J. R. Statist. Soc. B (2000)
62, *Part* 2, *pp.* 399–412

Inference for multivariate normal hierarchical models

Philip J. Everson

Swarthmore College, USA

and Carl N. Morris

Harvard University, Cambridge, USA

[Received October 1997. Revised August 1999]

Summary. This paper provides a new method and algorithm for making inferences about the parameters of a two-level multivariate normal hierarchical model. One has observed J p-dimensional vector outcomes, distributed at level 1 as multivariate normal with unknown mean vectors and with known covariance matrices. At level 2, the unknown mean vectors also have normal distributions, with common unknown covariance matrix $\mathbf{A}$ and with means depending on known covariates and on unknown regression coefficients. The algorithm samples independently from the marginal posterior distribution of $\mathbf{A}$ by using rejection procedures. Functions such as posterior means and covariances of the level 1 mean vectors and of the level 2 regression coefficient are estimated by averaging over posterior values calculated conditionally on each value of $\mathbf{A}$ drawn. This estimation accounts for the uncertainty in $\mathbf{A}$, unlike standard restricted maximum likelihood empirical Bayes procedures. It is based on independent draws from the exact posterior distributions, unlike Gibbs sampling. The procedure is demonstrated for profiling hospitals based on patients' responses concerning $p = 2$ types of problems (non-surgical and surgical). The frequency operating characteristics of the rule corresponding to a particular vague multivariate prior distribution are shown via simulation to achieve their nominal values in that setting.

Keywords: Constrained Wishart distribution; Importance weighting; Interval estimates; Medical profiling; Multivariate empirical Bayes procedures; Rejection sampling; Restricted maximum likelihood

1. Introduction and summarizing remarks

We make inferences simultaneously from data about J independent problems, each involving a p-dimensional parameter vector $\boldsymbol{\theta}_j$, $j = 1, \ldots, J$. Level 1 in each problem provides an unbiased estimate $\mathbf{Y}_j$ for $\boldsymbol{\theta}_j$, having a multivariate normal distribution with mean vector $\boldsymbol{\theta}_j$ and known $p \times p$ covariance matrix $\mathbf{V}_j$. Thus the level 1 model is

$$\mathbf{Y}_j | \boldsymbol{\theta}_j \overset{\text{indep}}{\sim} N_p(\boldsymbol{\theta}_j, \mathbf{V}_j), \qquad j = 1, \ldots, J. \tag{1}$$

In the application in Section 4, $\boldsymbol{\theta}_j$ is the vector of true problem rates for $p = 2$ different types of service problems in hospital j and $\mathbf{Y}_j$ is the average of observed rates for a sample of patients from that hospital. Alternatively, $\boldsymbol{\theta}_j$ might be the regression coefficient vector for predicting students' test scores for p observable student covariates in school j, with $\mathbf{Y}_j$ the

Address for correspondence: Philip J. Everson, Department of Mathematics and Statistics, Swarthmore College, Swarthmore, PA 19081, USA.
E-mail: peverso1@swarthmore.edu

 1369–7412/00/62399

least squares estimate from scores for a sample of students in that school (e.g. Bryk and Raudenbush (1992)). For other applications $\mathbf{Y}_j$ may be a more complicated vector function of within-group observations, unbiased for $\boldsymbol{\theta}_j$. The within-group covariance matrix $\mathbf{V}_j$ often would not be known exactly, typically being estimated from within-group data. Because there are many examples having abundant within-group observations, for which errors in estimating these covariances are not serious, we assume for simplicity that the $\mathbf{V}_j$ are known.

Level 2 of the hierarchical model connects the J conditionally independent inference problems. The p-vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J$ for the J problems follow conjugate multivariate normal distributions with means depending on q known level 2 covariates (including any constant term):

$$\boldsymbol{\theta}_j \overset{\text{indep}}{\sim} N_p(\mathbf{W}_j'\boldsymbol{\gamma}, \mathbf{A}), \qquad j = 1, \ldots, J. \tag{2}$$

In expression (2) the known $r \times p$ ($r \equiv pq$) matrix $\mathbf{W}_j$ is block diagonal with a q-dimensional covariate vector $\mathbf{w}_j$ repeated along the diagonal, i.e. $\mathbf{W}_j = \mathbf{I}_{p\times p} \otimes \mathbf{w}_j$ in Kronecker product notation. The (unknown) hyperparameter $\boldsymbol{\gamma}$ has length $r \equiv pq$, $q < J$, and $\mathbf{A}$ is the common (unknown) $p \times p$ covariance matrix for each $\boldsymbol{\theta}_j$.

Early developments of this multivariate ($p \geqslant 2$) two-level normal model include those of Lindley and Smith (1972) and Efron and Morris (1972) who investigated the operating characteristics. Hierarchical (or multilevel) models constitute a large literature that includes other perspectives from which to view these same models. Much of that work is summarized in Goldstein (1995) and Bryk and Raudenbush (1992), who take the multilevel and/or hierarchical modelling perspective, Carlin and Louis (1996), who take the empirical Bayes perspective, Longford (1993) for random-coefficient or random-effects models, and West and Harrison (1997), from the Bayesian perspective.

Faster computing now makes fitting such hierarchical models feasible, and these models are being used increasingly in ever more general forms for applications. Commercial software for fitting these models includes HLM (Bryk *et al.*, 1996) and MLwiN (Hox (1998) or see `http://www.ioe.ac.uk/multilevel/`). Available freeware includes the BUGS program (Gilks *et al.*, 1996) and other software reviewed in Normand (1995).

Meng and van Dyk (1998) have developed fast EM-type algorithms for computing maximum likelihood and restricted maximum likelihood (REML) estimates for a multivariate normal model similar to that of this paper. However, if the number of observations J is not large, a more fully Bayesian approach is usually necessary for good frequency operating characteristics, as demonstrated in Section 5, with J as large as 27. Prior distributions for the unknown hyperparameters complete our Bayesian model specification. We choose the distribution for $\boldsymbol{\gamma}$ to be flat on R^r, with both vague distributions and proper distributions for $\mathbf{A}$ introduced in Section 2. We believe that some of these choices of distributions provide good repeated sampling operating characteristics.

The unknown parameters are the $pq + Jp + p(p+1)/2$ elements of $\boldsymbol{\gamma}$, $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$ and $\mathbf{A}$. The special structure of the model defined by expressions (1) and (2) is exploited here to obtain independent samples from the exact distributions of $\mathbf{A}$, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, given $\mathbf{Y}$, in contrast with Gibbs sampling procedures that obtain dependent samples from distributions that converge to the exact posterior distributions (Gelfand and Smith, 1990).

Section 2 shows that the posterior density of $\mathbf{B}_0 = \mathbf{V}_0^{1/2}(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{V}_0^{1/2}$, a bounded transformation of $\mathbf{A}$, is exactly that of a constrained Wishart distribution in the special case of expressions (1) and (2) with all level 1 covariance matrices $\mathbf{V}_j$ equal to a common matrix $\mathbf{V}_0$. Section 3 shows further that, even when the $\mathbf{V}_j$ are unequal, a constrained Wishart density

envelops the posterior density of $\mathbf{B}_0$ given $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_J')'$, for an arbitrary choice of $\mathbf{V}_0$. This envelope density enables a rejection sampling procedure that accepts 100% of the time in the equal covariance case, and nearly 100% of the time when the $\mathbf{V}_j$ are nearly equal. Even in the $p = 2$ example of Section 4, where the $\mathbf{V}_j$ vary by a factor of more than 8, the acceptance rate is 89%.

The method of Section 3 provides an independent sample $\{\mathbf{A}_1, \ldots, \mathbf{A}_N\}$ from the distribution of $\mathbf{A}$ given $\mathbf{Y}$. These drawn values then provide independent samples from the joint distribution of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J, \boldsymbol{\gamma})$ given $\mathbf{Y}$ by sampling first from the conditional distribution of $\boldsymbol{\gamma}$ given $(\mathbf{Y}, \mathbf{A}_k)$ and then from the distribution of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$ given $(\mathbf{Y}, \boldsymbol{\gamma}, \mathbf{A}_k)$ for each value $\mathbf{A}_k$, $k = 1, \ldots, N$. The conditional distributions are displayed in expressions (14) and (15) of Appendix A.1. Both are normal distributions with known parameters when a uniform or multivariate normal prior distribution is specified for the level 2 regression coefficient $\boldsymbol{\gamma}$ (e.g. Bryk and Raudenbush (1992)). If functions of $\mathbf{A}$, such as the posterior means and covariances of the $\boldsymbol{\theta}_j$ and of $\boldsymbol{\gamma}$, are to be estimated, importance sampling can be used instead of rejection sampling to make use of every sampled constrained Wishart matrix.

