambda_{r..}^{(1)} = (\lambda_{1r}^{(1)}, ..., \lambda_{Gr}^{(1)})^T,$$

$$\nu_{ij}^{(1)} = (\nu_{1ij}^{(1)}, ..., \nu_{Gij}^{(1)})^T$$

Step 6: Update u_{rj}(r=1,...,R, j=1,...,J) from the following distribution: $p(u_{rj}) \sim N(\hat{u}_{rj}, D_{rj}^{(u)})$

where

$$D_{rj}^{(u)} = \left(\frac{n_j}{\sigma_{er}^2} + \frac{1}{\sigma_{ur}^2} \right)^{-1}$$

and

$$\hat{u}_{rj} = \frac{D_{rj}^{(u)}}{\sigma_{er}^2} \sum_{i=1}^{n_j} d_{rij}^{(u)}$$

where

$$d_{rij}^{(u)} = e_{rij} - u_{rj}$$

Step 7: Update σ_{ur}^2 from the following distribution: $p(\sigma_{ur}^2) \sim \Gamma^{-1}(\hat{a}_{ur}, \hat{b}_{ur})$ where $\hat{a}_{ur} = J/2 + a_{ur}^*$ and $\hat{b}_{ur} = \frac{1}{2}\sum_j u_{rj}^2 + b_{ur}^*$.

Step 8: Update σ_{er}^2 from the following distribution: $p(\sigma_{er}^2) \sim \Gamma^{-1}(\hat{a}_{er}, \hat{b}_{er})$ where $\hat{a}_{er} = N/2 + a_{er}^*$ and $\hat{b}_{er} = \frac{1}{2}\sum_{ij} e_{rij}^2 + b_{er}^*$.

Note that the level 1 residuals, e_{rij} can be calculated by subtraction at every step of the algorithm.

11.7 Unconstrained Factor Variance Matrices

In the general algorithm it was assumed that the factor variances are all constrained. Typically, one would fix the variances to equal 1 and the covariances

to equal 0 and have independent factors. This form permits the simplification of steps 4 and 5 of the algorithm to univariate normal updates for each factor separately. One may however wish to consider correlations between the factors. Here one will have to modify the algorithm to allow another special case where the variances are constrained to be 1 but the covariances can be freely estimated. Where the resulting correlations obtained are estimated to be close to 1 or –1 then one may be fitting too many factors at that particular level. As the variances are constrained to equal 1 the covariances between factors equal the correlations between the factors. This means that each covariance is constrained to lie between –1 and 1. Only the factor variance matrix at level 2 is considered here, as the step for the level 1 variance matrix simply involves changing subscripts. The following priors are used:

$$p(\Omega_{2,lm}) \sim Uniform(-1, 1)\ \forall\ l \neq m$$

Here $\Omega_{2,lm}$is the $l, m-$th element of the level 2 factor variance matrix. These covariance parameters are updated using a Metropolis step and a Normal random walk proposal (see Browne & Rasbash, in preparation) for more details on using Metropolis Hastings methods for constrained variance matrices).

Step 9 : At iteration t generate $\Omega_{2,lm}^* \sim \mathrm{N}(\Omega_{2,lm}^{(t-1)}, \sigma_{plm}^2)$ where σ_{plm}^2 is a proposal distribution variance that has to be set for each covariance. Form a proposed new matrix Ω_2^* by replacing the l, m-th element of $\Omega_2^{(t-1)}$ by $\Omega_{2,lm}^*$. If Ω_2^* is positive definite, set $\Omega_{2,lm}^{(t)} = \Omega_{2,lm}^*$ with probability $\min(1, p(\Omega_2^*|\nu_{fj}^{(2)})/p(\Omega_2^{(t-1)}|\nu_{fj}^{(2)})$, and $\Omega_{2,lm}^{(t)} = \Omega_{2,lm}^{(t-1)}$ otherwise, or if not positive definite.

Here $p(\Omega_2^*|\nu_{fj}^{(2)}) = \prod_j |\Omega_2^*|^{-1/2} \exp((\nu_{fj}^{(2)})^T (\Omega_2^*)^{-1} \nu_{fj}^{(2)})$and

$$p(\Omega_2^{(t-1)}|\nu_{fj}^{(2)}) = \prod_j |\Omega_2^{(t-1)}|^{-1/2} \exp((\nu_{fj}^{(2)})^T (\Omega_2^{(t-1)})^{-1} \nu_{fj}^{(2)})$$

This procedure is repeated for each covariance that is not constrained.

