Efficient Parametrisations for Normal Linear Mixed Models
Author(s): Alan E. Gelfand, Sujit K. Sahu and Bradley P. Carlin
Source: *Biometrika*, Vol. 82, No. 3 (Sep., 1995), pp. 479-488
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.org/stable/2337527
Accessed: 13-10-2015 09:45 UTC

Biometrika (1995), **82**, 3, *pp.* 479–88
Printed in Great Britain

Efficient parametrisations for normal linear mixed models

BY ALAN E. GELFAND
Department of Statistics, University of Connecticut, Box U-120, Storrs, Connecticut 06269-3120, U.S.A.

SUJIT K. SAHU
Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, U.K.

AND BRADLEY P. CARLIN
Division of Biostatistics, School of Public Health, University of Minnesota, Box 303 Mayo Memorial Building, Minneapolis, Minnesota 55455-0392, U.S.A.

SUMMARY

The generality and easy programmability of modern sampling-based methods for maximisation of likelihoods and summarisation of posterior distributions have led to a tremendous increase in the complexity and dimensionality of the statistical models used in practice. However, these methods can often be extremely slow to converge, due to high correlations between, or weak identifiability of, certain model parameters. We present simple hierarchical centring reparametrisations that often give improved convergence for a broad class of normal linear mixed models. In particular, we study the two-stage hierarchical normal linear model, the Laird–Ware model for longitudinal data, and a general structure for hierarchically nested linear models. Using analytical arguments, simulation studies, and an example involving clinical markers of acquired immune deficiency syndrome (AIDS), we indicate when reparametrisation is likely to provide substantial gains in efficiency.

Some key words: Gibbs sampler; Hierarchical model; Identifiability; Laird-Ware model; Markov chain Monte Carlo; Nested models; Random effects model; Rate of convergence.

1. INTRODUCTION

With the increasing availability of computing power, the scope of models which scientists attempt to study becomes ever more complex. Currently there is considerable interest in fitting 'structured random effects models'. We focus here on normal linear mixed effects models; a separate paper (Gelfand, Sahu & Carlin, 1995) considers nonnormal generalised linear mixed models (Breslow & Clayton, 1993). Such models can yield much behavioural insight, but fitting them often presents a formidable challenge. The associated likelihood or posterior surface can be very irregular, due to high correlation amongst the parameters, or weak identifiability of certain parameters. These problems often arise from vague prior specification for some model parameters, when the data carry inadequate information to 'separate out' all the parameters.

In these settings, fitting techniques for maximisation, e.g. by iteratively reweighted least squares or Fisher scoring, may well not converge. Markov chain Monte Carlo algorithms that update all the states simultaneously, such as a multivariate Metropolis–Hastings

algorithm, will generally be infeasible in the high-dimensional parameter spaces we envision, owing to the difficulty in finding a suitable candidate density. Markov chain Monte Carlo algorithms such as the Gibbs sampler that update component-wise, using lower-dimensional conditional distributions, will also converge very slowly, since high posterior correlations preclude large jumps across the parameter space. Liu, Wong & Kong (1994) describe the benefits of 'blocking', i.e. updating parameters in medium-dimensional groups, and 'collapsing', i.e. generating from partially marginalised distributions, but these strategies do not completely overcome the difficulty. Auxiliary variable techniques, see e.g. Besag & Green (1993), have also been recommended as a method of breaking correlations, but to date they have been successful only in certain stylised applications.

In exploring high dimensional parametric models, the choice of parametrisation can dramatically affect the shape of the surface, and hence its amenability to analysis. This paper attempts to shed light on the effects of various parametrisations for generalised linear mixed models. We offer an assortment of analytical and empirical arguments to suggest that in practice the use of hierarchical centring will generally result in better behaviour by the Markov chain Monte Carlo algorithm. We make no optimality claims for such centring. Instead we present some general rules built around centring which are easily followed and can be expected to yield a more efficient parametrisation in such problems.

The following example may help to convey the idea of hierarchical centring. Consider the model $Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk}$, with

$$\varepsilon_{ijk} \sim N(0, \sigma_e^2), \quad \beta_{ij} \sim N(0, \sigma_\beta^2), \quad \alpha_i \sim N(0, \sigma_\alpha^2), \quad \mu \sim N(\mu_0, \sigma_\mu^2),$$

all terms being independent. Assume that all the variance components are known. Denote the collection of data $\{Y_{ijk}\}$ by Y, the α_i's by α, and the β_{ij}'s by β. As an alternative parametrisation, replace α by η, where $\eta_i = \mu + \alpha_i$, and β by ρ, where $\rho_{ij} = \mu + \alpha_i + \beta_{ij}$. The first parametrisation has only two hierarchical levels, while the second has four. In particular, η is 'centred' about μ, and ρ is 'centred' about η. In subsequent sections, we shall argue that such centring will often result in better behaviour by the Markov chain Monte Carlo algorithm. 'Partially centred' parametrisations, using three hierarchical steps, include (μ, α, ρ) and (μ, η, β). In this example, to which we return in § 4, hierarchical centring is very natural. For other models it might be created artificially as described in § 3.