Section 4 demonstrates the algorithm for the bivariate ($p = 2$) hospital profiling application described above, showing how closely the constrained Wishart density approximates the target posterior density in this unequal covariance case. Posterior variances for the $\boldsymbol{\theta}_j$ are compared with those computed by the widely used REML procedure. The REML procedure, although faster than simulation methods, reports optimistically small standard errors that result in undercoverage of interval estimates.

Section 5 reports frequentist coverage rates from two simulations based on the example in Section 4 where intervals with 90% coverage for the elements of the $\boldsymbol{\theta}_j$ are sought. The nominal 90% REML intervals have actual coverage rates of 87% and 81% for the $p = 2$ components of the $\boldsymbol{\theta}_j$-vectors in the example. Our independent sampling procedure (using a uniform prior distribution for $\mathbf{A}$) has actual coverage rates of 90% for both components. In a second example, the actual coverage rates are 86% and 81% for REML and 90% and 92% for the independent sampling procedure.

2. Equal covariance case and the constrained Wishart distribution

For the special case of model (1)–(2) having all level 1 covariance matrices $\mathbf{V}_j = \mathbf{V}_0$, the equal covariance case of Efron and Morris (1972), we can sample easily from the posterior distribution of $\mathbf{A}$ given $\mathbf{Y}$. Define the variable

$$\mathbf{B}_0 = \mathbf{V}_0^{1/2}(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{V}_0^{1/2},$$

$\mathbf{V}_0^{1/2}$ being the symmetric matrix square root of $\mathbf{V}_0$. Because $\mathbf{A}$ is non-negative definite, $\mathbf{B}_0$ has all eigenvalues in (0, 1], i.e. $\mathbf{0} < \mathbf{B}_0 \leqslant \mathbf{I}$.

The equal covariance marginal likelihood function for $\mathbf{B}_0$, derived in Appendix A.2, is

$$L_0(\mathbf{B}_0) \propto |\mathbf{B}_0|^{(J-q)/2} \operatorname{etr}(-\mathbf{B}_0\mathbf{S}/2), \qquad \mathbf{0} < \mathbf{B}_0 \leqslant \mathbf{I}, \tag{3}$$

where

$$\mathbf{S} \equiv \mathbf{V}_0^{-1/2}\left\{\sum (\mathbf{Y}_j - \mathbf{W}_j'\hat{\boldsymbol{\gamma}})(\mathbf{Y}_j - \mathbf{W}_j'\hat{\boldsymbol{\gamma}})'\right\}\mathbf{V}_0^{-1/2}, \tag{4}$$

and

$$\hat{\boldsymbol{\gamma}} \equiv \left(\sum \mathbf{W}_j\mathbf{W}_j'\right)^{-1} \sum \mathbf{W}_j\mathbf{Y}_j. \tag{5}$$

Here etr is defined as $\text{etr}(\mathbf{X}) \equiv \exp\{\text{tr}(\mathbf{X})\}$. Likelihood (3) has the form of a constrained Wishart density, defined as follows.

Definition 1. A symmetric $p \times p$ random matrix $\mathbf{X}$ has a 'constrained Wishart distribution', written $\mathbf{X} \sim \text{CWish}_p(\nu, \boldsymbol{\Sigma}; \mathbf{Q})$, of dimension p, degrees of freedom $\nu > 0$, symmetric $p \times p$ scale matrix $\boldsymbol{\Sigma}$ and diagonal constraint matrix $\mathbf{Q}$, if the density function $f(\mathbf{X})$ satisfies

$$f(\mathbf{X}) \propto |\mathbf{X}|^{(\nu-p-1)/2} \,\text{etr}(-\mathbf{X}\boldsymbol{\Sigma}^{-1}/2), \qquad \mathbf{0} < \mathbf{X} \leqslant \mathbf{Q}. \tag{6}$$

Here $\mathbf{0} < \mathbf{X} \leqslant \mathbf{Q}$ means $\mathbf{X} > \mathbf{0}$ and $\mathbf{Q} - \mathbf{X} \geqslant \mathbf{0}$. Note that $\lambda_{\min} = \infty$ gives the usual (unconstrained) Wishart distribution, $\lambda_{\min}$ being the minimum eigenvalue of $\mathbf{Q}$.

The conjugate prior distributions for likelihood (3) take the form

$$p(\mathbf{B}_0)d\mathbf{B}_0 \propto |\mathbf{B}_0|^{(\nu_*-p-1)/2} \,\text{etr}(-\mathbf{B}_0\mathbf{S}_*/2)d\mathbf{B}_0, \qquad \mathbf{0} < \mathbf{B}_0 \leqslant \mathbf{I}. \tag{7}$$

These are proper constrained Wishart densities, $\mathbf{B}_0 \sim \text{CWish}(\nu_*, \mathbf{S}_*^{-1}; \mathbf{I})$, if $\nu_* > 0$ and $\mathbf{S}_* > \mathbf{0}$. But expression (7) also allows for improper prior distributions such as the uniform distribution on $\mathbf{A} \geqslant \mathbf{0}$, considered in later sections. This follows because the Jacobian for the transformation from $\mathbf{A}$ to $\mathbf{B}_0$ is $|\mathbf{B}_0|^{-(p+1)}$, and by comparing this with expression (7) for $\nu_* = -(p+1)$ and $\mathbf{S}_* = \mathbf{0}$. Improper prior distributions are permitted only if they provide proper posterior distributions. This uniform distribution for $\mathbf{A}$ is satisfactory provided that $J > q + p + 1$, and, more generally, proper posterior distributions arise from the class (7) only if $J > q - \nu_*$.

Theorem 1 summarizes and establishes the likelihood result (3) and the conjugacy of the distributions (7) for the equal covariance case, with the proof deferred to Appendix A.2.

Theorem 1. Assume the two-level normal hierarchical model defined by expressions (1) and (2) with all $\mathbf{V}_j = \mathbf{V}_0$, a known $p \times p$ symmetric positive definite matrix. Let $\mathbf{B}_0 = \mathbf{V}_0^{1/2}(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{V}_0^{1/2}$, with $\mathbf{V}_0^{1/2} = (\mathbf{V}_0^{1/2})'$, have a (possibly improper) prior distribution of the form (7) for some $p \times p$ symmetric matrix $\mathbf{S}_* \geqslant \mathbf{0}$, and with $\nu_* > -(J - q)$. Then, given $\mathbf{Y}$,

$$\mathbf{B}_0|\mathbf{Y} \sim \text{CWish}_p\{J - q + \nu_*, (\mathbf{S} + \mathbf{S}_*)^{-1}; \mathbf{I}\},$$

with the statistic $\mathbf{S}$ as in expression (4).

Theorem 1 states that, in the equal covariance case, we can sample independently from the distribution of $\mathbf{B}_0$ given $\mathbf{Y}$ by drawing independent Wishart matrices (e.g. by the method of Odell and Feiveson (1966)) and keeping only those matrices with maximum eigenvalue at most 1. A more efficient procedure for simulating from a constrained Wishart distribution appears in Everson and Morris (2000). An independent sample from the distribution of $\mathbf{A}$ given $\mathbf{Y}$ then is obtained through the inverse relationship $\mathbf{A} = \mathbf{V}_0^{1/2}(\mathbf{B}_0^{-1} - \mathbf{I})\mathbf{V}_0^{1/2}$, facilitating inferences about the unknown parameters $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and $\mathbf{A}$.

3. Unequal covariance case

This section provides an approximation to the unequal covariance posterior density of a bounded transformation of $\mathbf{A}$ and shows that this approximating density can be chosen to envelop the posterior density. The approximating density therefore can be used with rejection sampling procedures to draw independent samples from the distribution of $\mathbf{A}$ given $\mathbf{Y}$.