Missing Data. The example that is discussed in this chapter has the additional difficulty that individuals have different numbers of responses. This is not a problem for the MCMC methods if one is prepared to assume missingness is at random or effectively so by design. This is equivalent to giving the missing data a uniform prior. One then has to simply add an extra Gibbs sampling step to the algorithm to sample the missing values at each iteration. As an illustration, consider an individual who is missing response r. In a factor model the correlation between responses is explained in the factor terms and conditional on these terms the responses for an individual are independent and so the conditional distributions of the missing responses have simple forms.

Step 10: Update y_{rij} (r=1,...,R, i=1,...,n_j, j=1,...,J $\forall$ y_{rij} that are missing) from the following distribution, given the current values, $y_{rij} \sim N(\hat{y}_{rij}, \sigma^2_{er})$ where $\hat{y}_{rij} = \beta_r + \sum_{f=1}^{F} \lambda^{(2)}_{fr}\nu^{(2)}_{fj} + \sum_{g=1}^{G} \lambda^{(1)}_{gr}\nu^{(1)}_{gij} + u_{rj}$.

11.8 Example

The example uses a data set discussed by Goldstein (1995, Chapter 4) and consists of a set of responses to a series of 4 test booklets by 2439 pupils in 99 schools. Each student responded to a core booklet containing Earth Science, Biology and Physics items and to a further two booklets randomly chosen from three available. Two of these booklets were in Biology and one in Physics. As a result there are 6 possible scores, one in Earth Science, three in Biology and 2 in Physics, each student having up to five. A full description of the data is given in Goldstein (1995).

A multivariate 2-level model fitted to the data gives the following maximum likelihood estimates for the means and covariance/correlation matrices in Table 11.4. The model can be written as follows

$$y_{ijk} = \sum_{h=1}^{6} \beta_i x_{hjk} + \sum_{h=1}^{6} \gamma_i x_{hjk} z_{jk} + \sum_{h=1}^{6} u_{ijk} x_{hjk} + \sum_{h=1}^{6} v_{ik} x_{hjk}$$

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{pmatrix} \sim N(0, \Omega_u) \quad \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{pmatrix} \sim N(0, \Omega_v) \tag{11.5}$$

$x_{hjk} = 1$ *if* $h = i$, 0 *otherwise*
$z_{jk} = 1$ *if a girl*, $= 0$ *if a boy*
i indexes response variables,
j indexes students,
k indexes schools

Two 2 level factor models are now fit to these data, as shown in Table 11.5. Improper uniform priors are used for the fixed effects and loadings and $\Gamma^{-1}(\epsilon, \epsilon)$ priors for the variances with $\epsilon = 10^{-3}$. The fixed effects in Table 11.5 are omitted, since they are very close to those in Table 11.4. Model A has two factors at level 1 and a single factor at level 2. For illustration, all the variances are constrained to be 1.0, and covariance (correlation) between the level 1 factors are allowed to be estimated. Inspection of the correlation structure suggests a model where the first factor at level 1 estimates the loadings for Earth Science and Biology, constraining those for Physics to be

zero (the physics responses have the highest correlation), and for the second factor at level 1 to allow only the loadings for Physics to be unconstrained. The high correlation of 0.90 between the factors suggests that perhaps a single factor will be an adequate summary. Although the results are not presented in this chapter, a similar structure for two factors at the school level where the correlation is estimated to be 0.97 was also studied, strongly suggesting a single factor at that level.

Table 11.4. Science attainment estimates.

Fixed	Estimate (s.e.)
Earth Science Core	0.838 (0.0076)
Biology Core	0.711 (0.0100)
Biology R3	0.684 (0.0109)
Biology R4	0.591 (0.0167)
Physics Core	0.752 (0.0128)
Physics R2	0.664 (0.0128)
Earth Science Core (girls - boys)	-0.0030 (0.0059)
Biology Core (girls - boys)	-0.0151 (0.0066)
Biology R3 (girls - boys)	0.0040 (0.0125)
Biology R4 (girls - boys)	-0.0492 (0.0137)
Physics Core (girls - boys)	-0.0696 (0.0073)
Physics R2 (girls - boys)	-0.0696 (0.0116)

Level 2 (School)						
	E.Sc. core	Biol. Core	Biol R3	Biol R4	Phys. core	Phys. R2
E.Sc. core	0.0041					
Biol. core	0.68	0.0076				
Biol R3	0.51	0.68	0.0037			
Biol R4	0.46	0.68	0.45	0.0183		
Phys. core	0.57	0.90	0.76	0.63	0.0104	
Phys. R2	0.54	0.78	0.57	0.65	0.78	0.0095
Level 1 (Student)						
	E.Sc. core	Biol. Core	Biol R3	Biol R4	Phys. core	Phys. R2
E.Sc. core	0.0206					
Biol. core	0.27	0.0261				
Biol R3	0.12	0.13	0.0478			
Biol R4	0.14	0.27	0.20	0.0585		
Phys. core	0.26	0.42	0.11	0.27	0.0314	
Phys. R2	0.22	0.33	0.14	0.37	0.41	0.0449

Random: Variances on diagonal; correlations off-diagonal.