There is a small literature on the general problem of efficient reparametrisations (Hills & Smith, 1992; Geweke, 1995). Approximate orthogonalisation has been recommended to produce roughly uncorrelated parameters under the posterior. In general such a transformation requires the square root of an approximation to the joint covariance matrix. In high dimensions, obtaining an accurate approximation and computing its inverse can be problematic. However, approximate orthogonalisation nicely complements our hierarchical centring. The centring typically yields low correlation between low dimensional blocks of parameters. Transformation to diminish correlation within such blocks is then usually manageable. Strategies for adaptive approximate orthogonalisation as discussed by Müller (1995) thus become applicable.

In § 2 we consider individual level normal linear models centred around a population level model. Here, exact calculations can be carried out, to clarify the difference between centring and not centring. In § 3 we discuss the general Laird–Ware model (Laird & Ware, 1982), and offer an illustrative analysis of an AIDS data set. Section 4 looks at nested

random effects models. Finally, § 5 offers several concluding remarks, including some indication of how our centring methods may be sensibly applied to nonnested models.

2. Centring in the hierarchical normal linear model

Here we investigate analytically the effect of centring for a two-stage hierarchical normal linear model. Let Y_i be $n_i \times 1$, X_i be $n_i \times p$, η_i be $p \times 1$ and assume that

$$Y_i|\eta_i \sim N(X_i\eta_i, \sigma_i^2 I_{n_i}) \quad (i = 1, \ldots, m).$$

As a general rule we would first recentre and rescale any quantitative covariates over i, which will aid Bayesian fitting by reducing correlations in the likelihood surface. Let $\eta_i|\mu \sim N(\mu, D)$ where μ is $p \times 1$. That is, we assume individual models centred around a population model, for instance growth curves about a baseline population curve. For the moment, assume σ_i^2 and D known, and take a flat prior on μ. We refer to $(\mu, \eta_1, \ldots, \eta_m)$ as the centred parametrisation, while $(\mu, \alpha_1, \ldots, \alpha_m)$ with $\alpha_i = \eta_i - \mu$ is the uncentred parametrisation. In either case, the total number of parameters is $(m+1)p$ and the posterior distribution is multivariate normal (Lindley & Smith, 1972), so that the parametrisations can be compared via their posterior correlation structure, weak correlations being desirable. Our model specification is a special case of the Laird–Ware model (Laird & Ware, 1982) which sets $E(Y_i) = X_i\mu + Z_i\alpha_i$. Here we assume $X_i = Z_i$. Centring for the general Laird–Ware model will be discussed in the next section.

The Y_i, given μ alone, are conditionally independent with distribution $Y_i|\mu \sim N(X_i\,\mu, \Sigma_i)$, where $\Sigma_i = \sigma_i^2 I_{n_i} + X_i D X_i^{\mathrm{T}}$. Thus $\mu|Y \sim N\{\hat{\mu}, (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}\}$, where

$$X^{\mathrm{T}} = (X_1^{\mathrm{T}}, \ldots, X_m^{\mathrm{T}}), \quad \Sigma = \mathrm{diag}\,(\Sigma_1, \ldots, \Sigma_m), \quad \hat{\mu} = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}Y$$

with $Y^{\mathrm{T}} = (Y_1^{\mathrm{T}}, \ldots, Y_m^{\mathrm{T}})$. Note that

$$X^{\mathrm{T}}\Sigma^{-1}X = \sum X_i^{\mathrm{T}}\Sigma_i^{-1}X_i, \quad X^{\mathrm{T}}\Sigma^{-1}Y = \sum X_i^{\mathrm{T}}\Sigma_i^{-1}Y_i.$$

Let $A_i = X_i^{\mathrm{T}}\Sigma_i^{-1}X_i$ and $A = X^{\mathrm{T}}\Sigma^{-1}X$. Henceforth we assume A^{-1} exists, although A_i^{-1} need not if, for instance, $n_i < p$.

Standard calculations yield $\eta_i|\mu,\ Y \sim N(B_i b_i, B_i)$, where

$$B_i = (\sigma_i^{-2}X_i^{\mathrm{T}}X_i + D^{-1})^{-1}, \quad b_i = \sigma_i^{-2}X_i^{\mathrm{T}}Y_i + D^{-1}\mu.$$

Marginalising over μ gives that $\eta|Y$ is normal with $E(\eta_i|Y) = B_i\hat{b}_i$ and

$$\mathrm{var}\,(\eta_i|Y) = B_i + B_i D^{-1}A^{-1}D^{-1}B_i, \tag{2·1}$$

where $\hat{b}_i = \sigma_i^{-2}X_i^{\mathrm{T}}Y_i + D^{-1}\hat{\mu}$. Moreover,

$$\mathrm{cov}\,(\eta_i, \mu|Y) = B_i D^{-1}A^{-1}, \quad \mathrm{cov}\,(\eta_i, \eta_j|Y) = B_i D^{-1}A^{-1}D^{-1}B_j \quad (i \neq j). \tag{2·2}$$

