Bayesian Analysis (2006) **1**, Number 3, pp. 535–542

A Default Conjugate Prior for Variance Components in Generalized Linear Mixed Models (Comment on Article by Browne and Draper)

Robert E. Kass[*] and Ranjini Natarajan[†]

Abstract. For a scalar random-effect variance, Browne and Draper (2005) have found that the uniform prior works well. It would be valuable to know more about the vector case, in which a second-stage prior on the random effects variance matrix **D** is needed. We suggest consideration of an inverse Wishart prior for **D** where the scale matrix is determined from the first-stage variance.

Keywords: Choice of prior, hierarchical models, noninformative priors, random effects

1 Comments

There is no standard solution to the problem of choosing a prior on the random-effects variance in random-effects models, or mixed models, or what Bayesian analysts usually call "hierarchical models." In the case of a scalar random effect, Browne and Draper (2005) investigated the frequentist behavior of posterior estimates based on a uniform prior and an inverted-gamma prior. They also compared the Bayesian methods to likelihood and quasi-likelihood alternatives.

The main Bayesian messages we take home from Browne and Draper's study are that, in the case of a scalar random effect, (1) a uniform prior on the variance produces posterior distributions with very good operating characteristics: the coverage probabilities remain close to .95 for all of their simulations; and (2) the uniform prior is a bit better than a quasi-uniform inverted-gamma prior. Though the situations for Normal and non-Normal models seem to us different in principle, with some kind of correction seeming necessary before prior rules for non-Normal models match those for the Normal models, the work by Browne and Draper strengthens an already strong case for the uniform prior becoming the "standard solution." The main general statistical message seems to be that this Bayesian method works well. We would underscore the additional general comment made by Browne and Draper, and many before them, that estimates of fixed effects remain very good in the presence of modest errors in estimation of the variance components. This is part of what makes generalized estimating equation estimators so effective (Diggle et al., 2002).

[*] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, http://www.stat.cmu.edu/~kass

[†] Unaffiliated

What happens in the vector case? As the dimensionality increases, one anticipates degradation of performance: the choice of prior is likely to matter much more, and one may expect trouble in estimating fixed effects, as well. It would be good to have results like those of Browne and Draper's so that we would know more precisely when to worry, and it would also be very valuable if the field could settle on a reasonable default prior for the non-worrisome and not-very-worrisome situations. The tradition in statistical research is to report results of the form "method A (often the authors' method) works better than method B." This is useful, but statisticians too rarely give practical guidance as to when a method breaks down.

Perhaps future studies of priors for random effects in the vector case will be undertaken. If so, we would like to make one more suggestion: it may be worthwhile to evaluate yet another prior, one we call a "default conjugate prior." In the remainder of our commentary we will describe this prior and indicate why we think it may be of use.

2 A Default Conjugate Prior

In the vector case, under the assumption of a Normal distribution for the random effects (the second stage of the hierarchical model), the uniform prior remains a reasonable candidate. It is also possible to use an inverted Wishart prior on the random effects variance matrix $\mathbf{D}$, which requires the specification of a scale matrix typically considered to be a guess at the value of $\mathbf{D}$. There is, however, rarely good scientific information on which to base this guess. A frequently-applied procedure is to set the scale matrix equal to the maximum likelihood estimator (MLE) of $\mathbf{D}$. Natarajan and Kass (2000) reported simulations indicating that posterior distributions based on this procedure can lead to poor estimates of $\mathbf{D}$, and we also gave a real-data example where scientific inferences are seriously affected. In that paper we also proposed an alternative — the "approximate uniform shrinkage" prior — and showed it to lead to better-behaved posteriors. That prior is easy enough to use, but has not caught on. We here draw attention to yet another alternative, namely the "default conjugate prior." Rather than using the MLE as the scale matrix of the inverse Wishart prior, it may be preferable to base a "guess" at the value of $\mathbf{D}$ on the first-stage data variability. Although the method uses first-stage data both for formulation of the second-stage prior and for computation of the posterior, we note that this particular re-use of the data has asymptotically negligible effects on the posterior.