Let $\mathbf{V}_0$ be any symmetric positive definite $p \times p$ matrix and define $\mathbf{B}_0 = \mathbf{V}_0^{1/2}(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{V}_0^{1/2}$. Theorem 2, with the proof deferred to Appendix A.3, shows that, with an appropriate choice of degrees of freedom ν, a constrained Wishart density 'envelops' the posterior density $f(\mathbf{B}_0|\mathbf{Y})$. A density function f_0 is said to envelop a density function f_1 if the ratio $f_1(x)/f_0(x)$ is bounded (e.g. Ripley (1987)).

Theorem 2. For a given symmetric $p \times p$ matrix $\mathbf{V}_0 > \mathbf{0}$, $f(\mathbf{B}_0|\mathbf{Y})$ is the posterior density function for $\mathbf{B}_0 = \mathbf{V}_0^{1/2}(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{V}_0^{1/2}$, assuming some prior density $p(\mathbf{B}_0)$ of the form (7) with $\nu_* > -(J-q)$. Let f_0 be the density function for any CWish(ν, $\mathbf{\Sigma}$; $\mathbf{I}$) distribution. Then f_0 envelops $f(\mathbf{B}_0|\mathbf{Y})$ for any $\mathbf{\Sigma} > \mathbf{0}$ and for any $\nu \leqslant J - q + \nu_*$.

Theorem 2 establishes the existence of a finite constant M that bounds the ratio $f(\mathbf{B}_0|\mathbf{Y})/f_0(\mathbf{B}_0)$. This validates rejection procedures by using a constrained Wishart density as an envelope for $f(\mathbf{B}_0|\mathbf{Y})$, with acceptance rates or mean importance weights equalling $1/M$. In practice we choose $\nu = J - q + \nu_*$ as the degrees of freedom, yielding the least dispersed constrained Wishart density that envelops $f(\mathbf{B}_0|\mathbf{Y})$. If all the $\mathbf{V}_j$ are equal to $\mathbf{V}_0$ and we choose $\nu = J - q + \nu_*$, then $f_0(\mathbf{B}_0) = f(\mathbf{B}_0|\mathbf{Y})$ (by theorem 1) so $M = 1$. For unequal $\mathbf{V}_j$ we choose $\mathbf{V}_0$ to represent a 'typical' $\mathbf{V}_j$, such as their arithmetic mean or harmonic mean, in an attempt to mimic the equal covariance case. With unequal $\mathbf{V}_j$ we have $M > 1$, with M increasing for larger differences in the $\mathbf{V}_j$. In the bivariate example treated in Section 4 with a uniform prior distribution assumed for $\mathbf{A}$, the estimated value of M is 1.12, for an acceptance rate over 89%, even though the ratio of the maximum to minimum $\mathbf{V}_j$ exceeds 8.

Theorem 2 implies that we are free to choose the parameter $\mathbf{\Sigma} > \mathbf{0}$ to improve the fit of our envelope density, because the enveloping property of the truncated Wishart density depends only on ν. The mode of a Wishart$_p(\nu, \mathbf{\Sigma})$ density is $(\nu - p - 1)\mathbf{\Sigma}$. Choosing $\mathbf{\Sigma} = \hat{\mathbf{B}}_0/(\nu - p - 1)$, with $\hat{\mathbf{B}}_0$ the posterior mode of $\mathbf{B}_0$, produces an envelope density f_0 with the same mode as $f(\mathbf{B}_0|\mathbf{Y})$. Dempster *et al.* (1981) described an EM algorithm to locate the REML estimate $\hat{\mathbf{A}}$ in this problem. If we specify a uniform prior distribution for $\mathbf{B}_0$, then $\hat{\mathbf{B}}_0 = \mathbf{V}_0^{1/2}(\mathbf{V}_0 + \hat{\mathbf{A}})\mathbf{V}_0^{1/2}$ would maximize $f(\mathbf{B}_0|\mathbf{Y})$. Section 4.4 of Everson (1995) describes an adjusted EM algorithm to locate the posterior mode of $\mathbf{B}_0$ for the more general prior specification (7), a special case of which is the uniform prior distribution for $\mathbf{A}$ used in the application in Section 4 and in the simulations in Section 5.

The parameters ν and $\mathbf{\Sigma}$ define a truncated Wishart envelope density $f_0(\mathbf{B}_0)$. The algorithm proposed samples Wishart matrices $\mathbf{B}_{01}, \ldots, \mathbf{B}_{0N}$ with these parameters and with all eigenvalues constrained to be less than 1. Rejection sampling accepts each constrained Wishart matrix $\mathbf{B}_{0i}$ with probability α_i/m, yielding a sample from the distribution of $\mathbf{B}_0|\mathbf{Y}$ and hence from that of $\mathbf{A}$ given $\mathbf{Y}$, with $\mathbf{A}_i = \mathbf{V}_0^{1/2}(\mathbf{B}_{0i}^{-1} - \mathbf{I})\mathbf{V}_0^{1/2}$. Here

$$\alpha_i = c f(\mathbf{B}_{0i}|\mathbf{Y})/f_0(\mathbf{B}_{0i}),$$

the ratio of the target density to the envelope density at $\mathbf{B}_{0i}$ calculated up to a constant c, and $m = cM$, where

$$M = \sup\{f(\mathbf{B}_0|\mathbf{Y})/f_0(\mathbf{B}_0)\} \geqslant 1.$$

To find m we must consider all possible values $\mathbf{B}_0$, not just those in our sample.

Alternatively, sampling–importance resampling (Rubin, 1987) resamples from among the $\mathbf{B}_{0i}$ with sampling weights proportional to the α_i, yielding an independent sample from approximately the distribution desired. The advantage of the sampling–importance resampling

algorithm is that m need not be evaluated. A much larger initial sample size N is required for reliable results, however.

If the goal of inference is merely to estimate functions of the unknown parameters, such as their posterior means and covariances, importance sampling is an alternative that does not require that m be evaluated. Importance sampling makes use of all sampled constrained Wishart matrices to estimate functions of $\mathbf{A}$. For example, we estimate the posterior mean $\boldsymbol{\gamma}^*$ of the level 2 regression coefficient $\boldsymbol{\gamma}$ by averaging over $\boldsymbol{\gamma}^*_{(\mathbf{A}_i)} = E(\boldsymbol{\gamma}|\mathbf{Y}, \mathbf{A}_i)$ with importance weights α_i:

$$\boldsymbol{\gamma}^* = E(\boldsymbol{\gamma}|\mathbf{Y}) = E(\boldsymbol{\gamma}^*_{(\mathbf{A})}|\mathbf{Y}) \doteq \sum_{i=1}^{N} (\alpha_i \boldsymbol{\gamma}^*_{(\mathbf{A}_i)}) \Big/ \sum_{i=1}^{N} \alpha_i. \tag{8}$$

To estimate the posterior covariance of $\boldsymbol{\gamma}$ we average over $D^*_{\boldsymbol{\gamma}(\mathbf{A}_i)} = \mathrm{cov}(\boldsymbol{\gamma}|\mathbf{Y}, \mathbf{A}_i)$ and add the covariance of the means conditional on each $\mathbf{A}_i$:

$$\begin{aligned} \mathbf{D}^*_{\boldsymbol{\gamma}} &= \mathrm{cov}(\boldsymbol{\gamma}|\mathbf{Y}) = E(\mathbf{D}^*_{\boldsymbol{\gamma}(\mathbf{A})}|\mathbf{Y}) + \mathrm{cov}(\boldsymbol{\gamma}^*_{(\mathbf{A})}|\mathbf{Y}) \\ &\doteq \sum_{i=1}^{N} \alpha_i \{\mathbf{D}^*_{\boldsymbol{\gamma}(\mathbf{A}_i)} + (\boldsymbol{\gamma}^*_{(\mathbf{A}_i)} - \boldsymbol{\gamma}^*)(\boldsymbol{\gamma}^*_{(\mathbf{A}_i)} - \boldsymbol{\gamma}^*)'\} \Big/ \sum_{i=1}^{N} \alpha_i. \end{aligned} \tag{9}$$

An analogous procedure estimates $\boldsymbol{\theta}^*_j = E(\boldsymbol{\theta}_j|\mathbf{Y})$ and $\mathbf{V}^*_j = \mathrm{cov}(\boldsymbol{\theta}_j|\mathbf{Y})$, $j = 1, \ldots, J$. Expressions for $\boldsymbol{\gamma}^*_{(\mathbf{A})}$, $\mathbf{D}^*_{\boldsymbol{\gamma}(\mathbf{A})}$, $\boldsymbol{\theta}^*_{j(\mathbf{A})}$ and $\mathbf{V}^*_{j(\mathbf{A})}$ appear in expressions (14) and (16) in Appendix A.1. Although it is not necessary to locate the maximum density ratio and to calculate m to use importance sampling, it is still useful to do so. The average $\bar{\alpha}/m$ is unbiased for the mean importance weight $1/M$ and suggests a sample size N that will provide a sufficiently large effective sample size N/M.

