The Canadian Journal of Statistics
Vol. 34, No. 3, 2006, Pages 377–390
La revue canadienne de statistique

Conservative prior distributions for variance parameters in hierarchical models

Paul GUSTAFSON, Shahadut HOSSAIN, and Ying C. MACNAB

Key words and phrases: Bayesian analysis; hierarchical model; linear mixed model; prior distribution; variance component.

MSC 2000: Primary 62F15; secondary 62P10.

Abstract: Bayesian hierarchical models typically involve specifying prior distributions for one or more variance components. This is rather removed from the observed data, so specification based on expert knowledge can be difficult. While there are suggestions for "default" priors in the literature, often a conditionally conjugate inverse-gamma specification is used, despite documented drawbacks of this choice. The authors suggest "conservative" prior distributions for variance components, which deliberately give more weight to smaller values. These are appropriate for investigators who are skeptical about the presence of variability in the second-stage parameters (random effects) and want to particularly guard against inferring more structure than is really present. The suggested priors readily adapt to various hierarchical modelling settings, such as fitting smooth curves, modelling spatial variation and combining data from multiple sites.

Lois a priori conservatrices pour les paramètres de variance de modèles hiérarchiques

Résumé : Les modèles bayésiens hiérarchiques comportent généralement une ou des composantes de variance que l'on doit doter de lois a priori. Le choix de ces lois est délicat car la variation est un aspect des données difficile à cerner. De toutes les lois a priori "par défaut," une loi conjuguée inverse-gamma conditionnelle est la plus souvent employée, malgré ses inconvénients. Les auteurs proposent des lois a priori "conservatrices" pour les composantes de la variance qui privilégient les petites valeurs. Elles conviennent bien aux situations où le chercheur s'interroge sur la présence réelle de variabilité dans les paramètres de deuxième degré (effets aléatoires) et qu'il veut éviter d'imposer une structure artificielle. Les lois a priori suggérées s'adaptent à diverses situations propices à la modélisation hiérarchique, notamment l'ajustement de courbes lisses et la modélisation de variation spatiale ou de données issues de nombreux sites.

1. INTRODUCTION

Consider the following template for a hierarchical model based on normal distributions and linear relationships, often referred to as a linear mixed model (LMM) in non-Bayesian contexts. Say that the observable vector Y having n components is distributed as

$$Y \mid \theta_f, \theta_r, \omega, \lambda \quad \sim \quad \mathsf{N}_n(A_r\theta_r + A_f\theta_f, \omega V_1). \tag{1}$$

Here θ_r and θ_f are vectors of random and fixed effects with dimensions p and q respectively, while A_r $(n \times p)$ and A_f $(n \times q)$ are the corresponding design matrices. The scalar parameter ω is typically unknown, whereas the $n \times n$ matrix V_1 is taken as known. Often V_1 is specified as a correlation matrix (the identity matrix in many instances), so that ω is a variance component—the error variance for each element of Y. (For clearer discussion of prior distributions on variance components we do *not* follow the convention of denoting variances as squared quantities.) Note also that λ is included as a conditioning argument in (1), while in fact being absent on the right-hand side. Whereas (1) constitutes the first-stage of the hierarchical model, the variance component λ governs the variation in the random effects θ_r at the second-stage.

More precisely, the second-stage describes distributions of the fixed and random effects given the variance components, i.e., the distribution of $(\theta_f, \theta_r \mid \omega, \lambda)$. Commonly it is assumed that

$$\pi(\theta_f, \theta_r \mid \omega, \lambda) \quad = \quad \pi(\theta_f)\pi(\theta_r \mid \lambda),$$

with the second term being of more concern than the first. That is, typically inferences are not sensitive to the choice of prior on the fixed effects, and some kind of diffuse prior, perhaps improper, is postulated. For the random effects we consider

$$\theta_r \,|\, \lambda \quad \sim \quad \mathrm{N}_p(0, \lambda V_2), \tag{2}$$

where V_2 is a known $p \times p$ matrix, while the scalar parameter λ is an unknown variance component governing the magnitude of variability in the random effects. Often an appropriate form for V_2 is readily apparent from the modelling context at hand.

At the third stage of the hierarchy, a prior distribution for the variance components (ω, λ) must be specified. The present article focuses on this choice of distribution. We think of the prior on both variance components jointly in terms of

$$\pi(\lambda, \omega) \quad = \quad \pi(\lambda \,|\, \omega)\pi(\omega). \tag{3}$$

Here again the more subtle specification is typically for $(\lambda \,|\, \omega)$, with diffuse priors for ω being commonly employed.

Taken together the three stages of the hierarchical model define a joint distribution over $(Y, \theta_f, \theta_r, \omega, \lambda)$, and consequently a posterior distribution over all unobserved quantities $(\theta_f, \theta_r, \omega, \lambda)$ given the observed Y. In the absence of closed-form expressions, Markov chain Monte Carlo (MCMC) methods can be used to simulate samples from the posterior distribution.

Three-stage hierarchical models of this or similar form have found widespread application. As a first example, consider a multi-centre medical study conducted at p sites, or a meta-analysis which aims to aggregate p different studies of the same treatment-outcome relationship. Here $\theta_{r,j}$ is the random effect associated with the jth site or study. The random effects are usually regarded as conditionally independent and identically distributed given the variance components, i.e., $V_2 = I_p$, with λ governing the magnitude of the site-to-site variation. Via the posterior distribution of λ one obtains a data-driven compromise between pooling the data without regard to site, versus separate analyses for each site. There is a large literature on such applications of hierarchical modelling and their attendant benefits. See, for instance, Stangl & Berry (2000).

As a second example, much recent work on spatial modelling in general, and disease-mapping in particular, falls into the framework of a three-stage hierarchical model. Consider, for instance, disease-rate data for p contiguous geographical regions. Take $n = p$ and $q = 1$, and let Y_j be a suitably transformed empirical disease rate in the jth region, whereas $\theta_f + \theta_{r,j}$ corresponds to the true underlying disease rate. In this context, V_2 might be chosen such that underlying rates for adjacent regions are likely more similar than rates for distant regions. Then λ governs the overall spatial variation in the underlying disease rate. Again the hierarchical model yields a data-driven compromise, this time between the extremes of (i) estimating the disease rate to be constant across space, and (ii) providing p unrelated estimates for the regions separately. In many instances (i) is unrealistic, while (ii) typically suffers from high variance estimates and an inferred spatial distribution of rates which is implausibly rough.