2.1 The Two-Stage Hierarchical Model

Let us consider the following class of two-stage models:

$$\begin{aligned}
\mathbf{Y}_i|\mathbf{b}_i &\sim \prod_{j=1}^{n_i} f\left(Y_{ij}|\mathbf{b}_i,\boldsymbol{\beta}\right), \qquad i=1,\ldots,k,\ j=1,\ldots,n_i, \\
\mathbf{b}_i &\sim \mathrm{N}_q\left(\mathbf{0},\,\mathbf{D}\right), \qquad (1)
\end{aligned}$$

where $\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, \ldots, Y_{in_i}\right)^{\mathrm{T}}$ is a vector of observed responses for the i^{th} experimental unit (cluster), $\mathbf{b}_i$ is a $q \times 1$ vector of unobserved cluster-specific random effects and $f\left(.\right)$ is an exponential family density function with dispersion parameter ϕ assumed known. The conditional mean of Y_{ij} is assumed to satisfy $\mu_{ij}^{\mathbf{b}} = h\left(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i\right)$, where $\mathbf{x}_{ij}\ (p \times 1)$ and $\mathbf{z}_{ij}\ (q \times 1)$ are design vectors corresponding to the fixed effects $\boldsymbol{\beta}$ and the random effects $\mathbf{b}_i$ respectively and $h\left(.\right)$ is a known link function with inverse $g\left(.\right)$. Such models belong to the family of generalized linear mixed models (GLMMs). By way of notation we let $\mathbf{X}_i\ (n_i \times p)$ and $\mathbf{Z}_i\ (n_i \times q)$ denote full-rank matrices with rows $\mathbf{x}_{ij}^{\mathrm{T}}$ and $\mathbf{z}_{ij}^{\mathrm{T}}$, respectively.

2.2 Definition and motivation

In this section we assume the prior on $\boldsymbol{\beta}$ will be diffuse (in implementation, typically a multivariate Normal with large variances), and consider the problem of specifying the $q \times q$ scale matrix $\mathbf{R}$ of an inverted Wishart prior for $\mathbf{D}$. Specifically, a random positive-definite symmetric matrix $\mathbf{D}$ is distributed according to an inverted Wishart distribution with $\rho(> q-1)$ degrees of freedom and scale matrix $\mathbf{R}$ if its probability density function is proportional to $\det\left(\mathbf{D}\right)^{-(\rho+q+1)/2} exp\left(-\frac{\rho}{2}\mathrm{tr}\left(\mathbf{R}\mathbf{D}^{-1}\right)\right)$. We denote this inverted Wishart distribution by $\mathrm{IW}\left(\rho, \rho\mathbf{R}\right)$. Note that when $q = 1$, the inverted Wishart reduces to an inverted gamma distribution and ρ is typically referred to as the shape parameter. We will denote the inverted gamma by IG. Conventional wisdom dictates that a good default specification is one for which ρ is taken to be small and $\mathbf{R}$ is a “minimally informative” prior guess of $\mathbf{D}$.

We now define a default Wishart prior for $\mathbf{D}$ with

$$\begin{aligned} \rho &= q, \\ \tilde{\mathbf{R}} &= c \cdot \left(\frac{1}{k}\sum_{i=1}^{k}\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\left(\boldsymbol{\beta}\right)\mathbf{Z}_i\right)^{-1}, \end{aligned}$$

where $\mathbf{W}_i\left(\boldsymbol{\beta}\right)(n_i \times n_i)$ denotes the usual diagonal GLM weight matrix with diagonal elements $\left\{\phi v\left(\mu_{ij}^{\mathbf{0}}\right)\left[\partial g\left(\mu_{ij}^{\mathbf{0}}\right)/\partial\mu_{ij}^{\mathbf{0}}\right]^2\right\}^{-1}$, $v\left(.\right)$ is the known variance function based on the density $f\left(.\right)$ and the superscript zeros indicate the substitution of $\mathbf{b}_i$ with zero in these quantities. The value of c is an inflation factor representing the amount by which the within-cluster variability should be increased in determining R^*. In our simulation we used $c = 1$. Note that the inverse of $\frac{1}{k}\sum_{i=1}^{k}\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\left(\boldsymbol{\beta}\right)\mathbf{Z}_i$ exists by the full-rank assumption on $\mathbf{Z}_i$. Thus, calculation of $\tilde{\mathbf{R}}$ is straightforward, requiring only a few matrix operations and knowledge of the form of the weight matrix $\mathbf{W}_i$ for the particular exponential family under consideration McCullagh and Nelder (1989), pp. 30.