4. Hospital profiling application

The data of Table 1 are developed from interviews of 1869 adult surgery patients from 27 hospitals (Cleary *et al.*, 1991). These are the 27 teaching hospitals and private academic

Table 1. Problem rate data†

Hospital j	*NSrg* Y_{1j}	*Srg* Y_{2j}	*Severity* w_j	*Size* n_j	*Hospital j*	*NSrg* Y_{1j}	*Srg* Y_{2j}	*Severity* w_j	*Size* n_j
1	10.18	15.06	0.75	24	15	15.80	11.50	0.26	61
2	11.55	17.97	0.62	32	16	14.81	20.56	0.56	62
3	16.21	12.50	0.66	32	17	11.14	13.02	0.02	62
4	12.31	14.88	0.26	43	18	17.12	14.60	0.41	66
5	12.88	15.21	0.96	44	19	16.93	16.28	0.56	68
6	11.84	17.69	0.44	45	20	11.02	13.52	0.34	68
7	14.82	16.96	0.44	48	21	14.69	16.49	0.56	72
8	13.05	15.07	0.55	49	22	10.48	14.24	0.79	77
9	12.43	12.01	0.33	51	23	15.82	15.13	0.47	87
10	8.35	9.43	0.47	53	24	12.66	14.99	0.71	122
11	17.97	26.82	0.48	56	25	10.41	17.25	0.45	124
12	11.84	15.64	0.34	58	26	10.32	10.13	0.05	149
13	12.43	13.94	0.28	58	27	13.72	18.18	0.77	198
14	14.73	15.40	0.63	60	Average	13.07	15.35	0.49	69.2

†Patient-reported problem rates for surgical patients in 27 hospitals: Y_{1j} and Y_{2j} represent average non-surgical and surgical problem rates for the n_j patients in hospital j. Larger values of Y indicate greater frequencies of problems. The hospital severity index is given by w_j, ranging from 0.016 to 0.955, with larger values of w_j indicating sicker patients on average. The observations are sorted on n_j.

health centres sampled having at least 20 surveyed surgery patients, the actual sample sizes n_j ranging from 24 to 198. Each patient indicated whether he or she had experienced each of a number of potential problems, classified as non-surgical (NSrg) or surgical (Srg) in nature. With reasonably large within-hospital sample sizes n_j, the averages of the vectors of non-surgical percentages and surgical percentages over the n_j patients in each hospital follow an approximate bivariate normal distribution, even without transforming the proportions. The sample size restriction may introduce a sampling bias, but we choose to ignore this and other possible problems. This example is intended primarily to provide a realistic setting in which to demonstrate and evaluate the procedure of Section 3.

The value Y_{1j} is the average percentage of non-surgical (NSrg) problems and Y_{2j} the average percentage for surgical (Srg) problems reported by the n_j patients in hospital j. All 1869 patient outcomes are used to estimate a common 2×2 within-hospital covariance matrix $\mathbf{\Sigma}_0$, so that $\mathbf{V}_j = \mathbf{\Sigma}_0 / n_j$:

$$\mathbf{\Sigma}_0 = \begin{pmatrix} 148.87 & 140.43 \\ 140.43 & 490.60 \end{pmatrix} \doteq \begin{pmatrix} (12.20)^2 & (0.52)(12.2)(22.15) \\ (0.52)(12.2)(22.15) & (22.15)^2 \end{pmatrix}. \tag{10}$$

The $p = 2$ vector $\mathbf{Y}_j = (Y_{1j},\ Y_{2j})'$ is assumed to follow the level 1 model (1), $j = 1, \ldots, 27$, with the $\mathbf{V}_j$ treated as known owing to the large sample used to estimate $\mathbf{\Sigma}_0$.

The severity w_j measures the average health index for the n_j surveyed patients in hospital j, higher values indicating worse average health. The mean values $\boldsymbol{\theta}_j$ follow the level 2 distribution (2) with $E(\boldsymbol{\theta}_j) = (E(\theta_{1j}),\ E(\theta_{2j}))' = (\gamma_{10} + \gamma_{11} w_j,\ \gamma_{20} + \gamma_{21} w_j)' = \mathbf{W}_j' \boldsymbol{\gamma}$. Thus the 4×2 level 2 covariate matrix is $\mathbf{W}_j = \mathbf{I} \otimes \mathbf{w}_j$, where $\mathbf{w}_j = (1,\ w_j)'$, and $\mathbf{I}$ is the 2×2 identity matrix. The unknown 4×1 level 2 regression coefficient is $\boldsymbol{\gamma} = (\gamma_{10},\ \gamma_{11},\ \gamma_{20},\ \gamma_{21})'$. Hospitals are to be compared, the better hospitals having smaller values of $\boldsymbol{\theta}_j = (\theta_{1j},\ \theta_{2j})'$. We make inferences about the unknown $\{\boldsymbol{\theta}_j\}$, $\boldsymbol{\gamma}$ and $\mathbf{A}$, on the basis of the observed values $\{\mathbf{Y}_j,\ \mathbf{V}_j,\ \mathbf{w}_j\}$.

Implementation of the procedure of Section 3 requires a value $\mathbf{V}_0$ to define the sampling variable $\mathbf{B}_0 = \mathbf{V}_0^{1/2} (\mathbf{V}_0 + \mathbf{A})^{-1} \mathbf{V}_0^{1/2}$. Here we choose $\mathbf{V}_0$ as the average of the $\mathbf{V}_j$ to represent a 'typical' $\mathbf{V}_j$, approximating the equal covariance case of Section 2. Other choices are possible, and determining an optimal choice for $\mathbf{V}_0$ is a topic of continuing research.

Specifying the uniform prior distribution for $\mathbf{A}$ implies the prior density

$$p(\mathbf{B}_0) d\mathbf{B}_0 \propto |\mathbf{B}_0|^{-(p+1)} d\mathbf{B}_0$$

for $\mathbf{B}_0$, equivalent to $\nu_* = -(p+1) = -3$, and $\mathbf{S}_* = \mathbf{0}$ in expression (7). For this prior distribution, theorem 2 states that the CWish(ν, $\mathbf{S}_0^{-1}$; $\mathbf{I}$) density will envelop the density $f(\mathbf{B}_0|\mathbf{Y})$ for any choice of $\mathbf{V}_0$, provided that $\nu \leqslant 27 - 2 - 3 = 22$. We choose $\nu = 22$ for the least dispersed enveloping constrained Wishart density and select $\mathbf{S}_0^{-1} = \hat{\mathbf{B}}_0 / (22 - 2 - 1)$, where $\hat{\mathbf{B}}_0$ is the mode of $f(\mathbf{B}_0|\mathbf{Y})$, displayed in equation (11). This defines an envelope density $f_0(\mathbf{B}_0)$ with the same mode as $f(\mathbf{B}_0|\mathbf{Y})$:

$$\hat{\mathbf{B}}_0 = \begin{pmatrix} 0.29 & 0.03 \\ 0.03 & 0.64 \end{pmatrix}. \tag{11}$$

Fig. 1 shows how closely the constrained Wishart density $f_0(\mathbf{B}_0)$ approximates the true posterior density $f(\mathbf{B}_0|\mathbf{Y})$ in this problem by plotting the two curves for each of the three elements of $\mathbf{B}_0$ with the other two elements fixed at their values from the posterior mode (11). The fit for the joint density of the three elements is not quite as precise as for these profile plots, but the value $M = 1.12$ is sufficiently large that $M f_0(\mathbf{B}_0) \geqslant f(\mathbf{B}_0|\mathbf{Y}), \mathbf{0} < \mathbf{B}_0 \leqslant \mathbf{I}$, which implies an acceptance rate $1/M \doteq 89\%$ when performing rejection sampling using f_0 as an envelope density.