For model B the three topics of Earth Science, Biology and Physics were separated in order to separately have non-zero loadings on three corresponding factors at the student level. This time the high inter-correlation is that between the Biology and Physics booklets with only moderate (0.49, 0.55) correlations between Earth Science and Biology and Physics. This suggests that one needs at least two factors to describe the student level data and that the preliminary analysis suggesting just one factor can be improved. Since the analyses are for illustrative purposes no further possibilities were pursued with these data. Note that Model B at level 1 is strictly non-identified, given uniform priors. The constraint of a positive definite factor covariance matrix provides estimates of the factor correlations, and loadings, which are means over the respective feasible regions.

11.9 Discussion

This chapter has shown how factor models can be specified and fitted. The MCMC computations allow point and interval estimation with an advantage over maximum likelihood estimation in that full account is taken of the uncertainty associated with the estimates. In addition, it allows full Bayesian modelling with informative prior distributions which may be especially useful for identification problems. As pointed out in the introduction, the MCMC algorithm is readily extended to handle the general structural equation case, and further work is being carried out along the following lines. For simplicity the single level model case is considered to illustrate the procedure.

One kind of fairly general, single level, structural equation model can be written in the following matrix form (see McDonald, 1985 for some alternative representations)

$$\begin{aligned} A_1 v_1 &= A_2 v_2 + W \\ Y_1 &= \Lambda_1 v_1 + U_1 \\ Y_2 &= \Lambda_2 v_2 + U_2 \end{aligned} \tag{11.6}$$

Where Y_1, Y_2 are observed multivariate vectors of responses, A_1 is a known transformation matrix, often set to the identity matrix, A_2 is a coefficient matrix which specifies a multivariate linear model between the set of transformed factors, v_1, and v_2, Λ_1, Λ_2 are loadings, U_1, U_2 are uniquenesses, W is a random residual vector and W, U_1, U_2 are mutually independent with zero means. The extension of this model to the multilevel case follows that of the factor model and the discussion is restricted to sketching how the MCMC algorithm can be applied to Equation (11.6). Note, that as before one can add covariates and measured variables multiplying the latent variable terms as shown in Equation (11.6). Note that A_2 can be written as the vector A_2^* by stacking the rows of A_2. For example if

Table 11.5. Science attainment MCMC factor model estimates.

Parameter	A Estimate (s.e.)	B Estimate (s.e.)
Level 1: factor 1 loadings		
E.Sc. core	0.06 (0.004)	0.11 (0.02)
Biol. core	0.11 (0.004)	0*
Biol R3	0.05 (0.008)	0*
Biol R4	0.11 (0.009)	0*
Phys. core	0*	0*
Phys. R2	0*	0*
Level 1: factor 2 loadings		
E.Sc. core	0*	0*
Biol. core	0*	0.10 (0.005)
Biol R3	0*	0.05 (0.008)
Biol R4	0*	0.10 (0.009)
Phys. core	0.12 (0.005)	0*
Phys. R2	0.12 (0.007)	0*
Level 1: factor 3 loadings		
E.Sc. core	-	0*
Biol. core	-	0*
Biol R3	-	0*
Biol R4	-	0*
Phys. core	-	0.12 (0.005)
Phys. R2	-	0.12 (0.007)
Level 2: factor 1 loadings		
E.Sc. core	0.04 (0.007)	0.04 (0.007)
Biol. core	0.09 (0.008)	0.09 (0.008)
Biol R3	0.05 (0.009)	0.05 (0.010)
Biol R4	0.10 (0.016)	0.10 (0.016)
Phys. core	0.10 (0.010)	0.10 (0.010)
Phys. R2	0.09 (0.011)	0.09 (0.011)
Level 1: residual variances		
E.Sc. core	0.017 (0.001)	0.008 (0.004)
Biol. core	0.015 (0.001)	0.015 (0.001)
Biol R3	0.046 (0.002)	0.046 (0.002)
Biol R4	0.048 (0.002)	0.048 (0.002)
Phys. core	0.016 (0.001)	0.016 (0.001)
Phys. R2	0.029 (0.002)	0.030 (0.002)
Level 2: residual variances		
E.Sc. core	0.002 (0.0005)	0.002 (0.0005)
Biol. core	0.0008 (0.0003)	0.0008 (0.0003)
Biol R3	0.002 (0.0008)	0.002 (0.0008)
Biol R4	0.010 (0.002)	0.010 (0.002)
Phys. core	0.002 (0.0005)	0.002 (0.0005)
Phys. R2	0.003 (0.0009)	0.003 (0.0009)
Level 1 correlation factors 1 & 2	0.90 (0.03)	0.55 (0.10)
Level 1 correlation factors 1 & 3	-	0.49 (0.09)
Level 1 correlation factors 2 & 3	-	0.92 (0.04)