Hence in the μ-α space, we have $\alpha|Y$ normal with $E(\alpha_i|Y) = B_i\hat{b}_i - \hat{\mu}$ and

$$\mathrm{var}\,(\alpha_i|Y) = B_i + B_i D^{-1}A^{-1}D^{-1}B_i + A^{-1} - 2B_i D^{-1}A^{-1}. \tag{2·3}$$

Finally, direct calculation yields

$$\begin{aligned} \mathrm{cov}\,(\alpha_i, \mu|Y) &= B_i D^{-1}A^{-1} - A^{-1}, \\ \mathrm{cov}\,(\alpha_i, \alpha_j|Y) &= B_i D^{-1}A^{-1}D^{-1}B_j - (B_i + B_j)D^{-1}A^{-1} + A^{-1} \quad (i \neq j). \end{aligned} \tag{2·4}$$

Using a standard matrix identity (Rao, 1973, p. 33) we have

$$B_i D^{-1} + DA_i = I_{n_i}. \tag{2·5}$$

In (2·5), by assumption B_iD^{-1} is positive definite, while DA_i is positive semidefinite. We suggest that B_iD^{-1} and DA_i measure the relative contribution of the first stage or error variance and the second stage or random effect variance respectively. To clarify this, note that

$$\text{tr}\,(DA_i) = \text{tr}\,\{(\sigma_i^2 I + X_i D X_i^{\mathrm{T}})^{-1}(X_i D X_i^{\mathrm{T}})\},$$

reflecting the contribution of the random effects to the total 'variance' for Y_i. Alternatively,

$$|B_iD^{-1}| = (|\sigma_i^{-2} D X_i^{\mathrm{T}} X_i + I|^{-1}) \leqslant (|D|\,|\sigma_i^{-2} X_i^{\mathrm{T}} X_i|)^{-1}.$$

If the generalised variance $|D|$ is large or if σ_i^2 is small, $|B_iD^{-1}|$ is also small, reflecting the relative contribution of the random effects. For instance, in the elementary model

$$Y_i = \mu + \alpha_i + \varepsilon_i \quad (i = 1, \ldots, m),$$

where $\varepsilon_i \sim N(0, \sigma_e^2)$, $\alpha_i \sim N(0, \sigma_\alpha^2)$ and μ has a flat prior, we find $B_iD^{-1} = (\sigma_e^2 + \sigma_\alpha^2)^{-1}\sigma_e^2$.

In fact (2·5) implies that $|B_iD^{-1}| < 1$. We argue below that if $|B_iD^{-1}|$ is near zero the centring parametrisation is efficient, while if $|B_iD^{-1}|$ is near one the uncentred parametrisation will be preferred. To permit imprecise second stage specification compared with that of the first stage we write $D = \delta D_0$ and let $\delta \to \infty$; as δ varies we retain the same correlation structure. Also, as $\delta \to \infty$ all $|B_iD^{-1}| \to 0$. Appendix 1 shows that, as $\delta \to \infty$ with σ_i^2 fixed,

$$\text{corr}\,\{(\eta_{ir}, \mu_s)|\,Y\} \to 0, \quad \text{corr}\,\{(\eta_{ir}, \eta_{js})|\,Y\} \to 0,$$

but corr $\{(\alpha_{ir}, \mu_s)|\,Y\}$ and corr $\{(\alpha_{ir}, \alpha_{js})|\,Y\}$ do not. Next, let $\sigma_0^2 = \min\{\sigma_i^2\}$. Then as $\sigma_0^2 \to \infty$, all $|B_iD^{-1}| \to 1$. Appendix 1 shows further that, as $\sigma_0^2 \to \infty$ with δ fixed,

$$\text{corr}\,\{(\alpha_{ir}, \mu_s)|\,Y\} \to 0, \quad \text{corr}\,\{(\alpha_{ir}, \alpha_{js})|\,Y\} \to 0,$$

but corr $\{(\eta_{ir}, \mu_s)|\,Y\}$ and corr $\{(\eta_{ir}, \eta_{js})|\,Y\}$ do not.

Appendix 1 also notes that, regardless of the asymptotic behaviour of δ and σ_0^2, corr $\{(\eta_{ir}, \eta_{is})|\,Y\}$ and corr $\{(\alpha_{ir}, \alpha_{is})|\,Y\}$ are $O(1)$. The proposed centring is not intended to handle correlation within the ith block of parameters, $i = 1, \ldots, m$. In practice p is usually much smaller than $(m+1)p$, so that using, say, a Gibbs sampler we would draw η_i or α_i as a vector, rather than by individual components, thus diminishing concern regarding within-block correlation structure. We return to this issue in § 3.

