Gelman, Andrew

Working Paper

Prior distributions for variance parameters in hierarchical models

EERI Research Paper Series, No. 6/2004

Provided in Cooperation with:
Economics and Econometrics Research Institute (EERI), Brussels

Suggested Citation: Gelman, Andrew (2004) : Prior distributions for variance parameters in hierarchical models, EERI Research Paper Series, No. 6/2004, Economics and Econometrics Research Institute (EERI), Brussels

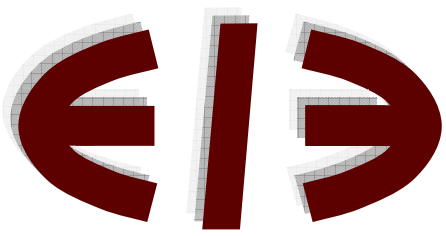
This Version is available at:
https://hdl.handle.net/10419/142500

EERI
Economics and Econometrics Research Institute

Prior Distributions for Variance Parameters in Hierarchical Models

Andrew Gelman

EERI Research Paper Series No 6/2004



EERI
Economics and Econometrics Research Institute
Avenue de Beaulieu
1160 Brussels
Belgium

Tel: +322 299 3523
Fax: +322 299 3523
www.eeri.eu

Prior distributions for variance parameters in hierarchical models*

Andrew Gelman†

February 5, 2004

Abstract

Various noninformative prior distributions have been suggested for scale parameters in hierarchical models. We construct a new folded-noncentral-t family of conditionally conjugate priors for hierarchical standard deviation parameters, and then consider noninformative and weakly informative priors in this family. We use an example to illustrate serious problems with the inverse-gamma family of "noninformative" prior distributions. We suggest instead to use a uniform prior on the hierarchical standard deviation, using the half-t family when the number of groups is small and in other settings where a weakly informative prior is desired.

Keywords: Bayesian inference, conditional conjugacy, folded-noncentral-t distribution, half-t distribution, hierarchical model, multilevel model, noninformative prior distribution, weakly informative prior distribution

1 Introduction

Hierarchical (multilevel) models are central to modern Bayesian statistics for both conceptual and practical reasons. On the theoretical side, hierarchical models allow a more "objective" approach to inference by estimating the parameters of prior distributions from data rather than requiring them to be specified using subjective information (see James and Stein, 1960, Efron and Morris, 1975, and Morris, 1983). At a practical level, hierarchical models are flexible tools for combining information and partial pooling of inferences (see, for example, Kreft and De Leeuw, 1998, Snijders and Bosker, 1999, Carlin and Louis, 2001, Raudenbush and Bryk, 2002, Gelman et al., 2003).

A hierarchical model requires hyperparameters, however, and these must be given their own prior distribution. In this paper, we discuss the prior distribution for hierarchical variance parameters. We consider some proposed noninformative prior distributions, including uniform and inverse-gamma families, in the context of an expanded conditionally-conjugate family.

*For *Bayesian Analysis*. We thank Rob Kass for inviting this paper, John Boscardin, John Carlin, Chuanhai Liu, Hal Stern, Francis Tuerlinckx, and Aki Vehtari for helpful suggestions, and the National Science Foundation for financial support.

†Department of Statistics and Department of Political Science, Columbia University, New York, U.S.A., gelman@stat.columbia.edu, www.stat.columbia.edu/∼gelman/

1.1 The basic hierarchical model

We shall work with a simple two-level normal model of data y_{ij} with group-level effects α_j:

$$\begin{aligned} y_{ij} &\sim \text{N}(\mu + \alpha_j, \sigma_y^2), \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J \\ \alpha_j &\sim \text{N}(0, \sigma_\alpha^2), \quad j = 1, \ldots, J. \end{aligned} \tag{1}$$

We briefly discuss other hierarchical models in Section 5.2.