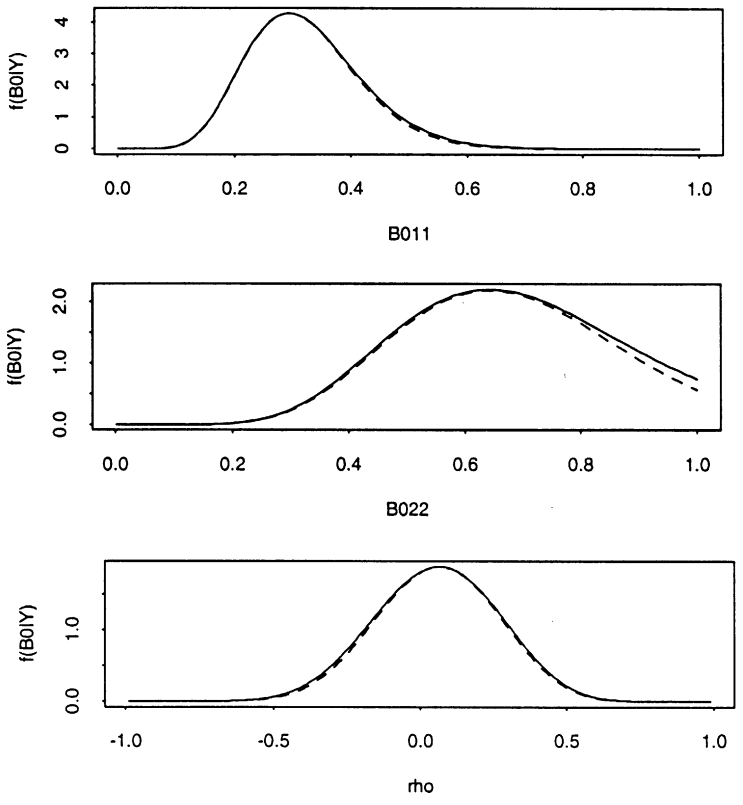

Fig. 1. Profile envelope plots of the elements of $\mathbf{B}_0$ for the bivariate problem rate example: the posterior density $f(\mathbf{B}_0|\mathbf{Y})$ (– – –) and a constrained Wishart density inflated by $M = 1.12$ to envelop $f(\mathbf{B}_0|\mathbf{Y})$ (———) are plotted as functions of each element $\mathbf{B}_0$ (B_{011}, B_{022} and $\rho = B_{012}(B_{011}B_{022})^{-1/2}$) with one element allowed to vary while the other two are held fixed at their values from the posterior mode $\hat{\mathbf{B}}_0$ of equation (11)

The remainder of Section 4 compares REML inferences with those based on the independent sampling procedure described in Section 3 for the hospital problem rate data of Table 1. The REML procedure estimates the posterior distributions of the $\boldsymbol{\theta}_j$ and of γ by the conditional distributions of $\boldsymbol{\theta}_j$ given $(\mathbf{Y}, \hat{\mathbf{A}})$, and of γ given $(\mathbf{Y}, \hat{\mathbf{A}})$, where $\hat{\mathbf{A}}$ is the marginal maximum likelihood estimate for the level 2 covariance matrix $\mathbf{A}$ (e.g. Bryk and Raudenbush (1992)). Although the two methods produce similar mean estimates for the $\boldsymbol{\theta}_j$, the REML variance estimates do not account for the uncertainty about $\mathbf{A}$ and hence are optimistically small. The procedure of Section 3 averages the conditional distributions given $\mathbf{A}$ over a large sample from the distribution of $\mathbf{A}$ given $\mathbf{Y}$ to account for this uncertainty. The REML standard deviation estimates are smaller than those for the sampling procedure for both elements of $\boldsymbol{\theta}_j$ in all 27 hospitals, as displayed in Fig. 2, averaging 13% smaller variances for the non-surgical rates and 39% smaller variances for the surgical rates. REML also reports smaller variance estimates of the four estimated components of γ, with 27% smaller variances for the non-surgical intercept and slope components and 33% smaller variances for the surgical components. The simulations of Section 5 demonstrate that the smaller REML uncertainty estimates produce a substantial undercoverage of interval estimates of the elements of the $\boldsymbol{\theta}_j$.

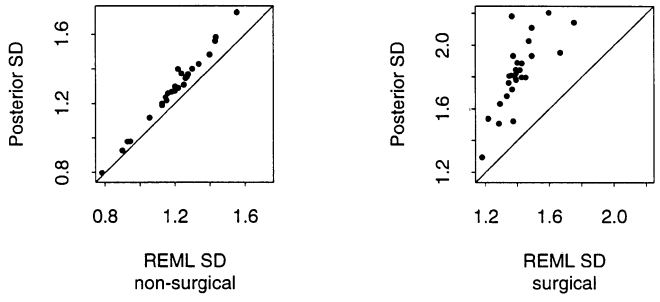

Fig. 2. Uncertainty estimates—comparisons of posterior standard deviations estimated by the independent sampling procedures with those reported by REML for estimates of mean non-surgical and surgical problem rates in 27 hospitals: the sampling estimates assume the uniform prior distribution on the between-hospitals covariance matrix **A**; all points lie above the line $y = x$ in both plots, indicating that REML reports smaller variability estimates for both components of all 27 θ_j

5. Absolute and relative performance of procedures

Sections 2 and 3 provide fast ways to compute the posterior distribution for certain prior distributions on $(\gamma, \mathbf{A})$. We believe that some of these prior specifications, particularly that with $\mathbf{A}$ uniformly distributed, provide point and interval estimates that perform well in repeated sampling for every level 2 distribution in the class defined by expression (2), i.e. for each and every value of $(\gamma, \mathbf{A})$. For a given loss function, the 'classical' risk evaluations use averages of the loss function over $\mathbf{Y}$ in expression (1) with $\boldsymbol{\theta}$ fixed. For example, the James–Stein estimator (James and Stein, 1961) is minimax for squared error loss in this sense. The performance characteristics of this section make a different risk calculation, one more appropriate for two-level models (Morris, 1983a). 'Two-level risk' evaluations average over the joint distribution of the level 1 data $\mathbf{Y}$ in expression (1) and over the level 2 parameter $\boldsymbol{\theta}$ in expression (2), for fixed parameters $(\gamma, \mathbf{A})$, so the two-level risk is a function of $(\gamma, \mathbf{A})$.

Very little is known about coverages of interval estimators in the two-level setting with J fixed. In the equal variance case for $p = 1$ and with $J \geqslant 4$, Casella and Hwang developed J-dimensional confidence spheres, based on the James–Stein estimator, with smaller volumes than the usual confidence sphere, and that cover in the classical sense (and thus in the two-level sense as well) the vector $\boldsymbol{\theta}$ more often than does the usual confidence sphere (Hwang and Casella, 1982; Casella and Hwang, 1983). Morris (1983b) gave mathematical arguments in the equal variance $p = 1$ case to suggest that for all $(\gamma, \mathbf{A})$ the Bayes intervals of this paper (flat $\mathbf{A}$), even though they are narrower, cover the components $\boldsymbol{\theta}_j$ as often as do the usual classical confidence intervals. We shall not cite the substantial corresponding literature on two-level asymptotics as we are primarily concerned with the performance of procedures for small and moderate J.

For J fixed and with $p \geqslant 2$, or even with $p = 1$ and unequal variances, we know of no published mathematical results about the two-level coverages of interval estimates. Because of the difficulty of such work, even for $p = 1$, we use simulations here, taking the $J = 27$ and $p = 2$ hospital example in Section 4 to provide a meaningful context. The favourable results that follow typify kindred results obtained in several other settings, e.g. in other simulations we have investigated with $p = 1$, and in the simulations of Christiansen and Morris (1997) for the Poisson case. No contrary results have come to our attention.