For this model, with $\sigma^2 := (\sigma_1^2, \ldots, \sigma_m^2)$ and D assumed known, the decision as to whether or not to centre is not data-dependent, but emerges from the model specification. In applications, however, these parameters will be unknown, so that we should instead consider the joint distribution

$$p(\mu, \eta, \sigma^2, D|\,Y) = p(\mu, \eta|\sigma^2, D, Y)p(\sigma^2, D|\,Y).$$

Some preliminary Monte Carlo sampling from $p(\sigma^2, D|\,Y)$ would enable investigation of the posterior distributions of the $|B_iD^{-1}|$. In practice, the random effects are introduced to 'soak up' high variability associated with the population effects model. In such cases we expect that the σ_i^2 will not dominate the variability, so that centring can be recommended.

3. THE LAIRD–WARE NORMAL LINEAR MIXED MODEL

Consider the general Laird–Ware model:

$$Y_i = X_i\alpha + Z_i\beta_i + \varepsilon_i \quad (i = 1, \ldots, m), \tag{3·1}$$

where Y_i is $n_i \times 1$, X_i is $n_i \times r$, α is an $r \times 1$ vector of population effects, Z_i is $n_i \times p$ and β_i is a $p \times 1$ vector of individual random effects. An assortment of artificial hierarchical centrings can be developed for (3·1). For example, suppose we have

$$Y_{ij} = \mu + \alpha x_i + \beta_i + \varepsilon_{ij}.$$

Let $\rho_i = \mu + \alpha x_i + \beta_i$ and let $\eta_i = \mu + \beta_i$. Parametrisations of possible interest then are (μ, α, ρ), (μ, α, η) and (μ, α, β). The unpublished Ph.D. thesis of S. K. Sahu studies the effect of centring for this and other similar examples using the analytical approach of § 2.

In many applications Z_i arises as a subset of the columns of X_i, since the number of population effects would typically be larger than the number of random effects per individual. For example, X_i might repeat the columns of Z_i, multiplied by certain covariate values; i.e.

$$X_i = (Z_i | a_{i1} Z_i | a_{i2} Z_i | \ldots | a_{is} Z_i),$$

where $a_i^{\mathrm{T}} = (a_{i1}, \ldots, a_{is})^{\mathrm{T}}$ is a $1 \times s$ vector of covariates for the ith individual. Partitioning α^{T} as $(\alpha^{(0)\mathrm{T}} | \alpha^{(1)\mathrm{T}} | \alpha^{(2)\mathrm{T}} | \ldots | \alpha^{(s)\mathrm{T}})$ and letting β denote the set $(\beta_1, \ldots, \beta_m)$, (α, β) is clearly an uncentred parametrisation. The fully centred parametrisation (α, ρ) uses $\rho_i = \alpha^{(0)} + \sum a_{il} \alpha^{(l)} + \beta_i$. Obviously there are many intermediate artificial centrings.

To illustrate, we analyse a subset of $m = 10$ cases from a recently conducted AIDS clinical trial. Here, Y_{ij} represents the square root of the jth CD4 cell count measurement on patient i, where $n_i \leqslant 5$ for all i. For a patient with no missing data, the jth measurement was taken t_j months after baseline at study enrollment, where $t_j \in \{0, 2, 6, 12, 18\}$. Each patient was HIV-positive at baseline, but not all had yet received an AIDS diagnosis. We wish to fit a simple linear decay model to these longitudinal data, taking into account the important covariate of baseline AIDS diagnosis.

We thus fit model (3·1) with $X_i = (Z_i | a_i Z_i)$, where the jth row of Z_i is given by $z_{ij}^{\mathrm{T}} = (1 \;\; t_j)$, and where a_i indicates the baseline AIDS status of individual i (1 = yes, 0 = no). Thus Z_i is $n_i \times 2$ and X_i is $n_i \times 4$; our simplified model allows the covariate to modify the slope and intercept at the population level, but not at the individual level. In order to ensure model identifiability, we specify $\alpha \sim N(\mu_\alpha, \Sigma_\alpha)$ and $\beta_i \sim N(0, D)$, independently. We complete the prior specification by taking $\sigma_e^2 \sim \mathrm{IG}(\nu, \tau)$ and $D \sim \mathrm{IW}\{(\delta\Lambda)^{-1}, \delta\}$, where IG denotes the inverse gamma distribution and IW the inverse Wishart distribution. The hyperparameter values chosen for the priors on the variance components will considerably influence their magnitudes under the posterior, and hence the reparametrisation for most rapid convergence of our Markov chain Monte Carlo algorithm.

Suppose we partition α^{T} as $((\alpha^{(0)})^{\mathrm{T}} \; (\alpha^{(a)})^{\mathrm{T}})$, separating the overall intercept and slope from the AIDS intercept and slope. Define

$$\eta_i = \alpha^{(0)} + \beta_i, \quad \rho_i = \alpha^{(0)} + a_i \alpha^{(a)} + \beta_i$$

and write $\beta = (\beta_1, \ldots, \beta_m)$, $\eta = (\eta_1, \ldots, \eta_m)$ and $\rho = (\rho_1, \ldots, \rho_m)$. We compare three candidate parametrisations: uncentred $(\alpha^{(0)}, \alpha^{(a)}, \beta)$, partially centred $(\alpha^{(0)}, \alpha^{(a)}, \eta)$, and fully centred $(\alpha^{(0)}, \alpha^{(a)}, \rho)$. In all three cases, the full conditional distributions necessary for the implementation of the Gibbs sampler arise as normal, inverse gamma, or inverse Wishart forms due to the conjugacy of our random effects and prior specification with the normal likelihood. Details in the case of the standard uncentred parametrisation are provided by Lange, Carlin & Gelfand (1992).