As a third rather different application of a three-stage hierarchical model, consider flexible regression of a continuous response variable on a continuous predictor using cublic splines. The common approach of penalizing roughness according to the integrated squared second-derivative of the regression function (see, for instance, Green & Silverman 1994) can be exactly represented using (1), (2), and (3). The resulting interpretation is that λ specifies the strength of the roughness penalty. While a non-Bayesian approach might use cross-validation to let the data inform the choice of λ, the Bayesian approach of assigning a prior to λ yields a principled way to estimate the regression function and characterize uncertainty about its shape. Here the hierarchical model gives a compromise between (i) fitting a straight line to the data and (ii) fitting an arbitrarily rough curve to the data (through all the data points, for instance, if there is a knot at every data point).

In these examples, appropriate specifications for (1) and (2) tend to be evident from the nature of the problem at hand. Specification of (3) tends to be less obvious, however, particularly given the extent to which λ is removed from any observable quantities in the hierarchy. In fact, the choice of third-stage distributions has received a fair amount of discussion in the literature. Many of the issues are reviewed by Gelman, Carlin, Stern & Rubin (2004), Gelman (2006), and Spiegelhalter, Abrams & Myles (2003). A particular starting point for the present work is Daniels (1999) who discusses *uniform-shrinkage* priors. The idea here is that at least with simple forms of (1) through (3), Bayesian estimates of the random effects θ_r can be expressed as weighted combinations of a pooled estimator (i.e., corresponding to an absence of random effects with $\lambda = 0$) and an 'unrelated components' estimator (the random effects becoming 'fixed' as λ increases). A uniform-shrinkage prior for λ is selected to induce a uniform prior distribution on the weights in this combination. This is an attempt to treat the two extremes of pooled analysis and separate analyses in a fair manner a priori.

The focus of the present paper is that sometimes one may not wish to be fair in this way. We discuss the selection of priors which are deliberately *conservative*, in the sense of particularly trying to guard against substantial *over-estimation* of the random effect variability. The motivation for being more concerned about over-estimation than under-estimation arises as many applications of hierarchical modelling inherit the classical notion that a Type I error is worse than a Type II error. To wit, in a multi-centre trial it may be more damaging to incorrectly conclude that patient outcomes vary substantially across sites (thereby instilling unnecessary concern in the patient population), than to underestimate, or fail to detect, real variation. Similarly, incorrectly concluding that disease rates vary substantially across regions can lead to costly and unnecessary intervention. Finally, in the curve-fitting scenario it can be quite dangerous to infer 'bumps' and other nonlinearities that really do not exist. We feel that often fitted curves from various smoothing procedures look suspiciously rough relative to what is plausible in the scientific context at hand.

Another motivation for the priors we propose is as part of sensitivity analysis, whereby inferences are made under a few different choices of prior distributions. If these inferences are similar, then one is less concerned about the influence of the prior in the analysis. One particular proposal in a medical context of evaluating a new treatment is to do the analysis under three priors, one that is skeptical about efficacy of this treatment, one that is neutral, and one that is enthusiastic (see Spiegelhalter, Freedman & Parmar 1994). Analogously, the conservative priors we propose are appropriate for someone who is skeptical about the existence of random effect variation, whereas the uniform-shrinkage priors alluded to above might be thought of as neutral.

2. A SIMPLE CASE

To initiate our discussion of the proposed prior distributions, we reduce to a much simplified, or *prototype*, hierarchical model. In particular, consider a model without fixed effects and with known error variance ω in (1). Assume there is a single observation per random effect ($n = p$, $A_r = I_p$), each of which has the same error variance ($V_1 = I_p$). The random effects are taken to be exchangeable, (i.e., independent and identically distributed given the variance component λ), as is typical in a multi-centre trial context, for instance. Thus $V_2 = I_p$. With all these assumptions, the first two stages of the prototype model are simply

$$Y \mid \theta_r, \lambda \sim \mathsf{N}_p(\theta_r, \omega I_p), \tag{4}$$

$$\theta_r \mid \lambda \sim \mathsf{N}_p(0, \lambda I_p). \tag{5}$$

This is the same initial model considered by Daniels (1999).

From (4) and (5), marginalizing over the random effects gives $(Y \mid \lambda) \sim \mathsf{N}_p\{0, (\omega + \lambda) I_p\}$, so that the likelihood function is

$$L(\lambda) \propto \frac{1}{(\omega + \lambda)^{p/2}} \exp\left\{ -\frac{(\mathrm{SS}_p/2)}{\omega + \lambda} \right\},$$

where $SS_p = \sum_{i=1}^{p} y_i^2$. A family of proper conjugate prior densities indexed by $a > 0$ and depending on ω is then

$$\pi^{(a)}(\lambda \,|\, \omega) \;=\; \omega^{-1} g_a(\lambda/\omega), \tag{6}$$

where

$$g_a(z) \;=\; a(1+z)^{-(a+1)}. \tag{7}$$

Note that the $a = 0$ limit gives the (improper) Jeffreys prior for this problem, $a = 1/2$ gives the 'proper Jeffreys prior' proposed by Berger & Deely (1988), and $a = 1$ gives the uniform-shrinkage prior (Daniels 1999). Note also that a more general conjugate family obtains by replacing (7) with $g_{a,b}(z) \propto (1+z)^{-(a+1)} \exp\{-b(1+z)^{-1}\}$, with prior-to-posterior updating mapping from (a, b) to $(a + p/2, b + \omega^{-1} SS_p/2)$. For a reason to be mentioned shortly, however, we only consider the $b = 0$ case corresponding to (7).

To study the potential for over-estimation of λ, we focus on the posterior *mode* of λ, given by

$$\hat{\lambda} \;=\; \max\left\{ \frac{SS_p}{p + 2(a+1)} - \omega, 0 \right\}.$$

This choice, as opposed to say a posterior mean or median, requires some comment. We deliberately consider prior densities which are positive but finite at $\lambda = 0$, since the data never rule out the possibility that $\lambda = 0$, i.e., $L(0) > 0$ for any value of SS_p. Selecting a prior density for λ which vanishes at zero, as with the common choice of an inverse-gamma distribution, entails an unnecessarily strong assumption about the existence of random effect variation. With prior (6), the posterior density for λ will also be positive but finite at zero. Depending on the observed data though, this density will be either monotonically decreasing (so that $\hat{\lambda} = 0$) or unimodal (so that $\hat{\lambda} > 0$). This qualitative distinction seems useful and appropriate, but it is lost when say a posterior mean or median is used to estimate λ, since these estimators are necessarily positive for any data.