We now offer two heuristic justifications for $\tilde{\mathbf{R}}$. The first arises from the approximate shrinkage estimate of $\mathbf{b}_i$ — that is, $\widetilde{\mathbf{b}}_i = \mathbf{D}\mathbf{Z}_i^{\mathrm{T}}\left(\mathbf{W}_i^{-1}\left(\boldsymbol{\beta}\right) + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{\mathrm{T}}\right)^{-1}\left(\mathbf{Y}_i^* - h\left(\mathbf{X}_i\boldsymbol{\beta}\right)\right)$ where $\mathbf{Y}_i^*$ is the working dependent variable Breslow and Clayton (1993). After some matrix manipulations, it can be shown that $\widetilde{\mathbf{b}}_i$ may be expressed as

$$\widetilde{\mathbf{b}}_i \quad = \quad \mathbf{S}_i\mathbf{0} + (\mathbf{I} - \mathbf{S}_i)\,\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,(\mathbf{Y}_i^* - h\,(\mathbf{X}_i\boldsymbol{\beta}))\,,$$

where $\mathbf{I}$ is the $q \times q$ identity matrix and $\mathbf{S}_i = \mathbf{I} - \left(\mathbf{D}^{-1} + \mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$. The matrix $\mathbf{S}_i$ controls the relative contribution of the prior mean $\mathbf{0}$ and the data to the posterior update of $\mathbf{b}_i$, and thus offers a natural metric for evaluating the informativeness of a particular prior guess for $\mathbf{D}$. It ranges from $\mathbf{I} - \left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$ when $\mathbf{D} = \infty$, which corresponds to a flat prior for $\mathbf{b}_i$, to $\mathbf{I}$ when $\mathbf{D} = \mathbf{0}$, which corresponds to a point mass prior for $\mathbf{b}_i$ at zero. A prior guess of $\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$ for $\mathbf{D}$ would result in a weight of $\mathbf{I} - \frac{1}{2}\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$, which is exactly half-way between the weights accorded by the two extreme choices of $\mathbf{D}$. Thus, this seems like a reasonable guess for $\mathbf{D}$ in the absence of any other prior knowledge. However, since $\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$ varies with i, we suggest replacing it with its harmonic mean over clusters, which leads to our choice of $\widetilde{\mathbf{R}}$.

A second justification arises from considering a maximum likelihood-based Normal approximation to the GLMM in which the exponential family specification is replaced with

$$\widehat{\mathbf{b}}_i \quad \sim \quad \mathrm{N}_q\left(\mathbf{b}_i, \mathbf{I}\left(\widehat{\mathbf{b}}_i\right)\right),$$

where $\widehat{\mathbf{b}}_i$ is the ML estimator of $\mathbf{b}_i$ based on the first-stage likelihood $\prod_{j=1}^{n_i} f\,(Y_{ij}|\mathbf{b}_i, \boldsymbol{\beta})$, and $\mathbf{I}\left(\widehat{\mathbf{b}}_i\right)$ is the observed information evaluated at $\widehat{\mathbf{b}}_i$. It can be shown that $\mathbf{I}\left(\widehat{\mathbf{b}}_i\right) = \left(\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{W}}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$, where $\widehat{\mathbf{W}}_i\,(\boldsymbol{\beta})$ is the GLM weight matrix $\mathbf{W}_i$ defined previously but with $\widehat{\mathbf{b}}_i$ in place of zero. However, when $\widehat{\mathbf{W}}_i\,(\boldsymbol{\beta})$ is close to $\mathbf{W}_i\,(\boldsymbol{\beta})$, the within-cluster variance $\mathbf{I}\left(\widehat{\mathbf{b}}_i\right)$ will be approximated well by $\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$. Thus, a prior guess of $\left(\frac{1}{k}\sum_{i=1}^{k}\mathbf{Z}_i^{\mathrm{T}}\mathbf{W}_i\,(\boldsymbol{\beta})\,\mathbf{Z}_i\right)^{-1}$ for $\mathbf{D}$, corresponds roughly to an *a priori* belief that the between-cluster variance is equal to the harmonic mean of the within-cluster variance.