In order to completely specify the prior distribution, we must select values for μ_α, Σ_α, ν, τ, δ and Λ. The patients in our particular data set are very ill, having already failed or

emerged as intolerant to zidovudine (AZT) therapy. While healthy persons have CD4 counts well over 500, in our data set we expect baseline counts of around 100 for those without an AIDS diagnosis, and perhaps near 50 for those with AIDS. Our feelings with respect to slopes are more tentative: we expect a further decline in CD4 counts as time goes on, but their high variability and low initial level suggest that an increase is not impossible. Thus, before centring the covariate a_i, we select $\mu_\alpha^T = (10, 0, -3, 0)$ and $\Sigma_\alpha = \text{diag}(2^2, 1^2, 1^2, 1^2)$. Next, we set $\nu = 3$ and $\tau = 0{\cdot}02$, so that σ_e^2 has prior mean and standard deviation equal to 5^2. Finally, we take $\delta = 2$, the smallest value for which the prior will be proper and $\Lambda = \text{diag}(2^2, 1^2)$, so that the individual level intercepts and slopes have variability roughly equal to that specified for the overall intercept and slope, $\alpha^{(0)}$.

For each of our three parametrisations, we use an overdispersed initial distribution to select five starting points in corresponding parameter space. From these points we run five parallel Gibbs sampling chains each for 100 iterations, a number small enough to reveal differences among the parametrisations. Table 1 presents convergence summaries of the resulting sampled values for three sets of individual level random effects, β_6, β_7 and β_8, as well as the four components of the population parameter α. Patient 6 is AIDS-positive with a single observation ($n_6 = 1$), patient 7 is AIDS-positive with four observations, and patient 8 is AIDS-negative with three observations. A rapidly mixing chain should have Gelman & Rubin (1992) statistics near 1, due to overlap of the originally overdispersed chains, and sample autocorrelation near 0. The diagnostics for the random effects suggest that all three parametrisations are more or less equivalent, though some preference for the two centred parametrisations is evident. However, the diagnostics for the population effects show a much clearer preference for centring, and, in the case of the AIDS slope $\alpha_1^{(a)}$, for the full centring of the random effects around both the overall and AIDS main effects offered by parametrisation 3. This is not unexpected, since our prior specification for D was substantially more vague than that for σ_e^2. Analyses using priors that encourage even smaller values for σ_e^2 and larger values for D produce even more dramatic differences in convergence.

Table 1. *Convergence diagnostics for the* AIDS *data example*

	Parametrisation 1 (uncentred)			Parametrisation 2 (partially centred)			Parametrisation 3 (fully centred)		
	G&R quantiles		lag 1	G&R quantiles		lag 1	G&R quantiles		lag 1
	50%	95%	ACF	50%	95%	ACF	50%	95%	ACF
$\beta_{6,0}$	1·00	1·01	0·159	1·00	1·01	0·161	1·00	1·02	0·149
$\beta_{6,1}$	1·03	1·07	−0·060	1·03	1·07	0·075	1·02	1·06	0·139
$\beta_{7,0}$	1·04	1·12	0·082	1·00	1·02	0·100	1·00	1·03	0·074
$\beta_{7,1}$	1·13	1·34	0·030	1·05	1·14	−0·018	1·01	1·03	−0·015
$\beta_{8,0}$	1·13	1·33	0·515	1·06	1·18	0·440	1·03	1·08	0·368
$\beta_{8,1}$	1·25	1·59	0·397	1·07	1·19	0·562	1·01	1·04	0·306
$\alpha_0^{(0)}$	1·16	1·39	0·338	1·02	1·06	0·467	1·02	1·06	0·473
$\alpha_1^{(0)}$	1·56	2·22	0·549	1·00	1·01	0·348	1·00	1·01	0·350
$\alpha_0^{(a)}$	1·02	1·06	0·305	1·02	1·06	0·171	1·00	1·02	−0·010
$\alpha_1^{(a)}$	1·52	2·12	0·767	1·34	1·79	0·877	1·02	1·07	0·443

Columns provide the 50th and 95th percentiles of Gelman & Rubin's (1992) variance inflation factor (G&R), and the value of the lag 1 sample autocorrelation (ACF) computed from the first of the five sampled chains.

4. The nested random effects normal linear model

Suppose there are additional levels of the hierarchy. For instance, we might denote the response for the kth child in the jth class of the ith school by Y_{ijk} and model it as

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk}, \tag{4·1}$$

where $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, n_{ij}$. At each level of the hierarchy we set up a linear model relating the terms in (4·1) to explanatory variables, i.e.

$$\alpha_i = x_i^{\mathrm{T}}\alpha + \varepsilon_i, \quad \beta_{ij} = w_{ij}^{\mathrm{T}}\beta_i + \lambda_{ij}, \quad \gamma_{ijk} = z_{ijk}^{\mathrm{T}}\gamma_{ij} + \xi_{ijk}.$$

Here, x_i denotes school-level covariates, w_{ij} class-level covariates, and z_{ijk} student-level covariates. Such models are referred to as multilevel or nested mixed linear models, e.g. Goldstein (1986), Longford (1987). Our centring ideas can be relevant here, applied in sequence from the highest level down.

We illustrate using (4·1) with no covariates, assuming $\varepsilon_{ijk} \sim N(0, \sigma_e^2)$, $\beta_{ij} \sim N(0, \sigma_\beta^2)$ and $\alpha_i \sim N(0, \sigma_\alpha^2)$, all independent. Define $\rho_{ij} = \mu + \alpha_i + \beta_{ij}$, whence $Y_{ijk}|\rho_{ij} \sim N(\rho_{ij}, \sigma_e^2)$. Define $\eta_i = \mu + \alpha_i$ so that $\rho_{ij}|\eta_i \sim N(\eta_i, \sigma_\beta^2)$ and $\eta_i|\mu \sim N(\mu, \sigma_\alpha^2)$. We assume a flat prior for μ, and take σ_e^2, σ_β^2 and σ_α^2 as known. The centred parametrisation is thus (μ, η, ρ). Exact calculations are more complicated here since marginalisation produces dependence amongst the Y_{ijk}.

Let

$$Y_{ij}^{\mathrm{T}} = (Y_{ij1}, \ldots, Y_{ijn_{ij}}), \quad Y_i^{\mathrm{T}} = (Y_{i1}^{\mathrm{T}}, \ldots, Y_{iJ}^{\mathrm{T}}), \quad Y^{\mathrm{T}} = (Y_1^{\mathrm{T}}, \ldots, Y_I^{\mathrm{T}}).$$

Then, given μ, the Y_i are independent normal with $E(Y_i|\mu) = \mu 1_{n_{i+}}$ and variance

$$\mathrm{var}\,(Y_i|\mu) = \Sigma_i := \sigma_e^2 I_{n_{i+}} + X_i D X_i^{\mathrm{T}}. \tag{4·2}$$

Here $n_{i+} = \sum n_{ij}$, $D = \mathrm{diag}\,(\sigma_\alpha^2, \sigma_\beta^2, \ldots, \sigma_\beta^2)$, and X_i is $n_{i+} \times (J+1)$ with first column all 1's, and $(j+1)$th column $(j = 1, \ldots, J)$ having a 1 in each of the n_{ij} rows corresponding to Y_{ijk} $(k = 1, \ldots, n_{ij})$, and 0's elsewhere. From (4·2),

$$E(\mu|Y) = \hat{\mu} := (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}Y, \quad \mathrm{var}\,(\mu|Y) = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1} = (\textstyle\sum X_i^{\mathrm{T}}\Sigma_i^{-1}X_i)^{-1}.$$

Following § 2 we investigate the $|B_i D^{-1}|$, where $B_i = (\sigma_e^{-2}X_i^{\mathrm{T}}X_i + D^{-1})^{-1}$. In Appendix 2 we show that if $\sigma_\beta^2 \to \infty$ with σ_α^2 and σ_e^2 fixed, then $|B_i D^{-1}| \to 0$, so centring will be preferred. If on the other hand $\sigma_e^2 \to \infty$ with σ_α^2 and σ_β^2 fixed, then $|B_i D^{-1}| \to 1$ and the uncentred parametrisation will be better.

We now describe a simulation study of (4·1) with $I = J = 3$, $n_{ij} = 5$ and the true value of μ equal to 0. We are interested in comparing four possible parametrisations: uncentred (μ, α, β), α's centred (μ, η, β), β's centred (μ, α, ρ), and both α's and β's centred (μ, η, ρ). After selecting fixed values for σ_e^2, σ_β^2 and σ_α^2, designed to favour one parametrisation or another, we first generate 'true' parameter values from the assumed priors, then data given these parameter values. A single Gibbs sampling chain is then run using the uncentred parametrisation for 10 000 iterations. The output from this chain is used to estimate the posterior correlation matrix among the $IJ + I + 1 = 13$ parameters under each of the four parametrisations.

