The Canadian Journal of Statistics
Vol. 27, No. 3, 1999, Pages 567–578
La revue canadienne de statistique

A prior for the variance in hierarchical models

Michael J. DANIELS

Key words and phrases: Noninformative prior; shrinkage; multi-level models.

AMS 1991 subject classifications: Primary 62F15; secondary 62J12.

ABSTRACT

The choice of prior distributions for the variances can be important and quite difficult in Bayesian hierarchical and variance component models. For situations where little prior information is available, a 'noninformative' type prior is usually chosen. 'Noninformative' priors have been discussed by many authors and used in many contexts. However, care must be taken using these prior distributions as many are improper and thus, can lead to improper posterior distributions. Additionally, in small samples, these priors can be 'informative'. In this paper, we investigate a proper 'vague' prior, the uniform shrinkage prior (Strawderman 1971; Christiansen & Morris 1997). We discuss its properties and show how posterior distributions for common hierarchical models using this prior lead to proper posterior distributions. We also illustrate the attractive frequentist properties of this prior for a normal hierarchical model including testing and estimation. To conclude, we generalize this prior to the multivariate situation of a covariance matrix.

RÉSUMÉ

Le choix d'une loi a priori pour les variances peut s'avérer à la fois difficile et important dans le cadre d'une analyse bayésienne hiérarchique ou d'un modèle des composantes de la variance. En l'absence totale ou quasi-totale d'information a priori, l'emploi d'une loi 'non informative' est de mise. Plusieurs lois de ce type ont été proposées dans différents contextes, mais leur utilisation est délicate, puisque certaines d'entre elles sont impropres et peuvent conduire à des lois a posteriori non intégrables. Dans de petits échantillons, ces lois peuvent aussi se révéler 'informatives'. Cet article est consacré à l'étude d'une loi a priori à la fois vague et intégrable, la loi a priori à rétrécissement uniforme (Strawderman 1971; Christiansen & Morris, 1997). Certaines de ses propriétés sont évoquées, notamment le fait que les lois a posteriori auxquelles elle conduit dans certains modèles hiérarchiques classiques sont bel et bien intégrables. Ses propriétés fréquentistes sont également mises en valeur dans des situations d'estimation et de test au sein du modèle hiérarchique gaussien. On montre en outre comment elle peut être généralisée au cas multivarié d'une matrice de covariance.

1. INTRODUCTION

In a Bayesian analysis of hierarchical or variance component models, the choice of prior for the variances is important and may be difficult. The most commonly used prior for a variance component is the inverse Gamma distribution, which is the conjugate distribution of the normal distribution (e.g., Gelman, Carlin, Stern & Rubin 1995, p. 71), but in many applications little or no prior information about the variance components is available. As a result, one often desires a 'vague' or 'noninformative' prior distribution to reflect this uncertainty. One possibility is to place an inverse Gamma distribution with large variance on the variance component. However, as its variance approaches infinity (while the shape parameter is held fixed), this prior approaches an 'improper' flat prior on the variance component, which can lead to improper posterior distributions and tends to inflate the variance component for small samples. Other 'noninformative' priors have been discussed by many authors (e.g., Jeffreys 1961; Box & Tiao 1973; Berger & Deely 1988; Berger & Bernardo 1992). The purpose of this paper is to investigate a prior, the uniform shrinkage prior, first suggested by Strawderman (1971) and later generalized by Morris (e.g., Christiansen & Morris 1997).

In Section 2, we derive and discuss properties of the uniform shrinkage prior. We give examples of some common hierarchical regression models with corresponding shrinkage priors and show that the posteriors are proper in Section 3. Frequentist procedures based on this prior are explored in Section 4. Specifically, we examine testing (Bayes factors), point estimation, and interval estimation. Finally, we derive the multivariate analogue to the uniform shrinkage prior.

2. DEFINITION AND RELATION TO OTHER PRIORS

Consider the following normal-normal hierarchical model:

$$Y_i|\theta_i \sim \mathrm{N}(\theta_i, \sigma^2), \quad \theta_i|\tau^2 \sim \mathrm{N}(0, \tau^2), \quad \tau^2 \sim \pi(\tau^2).$$

Some of the common 'noninformative' prior choices for $\pi(\tau^2)$ appear in Table 1 and Figure 1, including Jeffreys prior (Jeffreys 1961) for the above model and a proper version of Jeffreys prior (Berger & Deely 1988).

TABLE 1: Common 'noninformative' priors for a variance component.

prior	form
flat	K (constant)
location-scale	$1/\tau^2$
right invariant Haar density	$1/\tau$
Jeffreys prior	$1/(\sigma^2 + \tau^2)$
proper Jeffreys	$\sigma/\{2(\sigma^2 + \tau^2)^{3/2}\}$
uniform shrinkage	$\sigma^2/\{(\sigma^2 + \tau^2)^2\}$
DuMouchel	$\sigma/\{2\tau(\sigma + \tau)^2\}$

An alternative proper prior, shown to give a minimax estimator for the k means for the above model by Strawderman (1971) under certain conditions, would be to place a uniform prior on the shrinkage parameter, $S = \sigma^2/(\sigma^2 + \tau^2)$. The shrinkage parameter here refers to the weight placed on the prior mean for the posterior mean of θ_i in the above model:

$$\mathrm{E}(\theta_i|\sigma^2, \tau^2, y) = \frac{\tau^2}{\sigma^2 + \tau^2}Y_i + \frac{\sigma^2}{\sigma^2 + \tau^2}0 = \frac{\tau^2}{\sigma^2 + \tau^2}Y_i.$$

This induces the following form for the prior for τ^2, $\pi(\tau^2) = \sigma^2/(\sigma^2 + \tau^2)^2$. This form looks similar to the prior, used by DuMouchel in his hierarchical meta-analysis models (DuMouchel 1994), $\pi(\tau) = \sigma/(\sigma + \tau)^2$ [$\pi(\tau^2) = \sigma/(\sigma + \tau)^2\{1/(2\tau)\}$]. His prior is derived from a log-logistic model; however, it possesses similar properties to the uniform shrinkage prior.