* indicates constrained parameter. A chain of length 20,000 with a burn in of 2000 was used. Level 1 is student, level 2 is school

$$A_2 = \begin{pmatrix} a_0 & a_1 \\ a_2 & a_3 \end{pmatrix}, \quad then \quad A_2^* = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

The distributional form of the model can be written as

$$\begin{aligned} &A_1 v_1 \sim MVN(A_2 v_2, \Sigma_3) \\ &v_1 \sim MVN(0, \Sigma_{v_1}), v_2 \sim MVN(0, \Sigma_{v_2}) \\ &Y_1 \sim MVN(\Lambda_1 v_1, \Sigma_1), Y_2 \sim MVN(\Lambda_2 v_2, \Sigma_2) \end{aligned}$$

with priors

$$A_2^* \sim MVN(\hat{A}_2^*, \Sigma_{A_2^*}), \quad \Lambda_1 \sim MVN(\hat{\Lambda}_1, \Sigma_{\Lambda_1}), \quad \Lambda_2 \sim MVN(\hat{\Lambda}_2, \Sigma_{\Lambda_2})$$

and $\Sigma_1, \Sigma_2, \Sigma_3$ having inverse Wishart priors. The coefficient and loading matrices have conditional normal distributions as do the factor values. The covariance matrices and uniqueness variance matrices involve steps similar to those given in the earlier algorithm. The extension to two levels and more follows the same general procedure as shown earlier.

The model can be generalised further by considering m sets of response variables, $Y_1, Y_2, ... Y_m$ in Equation (11.6) and several, linked, multiple group structural relationships with the k-th relationship having the general form

$$\sum_h V_h^{(k)} A_h^{(k)} = \sum_g V_g^{(k)} A_g^{(k)} + W^{(k)}$$

and the above procedure can be extended for this case. Note that the model for simultaneous factor analysis (or, more generally, structural equation model) in several populations is a special case of this model, with the addition of any required constraints on parameter values across populations.

The model can also be generalised to include fixed effects, responses at level 2 and covariates Z_h for the factors, which may be a subset of the fixed effects covariates X

$$\begin{aligned} &Y^{(1)} = X\beta + \Lambda_2^{(1)} v_2 Z_2^{(1)} + u^{(1)} + \Lambda_1^{(1)} v_1 Z_1 + e^{(1)} \\ &Y^{(2)} = \Lambda_2^{(2)} v_2 Z_2^{(2)} + u^{(2)} \\ &Y^{(1)} = \{y_{rij}\}, \quad Y^{(2)} = \{y_{rj}\} \\ &r = 1, ..., R \quad i = 1, ..., i_j \quad j = 1, ..., J \end{aligned} \tag{11.7}$$

The superscript refers to the level at which the measurement exists, so that, for example, y_{1ij}, y_{2j} refer respectively to the first measurement in the

i-th level 1 unit in the j-th level 2 unit (say students and schools) and the second measurement taken at school level for the $j-$th school.

Further work is currently being carried out on applying these procedures to non-linear models and specifically to generalised linear models. For simplicity consider the binomial response logistic model as anillustration. Write

$$\begin{aligned} &E(y_{ij}) = \pi_{ij} = [1 + \exp -(a_i + \lambda_i v_j)]^{-1} \\ &y_{ij} \sim \text{Bin}(\pi_{ij}, n_{ij}) \end{aligned} \tag{11.8}$$

The simplest model is the multiple binary response model ($n_{ij} = 1$) that is referred to in the psychometric literature as a unidimensional item response model (Goldstein & Wood, 1989; Bartholomew & Knott, 1999). Estimation for this model is not possible using a simple Gibbs sampling algorithm, but as in the standard binomial multilevel case (see Browne, 1998) any Gibbs steps that do not have standard conditional posterior distributions could be replaced with Metropolis Hastings steps.