Table 2(a) tabulates the resulting estimated correlations arising under the four parametrisations for the case $\sigma_e = 10$, $\sigma_\beta = 1$ and $\sigma_\alpha = 1$. Since the model variability is large relative to that in both stages of the prior, we would expect the uncentred parametrisation to be best. This is in fact the case, with 75 of the 78 sample correlations falling in $(-0{\cdot}1, 0{\cdot}1)$. Table 2(b) gives results for the case $\sigma_e = 1$, $\sigma_\beta = 10$ and $\sigma_\alpha = 1$. Here, as predicted by our

Table 2. *Absolute sample correlations for different parametrisations*

(*a*) $\sigma_e = 10$, $\sigma_\beta = 1$, $\sigma_\alpha = 1$

	Lower bin boundary					
	0·0	0·1	0·3	0·5	0·7	0·9
	Upper bin boundary					
Parametrisation	0·1	0·3	0·5	0·7	0·9	1·0
Uncentred	75	3	0	0	0	0
α's centred	72	0	0	3	3	0
β's centred	3	23	7	27	18	0
Both α's and β's centred	0	0	0	48	30	0

(*b*) $\sigma_e = 1$, $\sigma_\beta = 10$, $\sigma_\alpha = 1$

	Lower bin boundary					
	0·0	0·1	0·3	0·5	0·7	0·9
	Upper bin boundary					
Parametrisation	0·1	0·3	0·5	0·7	0·9	1·0
Uncentred	13	20	0	0	0	45
α's centred	0	0	0	0	0	78
β's centred	76	2	0	0	0	0
Both α's and β's centred	72	0	0	0	0	6

(*c*) $\sigma_e = 1$, $\sigma_\beta = 1$, $\sigma_\alpha = 10$

	Lower bin boundary					
	0·0	0·1	0·3	0·5	0·7	0·9
	Upper bin boundary					
Parametrisation	0·1	0·3	0·5	0·7	0·9	1·0
Uncentred	54	9	0	9	0	6
α's centred	60	0	0	9	9	0
β's centred	72	0	0	0	0	6
Both α's and β's centred	69	9	0	0	0	0

theoretical work, the parametrisation wherein only the β's are centred emerges as best. Finally, Table 2(c) considers the case $\sigma_e = 1$, $\sigma_\beta = 1$ and $\sigma_\alpha = 10$, where we have no directly applicable theory. We might have suspected that centring the α's alone would be best, but the table shows that this parametrisation suffers from some fairly large positive and negative correlations: among the β_{ij}'s for fixed i, and between η_i and the β_{ij}'s, respectively. Centring both the α's and β's is preferred. Our suggestion is to centre the effects with large prior variance relative to σ_e^2 and all effects at lower, i.e. more heavily subscripted, stages in the nested hierarchy.

5. CONCLUSION

All the models we have considered admit an unambiguous overall hierarchical centring which may be combined with further artificial centring. What of settings that are not naturally hierarchical, such as the simple two-way additive model $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$? Preliminary simulation work (Gelfand et al., 1995) indicates that, at a given stage in the hierarchy, centring the random effects with the largest posterior variance can lead to improved convergence, especially if this variability dominates that at all lower levels. Thus

in the two-way model we would centre either the α's or the β's, provided their variability dominated that at the data level. In summary, hierarchical centring appears to be a useful, simple, and practical tool for potentially improving convergence of maximisation and sampling-based algorithms for likelihood and Bayesian analysis of Gaussian linear mixed models.

Appendix 1

Limiting correlation behaviour for the two-stage normal linear model

We consider the behaviour of (2·1)–(2·4) as $\delta \to \infty$ or as the $\sigma_i^2 \to \infty$. Let $L_i := \sigma_i^{-2} D_0 X_i^{\mathrm{T}} X_i + \delta^{-1} I$. Then $B_i = L_i^{-1} D_0$ and $B_i D^{-1} = \delta^{-1} L_i^{-1}$. Moreover, as $\delta \to \infty$, $L_i^{-1} \to \sigma_i^2 (X_i^{\mathrm{T}} X_i)^{-1} D_0^{-1}$.

First let δ grow large with the σ_i^2 fixed. Then $(B_i)_{rs} = O(1)$ and $(B_i D^{-1})_{rs} = O(\delta^{-1})$. According to (2·5),

$$\sum B_l D^{-1} + DA = mI$$

so $A = mD^{-1} - D^{-1} \sum B_l D^{-1}$ and thus $(A^{-1})_{rs} = O(\delta)$. Furthermore,

$$(B_i D^{-1} A^{-1})^{-1} = B_i^{-1} DA = mB_i^{-1} - B_i^{-1} \sum B_l D^{-1}.$$

Hence $(B_i D^{-1} A^{-1})_{rs} = O(1)$. In fact, $B_i D^{-1} A^{-1} \to m^{-1} \sigma_i^2 (X_i^{\mathrm{T}} X_i)^{-1}$. Finally $(B_i D^{-1} A^{-1} D^{-1} B_j)_{rs} = O(\delta^{-1})$ regardless of whether of not $i = j$. We deduce

$$\mathrm{corr}\,\{(\eta_{ir}, \mu_s) | Y\} = O(\delta^{-\frac{1}{2}}), \quad \mathrm{corr}\,\{(\eta_{ir}, \eta_{js}) | Y\} = O(\delta^{-1}),$$

while corr $\{(\eta_{ir}, \eta_{is}) | Y\}$, corr $\{(\alpha_{ir}, \mu_s) | Y\}$, corr $\{(\alpha_{ir}, \alpha_{js}) | Y\}$ and corr $\{(\alpha_{ir}, \alpha_{is}) | Y\}$ are all $O(1)$.

Now let the σ_i^2 grow large with δ fixed. As $\sigma_i^2 \to \infty$ we have $B_i \to D$, $A_i = O(\sigma_i^{-2})$ and, if $\sigma_0^2 = \min(\sigma_i^2)$, then $(A^{-1})_{rs} = O(\sigma_0^2)$. Also

$$B_i D^{-1} A^{-1} - A^{-1} = -DA_i A^{-1} = O(1).$$

Hence

$$\mathrm{corr}\,\{(\alpha_{ir}, \mu_s) | Y\} = O(\sigma_0^{-1}), \quad \mathrm{corr}\,\{(\alpha_{ir}, \alpha_{js}) | Y\} = O(\sigma_0^{-1}),$$

while corr $\{(\alpha_{ir}, \alpha_{is}) | Y\}$, corr $\{(\eta_{ir}, \mu_s) | Y\}$, corr $\{(\eta_{ir}, \eta_{js}) | Y\}$ and corr $\{(\eta_{ir}, \eta_{is}) | Y\}$ are all $O(1)$.

Appendix 2

Limiting correlation behaviour for the nested normal linear model

From (2·5) we compute $B_i D^{-1} = I - DA_i = I - DX_i^{\mathrm{T}} \Sigma_i^{-1} X_i$. We may show that $\Sigma_i = W_i + \sigma_\alpha^2 1 1^{\mathrm{T}}$, where $W_i = \mathrm{diag}\,(W_{i1}, \ldots, W_{iJ})$ with $W_{ij} = \sigma_e^2 I + \sigma_\beta^2 1 1^{\mathrm{T}}$, an $n_{ij} \times n_{ij}$ matrix. We may then compute Σ_i^{-1} and thus $B_i D^{-1}$. We find

$$(B_i D^{-1})_{11} = (1 + \sigma_\alpha^2 q_i)^{-1}, \quad (B_i D^{-1})_{1j} = -\sigma_\alpha^2 n_{ij} e_{ij} (1 + \sigma_\alpha^2 q_i)^{-1},$$

$$(B_i D^{-1})_{j1} = -\sigma_\beta^2 n_{ij} e_{ij} (1 + \sigma_\alpha^2 q_i)^{-1}, \quad (B_i D^{-1})_{ll} = 1 - \sigma_\beta^2 n_{il} e_{il} \{1 - \sigma_\alpha^2 n_{il} e_{il} (1 + \sigma_\alpha^2 q_i)^{-1}\}$$

for $l > 1$, and

$$(B_i D^{-1})_{rs} = \sigma_\alpha^2 \sigma_\beta^2 n_{ir} n_{is} e_{ir} e_{is} (1 + \sigma_\alpha^2 q_i)^{-1}$$

for $r > 1$ and $s > 1$. Here, $e_{ij} = (\sigma_e^2 + \sigma_\beta^2 n_{ij})^{-1}$ and $q_i = \sum n_{ij} e_{ij}$. Suppose that $\sigma_\beta^2 \to \infty$ with σ_e^2 and σ_α^2 fixed. Note that $\sigma_\beta^2 n_{ij} e_{ij} \to 1$ and $\sigma_\beta^2 d_i \to J^{-1}$. Hence $(B_i D^{-1})_{rs} \to 0$ for $r > 1$ and $s > 1$, $(B_i D^{-1})_{ll} \to 0$ for $l > 1$, and $(B_i D^{-1})_{1j} \to 0$. Thus $|B_i D^{-1}| \to 0$. If $\sigma_e^2 \to \infty$ with σ_α^2 and σ_β^2 fixed, $(B_i D^{-1})_{11} \to 1$ and $(B_i D^{-1})_{ll} \to 1$ for $l > 1$; all other entries tend to 0. Hence $B_i D^{-1} \to I$ and thus $|B_i D^{-1}| \to 1$.

REFERENCES

BESAG, J. & GREEN, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. R. Statist. Soc.* B **55**, 25–37.

BRESLOW, N. E. & CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

GELFAND, A. E., SAHU, S. K. & CARLIN, B. P. (1995). Efficient parametrizations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. To appear. Oxford University Press.

GELMAN, A. & RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457–511.

GEWEKE, J. (1995). Bayesian reduced rank regression in econometrics. *J. Economet.* To appear.

GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43–56.

HILLS, S. E. & SMITH, A. F. M. (1992). Parametrization issues in Bayesian inference. In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 461–9. Oxford University Press.

LAIRD, N. M. & WARE, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–74.

LANGE, N., CARLIN, B. P. & GELFAND, A. E. (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers (with discussion). *J. Am. Statist. Assoc.* **87**, 615–32.

LINDLEY, D. V. & SMITH, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B **34**, 1–41.

LIU, J. S., WONG, W. H. & KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

LONGFORD, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 817–27.

MÜLLER, P. (1995). Metropolis posterior integration schemes. *Statist. Comp.* To appear.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications.* New York: Wiley.

[*Received December* 1993. *Revised November* 1994]