The uniform shrinkage prior has several attractive properties as discussed by various authors including:

- the prior is proper, which is crucial for comparison of models using Bayes Factors;
- the density for the variance parameter or component is maximized at zero and decreases with the variance parameter; this is advantageous in that prior weight is given to complete shrinkage to the prior mean;
- the density is rather diffuse and noninformative; the expectation of the variance parameter and its inverse are both infinite.
- the prior is easily generalized to many conjugate models (see below).

The uniform shrinkage prior can easily be derived for any conjugate model. For example, one has $\tau^2 \sim \pi(\tau^2)$ and $\boldsymbol{\beta} \sim d\boldsymbol{\beta}$ in the general two-level model

$$Y_i|\theta_i \sim f(y_i;\theta_i), \quad \theta_i|m_i,\tau^2 \sim g(\theta_i;m_i,\tau^2),$$

where $\mathrm{E}(\theta_i|m_i,\tau^2) = m_i = \ell^{-1}(\mathbf{X}_i'\boldsymbol{\beta})$ with ℓ any appropriate link function and g is a distribution conjugate to f. Using the conjugate structure of the first two levels of the model and conditioning on the regression parameter vector $\boldsymbol{\beta}$ and the variance parameter τ^2, we can write the posterior mean of the θ_i as

$$\mathrm{E}(\theta_i|\boldsymbol{\beta},\tau^2,y_i) = S(\tau^2)\hat{\theta}_i(y_i) + \{1 - S(\tau^2)\}\ell^{-1}(\mathbf{X}_i'\boldsymbol{\beta})$$

where $\hat{\theta}_i(y_i)$ is some function of the data y_i, and $S(\tau^2)$ is the shrinkage parameter, constrained to $(0,1)$, which is a function of τ^2. We place a uniform distribution on the shrinkage parameter $S(\tau^2)$, and then do the appropriate transformation to obtain the prior for τ^2, viz.

$$\pi(\tau^2) = dS^{-1}(\tau^2)/d\tau^2.$$

3. EXAMPLES AND PROPRIETY OF THE POSTERIOR

Often, the prior for τ^2 will involve other unknown parameters indexed by i, such as m_i in the above model or σ_i^2 for the normal-normal model. In this situation the shrinkage parameter will be a function of i, e.g., $\sigma_i^2/(\sigma_i^2 + \tau^2)$, and an appropriate value must be chosen to substitute for σ_i^2 or whatever parameter indexed by i is included in the prior. Choice of these constants for Poisson-Gamma models and for a generalized two-level conjugate structure is discussed in Christiansen & Morris (1997) and Daniels & Gatsonis (1999) respectively. For the normal model, the harmonic mean of the σ_i^2 is a simple choice (DuMouchel 1994). Table 2 presents the prior for several common situations and also, reasonable choices for the constants. In this table, the t_i stand for fixed exposures. Of interest would be whether the uniform shrinkage prior leads to a proper posterior for the standard exchangeable and nonexchangeable normal-normal model with a flat prior on the common mean (μ) or the regression coefficient (β) (Berger 1985, p. 183):

$$Y_i|\theta_i \sim \mathrm{N}(\theta_i,\sigma^2), \qquad \theta_i|\tau^2 \sim \mathrm{N}(\theta,\tau^2), \qquad \tau^2 \sim \pi(\tau^2), \qquad \theta \sim d\theta.$$

For this model, the posterior will be proper with no restrictions for the uniform shrinkage prior or the proper version of Jeffreys prior on τ^2. However, if we allow σ^2 to vary with i and set $\theta_i = \mathbf{X}_i'\beta$ (nonexchangeable model), then the posterior will be proper if there are $k - 1$ or more observations, where k is the dimension of β. The proof follows simply from Berger (1985, pp. 190–191). For the same model, $1/\tau^2$ will lead to an improper posterior, but the flat prior and $1/\tau$ will lead to proper posteriors (Hobert & Casella 1994). For the Beta-Binomial and Poisson models, the uniform shrinkage prior on τ^2 will lead to a proper posterior for a flat prior on the regression coefficients. This is proven in the appendix. Obviously, for non-normal hierarchical models with a proper prior on the regression coefficients, the uniform shrinkage prior will lead to a proper posterior as the uniform shrinkage prior is itself proper.

The propriety of the posterior is especially important for Gibbs sampling. As discussed in Hobert & Casella (1996), a model can have all its full conditional distributions proper, but still have an improper posterior. This can lead to problems as the Gibbs chain may not indicate the impropriety of the posterior. In the above, we have shown that this uniform shrinkage prior leads to a proper posterior distribution for several common models. The only disadvantage of the prior relative to the flat prior and $1/\tau$ prior for normal-normal models is computational. The full conditional distributions for τ^2 for the above two priors will be inverse gamma. However, recent advances discussed in Everson & Morris (2000) provide simple ways to generate τ^2 using the uniform shrinkage prior in the normal-normal model. For the non-normal models, a Metropolis step (Smith & Roberts 1993) will be required regardless of the prior.

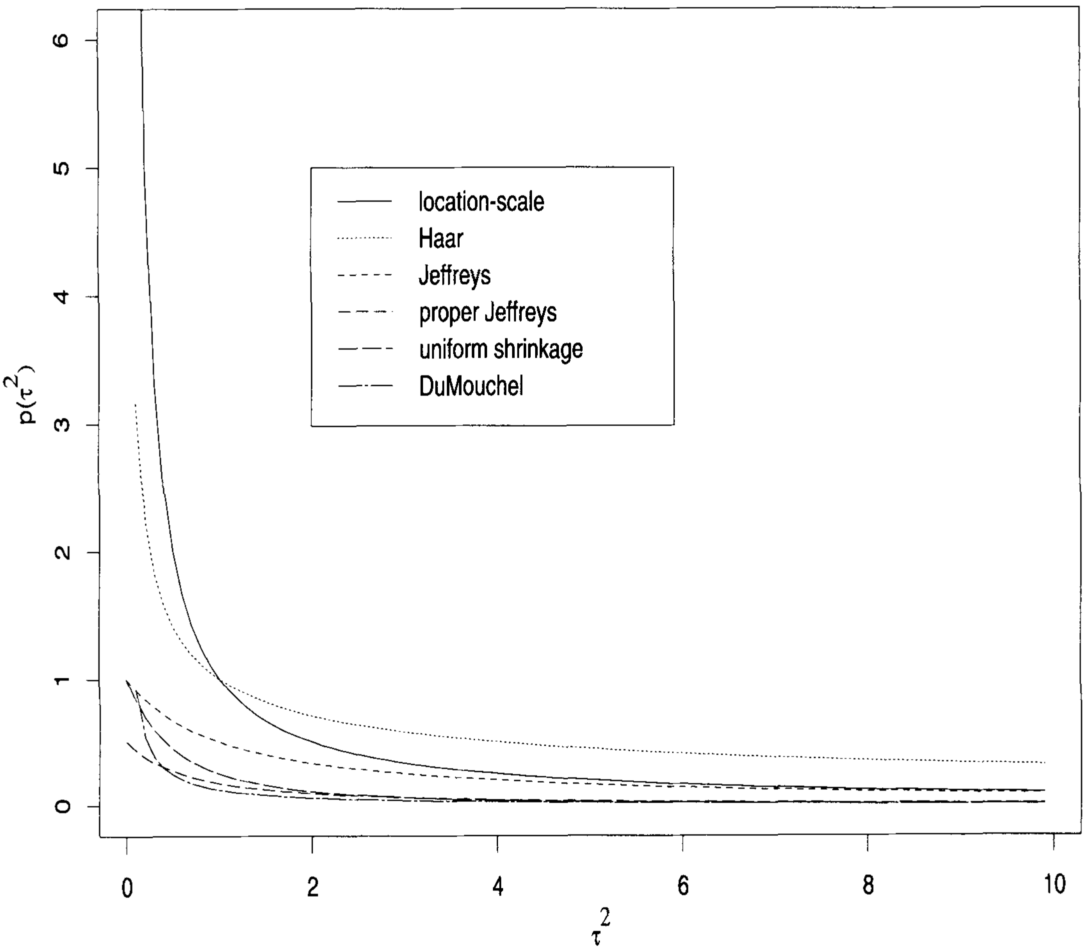

FIGURE 1. Behavior of common priors for variance components (with $\sigma^2 = 1$).

4. FREQUENTIST PROPERTIES OF THE UNIFORM SHRINKAGE PRIOR

The uniform shrinkage prior has been shown to have desirable theoretical properties. Of interest in addition would be to explore frequentist properties of the uniform shrinkage prior. First, we look at how the prior performs for testing and then for estimation.

4.1. *Testing* $\tau^2 = 0$.

To examine the properties of the uniform shrinkage prior for testing $H_0 : \tau^2 = 0$ vs $H_a : \tau^2 > 0$, we compute Bayes Factors (which require a proper prior for τ^2). Bayes factors (Jeffreys 1961) are defined as the ratio of the marginal likelihood under H_0 $\{m(y|\tau^2 = 0)\}$ to the marginal likelihood under H_a $\{m(y)\}$ and have the interpretation as the ratio of the posterior odds of H_0 to its prior odds. Here, $m(y)$ indicates the marginal distribution of the data after integrating out θ and τ^2. Simulations were run to look at the type I and type II error for the Bayes Factor (B) for testing $H_0 : \tau^2 = 0$ using the uniform shrinkage prior, the DuMouchel prior, and the proper version of Jeffreys prior for τ^2 using the normal-normal model above. The denominator of the Bayes Factor, $B = m(y|\tau^2 = 0)/m(y)$ was computed using quadrature (Gauss-Kronrod rules) to numerically integrate out $\log(\tau^2)$. In the table the $10/20/100$ correspond to the several sample sizes of the simulated dataset. Each simulation was run for 1000 times. For both types of errors, a correct decision was made if the Bayes Factor was on the correct side of 1 (Table 3).

TABLE 2: Prior for several common conjugate models based on the shrinkage parameter. (*Various choices discussed in Christiansen & Morris 1997.)

$Y_i\|\theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$	$\theta_i\|\beta, \tau^2 \sim \mathrm{N}(\mathbf{X}_i'\beta, \tau^2)$	$\frac{\sigma_c^2}{(\sigma_c^2+\tau^2)^2}$	harmonic mean
$Y_i\|\theta_i \sim \mathrm{Bin}(n_i, \theta_i)$	$\theta_i\|\beta, \tau^2 \sim \mathrm{Beta}(\tau^2 m_i, \tau^2(1-m_i))$	$\frac{n_c}{(n_c+\tau^2)^2}$	min, max, med(n_i)
$Y_i\|\theta_i \sim \mathrm{Poisson}(\theta_i)$	$\theta_i\|\beta, \tau^2 \sim \mathrm{Gamma}(\tau^2 m_i, \tau^2)$	$\frac{1}{(1+\tau^2)^2}$	-
$Y_i\|\theta_i \sim \mathrm{Poisson}(\theta_i)$	$\theta_i\|\beta, \tau^2 \sim \mathrm{Gamma}(\tau^2, \tau^2/m_i)$	$\frac{m_c}{(m_c+\tau^2)^2}$	sample mean
$Y_i\|\theta_i \sim \mathrm{Poisson}(\theta_i t_i)$	$\theta_i\|\beta, \tau^2 \sim \mathrm{Gamma}(\tau^2 m_i, \tau^2)$	$\frac{t_c}{(t_c+\tau^2)^2}$	*
$Y_i\|\theta_i \sim \mathrm{Poisson}(\theta_i t_i)$	$\theta_i\|\beta, \tau^2 \sim \mathrm{Gamma}(\tau^2, \tau^2/m_i)$	$\frac{t_c m_c}{(t_c m_c+\tau^2)^2}$	*
$Y_i\|\theta_i \sim \mathrm{Exp}(\theta_i)$	$\theta_i\|\beta, \tau^2 \sim \mathrm{Gamma}(\tau^2 m_i, \tau^2)$	$\frac{y_c}{(y_c+\tau^2)^2}$	sample mean

TABLE 3: Type I and II errors computed under sample sizes of 10, 20 & 100.

Error	σ^2/τ^2	uniform shrinkage prior	DuMouchel prior	Proper Jeffreys
		10/20/100	10/20/100	10/20/100
Type I	-	.11/.09/.05	.06/.07/.04	.18/.12 /.06
Type II	1/5	.02/.00/.00	.01/.00/.00	.02/.00 /.00
Type II	1	.37/.20/.00	.46/.26/.00	.31/.16 /.00
Type II	5	.76/.76/.64	.82/.82/.67	.67/.72/.62

The properties of the Bayes Factor here are indicative of the three priors. The DuMouchel prior places much more weight than the other two priors very close to zero; the shrinkage and proper Jeffreys prior place more weight than DuMouchel's on intermediate values and as $\tau^2 \to \infty$; the uniform shrinkage prior approaches 0 the quickest. The Bayes factors computed from all three priors appear to have reasonable frequentist properties. Note that in the last row of the table, $\sigma^2/\tau^2 = 5$; the sampling variability being as much as five times the size of the 'heterogeneity' is not a very common scenario in real world applications.

4.2. Estimation of τ^2.

Using the same normal-normal model, we simulated data assuming several values for τ^2 and estimated τ^2 (posterior mode) for several choices of the prior and computed the mean squared error (mse). We again ran simulations with sample sizes of 10, 20 and 100, and estimated τ^2 with the uniform shrinkage prior, the DuMouchel prior, the flat prior, the location/scale prior, the Jeffreys prior, and a proper version of the Jeffreys prior, with σ^2 fixed at 1 (Table 4).

TABLE 4: MSE for various priors for τ^2 with $\sigma^2 = 1$ fixed.
(*MSE computed under sample sizes of 10, 20 & 100.)

prior	$\tau^2 = 0$	$\tau^2 = .2$	$\tau^2 = 1$	$\tau^2 = 5$
	10/20/100*	10/20/100*	10/20/100*	10/20/100*
shrinkage	.016/.016/.005	.066/.054/.021	.595/.365/.080	7.37/3.66/.783
DuMouchel	.005/.003/.000	.067/.057/.040	.833/.636/.257	7.45/3.60/.779
location/scale	.026/.010/.000	.120/.077/.040	.827/.579/.200	7.10/3.57/.784
Jeffreys	.043/.032/.007	.115/.073/.022	.597/.341/.078	6.55/3.41/.775
proper Jeffreys	.026/.023/.006	.084/.062/.021	.581/.348/.079	6.83/3.48/.776
Flat	.120/.061/.009	.254/.113/.024	.827/.378/.080	7.75/3.77/.796

TABLE 5: 90% coverage probabilities for the uniform shrinkage prior for τ^2 with $\sigma^2 = 1$ fixed.

	Sample size 10			Sample size 20			Sample size 100		
σ^2/τ^2	upper	lower	total	upper	lower	total	upper	lower	total
1/5	.13	.02	.85	.10	.03	.87	.06	.03	.91
1	.01	.02	.97	.06	.02	.92	.07	.04	.89
5	.00	.03	.97	.00	.04	.96	.01	.04	.95

As expected, the DuMouchel and location/scale prior were best (in terms of mean squared error) at estimating τ^2 when $\tau^2 = 0$ with the shrinkage close behind. The worst for estimating $\tau^2 = 0$ was the flat prior which tends to skew estimates toward larger values of τ^2. The ability to detect the situation of no heterogeneity ($\tau^2 = 0$) is often important. For the common scenario of $\sigma^2/\tau^2 = 1$, that is the within-individual variation about the same magnitude as the between-individual variation, the shrinkage and Jeffreys priors performed well with the other priors doing considerably worse. For τ^2 large relative to σ^2 and vice versa, the Jeffreys and shrinkage performed best, respectively. Except for τ^2 and σ^2 of equal magnitude, the flat prior led to the worst estimators.

4.3. Interval Estimation.

Another frequentist quantity of interest is the coverage probabilities for interval estimates of τ^2 when using the uniform shrinkage prior. As above, using simulations of size 1000, we computed the 5% and 95% posterior quantiles for the posterior distribution of τ^2 using the uniform shrinkage prior and computed the percentage of 5% quantiles which were above the given τ^2 and the percentage of 95% quantiles which were below the given τ^2. Obviously, we would want both these quantities as close to 5% as possible. The results for various values of τ^2 and various sample sizes appear in Table 5. Note that here again, Gauss-Kronrod rules were used to integrate out $\log(\tau^2)$ numerically.

For samples of size 10 and 20, the coverage probabilities for the scenario of $\sigma^2/\tau^2 = 1/5$ were a bit low, .85 and .87 respectively, but overall the coverage probabilities tended to be conservative with coverage probabilities larger than .90. In general, the lower coverage probabilities were conservative while the upper coverage probabilties tended to be somewhat low. This is likely a remnant of the uniform shrinkage prior which places non-negligible weight on $\tau^2 = 0$. Based on these results, the uniform shrinkage prior once again appears to have reasonable frequentist properties.

5. MULTIVARIATE UNIFORM SHRINKAGE PRIORS

So far, only priors for independent univariate variance components have been discussed. For the multivariate case, a common choice for a prior for a variance matrix is the inverse Wishart prior, the conjugate prior for a multivariate normal model (Anderson 1984). Some common noninformative choices for a $p \times p$ covariance matrix would include the limiting case of the Wishart prior, the flat prior ($d\mathbf{D}$), Jeffreys prior ($|\mathbf{D}|^{-(p+1)/2}$) (Jeffreys 1961), and the reference prior (Yang & Berger 1994). As in the univariate case, a prior can be derived based on multivariate shrinkage. For example, consider the model:

$$\mathbf{Y}_i \sim \mathrm{N}_p(\mu_{\mathbf{i}}, \mathbf{\Sigma}_i), \boldsymbol{\mu}_i \sim \mathrm{N}_p(\mathbf{X}_i'\boldsymbol{\beta}, \mathbf{D}), \quad \boldsymbol{\beta} \sim d\boldsymbol{\beta}, \quad \mathbf{D} \sim \pi(\mathbf{D}).$$

The shrinkage factor here refers to the weight placed on the prior mean for $\boldsymbol{\mu}_i$: $\mathrm{E}(\boldsymbol{\mu}_i|\boldsymbol{\beta}, \mathbf{D}) = \mathbf{\Sigma}_i(\mathbf{\Sigma}_i + \mathbf{D})^{-1}\mathbf{X}_i'\boldsymbol{\beta} + \mathbf{D}(\mathbf{\Sigma}_i + \mathbf{D})^{-1}f(\mathbf{Y}_i)$. We place a uniform distribution on $\mathbf{\Sigma}_i(\mathbf{\Sigma}_i + \mathbf{D})^{-1}$ which means placing a uniform distribution on the space of all positive definite matrices with eigenvalues ranging from $(0, 1]$. To obtain the prior distribution for $\mathbf{D}$, we can compute the Jacobian which leads to $\pi(\mathbf{D}) \sim |\mathbf{\Sigma}_c||\mathbf{\Sigma}_c + \mathbf{D}|^{-(p+1)}I\{|\mathbf{D}| > 0\}$. For additional discussion, see Everson & Morris (2000). Choice of $\mathbf{\Sigma}_c$ can be made as in the univariate case, e.g., the harmonic mean of the $\mathbf{\Sigma}_i$. Note that for $p = 1$ this simplifies to the univariate uniform shrinkage prior. This prior leads to a proper posterior for the above model (see Proposition II in the appendix).

For generalized linear models with random effects (Zeger & Karim, 1991), we can also consider an approximate uniform shrinkage prior. For simplicity, let us assume all effects are random (random coefficients model). A random coefficient logistic model follows:

$$\mathrm{logit}(p_{ij}) = \mathbf{X}_i'\boldsymbol{\beta}_i, \quad \boldsymbol{\beta}_i \sim \mathrm{N}(\boldsymbol{\gamma}, \mathbf{D}), \quad \boldsymbol{\gamma} \sim d\boldsymbol{\gamma}, \quad \mathbf{D} \sim \pi(\mathbf{D}).$$

To form the approximate uniform shrinkage prior, replace the likelihood with a normal approximation to the maximum likelihood estimator of the regression coefficient at the first level and use the inverse of the observed information for the approximate variance, $\mathbf{\Sigma}_i$, $\hat{\boldsymbol{\beta}}_i \sim \mathrm{N}(\boldsymbol{\beta}_i, \mathbf{\Sigma}_i)$. Then proceed as in the normal case. A derivation of a uniform shrinkage prior for a mixed model can be found in Natarajan & Kass (1998).

As with the uniform shrinkage prior for a single variance component, the main disadvantage of the multivariate uniform shrinkage prior is computational as the full conditional distribution (used in Gibbs sampling) will no longer be Wishart as it is for the flat prior and Jeffreys prior. These computational issues are addressed in Daniels & Kass (1999) and more recently, a simplified sampling approach for the normal-normal model is discussed in Everson & Morris (2000).

6. SUMMARY

To conclude, the uniform shrinkage prior is a convenient, useful prior with desirable properties. The prior can be used in ’conjugate prior’ models by placing a uniform prior on the shrinkage parameter and is easily derived for many common situations. In addition, the fact that this is a relatively ’noninformative’ proper prior allows its use in computation of Bayes Factors for testing $\tau^2 = 0$, i.e., whether or not the variability at the second level of the model is 0. Based on simulations, this prior seems to perform well for both estimation and testing. In addition, it is easily generalizable to a prior for a covariance matrix.

Finally, two issues should be addressed. First, in very small samples, a truly noninformative prior does not exist. In this situation, the uniform shrinkage prior can be thought of as a proper prior which is a compromise between Jeffreys prior which pulls the estimate strongly towards zero and the flat prior which pulls the estimate away from zero. Secondly, placing a uniform distribution on the shrinkage weight implies that the data variance σ^2 and the prior variance τ^2 are of the same order of magnitude.

APPENDIX: PROPRIETY OF THE POSTERIOR

The purpose of this appendix is to show that the posterior distribution will be proper for several common hierarchical regression models with the uniform shrinkage prior on the variance parameter and a flat prior on the regression coefficients.

PROPOSITION 1a. *Consider the following Poisson hierarchical model in which*

1. $Y_i|\theta_i \sim P(\theta_i)$

2. $\theta_i|\delta, \eta \sim \text{Gamma}(\delta m_i, \delta)$ *where* $\log(m_i) = \mathbf{X}_i{}'\boldsymbol{\eta}$

3. $\eta \sim d\eta, \delta \sim \pi(\delta) = 1/(1+\delta)^2$.

Then $m(\mathbf{y}) = \int_{-\infty}^{\infty}\int_0^{\infty} f(y|\delta,\eta)\pi(\delta,\eta)d\delta d\eta < \infty$.

Proof. We will proceed to show that $m(\mathbf{y})$ is finite. To simplify calculations, let $X_i = 1$ and $m_i = m$. Thus, we are interested in evaluating the following integral:

$$m(\mathbf{y}) = \int_{-\infty}^{\infty}\int_0^{\infty} \left\{\prod_{i=1}^{n} \frac{\Gamma(y_i + \delta e^{\eta})}{\Gamma(\delta e^{\eta})} \frac{\delta^{\delta e^{\eta}}}{(\delta+1)^{y_i+\delta e^{\eta}}} \frac{1}{\Gamma(y_i+1)}\right\} \frac{1}{(1+\delta)^2} d\delta d\eta.$$

To evaluate this integral, we first make a variable transformation from (δ, η) to (α, β) where $\alpha = \delta e^{\eta}$ and $\beta = \delta$. The above expression then becomes

$$\int_0^{\infty}\int_0^{\infty} \left\{\prod_{i=1}^{n} \frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)} \frac{\beta^{\alpha}}{(\beta+1)^{y_i+\alpha}\Gamma(y_i+1)}\right\} \frac{1}{\alpha}\frac{1}{(1+\beta)^2} d\beta d\alpha$$

which is proportional to

$$\int_0^{\infty}\int_0^{\infty} \left\{\prod_{i=1}^{n} \frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)}\right\} \left(\frac{\beta}{\beta+1}\right)^{n\alpha} \left(\frac{1}{\beta+1}\right)^{T} \cdot \frac{1}{\alpha}\frac{1}{(1+\beta)^2} d\beta d\alpha$$

where $T = \sum y_i$. We now do one additional tranformation from (α, β) to (α, θ) where $\theta = \beta/(1+\beta)$. The above expression is then equal to

$$\int_0^{\infty}\int_0^{1} \left\{\prod_{i=1}^{n} \frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)}\right\} \theta^{n\alpha}(1-\theta)^{T}\frac{1}{\alpha} d\theta d\alpha.$$

By integrating out θ and doing some cancellations, one then obtains:

$$\int_0^\infty \left\{ \frac{\prod_{i=1}^n (y_i + \alpha - 1) \cdots (\alpha + 1)\alpha}{\alpha} \right\} \frac{\Gamma(n\alpha + 1)\Gamma(T + 1)}{\Gamma(n\alpha + T + 2)} d\alpha.$$

Define $y^\star = \max(y_1, \ldots, y_n)$. This last integral is bounded above by

$$\int_0^\infty \frac{(y^\star + \alpha - 1)^{T-1}}{(n\alpha + 1)^{T+1}} \Gamma(T + 1) d\alpha,$$

which is finite since it is proportional to $1/(k + \alpha)^2$ where k is a non-zero constant. □

PROPOSITION 1b. *Consider the Poisson hierarchical model in which*

1. $Y_i|\theta_i \sim P(\theta_i)$

2. $\theta_i|\delta, \eta \sim \text{Gamma}(\delta, \delta/m_i) \log(m_i) = \mathbf{X}_i{}'\boldsymbol{\eta}$

3. $\eta \sim d\eta, \delta \sim k/(k + \delta)^2$.

Then $m(\mathbf{y}) = \int_{-\infty}^{\infty} \int_0^\infty f(y|\delta, \eta)\pi(\delta, \eta) d\delta d\eta < \infty$.

Proof. We will again proceed to show that $m(\mathbf{y})$ is finite. To simplify calculations, let $X_i = 1$ and $m_i = m$. Then,

$$m(\mathbf{y}) = \int_{-\infty}^{\infty} \int_0^\infty \left\{ \prod_{i=1}^n \frac{\Gamma(y_i + \delta)\,(\delta/e^\eta)^\delta}{\Gamma(\delta)\,(\delta/e^\eta + 1)^{y_i+\delta}\,\Gamma(y_i + 1)} \right\} \cdot \frac{k}{(k + \delta)^2} d\delta d\eta$$

First, we make the transformation from (δ, η) to (α, β), where $\alpha = \delta$ and $\beta = \delta/(\exp\ \eta)$. The above expression is then proportional to

$$\int_0^\infty \int_0^\infty \left\{ \prod \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha)} \right\} \left(\frac{\beta}{\beta + 1} \right)^{n\alpha - 1} \left(\frac{1}{\beta + 1} \right)^{T+1} \frac{1}{(k + \alpha)^2} d\alpha d\beta,$$

where $T = \sum y_i$. Now, we make a second transformation from (α, β) to (α, θ), where $\theta = \beta/(\beta + 1)$. This yields

$$\int_0^\infty \int_0^1 \left\{ \prod \frac{\Gamma(y_i + \delta)}{\Gamma(\alpha)} \right\} \theta^{n\alpha - 1}(1 - \theta)^{T+1} \cdot \frac{1}{(1 - \theta)^2} \cdot \frac{1}{(k + \alpha)^2} d\theta d\alpha.$$

Upon integrating out θ and re-arranging terms, the above expression is proportional to

$$\int_0^\infty \frac{\prod_{i=1}^n (y_i + \alpha - 1) \cdots (\alpha + 1)\alpha}{(T + n\alpha - 1) \cdots (n\alpha + 1)(n\alpha)} \frac{1}{(k + \alpha)^2} d\alpha.$$

Letting $y^* = \max(y_1, \ldots, y_n)$ as before, the above is now less than or equal to

$$\int_0^\infty \frac{(y^* + \alpha - 1)^{T-1}}{(n\alpha + 1)^{T-1}} \cdot \frac{1}{n} \cdot \frac{1}{(k + \alpha)^2} d\alpha.$$

The first term in this integral can be bounded by a positive constant, M, and thus the following integral is finite as in Proposition 1a. □

PROPOSITION 1c. *Consider the beta-binomial model in which*

1. $Y_i|\theta_i \sim \text{Bin}\ (N_i, \theta_i)$

2. $\theta_i|\eta,\delta \sim \text{Beta}\{\delta m_i, \delta(1-m_i)\}$, $\text{logit}(m_i) = \mathbf{X}_i'\boldsymbol{\eta}$

3. $\eta \sim d\eta$, $\delta \sim \pi(\delta) = n_c/(n_c+\delta)^2$.

Then $m(\mathbf{y}) = \int_{-\infty}^{\infty}\int_0^{\infty} f(y|\delta,\eta)\pi(\delta,\eta)d\delta d\eta < \infty$.

Proof. To simplify calculations, let $X_i = 1$ and $m_i = m$. Then

$$m(\mathbf{y}) = \int_{-\infty}^{\infty}\int_0^{\infty}\left[\prod_{i=1}^{n}\frac{\Gamma(y_i+\delta\frac{e^\eta}{1+e^\eta})}{\Gamma(\delta\frac{e^\eta}{1+e^\eta})}\frac{\Gamma\{N_i-y_i+\delta(1-\frac{e^\eta}{1+e^\eta})\}}{\Gamma\{\delta(1-\frac{e^\eta}{1+e^\eta})\}}\frac{\Gamma(\delta)}{\Gamma(N_i+\delta)}\right]\frac{n_c}{(n_c+\delta)^2}d\delta d\eta.$$

First, we make the following variable transformation from (δ,η) to (α,β) where $\alpha = \delta\exp(\eta)/\{1+\exp(\eta)\}$ and $\beta = \delta[1-\exp(\eta)/\{1+\exp(\eta)\}]$. The above expression is now equal to

$$\int_0^{\infty}\int_0^{\infty}\prod_{i=1}^{n}\left\{\frac{(y_i+\alpha-1)\cdots(\alpha+1)\alpha\cdot(N_i-y_i+\beta-1)\cdots(\beta+1)\beta}{(N_i+\alpha+\beta-1)\cdots(\alpha+\beta+1)(\alpha+\beta)}\right\}\frac{\alpha+\beta}{\alpha\beta}\frac{n_c d\alpha d\beta}{(n_c+\alpha+\beta)^2}.$$

Let again $y^* = \max(y_1,\ldots,y_n)$ and set $n^* = \max(N_1-y_1,\ldots,N_n-y_n)$. Then, the former expression will be bounded by

$$\int_0^{\infty}\int_0^{\infty}\frac{(y^*+\alpha-1)^{\Sigma y_i-1}(n^*+\beta-1)^{\Sigma(N_i-y_i)-1}}{(\alpha+\beta+1)^{\Sigma N_i-1}}\cdot\frac{n_c}{(n_c+\alpha+\beta)^2}d\alpha d\beta.$$

Next, by rearranging terms and performing the transformation from (α,β) to (θ,ϕ) where

$$\theta = \frac{y^*+\alpha-1}{y^*+n^*+\alpha+\beta-2}$$

and $\phi = \alpha+\beta$, the above equals

$$\int_0^{\infty}\int_0^{1}\theta^{\Sigma y_i-1}(1-\theta)^{\Sigma(N_i-y_i)-1}\frac{(y^*+n^*+\phi-2)^{\Sigma N_i-1}}{(\phi+1)^{\Sigma N_i-1}}\frac{n_c}{(n_c+\phi)^2}d\theta d\phi.$$

We now integrate out θ and note that the first term in the integral can be bounded by a positive constant. Thus, similar to the previous proofs, the integral is bounded and finite. □

PROPOSITION 2. *Consider the model:*

1. $\mathbf{Y}_i \sim \text{N}(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i)$, *where* $\mathbf{Y}_i$ *is a* $p\times 1$ *vector with* $\boldsymbol{\Sigma}_i$ *known.*

2. $\boldsymbol{\theta}_i \sim \text{N}(\mathbf{X}_i'\boldsymbol{\beta}, \mathbf{D})$

3. $\boldsymbol{\beta} \sim d\boldsymbol{\beta}$, $\mathbf{D} \sim \pi(\mathbf{D}) = |\boldsymbol{\Sigma}_c||\boldsymbol{\Sigma}_c+\mathbf{D}|^{-(p+1)}$, *where* $\boldsymbol{\Sigma}_c$ *is as defined above.*

In this model, conditional on $\boldsymbol{\beta}$ *and* $\mathbf{D}$, $\mathbf{Y}_i$ *is normally distributed with mean* $\mathbf{X}_i'\boldsymbol{\beta}$ *and variance* $\boldsymbol{\Sigma}_i+\mathbf{D}$. *Then*

$$m(\mathbf{y}) = \int_{-\infty}^{\infty}\int_{|\mathbf{D}|>0} f(\mathbf{y}|\boldsymbol{\beta},\mathbf{D})\pi(\boldsymbol{\beta},\mathbf{D})d\boldsymbol{\beta}d\mathbf{D} < \infty.$$

Proof. To show that $m(\mathbf{y}) < \infty$, integrate out β analytically. This yields

$$\int_{|\mathbf{D}|>0} f(\mathbf{y}|\mathbf{D})\pi(\mathbf{D})d\mathbf{D} \propto \int_{|\mathbf{D}|>0}\frac{\exp\{-\frac{1}{2}\sum(\mathbf{Y}_i-\mathbf{X_i}'\hat{\beta})'(\boldsymbol{\Sigma}_i+\mathbf{D})^{-1}(\mathbf{Y}_i-\mathbf{X}_i'\hat{\boldsymbol{\beta}})\}\,d\mathbf{D}}{|\boldsymbol{\Sigma}_c+\mathbf{D}|^{p+1}\left(\prod_{i=1}^{n}|\boldsymbol{\Sigma}_i+\mathbf{D}|^{\frac{1}{2}}\right)|\sum\mathbf{X}_i'(\boldsymbol{\Sigma}_i+\mathbf{D})^{-1}\mathbf{X}_i|^{\frac{1}{2}}},$$

where $\hat{\boldsymbol{\beta}}$ is the Weighted Least Squares estimate of $\boldsymbol{\beta}$.

The exponential term is bounded by 1. Thus, the above expression is bounded by

$$\int_{|\mathbf{D}|>0}\left(|\boldsymbol{\Sigma}_c+\mathbf{D}|^{p+1}\sum \mathbf{X}_i'(\boldsymbol{\Sigma}_i+\mathbf{D})^{-1}\mathbf{X}_i|^{\frac{1}{2}}\prod_{i=1}^{n}|\boldsymbol{\Sigma}_i+\mathbf{D}|^{\frac{1}{2}}\right)^{-1} d\mathbf{D}.$$

Without loss of generality, let $\mathbf{X}_i = \mathbf{X}_1, \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_1$ in the middle term of the denominator. The above expression now becomes

$$\int_{|\mathbf{D}|>0}\left(|\boldsymbol{\Sigma}_c+\mathbf{D}|^{p+1}n^{\frac{p}{2}}|\mathbf{X}_1'(\boldsymbol{\Sigma}_1+\mathbf{D})^{-1}\mathbf{X}_1|^{\frac{1}{2}}\cdot|\boldsymbol{\Sigma}_1+\mathbf{D}|^{\frac{1}{2}}\prod_{i=2}^{n}|\boldsymbol{\Sigma}_i+\mathbf{D}|^{\frac{1}{2}}\right)^{-1} d\mathbf{D}.$$

Define the above integrand to be A and define c to be a positive constant. We now break the above integral into two pieces:

$$\int_{|\mathbf{D}|>0} A\, d\mathbf{D} = \int_{0<|\mathbf{D}|<c} A\, d\mathbf{D} + \int_{|\mathbf{D}|>c} A\, d\mathbf{D}.$$

We will proceed to show that the two terms are finite. Looking at the first term,

$$\int_{0<|\mathbf{D}|<c} |\boldsymbol{\Sigma}_c+\mathbf{D}|^{-(p+1)}\cdot\prod_{2}^{n}|\boldsymbol{\Sigma}_i+\mathbf{D}|^{-1/2}\cdot|\mathbf{X}_1'(\boldsymbol{\Sigma}_1+\mathbf{D})^{-1}\mathbf{X}_1|^{-1/2}|\boldsymbol{\Sigma}_1+\mathbf{D}|^{-1/2}d\mathbf{D},$$

we define the first piece to be A_1, the second piece A_2, and the third piece A_3. By inspection, the above integral is bounded and finite at c. Now, we examine the behavior of the integrand as $\mathbf{D}$ approaches 0. The three pieces of the integrand approach the following limits which are all finite:

$$A_1 \to |\boldsymbol{\Sigma}_c|^{-(p+1)}, \quad A_2 \to \prod_{i=2}^{n}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}, \quad A_3 \to |\mathbf{X}_1'\boldsymbol{\Sigma}_1^{-1}\mathbf{X}_1|^{\frac{1}{2}}|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}}.$$

Thus, since the integrand is bounded over a finite, continuous region $0 < |\mathbf{D}| < c$ and is a continuous function, we can conclude that the integral is finite. We now examine the second piece of the original integral, namely $\int_{|\mathbf{D}|>c} Ad\mathbf{D}$. We define $L(\mathbf{D}) = A_2A_3$. As $\mathbf{D}$ approaches infinity, A_2A_3 approaches $k = 0$. We also note that $L(\mathbf{D})$ is bounded over the entire region. As a result, the whole integral is bounded by $k\int_{|\mathbf{D}|>c}|\boldsymbol{\Sigma}_c+\mathbf{D}|^{-(p+1)}d\mathbf{D}$ and the proof is complete. □

ACKNOWLEDGEMENTS

The author thanks Robert Kass, Larry Wasserman and a referee for helpful comments.

REFERENCES

T. W. Anderson (1984). *An Introduction to Multivariate Statistical Analysis*, Second Edition. Wiley, New York.

J. O. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

J. O. Berger & J. M. Bernardo (1992). On the development of reference priors (with discussion). In *Bayesian Statistics–4* (J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds.), Oxford University Press, pp. 35–60.

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

G. E. P. Box & G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Wiley, New York.

C. L. Christiansen & C. N. Morris (1997). Hierarchical Poisson regression modeling. *Journal of the American Statistical Association*, 92, 618–632.

M. J. Daniels & C. G. Gatsonis (1999). Hierachical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, 94, 29–42.

M. J. Daniels & R. E. Kass (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. Technical Report no. 659, Carnegie Mellon University, Pittsburgh. To appear in *Journal of the American Statistical Association*.

W. E. DuMouchel (1994). Hierarchical Bayes linear models for meta-analysis. Technical Report no. 27, National Institute of Statistical Sciences, Research Triangle Park, NC.

P. J. Everson & C. N. Morris (2000). Inference for multivariate normal hierarchical models. Technical Report, Harvard University. To appear in *Journal of the Royal Statistical Society Series B*.

A. Gelman, B. P. Carlin, H. S. Stern & D. B. Rubin (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

J. P. Hobert & G. Casella (1994). Gibbs sampling with improper prior distributions. Technical Report, Biometrics Unit, Cornell University, Ithaca, NY.

J. P. Hobert & G. Casella (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461–1473.

H. Jeffreys (1961). *Theory of Probability*. Oxford Univ. Press.

R. Natarajan & R. E. Kass (1998). Reference Bayesian methods for generalized linear mixed models. Technical Report, Carnegie Mellon University, Pittsburgh.

A. F. M. Smith & G. O. Roberts (1993). Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B*, 55, 3–23.

W. E. Strawderman (1971). Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42, 385–388.

R. Yang & J. O. Berger (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 22, 1195–1211.

S. Zeger & M. Karim (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79–86.

Received 21 April 1997
Accepted 13 August 1998

Michael J. DANIELS: mdaniels@iastate.edu
Dept. of Statistics, Iowa State University of Science and Technology
102G Snedecor Hall, Ames, IA 50011-1210, USA