Model (1) has three hyperparameters—μ, σ_y, and σ_α—but in this paper we concern ourselves only with the last of these. Typically, enough data will be available to estimate μ and σ_y that one can use any reasonable noninformative prior distribution—for example, $p(\mu, \sigma_y) \propto 1$ or $p(\mu, \log \sigma_y) \propto 1$.

Various noninformative prior distributions have been suggested in Bayesian literature and software, including an improper uniform density on σ_α (Gelman et al., 2003) and proper distributions such as $p(\sigma_\alpha^2) \sim$ inv-gamma$(0.001, 0.001)$ (Spiegelhalter et al., 1994, 2003). In this paper, we explore and make recommendations for prior distributions for σ_α, beginning in Section 2 with conjugate families of proper prior distributions and then considering noninformative prior densities in Section 3. As we illustrate in Section 4, some of these prior distributions can unduly affect inferences, especially for problems where the number of groups J is small or the group-level variance σ_α^2 is close to zero. We conclude with recommendations in Section 5.

2 Conditionally-conjugate families

2.1 Inverse-gamma prior distribution for σ_α^2

The parameter σ_α^2 in model (1) does not have any simple family of conjugate prior distributions because its marginal likelihood depends in a complex way on the data from all J groups (Hill, 1965, Tiao and Tan, 1965). However, the inverse-gamma family is *conditionally conjugate*: that is, if σ_α^2 has an inverse-gamma prior distribution, then the conditional posterior distribution $p(\sigma_\alpha^2 \mid \alpha, \mu, \sigma_y, y)$ is also inverse-gamma. This conditional conjugacy allows σ_α^2 to be updated easily using the Gibbs sampler (see Gelfand and Smith, 1990) and also allows the prior distribution to be interpreted in terms of equivalent data (see, for example, Box and Tiao, 1973).

The inv-gamma(α, β) model for σ_α^2 can also be expressed as an inverse-χ^2 distribution with scale $s_\alpha^2 = \beta/\alpha$ and degrees of freedom $\nu_\alpha = 2\alpha$ (Gelman et al., 2003). The inverse-χ^2 parameterization can be helpful in understanding the information underlying various choices of proper prior distributions, as we discuss in Section 3.

2.2 Folded-noncentral-t prior distribution for σ_α

We can expand the family of conditionally-conjugate prior distributions by applying a redundant multiplicative reparameterization to model (1):

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\mu + \xi\eta_j, \sigma_y^2) \\ \eta_j &\sim \mathrm{N}(0, \sigma_\eta^2). \end{aligned} \tag{2}$$

The parameters α_j in (1) correspond to the products $\xi\eta_j$ in (2), and the hierarchical standard deviation σ_α in (1) corresponds to $|\xi|\sigma_\eta$ in (2). This "parameter expanded" model was originally constructed to speed up EM and Gibbs sampler computations (Liu, Rubin, and Wu, 1998, Liu and Wu, 1999, van Dyk and Meng, 2001, Gelman et al., 2004), and it is also been suggested that the additional parameter can increase the flexibility of applied modeling, especially in hierarchical regression models with several batches of varying coefficients (Gelman, 2004a). Here we merely note that this expanded model form allows conditionally conjugate prior distributions for both ξ and σ_η, and these parameters are independent in the conditional posterior distribution. There is thus an implicit conditionally conjugate prior distribution for $\sigma_\alpha = |\xi|\sigma_\eta$.

For simplicity we restrict ourselves to independent prior distributions on ξ and σ_η. In model (2), the conditionally-conjugate prior family for ξ is normal—given the data and all the other parameters in the model, the likelihood for ξ has the form of a normal distribution, derived from $\sum_{j=1}^J n_j$ factors of the form $(y_{ij} - \mu)/\eta_j \sim \mathrm{N}(\xi, \sigma_y^2/\eta_j^2)$. The conditionally-conjugate prior family for σ_η^2 is inverse-gamma, as discussed in Section 2.1.

The implicit conditionally-conjugate family for σ_α is then the set of distributions corresponding to the absolute value of a normal random variable, divided by the square root of a gamma random variable. That is, σ_α has the distribution of the absolute value of a noncentral-t variate (see, for example, Johnson and Kotz, 1972). We shall call this the *folded noncentral t distribution*, with the "folding" corresponding to the absolute value operator. The noncentral t in this context has three parameters, which can be identified with the mean of the normal distribution for ξ, and the scale and degrees of freedom for σ_η^2. (Without loss of generality, the scale of the normal distribution for ξ can be set to 1 since it cannot be separated from the scale for σ_η.)

The folded noncentral t distribution is not commonly used in statistics, and we find it convenient to understand it through various special and limiting cases. In the limit that the denominator is specified exactly, we have a folded normal distribution; conversely, specifying the numerator exactly yields the square-root-inverse-χ^2 distribution for σ_α, as in Section 2.1.

An appealing two-parameter family of prior distributions is determined by restricting the prior mean of the numerator to zero, so that the folded noncentral t distribution for σ_α becomes simply a

half-t—that is, the absolute value of a Student-t distribution centered at zero. We can parameterize this in terms of scale s_α and degrees of freedom ν:

$$p(\sigma_\alpha) \propto \left(1 + \frac{1}{\nu}\left(\frac{\sigma_\alpha}{s_\alpha}\right)^2\right)^{-(\nu+1)/2}.$$

This family includes, as special cases, the improper uniform density (if $\nu = -1$) and the proper half-Cauchy, $p(\sigma_\alpha) \propto \left(\sigma_\alpha^2 + s_\alpha^2\right)^{-1}$ (if $\nu = 1$).

The half-t family is not itself conditionally-conjugate—starting with a half-t prior distribution, you will still end up with a more general folded noncentral t conditional posterior—but it is a natural subclass of prior densities in which the distribution of the multiplicative parameter ξ is symmetric about zero.

3 Noninformative prior distributions

3.1 General considerations

Noninformative prior distributions are intended to allow Bayesian inference for parameters about which not much is known beyond the data included in the analysis at hand. Various justifications and interpretations of noninformative priors have been proposed over the years, including invariance (Jeffreys, 1961), maximum entropy (Jaynes, 1983), and agreement with classical estimators (Box and Tiao, 1973, Meng and Zaslavsky, 2002). In this paper, we follow the approach of Bernardo (1979) and consider so-called noninformative priors as "reference models" to be used as a standard of comparison or starting point in place of the proper, informative prior distributions that would be appropriate for a full Bayesian analysis (see also Kass and Wasserman, 1996).

We view any noninformative prior distribution as inherently provisional—after the model has been fit, one should look at the posterior distribution and see if it makes sense. If the posterior distribution does not make sense, this implies that additional prior knowledge is available that has not been included in the model, and it is appropriate to go back and include this in the form of an informative prior distribution.

3.2 Uniform prior distributions

We first consider uniform prior distributions while recalling that we must be explicit about the scale on which the distribution is defined. Various choices have been proposed for modeling variance parameters. A uniform prior distribution on $\log \sigma_\alpha$ would seem natural—working with the logarithm of a parameter that must be positive—but it results in an improper posterior distribution. The problem arises because the marginal likelihood, $p(y|\sigma_\alpha)$—after integrating over α, μ, σ_y in (1)—approaches a finite nonzero value as $\sigma_\alpha \to 0$. Thus, if the prior density for $\log \sigma_\alpha$ is uniform, the

posterior distribution will have infinite mass integrating to the limit $\log \sigma_\alpha \to -\infty$. To put it another way, in a hierarchical model the data can never rule out a group-level variance of zero, and so the prior distribution cannot put an infinite mass in this area.

Another option is a uniform prior distribution on σ_α itself, which has a finite integral near $\sigma_\alpha = 0$ and thus avoids the above problem. We generally use this noninformative density in our applied work (see Gelman et al., 2003), but it has a slightly disagreeable "bias" toward positive values, with its infinite prior mass in the range $\sigma_\alpha \to \infty$. With $J = 1$ or 2 groups, this actually results in an improper posterior density, essentially concluding $\sigma_\alpha = \infty$ and doing no shrinkage (see Gelman et al., 2003, Exercise 5.8). In a sense this is reasonable behavior, since it would seem difficult from the data alone to decide how much, if any, shrinkage should be done with data from only one or two groups—and in fact this would seem consistent with the work of Stein (1955) and James and Stein (1960) that unshrunken estimators are admissible if $J < 3$. However, from a Bayesian perspective it is awkward for the decision to be made ahead of time, as it were, with the data having no say in the matter. In addition, for small J, such as 4 or 5, we worry that the heavy right tail of the posterior distribution would tend to bias the estimates of σ_α and thus result in shrinkage that is less than optimal for estimating the individual α_j's.

We can interpret the various improper uniform prior densities as limits of conditionally-conjugate priors. The uniform prior distribution on $\log \sigma_\alpha$ is equivalent to $p(\sigma_\alpha) \propto \sigma_\alpha^{-1}$ or $p(\sigma_\alpha^2) \propto \sigma_\alpha^{-2}$, which has the form of an inverse-χ^2 density with 0 degrees of freedom and can be taken as a limit of proper conditionally-conjugate inverse-gamma priors.

The uniform density on σ_α is equivalent to $p(\sigma_\alpha^2) \propto \sigma_\alpha^{-1}$, an inverse-$\chi^2$ density with -1 degrees of freedom. This density cannot easily be seen as a limit of proper inverse-χ^2 densities (since these must have positive degrees of freedom), but it can be interpreted as a limit of the half-t family on σ_α, where the scale approaches ∞ (and any value of ν). Or, in the expanded notation of (2), one could assign any prior distribution to σ_η and a normal to ξ, and let the prior variance for ξ approach ∞.

Another noninformative prior distribution sometimes proposed in the Bayesian literature is uniform on σ_α^2. We do not recommend this, as it seems to have the positive bias described above, but more so, and also requires $J \geq 4$ groups for a proper posterior distribution.

3.3 Inverse-gamma(ϵ, ϵ) prior distributions

The inv-gamma(ϵ, ϵ) prior distribution is an attempt at noninformativeness within the conditionally conjugate family, with ϵ set to a low value such as 1 or 0.01 or 0.001 (the latter value being used in the examples in Bugs; see Spiegelhalter et al., 1994, 2003). A difficulty of this model is that in the

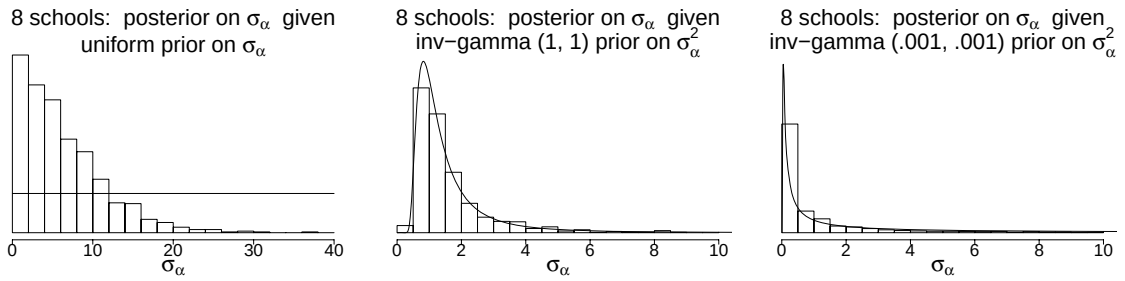


Figure 1: Histograms of posterior simulations of the between-school standard deviation, σ_α, from models with three different prior distributions: (a) uniform prior distribution on σ_α, (b) inverse-gamma$(1, 1)$ prior distribution on σ_α^2, (c) inverse-gamma$(0.001, 0.001)$ prior distribution on σ_α^2. The histograms are not all on the same scales. Overlain on each is the corresponding prior density function for σ_α. (For models (b) and (c), the density for σ_α is calculated using the gamma density function multiplied by the Jacobian of the $1/\sigma_\alpha^2$ transformation.) In models (b) and (c), posterior inferences are strongly constrained by the prior distribution. Adapted from Gelman et al. (2003, Appendix C).

limit of $\epsilon \to 0$ it yields an improper posterior density, and thus ϵ must be set to a reasonable value. Unfortunately, for datasets in which low values of σ_α are possible, inferences become very sensitive to ϵ in this model, and the prior distribution hardly looks noninformative, as we illustrate next.

4 Application to the 8-schools example

We demonstrate the properties of some proposed noninformative prior densities with a simple example of data from $J = 8$ educational testing experiments described in Gelman et al. (2003, Chapter 5 and Appendix C). Here, the parameters $\alpha_1, \ldots, \alpha_8$ represent the relative effects of Scholastic Aptitude Test coaching programs in eight different schools, and σ_α represents the between-school standard deviations of these effects. The effects are measured as points on the test, which was scored from 200 to 800; thus the largest possible range of effects could be 600 points, with a realistic upper limit on σ_α of 100, say.

4.1 Noninformative prior distributions for the 8-schools problem

Figure 1 shows the posterior distributions for the 8-schools model resulting from three different choices of prior distributions that are intended to be noninformative.

The leftmost histogram shows the posterior inference for σ_α (as represented by 6000 simulation draws from a model fit using Bugs) for the model with uniform prior density. The data show support for a range of values below $\sigma_\alpha = 20$, with a slight tail after that, reflecting the possibility of larger values, which are difficult to rule out given that the number of groups J is only 8—that is, not much more than the $J = 3$ required to ensure a proper posterior density with finite mass in the right tail.

In contrast, the middle histogram in Figure 1 shows the result with an inverse-gamma$(1, 1)$ prior distribution for σ_α^2. This new prior distribution leads to changed inferences. In particular, the posterior mean and median of σ_α are lower and shrinkage of the α_j's is greater than in the previously-fitted model with a uniform prior distribution on σ_α. To understand this, it helps to graph the prior distribution in the range for which the posterior distribution is substantial. The graph shows that the prior distribution is concentrated in the range $[0.5, 5]$, a narrow zone in which the likelihood is close to flat compared to this prior (as we can see because the distribution of the posterior simulations of σ_α closely matches the prior distribution, $p(\sigma_\alpha)$). By comparison, in the left graph, the uniform prior distribution on σ_α seems closer to "noninformative" for this problem, in the sense that it does not appear to be constraining the posterior inference.

Finally, the rightmost histogram in Figure 1 shows the corresponding result with an inverse-gamma$(0.001, 0.001)$ prior distribution for σ_α^2. This prior distribution is even more sharply peaked near zero and further distorts posterior inferences, with the problem arising because the marginal likelihood for σ_α remains high near zero.

In this example, we do not consider a uniform prior density on $\log \sigma_\alpha$, which would yield an improper posterior density with a spike at $\sigma_\alpha = 0$, like the rightmost graph in Figure 1, but more so. We also do not consider a uniform prior density on σ_α^2, which would yield a posterior distribution similar to the leftmost graph in Figure 1, but with a slightly higher right tail.

This example is a gratifying case in which the simplest approach—the uniform prior density on σ_α—seems to perform well. As detailed in Gelman et al. (2003, Appendix C), this model is also straightforward to program directly using the Gibbs sampler or in Bugs, using either the basic model (1) or slightly faster using the expanded parameterization (2).

The appearance of the histograms and density plots in Figure 1 is crucially affected by the choice to plot them on the scale of σ_α. If instead they were plotted on the scale of $\log \sigma_\alpha$, the inv-gamma$(0.001, 0.001)$ prior density would appear to be the flattest. However, the inverse-gamma(ϵ, ϵ) prior is not at all "noninformative" for this problem since the resulting posterior distribution remains highly sensitive to the choice of ϵ. As explained in Section 3.2, the hierarchical model likelihood does not constrain $\log \sigma_\alpha$ in the limit $\log \sigma_\alpha \to -\infty$, and so a prior distribution that is noninformative on the log scale will not work.

4.2 Weakly informative prior distribution for the 3-schools problem

The uniform prior distribution seems fine for the 8-school analysis, but problems arise if the number of groups J is much smaller, in which case the data supply little information about the group-level variance, and a noninformative prior distribution can lead to a posterior distribution that is improper

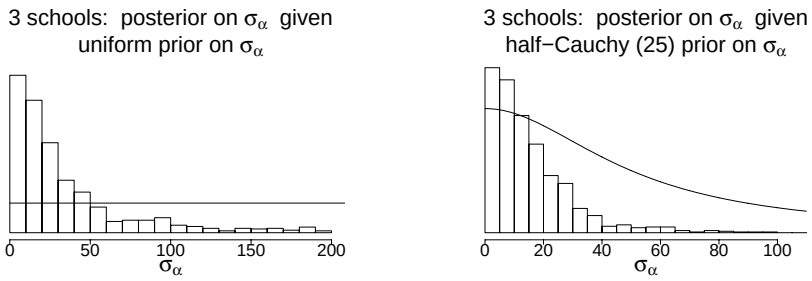


Figure 2: Histograms of posterior simulations of the between-school standard deviation, σ_α, from models for the 3-schools data with two different prior distributions on σ_α: (a) uniform $(0, \infty)$, (b) half-Cauchy with scale 25, set as a weakly informative prior distribution given that σ_α was expected to be well below 100. The histograms are not on the same scales. Overlain on each histogram is the corresponding prior density function. With only $J = 3$ groups, the noninformative uniform prior distribution is too weak, and the proper Cauchy distribution works better, without appearing to distort inferences in the area of high likelihood.

or is proper but unrealistically broad. We demonstrate by reanalyzing the 8-schools example using just the data from the first 3 of the schools.

Figure 2 displays the inferences for σ_α from two different prior distributions. First we continue with the default uniform distribution that worked well with $J = 8$ (as seen in Figure 1). Unfortunately, as the left histogram of Figure 2 shows, the resulting posterior distribution for the 3-schools dataset has an extremely long right tail, containing values of σ_α that are too high to be reasonable. This heavy tail is expected since J is so low (if J were any lower, the right tail would have an infinite integral), and using this as a posterior distribution will have the effect of undershrinking the estimates of the school effects α_j, as explained in Section 3.2.

The right histogram of Figure 2 shows the posterior inference for σ_α resulting from a half-Cauchy prior distribution of the sort described at the end of Section 2.2, with scale parameter 25. As the line on the graph shows, this prior distribution is close to flat over the plausible range of $\sigma_\alpha < 50$, falling off gradually beyond this point. We call this prior distribution "weakly informative" on this scale because, even at its tail, it has a gentle slope (unlike, for example, a half-normal distribution) and can let the data dominate if the likelihood is strong in that region. This prior distribution performs well in this example, reflecting the marginal likelihood for σ_α at its low end but removing much of the unrealistic upper tail.

This half-Cauchy prior distribution would also perform well in the 8-schools problem; however it was unnecessary because the default uniform prior gave reasonable results. With only 3 schools, we went to the trouble of using a weakly informative prior, a distribution that was not intended to represent our actual prior state of knowledge about σ_α but rather to constrain the posterior distribution, to an extent allowed by the data.

5 Recommendations

5.1 Prior distributions for variance parameters

In fitting hierarchical models, we recommend starting with a noninformative uniform prior density on standard deviation parameters σ_α. We expect this will generally work well unless the number of groups J is low (below 5, say). If J is low, the uniform prior density tends to lead to high estimates of σ_α, as discussed in Section 4.2. (This bias is an unavoidable consequence of the asymmetry in the parameter space, with variance parameters restricted to be positive. Similarly, there are no always-nonnegative classical unbiased estimators of σ_α or σ_α^2 in the hierarchical model.)

A user of a noninformative prior density might still like to use a proper distribution—reasons could include Bayesian scruple, the desire to perform prior predictive checks (see Box, 1980, Gelman, Meng, and Stern, 1996, and Bayarri and Berger, 2000) or Bayes factors (see Kass and Raftery, 1995, and O'Hagan, 1995, and Pauler, Wakefield, and Kass, 1999), or because computation is performed in Bugs, which requires proper distributions. For a noninformative but proper prior distribution, we recommend approximating the uniform density on σ_α by a uniform on a wide range (for example, $\mathrm{U}(0, 100)$ in the SAT coaching example) or a half-normal centered at 0 with standard deviation set to a high value such as 100. The latter approach is particularly easy to program as a $\mathrm{N}(0, 100^2)$ prior distribution for ξ in (2).

When more prior information is desired, for instance to restrict σ_α away from very large values, we recommend working within the half-t family of prior distributions, which are more flexible and have better behavior near 0, compared to the inverse-gamma family. A reasonable starting point is the half-Cauchy family, with scale set to a value that is high but not off the scale; for example, 25 in the example in Section 4.2.

Figure 1 illustrates the generally robust properties of the uniform prior density on σ_α. Many Bayesians have preferred the inverse-gamma prior family, possibly because its conditional conjugacy suggested clean mathematical properties. However, by writing the hierarchical model in the form (2), we see conditional conjugacy in the wider class of half-t distributions on σ_α, which include the uniform and half-Cauchy densities on σ_α (as well as inverse-gamma on σ_α^2) as special cases. From this perspective, the inverse-gamma family has nothing special to offer, and we prefer to work on the scale of the standard deviation parameter σ_α, which is typically directly interpretable in the original model.

5.2 Generalizations

The reasoning in this paper should apply to hierarchical regression models (including predictors at the individual or group levels), hierarchical generalized linear models (as discussed by Christiansen

and Morris, 1997, and Natarajan and Kass, 2000), and more complicated nonlinear models with hierarchical structure. The key idea is that parameters α_j—in general, group-level exchangeable parameters—have a common distribution with some scale parameter which we label σ_α. Some of the details will change—in particular, if the model is nonlinear, then the normal prior distribution for the multiplicative parameter ξ in (2) will not be conditionally conjugate, however ξ can still be updated using the Metropolis algorithm. In addition, when regression predictors must be estimated, more than $J = 3$ groups may be necessary to estimate σ_α from a noninformative prior distribution, thus requiring at least weakly informative prior distributions for the regression coefficients, the variance parameters, or both.

There is also room to generalize these distributions to variance matrices in multivariate hierarchical models, going beyond the commonly-used inverse-Wishart family of prior distributions (Box and Tiao, 1973), which has problems similar to the inverse-gamma for scalar variances. Noninformative or weakly informative conditionally-conjugate priors could be applied to structured models such as described by Barnard, McCulloch, and Meng (2000) and Daniels and Kass (1999, 2001), expanded using multiplicative parameters as in Liu (2001) to give the models more flexibility.

Further work needs to be done in developing the next level of hierarchical models, in which there are several batches of exchangeable parameters, each with their own variance parameter—the Bayesian counterpart to the analysis of variance (Sargent and Hodges, 1997, Gelman, 2004b). Specifying a prior distribution jointly on variance components at different levels of the model could be seen as a generalization of priors on the shrinkage factor, which is a function of both σ_y and σ_α (see Daniels, 1999, Natarajan and Kass, 2000, and Spiegelhalter, Abrams, and Myles, 2004, for an overview). In a model with several levels, it would make sense to give the variance parameters a parametric model with hyper-hyperparameters. This could be the ultimate solution to the difficulties of estimating σ_α for batches of parameters α_j where J is small, and we suppose that the folded-noncentral-t family could be useful here.

References

Barnard, J., McCulloch, R. E., and Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1311.

Bayarri, M. J. and Berger, J. (2000). P-values for composite null models (with discussion). *Journal of the American Statistical Association* **95**, 1127–1142.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion).

Journal of the Royal Statistical Society B **41**, 113–147.

Box, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A* **143**, 383–430.

Carlin, B. P., and Louis, T. A. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*, second edition. London: Chapman and Hall.

Christiansen, C., and Morris, C. (1997). Hierarchical Poisson regression models. *Journal of the American Statistical Association* **92**, 618–632.

Daniels, M. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics* **27**, 569–580.

Daniels, M. J., and Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association* **94**, 1254–1263.

Daniels, M. J., and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57**, 1173–1184.

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70**, 311–319.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A. (2004a). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, to appear.

Gelman, A. (2004b). Analysis of variance: why it is more important than ever. *Annals of Statistics*, to appear.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: Chapman and Hall.

Gelman, A., Huang, Z., van Dyk, D., and Boscardin, W. J. (2004). Transformed and parameter-expanded Gibbs samplers for multilevel linear and generalized linear models. Technical report, Department of Statistics, Columbia University.

Gelman, A., Meng, X. L., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6**, 733–807.

Hill, B. M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association* **60**, 806–825.

James, W., and Stein, C. (1960). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium* **1**, ed. J. Neyman, 361–380. Berkeley: University of California Press.

Jaynes, E. T. (1983). *Papers on Probability, Statistics, and Statistical Physics*, ed. R. D. Rosenkrantz. Dordrecht, Netherlands: Reidel.

Jeffreys, H. (1961). *Theory of Probability*, third edition. Oxford University Press.

Johnson, N. L., and Kotz, S. (1972). *Distributions in Statistics*, 4 vols. New York: Wiley.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association* **90**, 773–795.

Kass, R. E., and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343–1370.

Kreft, I., and De Leeuw, J. (1998). *Introducing Multilevel Modeling.* London: Sage.

Liu, C. (2001). Bayesian analysis of multivariate probit models. Discussion of "The art of data augmentation" by D. A. van Dyk and X. L. Meng. *Journal of Computational and Graphical Statistics* **10**, 75–81.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* **85**, 755–770.

Liu, J., and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94**, 1264–1274.

Meng, X. L., and Zaslavsky A. M. (2002). Single observation unbiased priors. *Annals of Statistics* **30**, 1345–1375.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *Journal of the American Statistical Association* **78**, 47–65.

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

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society B* **57**, 99–138.

Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999). Bayes factors for variance component models. *Journal of the American Statistical Association* **94**, 1242–1253.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models*, second edition. Thousand Oaks, Calif.: Sage.

Sargent, D. J., and Hodges, J. S. (1997). Smoothed ANOVA with application to subgroup analysis. Technical report, Department of Biostatistics, University of Minnesota.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Dover.

Snijders, T. A. B., and Bosker, R. J. (1999). *Multilevel Analysis.* London: Sage.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, section 5.7.3. Chichester: Wiley.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., and Lunn, D. (1994, 2003). BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England. `www.mrc-bsu.cam.ac.uk/bugs/`

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium* **1**, ed. J. Neyman, 197–206. Berkeley: University of California Press.

Tiao, G. C., and Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. I: Posterior distribution of variance components. *Biometrika* **52**, 37–53.

van Dyk, D. A., and Meng, X. L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10**, 1–111.