Table 2 summarizes two simulations, each fixing $\mathbf{V}$ and $\mathbf{W}$ at values from Table 1. The first simulation sets $\mathbf{A} = \mathbf{A}_1$ of equations (12), the estimated posterior mean of $\mathbf{A}$ based on the

data of Table 1. The second sets $\mathbf{A} = \mathbf{A}_2$ of equations (12), a value chosen to make the matrix $\mathbf{B}_0 = \mathbf{V}_0^{1/2}(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{V}_0^{1/2}$ further from diagonal:

$$\mathbf{A}_1 = \begin{pmatrix} 5.80 & 2.63 \\ 2.63 & 6.10 \end{pmatrix},$$

$$\mathbf{A}_2 = \begin{pmatrix} 3.38 & -0.77 \\ -0.77 & 2.55 \end{pmatrix}. \tag{12}$$

The choice of γ is irrelevant and arbitrarily set equal to $\mathbf{0}$ in both simulations. Each simulation generated 100 sets of $J = 27$ $\boldsymbol{\theta}_j$-vectors and $\mathbf{Y}_j$-vectors distributed according to expressions (1) and (2) to represent true and observed non-surgical and surgical problem rates. The 90% and 95% probability intervals for the elements θ_{1j} and θ_{2j} are approximated as the posterior mean estimate for each element plus or minus 1.645 or 1.960 posterior standard deviation estimates. Although only approximating the Bayesian intervals, this procedure mimics that used to construct the REML intervals for Table 2. A more precise Bayesian interval for $\boldsymbol{\theta}_j$ could be constructed by simulating a large sample from $\boldsymbol{\theta}_j|\mathbf{Y}$, and then finding the smallest interval that contains 90% or 95% of the simulated values.

The entries in Table 2 are non-coverage rates for the approximate Bayesian and REML (Section 4) interval estimates. Non-coverage for each of the $p = 2$ components of $\boldsymbol{\theta}_j$ is calculated as the overall percentage of 2700 intervals (100 simulations and $J = 27$) that do not contain their (simulated) target values. In summary, the independent sampling procedure with $\mathbf{A}$ uniformly distributed ('flat $\mathbf{A}$') produced intervals that yielded very nearly the desired frequentist coverage rates for three of the four cases, erring slightly on the conservative side in the fourth case, with a higher than nominal coverage rate. Although a Bayesian interval has the correct coverage (e.g. 90%) by construction when $\mathbf{A}$ is distributed according to the specified prior distribution, it would be remarkable if this coverage rate would hold for every individual value of $\mathbf{A}$, such as those used to generate Table 2. Thus our goal in choosing a prior distribution is to err on the conservative side when the coverage deviates from nominal. In contrast, the REML intervals are optimistically narrow, resulting in rather severe undercoverage in all four cases considered.

Using the observed rates $\mathbf{Y}_j$ and covariances $\mathbf{V}_j$ to construct classical confidence intervals produces appropriate coverages, as expected, but of course these intervals are much wider than the Bayesian intervals. The classical intervals are 25% wider for estimating the non-surgical (θ_{1j}) components of $\boldsymbol{\theta}_j$ and 51% wider for estimating the surgical (θ_{2j}) components. In

Table 2. Coverage comparisons†

		Results (%) for a nominal 10% error rate		*Results (%) for a nominal 5% error rate*	
		Flat **A**	*REML*	*Flat* **A**	*REML*
$\mathbf{A}_1$	NSrg	10.5 (0.2)	12.8 (0.5)	5.4 (0.2)	7.1 (0.4)
	Srg	9.9 (0.3)	19.0 (1.3)	5.0 (0.2)	12.8 (1.2)
$\mathbf{A}_2$	NSrg	9.9 (0.3)	13.8 (0.7)	5.0 (0.2)	8.1 (0.6)
	Srg	7.8 (0.3)	18.9 (1.2)	3.7 (0.2)	12.8 (1.1)
Average		9.5	16.1	4.8	10.2

†Average non-coverage rates (with simulation standard errors in parentheses) for interval estimates for the two elements (NSrg and Srg) of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{27}$ using the proposed sampling method assuming the uniform prior distribution on $\mathbf{A}$ (flat $\mathbf{A}$) and using REML. The results are based on 100 simulated data sets each for the two values $\mathbf{A}_1$ and $\mathbf{A}_2$ of equation (12).

addition, the weighted squared error loss, $\Sigma\,(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)'\mathbf{V}_j^{-1}(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)$, summed over all simulated data sets, is 43% smaller using the posterior mean estimates from the independent sampling procedure than when using $\mathbf{Y}_j$ to estimate $\boldsymbol{\theta}_j$.

6. Conclusions

The algorithm of Section 3 samples from the marginal distribution of the unknown $p \times p$ covariance matrix $\mathbf{A}$ of the group mean vectors $\boldsymbol{\theta}_j$, given $\mathbf{Y}$. The simple-to-use S-PLUS program `TLNise` ('two-level normal independent sampling estimation') implements the algorithm of this paper and is available at

`www.swarthmore.edu/NatSci/peverso1/tlnise.htm`

Our program, which has performed well in all the examples that we have considered, is potentially usable in a broad range of statistical applications.

Although the procedure described here works with a continuum of proper and improper distributions on the covariance matrix $\mathbf{A}$, we have emphasized the flat distribution on $\mathbf{A}$ particularly because we believe that the resulting inference procedures have good, or at least conservative, frequency properties in small and moderate samples. For example, the simulations in Section 5 show that the interval estimates of the Jp components of the group effect vectors $\boldsymbol{\theta}_j$ that ensue from this flat distribution on $\mathbf{A}$ have coverages that are close to their nominal values. In contrast, competing methods like methods based on maximum likelihood or on REML, as adopted for most commercial software, provide inadequate coverage if the number of groups J is not large.

Very flexible inferences, involving complicated functions of the unknown parameters, are possible from this and from Markov chain Monte Carlo (MCMC) methods because they sample from the posterior distributions of the unknowns. However, MCMC methods cannot compete with the relative speed of importance sampling methods that accept with high rates, as does the method developed here, because MCMC methods provide correlated draws, and because the usual worries about convergence attend their use. Our program's speed makes it relatively easy to evaluate operating characteristics with each use, a check that needs to be done much more often by those who develop or use multilevel models.

We have provided no numerical results here when p exceeds two dimensions. Our experience thus far with the procedures of this paper for p up to 5 suggests that over 40% of all constrained Wishart draws often will be accepted, allowing relatively fast computation. Of course high acceptance rates are guaranteed, even for high dimensions p, if the covariance matrices $\mathbf{V}_j$ are sufficiently similar, because in the limiting equal covariance case the constrained Wishart envelope distribution is precisely the posterior distribution. This procedure would still be of value in problems with lower acceptance rates, because even a single draw from the exact posterior distribution is sufficient to replace the 'burn-in' for an MCMC algorithm, removing any doubts about convergence. Additional independent exact draws will be valuable as starting-points for parallel MCMC sequences, also in equilibrium. The best solution is to obtain independent posterior samples, a goal which may be achievable in a broader class of problems.

Acknowledgements

The work of both authors was supported by Agency for Health Care Policy Research grant 1RO1HS0711801-01 at Harvard University. The second author is grateful for support

provided by the Center for Advanced Study of Behavioral Sciences at Stanford University 1993–1994, on National Science Foundation (NSF) grant SES-9022192, and also for NSF grants DMS-89-11562 and DMS-9705156. Both authors wish to thank the referees, who offered many helpful and insightful comments.

Appendix A: Proofs and details

A.1. Likelihood and conditional posterior distributions

Assuming the two-level model defined by expressions (1) and (2), the marginal likelihood function for the $p \times p$ level 2 covariance matrix $\mathbf{A}$ is

$$L(\mathbf{A}) = \left|\sum \mathbf{W}_j(\mathbf{V}_j + \mathbf{A})^{-1}\mathbf{W}_j'\right|^{-1/2} \left|\prod (\mathbf{V}_j + \mathbf{A})\right|^{-1/2} \exp\left\{-\sum (\mathbf{Y}_j - \mathbf{W}_j'\boldsymbol{\gamma}^*_{(\mathbf{A})})'(\mathbf{V}_j + \mathbf{A})^{-1}(\mathbf{Y}_j - \mathbf{W}_j\boldsymbol{\gamma}^*_{(\mathbf{A})})/2\right\}. \tag{13}$$