The issues that surround the specification and interpretation of single level factor and structural equation models are also present in multilevel versions. Parameter identification has already been discussed; another issue is the boundary "Heywood" case. Such solutions occur where sets of loading parameters tend towards zero or a correlation tends towards 1.0 and have been observed. A final important issue that only affects stochastic procedures is the problem of "flipping states". This means that there is not a unique solution even in a 1-factor problem as the loadings and factor values may all flip their sign to give an equivalent solution. When the number of factors increases there are greater problems as factors may swap over as the chains progress. This means that identifiability is an even greater consideration when using stochastic techniques.

For making inferences about individual parameters or functions of parameters one can use the chain values to provide point and interval estimates. These can also be used to provide large sample Wald tests for sets of parameters. Zhu and Lee (1999) propose a chi-square discrepancy function for evaluating the *posterior predictive* p-value, which is the Bayesian counterpart of the frequentist p-value statistic (Meng, 1994). In the multilevel case the α – level probability becomes

$$\begin{aligned} &\hat{p}_B(Y) = (\sum_{i=1}^{J} i_j)^{-1} \sum_{i=1}^{i_j} \chi_\alpha^2(i_j p) \geq D(Y_i | \theta^{(i)}, \hat{v}^{(i)}) \\ &D(Y_i | \theta^{(i)}, \hat{v}^{(i)}) = Y_i^T \Sigma_i^{-1} Y_i \end{aligned} \tag{11.9}$$

where Y_i is the vector of responses for the $i-$th level 2 unit and Σ_i is the (non-diagonal) residual covariance matrix.

References

Arbuckle, J.L. (1997). *AMOS: Version 3.6.* Chicago: Small Waters Corporation.

Bartholomew, D. J. and Knott, M. (1999). *Latent variable models and factor analysis.* (2^{nd} ed.). London: Arnold.

Browne, W. (1998). *Applying MCMC methods to multilevel models.* PhD thesis, University of Bath.

Browne, W. & Rasbash, J. (2001). MCMC algorithms for variance matrices with applications in multilevel modeling. (in preparation)

Blozis, S. A. & Cudeck, R. (1999). Conditionally linear mixed-effects models with latent variable covariates. *Journal of Educational and Behavioural Statistics*, 24, 245-270.

Everitt, B. S. (1984). *An introduction to latent variable models.* London: Chapman and Hall.

Geweke, J. & Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of International Financial Studies*, 9, 557-587.

Goldstein, H. (1995). *Multilevel statistical models.* London: Edward Arnold.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, 42, 139-167.

Longford, N., & Muthèn, B. O. (1992). Factor analysis for clustered observations. *Psychometrika*, 57, 581-597.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika*, 58, 575-585.

McDonald, R.P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*, 42, 215-232.

Meng, X. L. (1994). Posterior predictive p-values. *Annals of Statistics*, 22, 1142-1160.

Rabe-hesketh, S., Pickles, A. & Taylor, C. (2000). Sg129: Generalized linear latent and mixed models. *Stata Technical Bulletin*, 53, 47-57.

Rasbash, J., Browne, W., Goldstein, H., Plewis, I., Draper, D., Yang, M., Woodhouse, G., Healy, M., & Cameron, B. (2000). *A users guide to MLwiN* (second edition). London, Institute of Education.

Raudenbush, S. W. (1995). Maximum Likelihood estimation for unbalanced multilevel covariance structure models via the EM algorithm. *British Journal of Mathematical and Statistical Psychology*, 48, 359-70.

Rowe, K. J., & Hill, P. W. (1997). *Simultaneous estimation of multilevel structural equations to model students' educational progress.* Paper presented at the Tenth International Congress for School effectiveness and School Improvement, Memphis, Tennessee.

Rowe, K. J., & Hill, P.W. (in press). Modelling educational effectiveness in classrooms: The use of multilevel structural equations to model students' progress. *Educational Research and Evaluation.*

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47, 69-76.

Scheines, R., Hoijtink, H. & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, 64, 37-52.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data analysis.* London: Chapman and Hall.

Zhu, H.-T. & Lee, S.-Y. (1999). Statistical analysis of nonlinear factor analysis models. *British Journal of Mathematical and Statistical Psychology*, 52, 225-242.