Note that our specification for the prior on $\mathbf{D}$ depends on $\boldsymbol{\beta}$ through $\mu_{ij}^{\mathbf{b}}$, which appears in $\mathbf{W}_i\,(\boldsymbol{\beta})$, and is thus a specification of the conditional distribution of $\mathbf{D}$ given $\boldsymbol{\beta}$. A consequence of this appearance of $\boldsymbol{\beta}$ is that the full conditional distribution of $\boldsymbol{\beta}$ given the data and all other parameters will no longer be free of $\mathbf{D}$. Although this presents no substantial difficulties, the simplicity of the standard assumption of independence of $\boldsymbol{\beta}$ and $\mathbf{D}$ (together with a uniform or Normal prior on $\boldsymbol{\beta}$) enables particularly straightforward MCMC implementation via Gibbs sampling (Zeger and Karim, 1991). Thus, we propose a slight modification to the prior given above: we replace the family of conditional distributions of $\mathbf{D}$ given $\boldsymbol{\beta}$ by the single conditional distribution of $\mathbf{D}$ given $\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is an estimate of the regression coefficients from the GLM model obtained by pooling all the data and setting $\mathbf{b}_i = \mathbf{0}$, for all i. That is, we specify the default inverted Wishart by $\rho = q$, and

$$\mathbf{R}^* = c \cdot \left(\frac{1}{k} \sum_{i=1}^{k} \mathbf{Z}_i^{\mathrm{T}} \mathbf{W}_i \left(\widehat{\boldsymbol{\beta}} \right) \mathbf{Z}_i \right)^{-1} . \tag{2}$$

Note that $\mathbf{W}_i \left(\widehat{\boldsymbol{\beta}} \right)$ is a $O_p \left(k^{-1/2} \right)$ consistent estimator of $\mathbf{W}_i \left(\boldsymbol{\beta} \right)$, and that the estimate $\widehat{\boldsymbol{\beta}}$ may be obtained in a simple pre-calculation.

2.3 Asymptotic irrelevance of the data-dependence in the modified prior

Our modified default prior now depends on the data through the replacement of the conditional prior $\pi \left(\mathbf{D} | \boldsymbol{\beta} \right)$ with $\pi \left(\mathbf{D} | \widehat{\boldsymbol{\beta}} \right)$. It is possible for such a data-dependent substitution to yield very misleading inferences. For example, in the one-sample Normal problem using the conjugate family of prior distributions on the mean μ and variance σ^2: $\pi \left(\mu | \sigma^2 \right) = N \left(\mu_0, \lambda_0 \sigma^2 \right)$, $\pi \left(\sigma^2 \right) = \mathrm{IG} \left(\alpha_0, \beta_0 \right)$, one might take $\widehat{\sigma}$ to be the standard error of the sample mean and substitute it for σ in the Normal prior $\pi \left(\mu | \sigma^2 \right)$. This results in a prior whose informativeness is derived from the data; indeed, it would count the data twice, and is clearly an unreasonable procedure. The substitution we have made, however, is quite different: it does not carry the same amount of information as the full data set, but in fact carries less information than does a single observation (that is, a single cluster).

More formally, let $\boldsymbol{\lambda} = \left(\boldsymbol{\beta}, \mathbf{D} \right)$, $\pi_{def} \left(\boldsymbol{\lambda} \right)$ and $\pi_{mod} \left(\boldsymbol{\lambda} \right)$ be the original default conjugate prior and its modification, and $q \left(\boldsymbol{\lambda} \right)$ be any alternative non-data-dependent prior. Also, let $G \left(\boldsymbol{\lambda} \right)$ be a function to be estimated and let $E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, \pi_{def} \right)$ be the posterior expectation of $G \left(\boldsymbol{\lambda} \right)$ based on $\pi_{def} \left(\boldsymbol{\lambda} \right)$, and similarly for the other two priors. Then, as $k \to \infty$, we have

$$E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, \pi_{def} \right) = E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, q \right) \left(1 + O_p(k^{-1}) \right) , \tag{3}$$

which is one way of saying that, in large samples, the effect of changing the prior is roughly that of changing a single observation. If an informative data-dependent prior were used (analogous to that mentioned for the one-sample Normal) in place of $q \left(\boldsymbol{\lambda} \right)$, Equation (3) would no longer hold. Our modified prior produces

$$E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, \pi_{def} \right) = E \left(G \left(\boldsymbol{\lambda} \right) | \mathbf{Y}, \pi_{mod} \right) \left(1 + O_p(k^{-1}) \right) . \tag{4}$$