The distributions (14)–(16) assume an improper uniform prior distribution for the $r \times 1$ level 2 regression coefficient $\boldsymbol{\gamma}$:

$$\begin{aligned} \boldsymbol{\gamma}|\mathbf{Y}, \mathbf{A} &\sim N_r(\boldsymbol{\gamma}^*_{(\mathbf{A})}, \mathbf{D}^*_{\gamma(\mathbf{A})}), \\ \boldsymbol{\gamma}^*_{(\mathbf{A})} &= \left\{\sum \mathbf{W}_j(\mathbf{V}_j + \mathbf{A})^{-1}\mathbf{W}_j'\right\}^{-1} \sum \mathbf{W}_j(\mathbf{V}_j + \mathbf{A})^{-1}\mathbf{Y}_j, \\ \mathbf{D}^*_{\gamma(\mathbf{A})} &= \left\{\sum \mathbf{W}_j(\mathbf{V}_j + \mathbf{A})^{-1}\mathbf{W}_j'\right\}^{-1}; \end{aligned} \tag{14}$$

$$\begin{aligned} \boldsymbol{\theta}_j|\mathbf{Y}, \boldsymbol{\gamma}, \mathbf{A} &\sim N_p(\boldsymbol{\theta}^*_{j(\boldsymbol{\gamma},\mathbf{A})}, \mathbf{V}^*_{j(\boldsymbol{\gamma},\mathbf{A})}), \\ \boldsymbol{\theta}^*_{j(\boldsymbol{\gamma},\mathbf{A})} &= (\mathbf{I} - \mathbf{B}_j)\mathbf{Y}_j + \mathbf{B}_j\mathbf{W}_j'\boldsymbol{\gamma}, \\ \mathbf{V}^*_{j(\boldsymbol{\gamma},\mathbf{A})} &= (\mathbf{I} - \mathbf{B}_j)\mathbf{V}_j, \\ \mathbf{B}_j &= \mathbf{V}_j(\mathbf{V}_j + \mathbf{A})^{-1}; \end{aligned} \tag{15}$$

$$\begin{aligned} \boldsymbol{\theta}_j|\mathbf{Y}, \mathbf{A} &\sim N_p(\boldsymbol{\theta}^*_{j(\mathbf{A})}, \mathbf{V}^*_{j(\mathbf{A})}), \\ \boldsymbol{\theta}^*_{j(\mathbf{A})} &= (\mathbf{I} - \mathbf{B}_j)\mathbf{Y}_j + \mathbf{B}_j\mathbf{W}_j'\boldsymbol{\gamma}^*_{(\mathbf{A})}, \\ \mathbf{V}^*_{j(\mathbf{A})} &= (\mathbf{I} - \mathbf{B}_j)\mathbf{V}_j + \mathbf{B}_j\mathbf{W}_j'\mathbf{D}^*_{\gamma(\mathbf{A})}\mathbf{W}_j\mathbf{B}_j'. \end{aligned} \tag{16}$$

Also see Dempster *et al.* (1981).

A.2. Proof of theorem 1

Because any matrix is expressible as a Kronecker product of itself with the 1×1 matrix $\mathbf{1}$ (e.g. $(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{Y}_j = (\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{Y}_j \otimes \mathbf{1}$) and because the $pq \times p$ covariate matrix $\mathbf{W}_j$ satisfies $\mathbf{W}_j = \mathbf{I}_p \otimes \mathbf{w}_j$,

$$\begin{aligned} \sum \mathbf{W}_j(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{Y}_j &= \sum (\mathbf{I} \otimes \mathbf{w}_j)\{(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{Y}_j \otimes \mathbf{1}\} \\ &= \sum \{(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{Y}_j \otimes \mathbf{w}_j\}. \end{aligned} \tag{17}$$

Similarly,

$$\begin{aligned} \sum \mathbf{W}_j(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{W}_j' &= \sum (\mathbf{I} \otimes \mathbf{w}_j)\{(\mathbf{V}_0 + \mathbf{A})^{-1} \otimes \mathbf{1}\}(\mathbf{I} \otimes \mathbf{w}_j') \\ &= \sum \{(\mathbf{V}_0 + \mathbf{A})^{-1} \otimes \mathbf{w}_j\mathbf{w}_j'\} \\ &= (\mathbf{V}_0 + \mathbf{A})^{-1} \otimes \sum \mathbf{w}_j\mathbf{w}_j'. \end{aligned} \tag{18}$$

And, because the matrices $(\mathbf{V}_0 + \mathbf{A})^{-1}$ and $\Sigma\, \mathbf{w}_j\mathbf{w}_j'$ are $p \times p$ and $q \times q$ respectively,

$$\left|\sum \mathbf{W}_j(\mathbf{V}_0 + \mathbf{A})^{-1}\mathbf{W}_j'\right| = |\mathbf{V}_0 + \mathbf{A}|^{-q}\left|\sum \mathbf{w}_j\mathbf{w}_j'\right|^p. \tag{19}$$

(See Searle (1982).)

With all $\mathbf{V}_j$ equal to $\mathbf{V}_0$, we show that the conditional mean $\boldsymbol{\gamma}^*_{(\mathbf{A})}$ from distribution (14) reduces to $\hat{\boldsymbol{\gamma}}$ in expression (3) by using equations (17) and (18) as follows:

$$\begin{aligned}
\boldsymbol{\gamma}^*_{0(\mathbf{A})} &= \left\{\sum \mathbf{W}_j(\mathbf{V}_0+\mathbf{A})^{-1}\mathbf{W}_j'\right\}^{-1} \sum \mathbf{W}_j(\mathbf{V}_0+\mathbf{A})^{-1}\mathbf{Y}_j \\
&= \left\{(\mathbf{V}_0+\mathbf{A})\otimes\left(\sum \mathbf{w}_j\mathbf{w}_j'\right)^{-1}\right\}\sum\{(\mathbf{V}_0+\mathbf{A})^{-1}\mathbf{Y}_j\otimes\mathbf{w}_j\} \\
&= \sum\left\{(\mathbf{V}_0+\mathbf{A})\otimes\left(\sum \mathbf{w}_j\mathbf{w}_j'\right)^{-1}\right\}\{(\mathbf{V}_0+\mathbf{A})^{-1}\mathbf{Y}_j\otimes\mathbf{w}_j\} \\
&= \sum\left\{\mathbf{Y}_j\otimes\left(\sum \mathbf{w}_j\mathbf{w}_j'\right)^{-1}\mathbf{w}_j\right\} \\
&= \sum\left\{\mathbf{I}\otimes\left(\sum \mathbf{w}_j\mathbf{w}_j'\right)^{-1}\right\}(\mathbf{Y}_j\otimes\mathbf{w}_j) \\
&= \left\{\mathbf{I}\otimes\left(\sum \mathbf{w}_j\mathbf{w}_j'\right)^{-1}\right\}\sum(\mathbf{Y}_j\otimes\mathbf{w}_j) \\
&= \left\{\sum(\mathbf{I}\otimes\mathbf{w}_j)(\mathbf{I}\otimes\mathbf{w}_j')\right\}^{-1}\sum(\mathbf{I}\otimes\mathbf{w}_j)(\mathbf{Y}_j\otimes\mathbf{1}) \\
&= \left(\sum \mathbf{W}_j\mathbf{W}_j'\right)^{-1}\sum \mathbf{W}_j\mathbf{Y}_j = \hat{\boldsymbol{\gamma}}. \qquad (20)
\end{aligned}$$

Thus $\boldsymbol{\gamma}^*_{0(\mathbf{A})} = \hat{\boldsymbol{\gamma}}$ does not depend on $\mathbf{A}$. Substituting the equalities (19) and (20) into equation (13), the equal covariance likelihood $L_0(\mathbf{B}_0)$ for $\mathbf{B}_0 = \mathbf{V}_0^{1/2}(\mathbf{V}_0+\mathbf{A})^{-1}\mathbf{V}_0^{1/2}$ reduces to

$$\begin{aligned}
L_0(\mathbf{B}_0) &= \left|\sum \mathbf{w}_j\mathbf{w}_j'\right|^{-p/2}|\mathbf{V}_0+\mathbf{A}|^{-(J-q)/2}\,\mathrm{etr}\left\{-(\mathbf{V}_0+\mathbf{A})^{-1}\sum(\mathbf{Y}_j-\mathbf{W}_j'\hat{\boldsymbol{\gamma}})(\mathbf{Y}_j-\mathbf{W}_j'\hat{\boldsymbol{\gamma}})'/2\right\} \\
&\propto |\mathbf{B}_0|^{(J-q)/2}\,\mathrm{etr}(-\mathbf{B}_0\mathbf{S}/2). \qquad (21)
\end{aligned}$$

This establishes expression (3). The product of $L_0(\mathbf{B}_0)$ in expression (3) and $p(\mathbf{B}_0)$ in expression (7) is a proper $\mathrm{CWish}_p\{J-q+\nu_*, (\mathbf{S}+\mathbf{S}_*)^{-1}; \mathbf{I}\}$ density provided that $\nu_* > -(J-q)$.