The desirability of the posterior mode as an estimator of λ is admittedly tempered by reparameterization concerns. With a prior density on λ that is positive but finite at zero, both the prior and posterior densities of the standard deviation $\lambda^{1/2}$ are zero at zero, so that the posterior mode of $\lambda^{1/2}$ is necessarily positive for any dataset. Or, starting with a decreasing prior density for $\lambda^{1/2}$ that is positive but finite at zero yields prior and posterior modes of λ which are necessarily zero. In the random effect setting then, it makes most sense to consider the posterior mode in the same parameterization under which the prior density is positive but finite at zero. A related point is that the ratio of the posterior density to the prior density at zero is invariant under reparameterization. This is the so-called Savage–Dickey ratio which is the Bayes factor in favor of $\lambda = 0$; see Verdinelli & Wasserman (1995) for discussion.

Considering a literal sense of over-estimation, one can readily calculate that

$$\Pr(\hat{\lambda} > \lambda) \;=\; \Pr\{V_p > p + 2(a+1)\}, \tag{8}$$

where $V_p = (\omega + \lambda)^{-1} SS_p \sim \chi^2_p$. It is interesting to note that this probability does not vary with λ, and that this arguably desirable invariance property no longer obtains upon taking $b > 0$ in the more general family of conjugate priors $g_{a,b}(\,\cdot\,)$ mentioned earlier.

A first attempt to control over-estimation might involve choosing a to make (8) as small as desired. Of course this choice of prior would be heavily dependent on p, the dimension of the random effects. Put another way, for any fixed prior the probability of over-estimation will necessarily tend to 0.5 as p increases, in light of the consistency and asymptotic normality of Bayesian estimators. Since prior distributions which depend on the amount of data to be

collected are often criticized, we consider controlling the chance of over-estimation in a relative sense instead.

A useful formulation is to define over-estimation in terms of a fraction of the total variance $\mathrm{var}(Y_i) = \omega + \lambda$. Particularly,

$$\Pr\left\{\frac{\hat{\lambda}}{\lambda+\omega} > \frac{\lambda}{\lambda+\omega} + \varepsilon\right\} = \Pr\left\{\frac{V_p - p}{(2p)^{1/2}} > \frac{2(a+1)(1+\varepsilon)+\varepsilon p}{(2p)^{1/2}}\right\}$$
$$\approx 1 - \Phi\left(\frac{2(a+1)(1+\varepsilon)+\varepsilon p}{(2p)^{1/2}}\right). \tag{9}$$

In contrast to (8), (9) is not monotone in p. In particular, (9) is maximized when $p = 2(a+1)\varepsilon^{-1}(1+\varepsilon)$. Moreover, at least for many values of (a, ε), this is not a markedly peaked maximum. This raises the possibility of choosing a single prior which can guard against over-estimation regardless of the dimension of the random effects. Particularly,

$$\max_p \Pr\left\{\frac{\hat{\lambda}}{\lambda+\omega} > \frac{\lambda}{\lambda+\omega} + \varepsilon\right\} \approx 1 - \Phi\big(2(a+1)^{1/2}\varepsilon^{1/2}(1+\varepsilon)^{1/2}\big). \tag{10}$$

Thus for a given choice of ε and a maximum probability of over-estimation that will be tolerated, one can choose a commensurately on the basis of (10).

Table 1 gives values of (10) for selected values of a and ε. It is seen that Jeffreys' prior ($a = 0$) and the uniform-shrinkage prior ($a = 1$) are not conservative in the desired sense, and that a larger value of a is needed to obtain rather stringent control on the maximum probability of relative over-estimation. The table suggests that values of a in the range of 5 to 10 might be reasonable choices for a default conservative prior. We do not attempt to be more prescriptive than this by recommending a particular single value of a. We simply view Table 1 as a guide to the rough magnitude of a required in order to be conservative or skeptical about the existence of random effect variation.

TABLE 1: The maximum probability (10) for selected values of a and ε.

a	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
0	0.32	0.25	0.16
1	0.26	0.17	0.08
2.5	0.20	0.11	0.03
5	0.13	0.05	0.00
7.5	0.09	0.03	0.00
10	0.06	0.01	0.00

3. MORE GENERAL SITUATIONS

The conservative prior has been introduced as a distribution for λ when ω is known. Realistically ω is unknown, so we treat $\pi^{(a)}(\lambda \mid \omega)$ as a conditional prior distribution, and further assign a marginal prior distribution for ω. As mentioned earlier, prior information about ω is typically not crucial. Thus we tend to assign a proper but quite diffuse distribution for ω. A computationally convenient choice is the $\mathrm{IG}(\varepsilon, \varepsilon)$ distribution, for some small value of ε. When ω is unknown, the posterior mode of λ no longer has a closed form. Nonetheless, simulation studies in Hossain (2003) indicate that the extent of conservatism imparted by $\pi^{(a)}(\lambda \mid \omega)$ for various values of a is qualitatively unchanged upon moving to this more realistic scenario.

The bigger generalization is to general design matrices A_f and A_r and general covariance (or correlation) matrices V_1 and V_2. We argue that an appropriate generalization of $\pi^{(a)}$ is

$$\pi^{(a)}(\lambda \mid \omega) = \frac{c}{\omega \left| (A_r' V_1^{-1} A_r)^{-1} + (\lambda/\omega) V_2 \right|^{(a+1)/p}}, \tag{11}$$

where c is a normalizing constant. Note that with the given scale-parameter structure, c does not depend on ω. We claim that the choice of a in (11) mimics the choice in the prototype case of the previous section. That is, $a = 1$ corresponds to a (now weaker) sense of uniform-shrinkage and is neutral about the random effect variation, whereas larger values correspond to more conservative priors.

To elaborate, note that given fixed effects θ_f, the least-squares estimator of the random effects θ_r is $\hat{\theta}_{\mathrm{LS}} = (A_r' V_1^{-1} A_r)^{-1} A_r' V_1^{-1} \tilde{y}$, where $\tilde{y} = y - A_f \theta_f$. Conversely, the posterior mean, which shrinks toward the prior mean of zero, can be written as

$$\hat{\theta}_{\mathrm{SH}} = \left(A_r' V_1^{-1} A_r + \frac{\omega}{\lambda} V_2^{-1} \right)^{-1} (A_r' V_1^{-1} A_r) \hat{\theta}_{\mathrm{LS}}, \tag{12}$$

where the SH subscript reminds us that this is a *shrinkage* estimator.

To gain intuition, consider the case where both $A_r' V_1^{-1} A_r$ and V_2 are diagonal, and let $c_j = (A_r' V_1^{-1} A_r)_{jj}$, $d_j = (V_2^{-1})_{jj}$. In this case (12) can be expressed component-wise as

$$\hat{\theta}_{\mathrm{SH},j} = \frac{c_j}{c_j + d_j(\omega/\lambda)} \hat{\theta}_{\mathrm{LS},j}.$$

That is, each component is a linear combination of the corresponding component of the least-squares estimator and zero (the prior mean). To obtain a uniform prior distribution for the weights in this combination, we would choose the prior density

$$\pi_j^{(a)}(\lambda \mid \omega) = \frac{1}{c_j^{-1} d_j \omega} g_a \left(\frac{\lambda}{c_j^{-1} d_j \omega} \right),$$

with $a = 1$. Similarly, other choices of a yield the same favoritism for the prior mean in the estimation of $\theta_{r,j}$ as was exhibited in the simpler scenario of Section 2.

Clearly, if $c_j^{-1} d_j$ varies with j then so does $\pi_j^{(a)}$, and we cannot find a single prior for λ which imparts equal conservatism across estimation of different components of θ_r. The next best thing might be to select a prior which *in aggregate* imparts the desired degree of conservatism. An obvious possibility would be to choose a desired value of a and then amalgamate $\pi_1^{(a)}(\lambda \mid \omega), \ldots, \pi_p^{(a)}(\lambda \mid \omega)$ in some manner. Particularly, consider taking the normalized geometric mean, i.e., averaging the log-densities and then normalizing, as per

$$\begin{aligned} \pi^{(a)}(\lambda \mid \omega) &\propto \exp\left\{ p^{-1} \sum_{j=1}^{p} \log \pi_j^{(a)}(\lambda \mid \omega) \right\} \\ &\propto \frac{1}{\omega \prod_{j=1}^{p} \{ c_j^{-1} d_j + (\lambda/\omega) \}^{(a+1)/p}}. \end{aligned} \tag{13}$$

In fact it is easy to check that the general prior form (11) specializes precisely to (13) when $A_r' V_1^{-1} A_r$ and V_2 are both diagonal. In turn this lends credence to (11) as an appropriate prior in the more general setting, with the interpretation of the hyperparameter a retained. Generally we will refer to (11) with $a = 1$ as an *aggregate uniform-shrinkage* (AUS) prior, while (11) with a in the range of 5 to 10 say is a conservative prior.

We can give some further interpretation to the prior (11) beyond the case where $A_r' V_1^{-1} A_r$ and V_2 are diagonal. Note that (11) can be written as

$$\pi^{(a)}(\lambda \mid \omega) \propto \frac{1}{\omega \left| (A_r' V_1^{-1} A_r)^{-1} V_2^{-1} + (\lambda/\omega) I_p \right|^{(a+1)/p}} \propto \frac{1}{\omega \prod_{j=1}^{p} \{k_j + (\lambda/\omega)\}^{(a+1)/p}},$$

where $k_1, \ldots, k_p$ are the eigenvalues of $(A_r' V_1^{-1} A_r)^{-1} V_2^{-1}$, so that $k_1 + z, \ldots, k_p + z$ are the eigenvalues of $(A_r' V_1^{-1} A_r)^{-1} V_2^{-1} + z I_p$. Thus our prior is the normalized geometric mean of p priors of the form $g_a(\cdot)$, with the respective scale parameters being $k_1\omega, \ldots, k_p\omega$.

In terms of MCMC computation of posterior quantities arising from the prior (11), note that the posterior full conditional distributions for λ and ω do not correspond to standard distributions. In fact, both these full conditional densities can be expressed as members of the family

$$f(x \mid c, d, k, M) \propto \frac{1}{x^{c+1}} \exp(-d/x) \frac{1}{|M + x I_p|^k}. \tag{14}$$

This is not a log-concave density function, so the commonly used adaptive rejection sampling algorithm of Gilks & Wild (1992) cannot be applied. However, in the Appendix we describe a recursive and self-tuning rejection sampling algorithm to sample from (14). This algorithm to update λ and ω, along with straightforward multivariate normal draws from the full conditionals of θ_f and θ_r, comprise a readily-implemented Gibbs sampler for this class of models. We also note in passing then in some instances *hierarchical centering* can greatly improve MCMC sampler performance in hierarchical models (Gelfand, Sahu & Carlin 1995), and we do take advantage of this in the examples of Section 4. From a practical perspective then, using the prior (11) is hardly more difficult than using conditionally conjugate priors, such as inverse gamma distributions for both λ and ω.

4. EXAMPLE: SPATIAL MODELLING

4.1. Framework.

Consider a geographic jurisdiction partitioned into p regions or areas. Let $\theta_f + \theta_{r,j}$ be the underlying mean response of interest in the jth region, i.e., the random effects reflect spatial variability across the jurisdiction. A common choice of second-stage variance matrix which reflects spatial smoothness is $V_2 = \{(1-\rho)I + \rho Q\}^{-1}$, where Q is a neighbourhood matrix. Particularly, Q_{jj} is the total number of neighbouring regions to the jth region, and for $j \neq k$, Q_{jk} takes the value -1 if regions j and k are neighbours, and 0 otherwise. The parameter ρ reflects the extent to which the random effect variation is spatially smooth. For further discussion of such conditionally autoregressive hierarchical models for spatial modelling, see, for instance, Banerjee, Carlin & Gelfand (2004), MacNab (2003). If ρ is taken as known, then this specification fits exactly into the hierarchical model framework of (1) through (3), and the prior (11) can be used as discussed. If ρ is unknown, we can view (11) as a prior for $(\lambda \mid \omega, \rho)$, and then assign marginal priors $\pi(\omega, \rho) = \pi(\omega)\pi(\rho)$.

4.2. Simulated data example.

As a simple example, consider a square jurisdiction partitioned into a 5 by 5 grid of square regions, i.e., $p = 25$. Thus interior regions have 8 neighbouring regions (including those diagonally adjacent), while regions on the area boundary have 5 neighbours, except for the corner regions which have three neighbours. Say there are two replicate observations of Y for each region, so that $n = 2p$, and ω is clearly identified from the data. Three sets of (Y, θ_r) values are simulated under true parameter values $\theta_f = 0$, $\omega = (0.3)^2$, $\rho = 0.95$, and three different values of λ,

namely $\lambda = 0$, $\lambda = (0.2)^2$, and $\lambda = (0.4)^2$. (In fact the three datasets are created from the same vector of residuals and different multiples of the same vector of random effects, for the sake of comparability.) Posterior distributions are computed under prior (11) for both $a = 1$ and $a = 7.5$, taking the value of ρ as known.

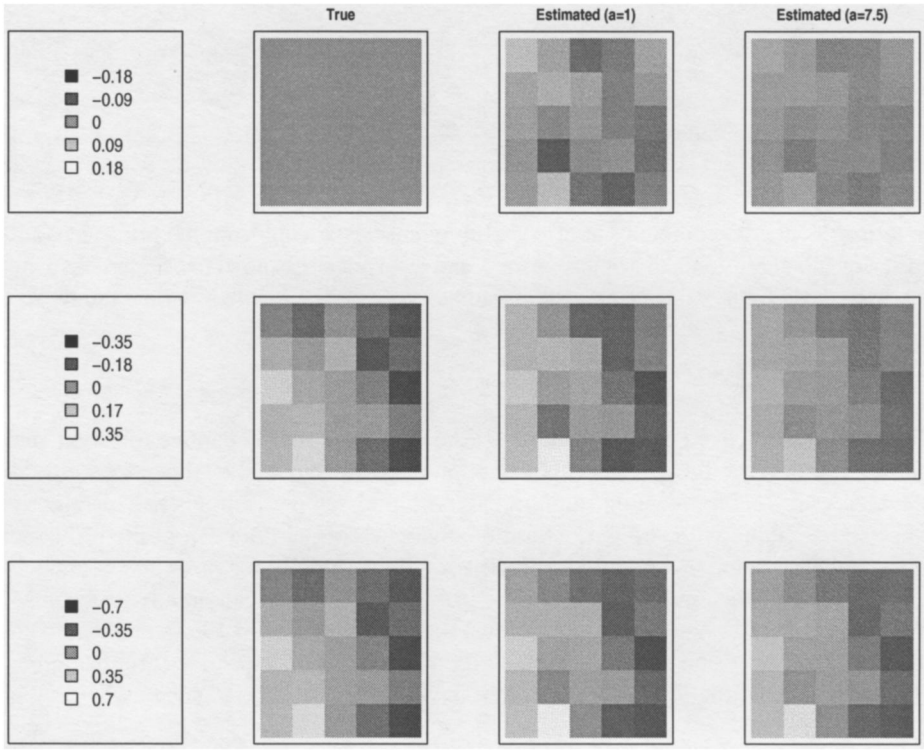

FIGURE 1: True values and posterior estimates of the spatial mean in the example of Section 4. The three rows of plots correspond to $\lambda = 0$, $\lambda = (0.2)^{1/2}$, $\lambda = (0.4)^{1/2}$. In each row, the left panel gives true values, the middle panel gives the posterior estimate using the AUS prior ($a = 1$) and the right panel gives the posterior estimate using the conservative prior ($a = 7.5$).

Figure 1 gives greyscale representations of both the true region means $\theta_f + \theta_{r,j}$ and the posterior mean estimates under both priors for each of the three underlying values of λ. The conservative prior estimates are visually smoother than their AUS counterparts in each instance. This is as desired when there is no real spatial variation ($\lambda = 0$). In the other two cases, the conservative estimates are somewhat over-smoothed relative to the truth, but by the same token the AUS estimates are somewhat under-smoothed. Thus we do not seem to pay a heavy price for being conservative when there is in fact spatial variation.

To investigate further we generated 50 datasets under each value of λ, with Figure 2 displaying the resulting point estimates of $\lambda^{1/2}$, the random effect standard deviation. In light of the earlier discussion of the shape of prior and posterior distributions for variance components, both the posterior mean of $\lambda^{1/2}$ and the square root of the posterior mode of λ are considered. Computing a posterior mode from MCMC output is not automatic, as some kind of smoothing procedure must be applied. We simply take the midpoint of the shortest interval containing proportion $\gamma = 0.02$ of the sampled values, applied to a version of the posterior sample which is 'folded-over' zero (so that the computed mode may take the value zero for a posterior sample concentrated at zero).

Figure 2 illustrates the extent to which estimates under the conservative prior are smaller

than those under the AUS prior. When there is no underlying spatial variation, the sampling distribution of the posterior mean estimator is substantially closer to zero when the conservative prior is applied. The contrast is even more marked for the posterior mode estimator, which is zero or very small for most datasets under the conservative prior, but not under the AUS prior. Thus again the conservative prior is fulfilling its mandate.

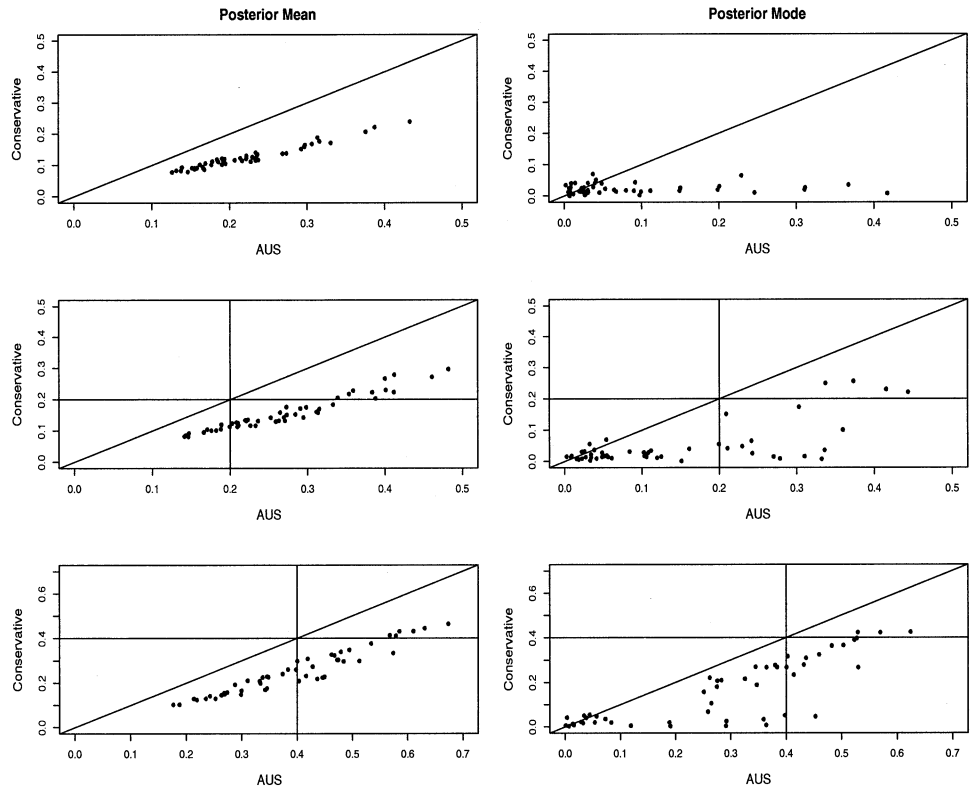

FIGURE 2: Point estimates of the random effect SD ($\lambda^{1/2}$) in the spatial modelling example. The posterior mean of $\lambda^{1/2}$ (left panels) and the root posterior mode of λ (right panels) are displayed for 50 simulated datasets under three different true values of λ ($\lambda = 0$, $\lambda = 0.2^2$, $\lambda = 0.4^2$ in the first through third rows of panels respectively). Within each panel, the estimates under the conservative prior ($a = 7.5$) are plotted against the estimates under the AUS prior ($a = 1$).

When the true variance component is $\lambda = 0.2^2$, we do see the price that is paid for being conservative, in that the posterior mode estimator is still at or near zero for most datasets under the conservative prior, but only for a minority of datasets under the AUS prior. If we do regard the posterior mode as summarizing evidence for the existence of random effect variability, then necessarily we lose power to detect such variation when we adopt the conservative prior. On the other hand, the posterior mean estimator actually performs somewhat better under the conservative prior than under the AUS prior. There is a clear negative estimation bias arising from the conservative prior, but there is also a positive bias with the AUS prior, and moreover the sampling variability of the estimator is larger under the AUS prior. Thus the empirical mean-squared-error is actually smaller when the conservative prior is adopted, even though the true variance component is positive.

When the true variance component is larger still ($\lambda = 0.4^2$), we see a relatively small and systematic difference between the posterior mean estimates arising from the two prior distributions. The conservative prior is still doing its job in the sense of yielding only a small chance of at most slightly overestimating the variance component, at the cost of increased bias relative to

the AUS-based estimator. For the posterior mode estimator, there is now enough spatial 'signal' to pull the estimator away from zero for most of the datasets when the AUS prior is used, and for about half the datasets when the conservative prior is used. The overall impression from Figure 2 is that the conservative prior seems to induce reasonable estimators, when in fact guarding against over-estimation is important.

4.3. Mortality data example.

We consider data on all-cause mortality in 1993, from the province of British Columbia, Canada. For the jth of 79 local health areas, these data are reported as a mortality count Z_j and a number-at-risk n_j. The number-at-risk ranges from 710 to 507291 across areas, while the mortality count ranges from 4 to 4271. We take $Y_j = \log Z_j - \log n_j$, and then model $Y_j \sim \mathrm{N}(\theta_f + \theta_{r,j}, \nu_j^2)$. Thus $\exp(\theta_f)$ is the overall mortality rate for a 'typical' region, while $\exp(\theta_{r,j})$ represents a multiplicative deviation in rate for the jth region compared to typical. We do not have replicate observations within regions, but regarding the mortality counts as Poisson distributed leads to the approximate procedure of taking $\nu_j^2 = n_j^{-1}$ as known. Placing this in the general framework of (1), we take $\omega = 1$ as known and let V_1 be diagonal with $V_{1,jj} = n_j^{-1}$. We fix $\rho = 0.8$ in the specification of the random effect variance matrix based on neighbourhood adjacency.

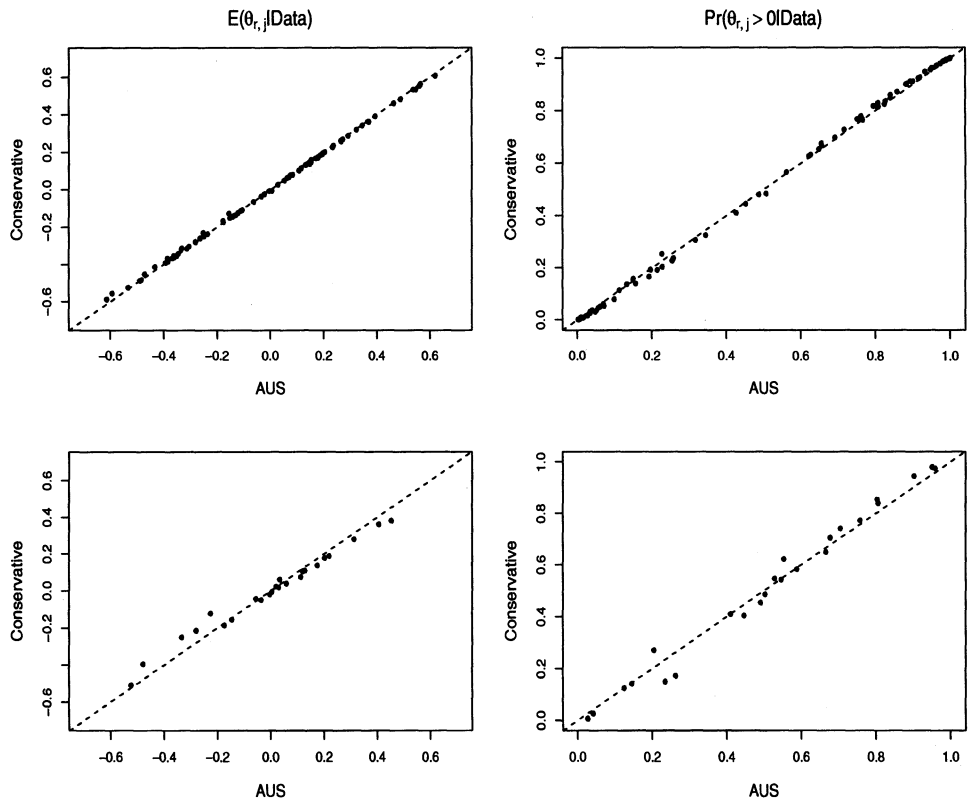

FIGURE 3: Posterior means and probabilities for the mortality example. $\mathrm{E}(\theta_{r,j} \mid \text{data})$ and $\Pr(\theta_{r,j} \mid \text{data})$, under conservative versus AUS prior distributions, are plotted. The top panels give results based on data from all 79 regions. The bottom panels correspond to data on only 24 regions.

For these data, in comparing posterior inference under a conservative prior ($a = 7.5$) versus an AUS prior ($a = 1$), we do see a reduction in the estimated variance component (on the order of a 15% reduction in both the posterior mean and median of λ). However, posterior mean estimates of the $\theta_{r,j}$ are virtually the same under the two priors, as witnessed in the upper-left

panel of Figure 3. The same finding obtains for the posterior probabilities of $(\theta_{r,j} > 0)$ (upper-right panel of Figure 3). It seems that data from 79 areas contain sufficient information about spatial variation to make inferences about area-specific effects robust to the choice of prior for the variance component.

To investigate further, we consider fitting the model to a subset of the data only. Particularly, we fit the model to the data from the first 24 regions only, which comprise a contiguous region in the Southeast corner of British Columbia. With these data we see a marked difference in the estimation of the variance component, with point estimates of λ under the conservative prior being roughly half as big as their AUS-prior counterparts. Commensurately, point estimates of the random effects themselves now show some discrepancies across the two priors (bottom-left panel of Figure 3) as do the corresponding posterior probabilities (bottom-right panel). To emphasize the extent to which there is more spatial smoothing under the conservative prior, note that $\hat{\theta}_r' V_2^{-1} \hat{\theta}_r$ is a summary of the estimated spatial variability, where $V_2^{-1} = (1-\rho)I + \rho Q$. This statistic takes the value 3.32 using posterior means under the conservative prior as $\hat{\theta}_r$, as opposed to 5.06 under the AUS prior.

5. DISCUSSION

The primary contribution of the present paper is the suggestion of a 'conservative' prior distribution for hierarchical models. This adds to the recent literature discussing the pros and cons of different prior distributions for variance parameters in hierarchical models. Note that our suggested prior choice is somewhat specific. Setting a in the range of $5 - 10$ corresponds to conservatism, in contrast to the AUS case of $a = 1$. In contrast, some of the literature on priors for variance parameters discusses different shapes for prior densities, without making explicit recommendations for the scale. As part of the contribution here, we suggest a natural generalization of the prior from the setting of a variance for independent and identically distributed random effects to the setting of a scalar variance component which multiplies a known (or at least highly-parameterized) correlation matrix. This is quite different from the generalization to a prior on the variance matrix for independent and identically distributed random effect vectors (Daniels 1999; Everson & Morris 2000). Whereas it is possible to obtain a uniform-shrinkage prior for an unknown variance matrix, we can only get an aggregate uniform-shrinkage prior for a scalar variance parameter which multiplies a known correlation matrix, as discussed in Section 3. We emphasize that our suggested prior can be used in an 'off-the-shelf' manner in a wide variety of settings. The spatial modelling setting of Section 4 is one example of this. We have also experimented with its use in the curve-fitting scenario described in Section 1. Supplemental material describing our findings is available (http://www.stat.ubc.ca/people/gustaf).

We have suggested prior distributions in the context of a hierarchical model based on normal distributions, i.e., a linear mixed model. However, we foresee generalizing the prior to models with a non-normal first-stage distribution, in line with Natarajan & Kass (2000) who generalized a uniform-shrinkage prior on a variance matrix to this setting. Explicitly, consider a generalized linear mixed model (GLMM) with $\mu_i = \mathrm{E}(Y_i)$, and link function $g(\cdot)$ such that

$$\begin{pmatrix} g(\mu_1) \\ \vdots \\ g(\mu_n) \end{pmatrix} = A_f \theta_f + A_r \theta_r.$$

Let $v(\cdot)$ denote the variance function, i.e., $\mathrm{var}(Y_i) = v(\mu_i)$. If θ_f is known, then standard generalized linear model results give the expected second derivative of the log-likelihood for θ_r as $\mathrm{E}\{-l''(\theta_r)\} = A_r' W A_r$, where the diagonal weight matrix W is given by $W_{ii}^{-1} = v(\mu_i)\{g'(\mu_i)\}^2$. That is, $A_r' W A_r$ in the generalized linear mixed model is analogous to $A_r'\{\omega V_1\}^{-1} A_r$ in the linear mixed model, describing the information in the data about the ran-

dom effects. This suggests that we modify (11) to

$$\pi^{(a)}(\lambda) \propto \frac{1}{\left|\left(A_r'\widetilde{W}A_r\right)^{-1}+\lambda V_2\right|^{(a+1)/p}},$$

where W, which depends on (θ_f, θ_r), has been replaced with some fixed weight matrix $\widetilde{W}$. One simple choice for $\widetilde{W}$ would be the weight matrix W arising when θ_f is set to a maximum likelihood estimate from the model without random effects, while θ_r is set to zero. As discussed by Natarajan and Kass (2000), this sort of argument leads to a data-dependent prior distribution. However, since the weight matrix tends to vary slowly over the parameter space in most instances, this data-dependence will usually be mild. The implementation and evaluation of this prior distribution for GLMM contexts is a topic of ongoing research.

APPENDIX

Here we give details of a rejection sampling algorithm for the full conditional distributions of the variance components under prior (11) for $(\lambda \mid \omega)$ along with an inverse gamma prior for ω. The posterior full conditional densities for both λ and ω take the form

$$f(x) \propto \frac{1}{x^{c+1}}\exp(-d/x)\frac{1}{|M+xI|^k},$$

for appropriate values of (c,d,k,M). This can be re-expressed as

$$f(x) \propto \frac{1}{x^{c+1}}\exp(-d/x)\frac{1}{\prod_{i=1}^{p}(m_i+x)^k}, \tag{15}$$

where $m_1 \leq \cdots \leq m_p$ are the eigenvalues of M. We refer to the distribution defined by (15) as the $F_p(c, d, k, m_1, \ldots, m_p)$ distribution.

As a recursive algorithm to simulate from F_p, consider first the case that $p = 1$. To generate $X \sim F_1(c, d, k, m_1)$:

1. Assign $\tilde{x} \leftarrow (c + k/2)^{-1}d$.
2. If $\tilde{x} \leq m_1$ then generate X by rejection sampling with $\mathrm{IG}(c, d)$ as the candidate distribution, giving the acceptance probability as $\{m_1/(m_1 + x)\}^k$.
3. Otherwise generate X by rejection sampling with $\mathrm{IG}(c+k, d)$ as the candidate distribution, giving the acceptance probability as $\{x/(m_1 + x)\}^k$.

Note that $\tilde{x}$ is a rough measure of location for whichever candidate distribution is used, and the candidate distribution is chosen to maximize the acceptance probability evaluated at $\tilde{x}$. Moreover, the maximized value cannot be less than $(1/2)^k$. Similar rejection sampling schemes for MCMC updating of variance components are described by Gustafson (1997, 1998).

In general to generate $X \sim F_p(c, d, k, m_1, \ldots, m_p)$ for $p > 1$ we use rejection sampling recursively. That is,

1. Assign $\tilde{x} \leftarrow (c + k/2)^{-1}d$.
2. If $m_p/(m_p + \tilde{x}) \geq \tilde{x}/(m_1 + \tilde{x})$ then generate X by rejection sampling with the candidate distribution $F_{p-1}(c, d, k, m_1, \ldots, m_{p-1})$, giving the acceptance probability as $\{m_p/(m_p + x)\}^k$.
3. Otherwise generate X by rejection sampling with $F_{p-1}(c + k, d, k, m_2, \ldots, m_p)$ as the candidate distribution, giving the acceptance probability as $\{x/(m_1 + x)\}^k$.

Again here we regard $\tilde{x}$ as the rough location of whichever candidate distribution is chosen (or more particularly, the rough location based on the inverse-gamma portion of the density only). Then the choice of candidate distribution is again based on maximizing the acceptance probability evaluated at $\tilde{x}$. Note that the ordering of the eigenvalues is used to advantage. The worst-case value of $\tilde{x}$ is such that $m_p/(m_p+\tilde{x}) = \tilde{x}/(m_1+\tilde{x})$, i.e., $\tilde{x} = (m_1 m_p)^{1/2}$. Even in this worst case, the acceptance probability evaluated at $\tilde{x}$ is $\left[1/\{1+(m_1/m_p)^{1/2}\}\right]^k$, which will be much larger than $(1/2)^k$ when the ratio of largest to smallest eigenvalues is considerable.

In practice this algorithm works quite well, even when p is moderately large so that a long string of recursive acceptances is needed to generate from the target distribution. Particularly, for the prior distributions considered here, $k = (a+1)/p$. Thus increasing p pushes the individual acceptance probabilities towards one, making longer strings of acceptances feasible.

When p is very large, this algorithm can be slow due to the very deep nesting of the recursion. We have found that a blocked version of the algorithm is useful in this circumstance, simulating from F_p using a candidate drawn from F_{p-r}, for some choice of r with $1 \le r < p$. Particularly, to generate $X \sim F_p(c, d, k, m_1, \ldots, m_p)$:

1. Assign $\tilde{x} \leftarrow (c + rk/2)^{-1} d$.
2. If $m_{p+1-r}/(m_{p+1-r}+\tilde{x}) \ge \tilde{x}/(m_r+\tilde{x})$ then generate X by rejection sampling with the candidate distribution $F_{p-r}(c, d, k, m_1, \ldots, m_{p-r})$. The resulting acceptance probability is $\prod_{j=p+1-r}^{p} \{m_j/(m_j+x)\}^k$.
3. Otherwise generate X by rejection sampling with $F_{p-r}(c+rk, d, k, m_{r+1}, \ldots, m_p)$ as the candidate distribution, giving the acceptance probability as $\prod_{j=1}^{r} \{x/(m_j+x)\}^k$.

Again there is a clear trade-off. Making r bigger means less deeply nested recursion, but smaller acceptance probabilities.

ACKNOWLEDGEMENTS

This research was supported by grants from the Natural Science and Engineering Research Council of Canada and the Canadian Institutes of Health Research.

REFERENCES

S. Banerjee, B. P. Carlin & A. E. Gelfand (2004). *Hierarchical Modelling and Analysis for Spatial Data*. Chapman & Hall/CRC Press, Boca Raton.

J. O. Berger & J. J. Deely (1988). A Bayesian approach to ranking and selection of related means with alternative to analysis-of-variance methodology. *Journal of the American Statistical Association*, 83, 364–373.

M. J. Daniels (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, 27, 567–578.

P. J. Everson & C. N. Morris (2000). Inference for multivariate normal hierarchical models. *Journal of the Royal Statistical Society Series B*, 62, 399–412.

A. E. Gelfand, S. K. Sahu & B. P. Carlin (1995). Efficient parameterizations for normal linear mixed models. *Biometrika*, 82, 479–488.

A. Gelman (2006). Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis*, 1, 515–534.

A. Gelman, J. B. Carlin, H. S. Stern & D. B. Rubin (2004). *Bayesian Data Analysis*, Second edition. Chapman & Hall/CRC, Boca Raton.

W. R. Gilks & P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics: Journal of the Royal Statistical Society, Series C*, 41, 337–348.

P. J. Green & B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London.

P. Gustafson (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics*, 53, 230–242.

P. Gustafson (1998). Flexible Bayesian modelling for survival data. *Lifetime Data Analysis*, 4, 281–299.

S. Hossain (2003). *A Conservative Prior for Bayesian Hierarchical Models in Biostatistics.* Unpublished M.Sc. thesis, Department of Statistics, University of British Columbia, Vancouver.

Y. C. MacNab (2003). Hierarchical Bayesian modelling of spatially correlated health service outcome and utilization rates. *Biometrics*, 59, 305–316.

R. Natarajan & R. E. Kass (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95, 227–237.

D. J. Spiegelhalter, K. R. Abrams & J. P. Myles (2003). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Wiley, Chichester.

D. J. Spiegelhalter, L. S. Freedman & M. K. Parmar (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society Series A*, 157, 357–416.

D. Stangl & D. Berry (eds.) (2000). *Meta-Analysis in Medicine and Health Policy.* Marcel Dekker, New York.

I. Verdinelli & L. Wasserman (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, 90, 614–618.

Received 20 June 2005
Accepted 11 January 2006

Paul GUSTAFSON: gustaf@stat.ubc.ca
Shahadut HOSSAIN: shahadut@stat.ubc.ca
Department of Statistics
University of British Columbia
Vancouver, British Columbia, Canada V6T 1Z2

Ying C. MACNAB: ymacnab@interchange.ubc.ca
Department of Health Care and Epidemiology
University of British Columbia
Vancouver, British Columbia, Canada V6T 1Z3