This result may be obtained from asymptotic expansions, as in Kass and Steffey (1989), Equation (3.14), using the MLE-based version mentioned just after that equation). The essential observation is that for any $\boldsymbol{\lambda}$ within an order $O(k^{-1/2})$ neighborhood of the true value (toward which a $\sqrt{k}$-consistent estimator will converge) the ratio of the original to modified priors satifies $\pi_{def} \left(\boldsymbol{\lambda} \right) / \pi_{mod} \left(\boldsymbol{\lambda} \right) = 1 + O_p \left(k^{-1/2} \right)$.

Kass and Steffey (1989) pointed out that when the empirical Bayes substitution of an MLE of $\boldsymbol{\lambda}$ is made, the resulting posterior variance of a random effect is too small, and no longer approximates to order $O_p(k^{-1})$ the correct posterior variance. This is another example of the use of data-dependent priors that may have strong, undesirable effects on inference. It is worth noting, again by way of contrast, that an expression analogous to (4) holds for posterior variances:

$$\text{var}\left(G\left(\boldsymbol{\lambda}\right)|\mathbf{Y},\pi_{def}\right)=\text{var}\left(G\left(\boldsymbol{\lambda}\right)|\mathbf{Y},\pi_{mod}\right)\left(1+O_p(k^{-1})\right).$$

2.4 Simulation study

We ran three simulations, with generally similar results, and report the most dramatic of them here. Unfortunately, while this illustrates the potential value of the default conjugate prior, it is yet again a scalar example.

We compared the performance of the default conjugate prior with three other priors: an inverted Wishart with $\rho = q$ and $\mathbf{R}$ given by the MLE of $\mathbf{D}$, an "ideal" inverted Wishart with $\rho = q$ and $\mathbf{R}$ given by the true value of $\mathbf{D}$, and the approximate uniform shrinkage prior π_{us} (Natarajan and Kass, 2000). The ideal prior provides an unattainable target for the other Wishart priors.

All priors were used in conjunction with a uniform prior for $\boldsymbol{\beta}$. The conditions under which this gives a proper posterior for GLMMs has been derived by Natarajan and Kass (2000), and were verified for the data here. Inferences for the four priors were based on 2,000 samples generated from their posterior distributions for each data set. Posterior sampling was performed using the Gibbs sampler and followed the implementation described by Zeger and Karim (1991) for the inverted Wishart priors, and Natarajan and Kass for π_{us}.

Breslow (1984) presented mutagenicity assay data on the number of revertant colonies of TA98 Salmonella (Y) at six doses of quinoline $(x = 0, 10, 33, 100, 333, 1000)$. Three plates were processed at each of the six dose levels resulting in a total of 18 observations. He considered the following Poisson GLMM for these data:

$$\begin{aligned} Y_i|b_i &\sim \text{Poisson}\left(\mu_i^b\right), \quad i=1,\ldots,18,\\ b_i &\sim \text{N}\left(0,\theta\right), \end{aligned} \tag{5}$$

with $\mu_i^b = exp\left(\beta_0+\beta_1\ln\left(x_i+10\right)+\beta_2 x_i+b_i\right)$. The single variance component θ captures the overdispersion due to plate-to-plate variability. The default conjugate prior is $\text{IG}\left(1,R^*\right)$ where $R^* = 18/\sum_{i=1}^{18} w_i$, $w_i = exp\left(\widehat{\beta}_0+\widehat{\beta}_1\ln(x_i+10)+\widehat{\beta}_2 x_i\right)$ and we estimated $\widehat{\boldsymbol{\beta}}$ from the first-stage Poisson likelihood function with $b_i = 0$. The approximate uniform shrinkage prior is $\pi_{us}\left(\theta\right)\propto 1/\left(1+\theta/R^*\right)^2$.

We generated 1,000 data sets from (5) with $\beta_0 = 2.203$, $\beta_1 = .311$, $\beta_2 = -.001$ and $\theta = .040$. These values were chosen because they are close to the estimates obtained for the salmonella data. The estimators of $\boldsymbol{\beta}$ and θ from the four priors were evaluated according to posterior risk and noncoverage probabilities for 95% posterior intervals (the noncoverage probabilities would, ideally, equal .05). The posterior risk was calculated under the squared-error loss function $L\left(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}\right)=\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ for $\boldsymbol{\beta}$, and the entropy loss function $L\left(\widehat{\theta},\theta\right) = \left(\widehat{\theta}/\theta - \ln|\widehat{\theta}/\theta| - 1\right)$ for θ. The Bayes estimators corresponding to these loss functions are the posterior mean and harmonic mean respectively. Note that the entropy loss function penalizes underestimation more severely than overestimation in cases when the true value of θ is close to zero. Thus, we would expect the prior π_{us} to have a slightly worse risk than the other priors since it places non-zero mass at zero.

Operating Characteristics	IW $(1,\theta)$	IW $\left(1,\widehat{\theta}\right)$	IW $(1,R^*)$	π_{us}
Risk				
$\boldsymbol{\beta}$	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
θ	.09 ± .00	.89 ± .05	.12 ± .00	.62 ± .02
Noncoverage				
β_0	.046 ± .007	.076 ± .008	.056 ± .007	.070 ± .008
β_1	.054 ± .007	.086 ± .009	.059 ± .007	.067 ± .008
β_2	.051 ± .007	.081 ± .009	.060 ± .007	.075 ± .008
θ	.011 ± .003	.194 ± .012	.007 ± .003	.037 ± .006

Table 1: Simulation results: risk and noncoverage probability. $IW(1,\theta)$ denotes the ideal diffuse conjugate prior based on the unknown true value $\theta = .04$. Note that in this one-dimensional case the inverse-Wishart becomes an inverse-gamma. $IW(1,\hat{\theta})$ denotes the diffuse conjugate prior based on the MLE $\hat{\theta}$. $IW(1,R^*)$ denotes the diffuse conjugate prior based on Equation (2), with $c = 1$. The average value of MLE, across the 1,000 data sets, was $\hat{\theta} = .027$ while the average value of R^* was $R^* = .033$.

Table 1 displays these results for $\boldsymbol{\beta}$ and θ under the four priors. An examination of the results shows that the inverted Wishart prior (here, an inverted gamma) centered at the MLE $\widehat{\theta}$ is dominated by the other priors, both in terms of risk and coverage probabilities. The poor risk of this prior is a consequence of the tendency of the MLE to underestimate the true value, while the worse coverage probabilities are due to its failure to account for the extra variability induced by plugging in $\widehat{\theta}$. The default conjugate prior is fairly competitive with the ideal prior and offers slightly better inferences than π_{us} for the regression coefficients.

2.5 Conclusions

There is not much knowledge about the performance of posteriors based on alternative priors for the matrix $\mathbf{D}$ in models of the form (1). The very limited results we have managed to present here are intended to offer the default conjugate prior in (2) as a plausible choice, and we would expect good results for this prior when c is chosen well. Possibly c could be estimated from the data. We hope the future will bring practical guidance as to when posteriors based on priors for $\mathbf{D}$, including the uniform prior, the default conjugate prior, or other interesting choices such as that recommended by Gelman (2005), are likely to have good frequentist operating characteristics.

References

Breslow, N. E. (1984). "Extra-Poisson variation in log-linear models." *Applied Statistics*, 33: 38–44. 540

Breslow, N. E. and Clayton, D. G. (1993). "Approximate Inference in Generalized Linear Mixed Models." *JASA*, 88: 9–25. 537

Browne, W. J. and Draper, D. (2005). "A comparison of Bayesian and likelihood-based methods for fitting multilevel models." *Bayesian Analysis*. To appear. 535

Diggle, P. J., Haegerty, P., Liang, K. Y., and Zeger, S. L. (2002). *The analysis of Longitudinal Data*. Oxford University Press, 2nd edition.

Gelman, A. (2005). "Prior distributions for variance parameters in hierarchical models (Comment on an article by Browne and Draper)." *Bayesian Analysis*. To appear. 542

Kass, R. E. and Steffey, D. (1989). "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)." *JASA*, 84: 717–726. 539

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd edition. 537

Natarajan, R. and Kass, R. E. (2000). "Reference Bayesian methods for generalized linear mixed models." *JASA*, 95: 227–237. 536, 540

Zeger, S. L. and Karim, M. R. (1991). "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach." *JASA*, 86: 79–86. 540