A.3. Proof of theorem 2

Without essential loss of generality, assume that $\mathbf{V}_0 = \mathbf{I}_p$ (this can be achieved by defining the equivalent problem with $\mathbf{Y}_j^* = \mathbf{V}_0^{-1/2}\mathbf{Y}_j$). Choose constants λ_0 and λ_1 with $0 < \lambda_0 \leqslant 1 \leqslant \lambda_1 < \infty$, so that $\lambda_0\mathbf{I}_p \leqslant \mathbf{V}_j \leqslant \lambda_1\mathbf{I}_p$ for all $j = 1, \ldots, J$. This is satisfied if λ_1 exceeds the largest of the Jp eigenvalues of $\mathbf{V}_1, \ldots, \mathbf{V}_J$, and if λ_0 is smaller than any of these Jp eigenvalues. Then, for all $j = 1, \ldots, J$, $\lambda_0(\mathbf{I}+\mathbf{A}) \leqslant \mathbf{V}_j+\mathbf{A} \leqslant \lambda_1(\mathbf{I}+\mathbf{A})$. Define

$$L_0(\mathbf{A}) \equiv |\mathbf{I}+\mathbf{A}|^{-\nu_0/2}\,\mathrm{etr}\,\{-(\mathbf{I}+\mathbf{A})^{-1}\mathbf{S}_0/2\}, \qquad \mathbf{0} < \mathbf{B}_0 \leqslant \mathbf{I}, \qquad (22)$$

so that $f_0(\mathbf{B}_0) = c_0\,L_0(\mathbf{A})\,p(\mathbf{B}_0)$ is a $\mathrm{CWish}(\nu, \boldsymbol{\Sigma}; \mathbf{I})$ density with $\nu = \nu_0+\nu_*$, $\boldsymbol{\Sigma} = (\mathbf{S}_0+\mathbf{S}_*)^{-1}$ and c_0 a normalizing constant that does not depend on $\mathbf{B}_0$. The posterior density of $\mathbf{B}_0$ is $f(\mathbf{B}_0|\mathbf{Y}) = c_1\,L(\mathbf{A})\,p(\mathbf{B}_0)$, for $L(\mathbf{A})$ from equation (13) and c_1 a normalizing constant, so the ratio to be bounded to prove theorem 2 is

$$\frac{f(\mathbf{B}_0|\mathbf{Y})}{f_0(\mathbf{B}_0)} = \frac{c_1\,L(\mathbf{A})}{c_0\,L_0(\mathbf{A})}\frac{p(\mathbf{B}_0)}{p(\mathbf{B}_0)} = \frac{c_1\,L(\mathbf{A})}{c_0\,L_0(\mathbf{A})}.$$

Because the exponential term in $L(\mathbf{A})$ must be negative or 0, we have

$$\begin{aligned}
L(\mathbf{A}) &\leqslant \left|\sum \mathbf{W}_j(\mathbf{V}_j+\mathbf{A})^{-1}\mathbf{W}_j'\right|^{-1/2}\prod|\mathbf{V}_j+\mathbf{A}|^{-1/2} \\
&\leqslant \left|\sum \mathbf{W}_j\frac{1}{\lambda_1}(\mathbf{I}+\mathbf{A})^{-1}\mathbf{W}_j'\right|^{-1/2}\prod|\lambda_0(\mathbf{I}+\mathbf{A})|^{-1/2} \\
&= \frac{\lambda_1^{pq/2}}{\lambda_0^{Jp/2}}|\mathbf{I}+\mathbf{A}|^{-(J-q)/2}\left|\sum \mathbf{w}_j\,\mathbf{w}_j'\right|^{-p/2}.
\end{aligned}$$

The last equality is justified by equation (19) in the proof of theorem 1. Also,

$$L_0(\mathbf{A}) \geqslant |\mathbf{I}+\mathbf{A}|^{-\nu_0/2}\,\mathrm{etr}(-\mathbf{S}_0/2),$$

the lower bound being at $\mathbf{A} = \mathbf{0}$. Thus, for

$$M = \frac{c_1 \lambda_1^{pq/2}}{c_0 \lambda_0^{Jp/2}} \left| \sum \mathbf{w}_j \mathbf{w}_j' \right|^{-p/2} \operatorname{etr}\left(-\frac{\mathbf{S}_0}{2} \right) < \infty,$$

$$\frac{f(\mathbf{B}_0|\mathbf{Y})}{f_0(\mathbf{B}_0)} \leqslant M|\mathbf{I} + \mathbf{A}|^{-(J-q-\nu_0)/2} \leqslant M$$

for all $\mathbf{A} \geqslant \mathbf{0}$ and hence for all $\mathbf{0} < \mathbf{B}_0 \leqslant \mathbf{I}$, provided that $\nu_0 \leqslant J - q$. This implies that the CWish(ν, $\mathbf{\Sigma}$; $\mathbf{I}$) density f_0 envelops $f(\mathbf{B}_0|\mathbf{Y})$ if $\nu \leqslant J - q + \nu_*$.

References

Bryk, A. S. and Raudenbush, S. W. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park: Sage.

Bryk, A. S., Raudenbush, S. W. and Congdon, R. T. (1996) *HLM: Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*. Chicago: Scientific Software International.

Carlin, B. P. and Louis, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman and Hall.

Casella, G. and Hwang, J. T. (1983) Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *J. Am. Statist. Ass.*, **78**, 688–698.

Christiansen, C. L. and Morris, C. N. (1997) Hierarchical Poisson regression modeling. *J. Am. Statist. Ass.*, **92**, 618–632.

Cleary, P., Edgman-Levitan, S., Roberts, M., Moloney, T., McMullen, W., Walker, J. and Delbanco, T. (1991) Data watch. *Hlth Aff.*, **10**, 254–267.

Dempster, A., Rubin, D. and Tsutakawa, R. (1981) Estimation in covariance components models. *J. Am. Statist. Ass.*, **76**, 341–353.

Efron, B. and Morris, C. (1972) Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika*, **59**, 335–347.

Everson, P. (1995) Inference for multivariate normal hierarchical models. *PhD Dissertation*. Department of Statistics, Harvard University, Cambridge.

Everson, P. and Morris, C. (2000) Simulation from Wishart distributions with eigenvalue constraints. *J. Comput. Graph. Statist.*, to be published.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo Methods in Practice*. London: Chapman and Hall.

Goldstein, H. (1995) *Multilevel Statistical Models*, 2nd edn. Chichester: Wiley.

Hox, J. (1998) Multilevel modeling in windows: a review of MLWin. *Multilev Modllng Newslett.*, **10**, no. 2, 2–5.

Hwang, J. T. and Casella, G. (1982) Minimax confidence sets for the mean of a multivariate normal distribution. *Ann. Statist.*, **10**, 868–881.

James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 361–379. Berkeley: University of California Press.

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

Longford, N. T. (1993) *Random Coefficient Models*. Oxford: Oxford University Press.

Meng, X.-L. and van Dyk, D. (1998) Fast EM-type implementations for mixed effects models. *J. R. Statist. Soc.* B, **60**, 559–578.

Morris, C. N. (1983a) Parametric empirical Bayes inference: theory and applications. *J. Am. Statist. Ass.*, **78**, 47–55.

———(1983b) Parametric empirical Bayes confidence intervals. In *Scientific Inference, Data Analysis, and Robustness* (eds G. E. P. Box, T. Leonard and C. F. Wu), pp. 25–50. New York: Academic Press.

Normand, S. L. (1995) Meta-analysis software: a comparative review. *Am. Statistn*, **49**, 298–309.

Odell, P. L. and Feiveson, A. H. (1966) A numerical procedure to generate a sample covariance matrix. *J. Am. Statist. Ass.*, **61**, 199–203.

Ripley, B. (1987) *Stochastic Simulation*. Chichester: Wiley.

Rubin, D. B. (1987) Comment on 'The calculation of posterior distributions by data augmentation' (by M. A. Tanner and W. H. Wong). *J. Am. Statist. Ass.*, **82**, 543–546.

Searle, S. R. (1982) *Matrix Algebra Useful for Statistics*. New York: Wiley.

West, M. and Harrison, P. J. (1997) *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer.