An MCMC algorithm for problems involving 'constrained' variance matrices with applications in multilevel modelling

by

Dr William Browne,
Centre for Multilevel Modelling
Institute of Education, London.

Summary

- Progression in Bayesian Software.
- Metropolis Hastings Sampling.
- Methods for a single variance.
- Methods for variance matrices.
- Constraints between and within variance matrices
- Several Examples.
- Inefficiencies in single-site updating.

Progress in General- Purpose Bayesian Software

- Conjugate priors allow the use of Gibbs sampling from standard distributions for simple problems.

- The AR sampler (Gilks and Wild 1992) as used in early versions of BUGS fits log-concave posteriors.

- The slice sampler (Neal 1997) also used in WinBUGS for more general restricted range posteriors.

- Univariate random-walk Metropolis used in WinBUGS for more general unrestricted range posteriors. Used in MLwiN for all parameters that do not have standard posteriors.

- BayesX uses methods described in Rue (2000) and Gamerman (1997) for more efficient sampling for certain classes of model that can already be fit by less efficient methods.

- MLwiN (development version) fits models with 'constrained' variance matrices using a Metropolis Hastings sampling method (Browne 2002).

- Some such models previously fitted by Chib and Greenberg, for example multivariate probit models (see later) but not previously implemented into a widely available software package.

Metropolis Hastings (MH) Sampling Step

Algorithm due to Metropolis et al. (1953) and Hastings (1970). Consider a situation where we have data y and parameters θ and we are interested in evaluating the joint posterior distribution, $p(\theta|y)$. Then to use MCMC sampling we split θ into several groups θ_i and sample from the conditional posterior distributions for each group in turn. So for example if we are interested in parameter θ_1 then we would like to take a sample from $p(\theta_1|y, \phi_1)$ where $\phi_1 = \{\theta\}/\{\theta_1\}$

So at time t we update θ_1 as follows:

1. Sample θ_1^* from $p_t(\theta_1^*|\theta_1^{(t-1)})$

where $p_t()$ is the 'proposal distribution' at step t.

2. Calculate $hr = \frac{p_t(\theta_1^{(t-1)}|\theta_1^*)}{p_t(\theta_1^*|\theta_1^{(t-1)})}$

which is known as the Hastings ratio and which equals 1 for symmetric proposals as found in pure Metropolis Sampling.

3. Calculate $r_t = hr * \frac{p(\theta_1^*|y,\phi_1)}{p(\theta_1^{(t-1)}|y,\phi_1)}$

4. Let $\theta_1^{(t)} = \theta_1^*$ with probability $min(1, r_t)$ otherwise let $\theta_1^{(t)} = \theta_1^{(t-1)}$

In a general MCMC algorithm we would now update the other parameters in a similar way or perhaps by a different MCMC method, for example Gibbs sampling. In the descriptions of the various methods that follow later we will simply give the form of proposal distribution and the Hastings ratio.

Updating a simple variance

To avoid the introduction of other unknown parameters we will consider the following simple problem (Browne 1998). Assume we generate 100 observations from a normal distribution with known mean 0 and unknown variance, σ^2. Now assume we have a conjugate $SI\chi^2$ prior with parameters τ_0 and σ_0^2 for σ^2. Then the posterior distribution for σ^2 is

$$\sigma^2 \sim SI\chi^2(\tau_0 + 100, (\tau_0\sigma_0^2 + 100V)/(\tau_0 + 100))$$

where V is the variance of the observations. In our example the data where simulated with variance 4, $\tau_0 = 3$ and $\sigma_0^2 = 6$. This leads to the posterior distribution $\sigma^2 \sim SI\chi^2(103, 4.46)$ which gives a posterior mean for σ^2 of 4.55.

A Gibbs sampling algorithm will simply sample independent draws from this distribution. We are interested in how well MH sampling methods will do in comparison. We will consider 4 approaches.

MH Methods

1. Proposal of same form as posterior.

Here we use

$$p_t(\sigma_*^2) \sim SI\chi^2(\omega + 2, \omega\sigma^2_{(t-1)}/(\omega + 2))$$

which is designed to have expectation $\sigma^2_{(t-1)}$ and has tuning parameter ω. This proposal is not symmetric so hr must be calculated.

$$hr = \left(\frac{\sigma_*^2}{\sigma^2_{(t-1)}}\right)^{\omega+3} exp\left(\frac{\omega}{2}\left(\frac{\sigma_*^2}{\sigma^2_{(t-1)}} - \frac{\sigma^2_{(t-1)}}{\sigma_*^2}\right)\right)$$

2. Transformation of the parameter

If we consider $log(\sigma^2)$ instead of σ^2 then we can use a random-walk normal proposal

$$p_t(log(\sigma_*^2)) \sim N(log(\sigma^2_{(t-1)}), s_p^2)$$

where s_p^2, the proposal variance is a tuning parameter. Note that a Jacobian of the transformation must be built into the prior in this case.

MH Methods

3. Truncation of the proposal to the restricted range

Here we use

$$p_t(\sigma_*^2) \sim N(\sigma_{(t-1)}^2, s_p^2)$$

with a lower truncation point at 0. s_p^2, the proposal variance is a tuning parameter. This results in a non-symmetric proposal so hr must be calculated.

$$hr = \frac{1-\Phi(-\sigma_{(t-1)}^2/s_p)}{1-\Phi(-\sigma_*^2/s_p)}$$

4. Assumption of zero likelihood for invalid proposals

Here again we use

$$p_t(\sigma_*^2) \sim N(\sigma_{(t-1)}^2, s_p^2)$$

but this time we do not truncate. s_p^2, the proposal variance is a tuning parameter. If a negative variance is proposed this is given likelihood 0 and so is automatically rejected.

Results

Define the autocorrelation time $\kappa = 1+2\sum_{k=1}^{\infty}\rho(k)$ where $\rho(k)$ is the autocorrelation at lag k (Kass et al. 1998). This we will approximate by summing until the first value of $k > 5$ such that $\rho(k) < 0.1$. Now $Eff = 100\%/\kappa$ gives a measure of efficiency of the MCMC method as independent sampling gives $Eff = 100\%$. The results in the table give approximately the best achievable Eff for each method based on observation. For the MH methods this occurs when tuning constants are set such that the acceptance rate is 44% (See Gelman, Roberts and Gilks 1995). The Normal proposals generated negative variances 22,004 times in 10 million iterations.

Method	Time for 10 M. updates	Eff
Gibbs	16.7s	100%
MH $SI\chi^2$	21.6s	24.5%
MH Log Normal	19.7s	24.4%
MH tr. Normal	18.5s	23.2%
MH Normal	12.1s	23.2%

Updating a Variance Matrix

Again to avoid the introduction of other unknown parameters we will consider the following simple problem (Browne 1998). Assume we generate 100 observations from a bivariate normal distribution with known mean vector and unknown variance matrix, Ω. Now assume we have a conjugate inverse Wishart prior with parameters ν_0 and Ω_0. Then the posterior distribution for Ω is

$$p(\Omega|y) \sim IW(\nu_0 + 100, \Omega_0 + 100V)$$

where V is the variance matrix of the observations. In our example the data where simulated with variance matrix, $\Omega = \begin{pmatrix} 2.0 & -0.2 \\ -0.2 & 1.0 \end{pmatrix}$. Our prior has parameters $\nu_0 = 3$ and $\Omega_0 = \begin{pmatrix} 5.0 & -0.5 \\ -0.5 & 2.0 \end{pmatrix}$.

This leads to the posterior distribution

$\Omega \sim IW\left(103, S = \begin{pmatrix} 196.02 & -17.54 \\ -17.54 & 113.12 \end{pmatrix}\right)$ which gives a posterior mean matrix

$$\Omega = \begin{pmatrix} 1.96 & -0.18 \\ -0.18 & 1.13 \end{pmatrix}.$$

MH Methods

1. Proposal of same form as posterior.

Here we use

$$p_t(\Omega_*) \sim IW(\omega + 3, \omega\Omega_{(t-1)})$$

which is designed to have expectation $\Omega_{(t-1)}$ and has tuning parameter ω. This proposal is not symmetric so hr must be calculated.

$$hr = \frac{|\Omega_*|^{\frac{2\omega+9}{2}}}{\left|\Omega_{(t-1)}\right|^{\frac{2\omega+9}{2}}} exp(\frac{\omega}{2}(tr(\Omega_{(t-1)}\Omega_*^{-1}) - tr(\Omega_*\Omega_{(t-1)}^{-1})))$$

2. Transformation of the parameter

Here consider the two variances, σ_0^2 and σ_1^2 and the correlation, $\rho_{0,1}$. These could be transformed to $log(\sigma_0^2)$, $log(\sigma_1^2)$ and either log $\left(\sqrt{\frac{1+\rho_{0,1}}{1-\rho_{0,1}}}\right)$ or alternatively $\frac{\rho_{0,1}}{\sqrt{1-\rho_{0,1}^2}}$. The problem is to calculate the Jacobian for the inverse Wishart prior but if instead we changed our prior to independent priors for the 2 variances and the correlation then we could use this approach.

MH Methods

3. Truncation of the proposal to the restricted range

Here we use truncated normal proposals for the three parameters, σ_0^2, σ_1^2 and $\sigma_{0,1}$.

The truncation points at time t are for σ_0^2,

$$\sigma_0^2 > \frac{(\sigma_{0,1(t-1)})^2}{\sigma_{1(t-1)}^2}$$

and then for $\sigma_{0,1}$

$$-\sqrt{\sigma_{0(t)}^2 \sigma_{1(t-1)}^2} < \sigma_{0,1} < \sqrt{\sigma_{0(t)}^2 \sigma_{1(t-1)}^2}$$

The truncation point for σ_1^2 has the same form as for σ_0^2. Here hr must be calculated for each step.

4. Assumption of zero likelihood for invalid proposals

Here we use normal proposals for the three parameters, σ_0^2, σ_1^2 and $\sigma_{0,1}$. If an invalid parameter value is proposed then the matrix Ω will no longer be positive definite and this can be established by attempting to find a Cholesky decomposition of Ω.

Results

In the table below we compare three of the four MH methods with Gibbs sampling. The transformation method requires using different prior distributions. We get an Eff value for each of the three parameters. The results in the table give approximately the best achievable Eff for each parameter for each method based on observation. For the MH methods this occurs when tuning constants are set such that the acceptance rate is 31% for the inverse Wishart method and 44% for the others (See Gelman, Roberts and Gilks 1995).

Method	Time for 10M.	Eff σ_0^2	Eff $\sigma_{0,1}$	Eff σ_1^2
Gibbs	3m29s	100%	100%	100%
MH IW	4m5s	10.2%	10.2%	10.2%
MH tr. Norm.	2m53s	20.5%	22.5%	20.5%
MH Normal	2m27s	20.5%	22.6%	20.4%

Constrained Variance Matrices

Here we are concerned with variance matrices that have additional constraints imposed upon them. We consider two types of constraints

Constraints between matrices

For example assume we have two variance matrices

$\Omega_1 = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_3 \end{pmatrix}$ and $\Omega_2 = \begin{pmatrix} \theta_1 & \theta_4 \\ \theta_4 & \theta_5 \end{pmatrix}$. Then these matrices are constrained by the fact they share a common parameter.

Constraints within matrices

For example assume we have the variance matrix $\Omega_3 = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & 1 \end{pmatrix}$. Then this matrix has its bottom left element constrained to equal 1.

Constraints of both types occur in practice, although problems with the second type of constraint have been studied more often.

MCMC Steps for Constrained Variance Matrices

- Except in special cases a Gibbs sampling update does not exist for the parameters of the matrices.

- A proposal of the same form as the posterior doesn't exist.

- Transformation methods are sometimes useful. (see Rats example at end.)

- Truncated normal and normal updates are possible as these are both single-parameter (single site) updating routines.

- Evaluating the truncation points is problematic in high dimensions. An $N * N$ matrix has $2^N - 1$ constraints of which each variance parameter is involved in 2^{N-1} and each covariance in 2^{N-2}.

- To check a Normal update is valid simply involves performing a Cholesky decomposition on each variance matrix.

Between Matrices Constraints Example 1

Heteroscedasticity in a Gaussian multilevel model

Browne et al. (2002) consider how to fit multilevel models where the error variance is a function of predictor variables using MCMC. An example from education is the following:

$$exam_{ij} = \beta_0 + \beta_1 LRT_{ij} + u_{0j} + u_{1j} LRT_{ij} + e_{ij}$$

$u_j \sim MVN(0, \Omega_u), e_{ij} \sim N(0, \sigma^2_{ij})$, where

$$\sigma^2_{ij} = \theta_0 + \theta_1 LRT_{ij} + \theta_2 LRT^2_{ij}$$

Here $exam_{ij}$ is the (normalised) total exam score at age 16 for pupil i in school j, LRT_{ij} is the same pupil's (standardised) mark in a reading test at age 11. A quadratic relationship with reading test score is assumed at level 1 for the variance function. Uniform priors were used for β, an inverse Wishart prior for Ω_u and a joint Uniform prior for θ subject to constraints.

Heteroscedasticity in a Gaussian multilevel model

The model was estimated using Gibbs sampling apart from for the θ parameters. Here there is very little difference between the truncated normal proposal and the normal proposal MH sampling. For 50,000 iterations the approaches took 6m38s and 6m28s respectively and the estimates were similar. Below are the estimates and Eff values for the normal proposal methods

Parameter	Estimate	Eff
β_0	-0.012 (0.041)	4.3%
β_1	0.558 (0.020)	15.7%
Ω_{u00}	0.097 (0.020)	57.5%
Ω_{u01}	0.020 (0.007)	34.6%
Ω_{u11}	0.015 (0.005)	19.4%
θ_0	0.554 (0.015)	11.3%
θ_1	-0.015 (0.006)	20.9%
θ_2	0.002 (0.009)	11.6%

Here we see that in fact a linear relationship at level 1 may fit equally well.

Between Matrices Constraints Example 2

Heteroscedasticity in a Bivariate Gaussian multilevel model

The concept of heteroscedasticity can be extended to multivariate response models. Here any/or all elements in the variance matrix can be functions of predictor variables. We will consider the following model fitted to an educational dataset analysed in Rasbash et al. (2000). The data consists of written and coursework marks for 1905 students sitting science GCSE in 73 schools in 1985. The model is as follows:

$$written_{ij} = \beta_0 + \beta_2 girl_{ij} + u_{0j} + e_{0ij}$$
$$csework_{ij} = \beta_1 + \beta_3 girl_{ij} + u_{1j} + e_{1ij}$$
$$u_j \sim MVN(0, \Omega_u), e_{ij} \sim MVN(0, \Omega_{eij})$$
$$\Omega_{eij} = \begin{pmatrix} \Omega_{m00} + \delta_{00} girl_{ij} & \Omega_{m01} + \delta_{01} girl_{ij} \\ \Omega_{m01} + \delta_{01} girl_{ij} & \Omega_{m11} + \delta_{11} girl_{ij} \end{pmatrix}$$

So here at level 1 we are estimating a variance matrix for boys and 3 difference parameters between boys and girls. We could in this model reparameterise to a variance matrix for boys and one for girls and use Gibbs sampling but we may want to constrain some of the δ parameters to zero.

Heteroscedasticity in a Bivariate multilevel model

The model was estimated using Gibbs sampling apart from for the δ and Ω_m parameters. There are some missing responses but these are dealt with as an additional Gibbs sampling step in the algorithm. Here there is very little difference between the truncated normal proposal and the normal proposal MH sampling. For 50,000 iterations the approaches took 14m52s and 15m22s respectively and the estimates were similar. Below are the estimates and Eff values for the normal proposal method.

Parameter	Estimate	Eff
β_0	49.46 (0.94)	5.6%
β_1	69.72 (1.20)	5.2%
β_2	-2.51 (0.56)	74.1%
β_3	6.72 (0.68)	74.2%
Ω_{u00}	49.93 (10.15)	49.6%
Ω_{u01}	25.67 (9.55)	51.2%
Ω_{u11}	78.24 (15.54)	54.2%
Ω_{m00}	116.6 (6.50)	2.1%
Ω_{m01}	69.55 (6.71)	1.3%
Ω_{m11}	200.0 (11.43)	1.7%
δ_{00}	15.7 (9.14)	2.0%
δ_{01}	7.32 (8.80)	1.2%
δ_{11}	-30.6 (14.1)	1.5%

Latent variable probit trick

Used in Albert and Chib (1993) for univariate models, and then Chib and Greenberg (1998) consider using latent variables and a probit link to fit multi-variate binary data.

Basic idea:

$y_{ij} \sim Bernouilli(\pi_{ij})$

$probit(\pi_{ij}) = X_{ij}\beta$

is equivalent to

$y^*_{ij} \sim Normal(X_{ij}\beta, 1)$

where y^*_{ij} is a latent variable constrained to be positive when $y_{ij} = 1$ and negative when $y_{ij} = 0$.

Within Matrices Constraints Example

Bivariate Mixed Response model

We will consider a 2 response educational dataset with 1 'continuous' point score in English and a behaviour rating binary response (0 badly behaved, 1 not badly behaved) for 1119 pupils from 47 schools in year 3 of their schooling. We use the probit link and instead model $behaviour^*$ the latent variable as described on the last slide.

The model is as follows:

$$English_{ij} = \beta_0 + \beta_2 x_{1ij} + \beta_4 x_{2ij} + \beta_5 x_{3ij} + u_{0j} + e_{0ij}$$
$$behaviour^*_{ij} = \beta_1 + \beta_3 x_{1ij} + \beta_6 x_{3ij} + u_{1j} + e_{1ij}$$
$$u_j \sim MVN(0, \Omega_u), e_{ij} \sim MVN(0, \Omega_{eij})$$
$$\Omega_{e11} = 1$$

Here x_1 is gender, x_2 is an indicator of English fluency at year 1 and x_3 is a test score at year 1.

Bivariate mixed response model

The model was estimated using Gibbs sampling apart from for the Ω_e parameters. The values of $behaviour^*$ are generated from an additional Gibbs sampling step in the algorithm. Here again there is very little difference between the truncated normal proposal and the normal proposal MH sampling. For 50,000 iterations the approaches took 4m32s and 4m31s respectively and the estimates were similar. Below are the estimates and Eff values for the normal proposal method.

Parameter	Estimate	Eff
β_0	-9.18 (3.28)	56.7%
β_1	-0.36 (0.19)	29.5%
β_2	-6.26 (1.04)	90.2%
β_3	-0.42 (0.09)	28.7%
β_4	6.38 (1.28)	73.9%
β_5	1.66 (0.09)	78.5%
β_6	0.057 (0.008)	27.3%
Ω_{u00}	41.23 (11.61)	38.0%
Ω_{u01}	0.058 (0.429)	16.3%
Ω_{u11}	0.066 (0.031)	9.1%
Ω_{e00}	291.6 (12.52)	14.1%
Ω_{e01}	6.21 (0.733)	6.8%
Ω_{e11}	1.00 (0.0)	-

The correlation at level 1, $\rho_{e01} = 0.36$.

Rats Dataset : Extremely correlated responses

Gelfand et al. (1990) fitted a two-level random slopes regression model to a dataset consisting of 5 repeated measurements of weight for a group of 30 laboratory rats, weighed at ages 8,15,22,29 and 36 days. We could of course fit a multivariate 5 response model to the dataset. One such model would consist of fitting a mean for each age and a 5*5 covariance matrix. We could then estimate such a model using MCMC (Gibbs sampling) and an inverse-Wishart prior for the matrix.

We find that fitting the model using Gibbs gives correlations between consecutive measurements of 0.92, 0.92, 0.95 and 0.92 respectively.

Attempting to fit such a model using single-site Metropolis causes incredibly correlated chains. We ran the normal update method for 500,000 iterations storing every 100th and still only got efficiencies of $< 1\%$ for most of the elements of the matrix and estimates very different to the Gibbs estimates.

Rats Dataset : Autocorrelated residuals

One possible alternative method when we have highly correlated residuals, is to add some structure to the covariance or correlation matrix. In the case of the rats dataset we have the weights of each rat increasing as time goes on and we observe that the between rats variance is increasing with time.

We will therefore impose constraints on the correlation matrix and use an AR(1) formulation i.e. $\rho_{eij} = \tau^{|i-j|}$, where ρ_{eij} is the ijth element of the correlation matrix and τ is the 1 parameter used to describe the whole correlation matrix.

We now use a variation of the transformation methods used earlier and use Metropolis steps to update the 5 variances and the correlation parameter τ. We use independent uniform priors for all 5 variances and τ. We get an estimate $\tau = 0.94$ which being larger than 3 of the 4 1-step correlations suggests that maybe an additional term and an AR(2) model is required.

Useful Web pages

- Slides of this talk available at http://multilevel.ioe.ac.uk/team/billtalk.html

- My Publications available at http://multilevel.ioe.ac.uk/team/bill.html

- Centre for multilevel modelling homepage : http://multilevel.ioe.ac.uk/index.html

- Latest 'development' version of MLwiN plus documentation: http://multilevel.ioe.ac.uk/dev/index.html