Bayesian Analysis (2014) **9**, Number 1, pp. 25–28

Comment on Article by Rubio and Steel

James G. Scott [*]

1 Objective priors for skewed distributions

The topic of default priors always brings to mind an old line of Poincare's: "If we were not ignorant there would be no probability, there could only be certainty. But our ignorance cannot be absolute, for then there would be no longer any probability at all." Rubio and Steel's paper on inference for skewed distributions should give us all cause to re-affirm an important guideline for practical Bayesian inference in the face of ignorance: use off-the-shelf priors only with care, especially in multi-parameter problems. (One should use subjective priors with even greater care, but that's for another day.)

I offer congratulations to the authors on a stimulating paper, and thank them for adding another example to the catalogue of simple multi-parameter problems where Jeffreys' rule fails to give a sensible answer. The most famous is the so-called Neyman–Scott problem (Neyman and Scott 1948), involving a vector of means and two observations per coordinate:

$$y_{ij} = \theta_i + e_{ij} \, , \quad e_{ij} \sim \mathrm{N}(0, \sigma^2) \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, 2.$$

Let $\theta = \theta_1, \ldots, \theta_n$. The Fisher information matrix for (θ, σ^2) is easily derived as

$$I(\theta, \sigma^2) = \mathrm{diag}(2/\sigma^2, \ldots, 2/\sigma^2, \{2n\}/\sigma^4) \, .$$

Thus the Jeffreys'-rule prior—as opposed to the prior that, one imagines, Jeffreys himself would have used—is $\pi(\theta, \sigma^2) \propto \sigma^{-n-2}$. The corresponding marginal posterior distribution for σ^2 is

$$\pi(\sigma^2 \mid y) \propto \frac{1}{(\sigma^2)^{n+1}} \exp \left\{ - \frac{\sum_{i=1}^n (x_{i1} - x_{i2})^2}{4\sigma^2} \right\} .$$

This converges to a point mass at $\sigma^2/2$ as $n \to \infty$, and thus is strongly inconsistent.

Rubio and Steel have given us a different, presumably rarer beast: a simple example where the multiparameter Jeffreys'-rule prior fails to give any answer at all, even a silly one. They go on to provide a recommended modification that does behave sensibly. I have two questions on the problem itself, and three on the solution.

2 The problem itself

First, Rubio and Steel's Equation (1) makes it clear that the distribution of interest is a two-component discrete mixture model, albeit one whose means, variances, and weights are coupled. Bayesian folklore holds that, in general, one should not use improper priors

[*] University of Texas at Austin james.scott@mccombs.utexas.edu

DOI:10.1214/14-BA867

in discrete mixture models (e.g. Wasserman 2000). For example, Bernardo and Girón (1988) derive the reference prior for a generic mixture model

$$p(x) = w \cdot p_1(x \mid \theta_1) + (1-w) \cdot p_2(x \mid \theta_2)\,,$$

and find that it is always proper.

I have never thought too carefully about this point: it always seemed like a straightforward extension of the generic prohibition against using improper priors in Bayes-factor computations, which arise implicitly in fitting mixture models. (Although see Berger et al. 1998, for situations in which such improper priors are justifiable.) My own understanding of this problem is thus quite shallow. I hope the authors can provide some insight on the relationship between this general pathology of improper priors in mixture models, and the specific case of two-piece location-scale models.

Second, do the authors know whether a similar result would obtain for other families of skewed distributions? Here is a reason to suspect that it might. One nice construction along these lines is the family of normal variance-mean mixtures, where each observation y_i is assumed to take the form

$$y_i = a + bv_i + s\sqrt{v_i}x_i\,, \quad v_i \sim G\,, \quad x_i \sim \mathrm{N}(0,1)$$

for parameters (a, b, s) and mixing distribution G. For standard choices of the mixing measure G (e.g. exponential or inverse-gamma), these skewed distributions play nicely inside larger, conditionally Gaussian hierarchical models (e.g. Polson and Scott 2013).

Alas, deriving priors for (a, b, s) by formal rules within this class of models is likely to be tedious outside of special cases, because it requires explicitly integrating over G. Nonetheless, we may say at least a few interesting things. In particular, any data set $y_1, \ldots, y_n$ arising from such a model will be consistent with the hypothesis that $s = 0$, as long as G has support on all of $\mathcal{R}^+$. To see this, simply let $b > 0$ and $a < y^{(1)}$, the smallest observation, in which case every residual may be attributed to a positive v_i. Thus the marginal likelihood of the data at $s = 0$ is nonzero, and we reach a simple conclusion: if the prior puts infinite mass in a neighborhood of $s = 0$, then so will the posterior. Therefore, either the Jeffreys-rule prior grows more slowly than s^{-1} near zero, or it will lead to an improper posterior for this model. This is nearly identical to the intuitive explanation offered in Section 3.3, and suggests that there may be something more general here than Rubio and Steel's conclusion about the two-piece model.

3 The proposed solution

Third, the authors have shown that, for the case of the two-piece location-scale model, the independence Jeffreys-rule prior yields a proper posterior even though the formal Jeffreys-rule prior does not. Lest anyone generalize too readily from this finding, allow me to highlight a case where the exact opposite occurs. (I thank Jim Berger for pointing out the following example to me several years ago.) Let $Y(x)$ be a spatial process observed at points $x_i \in \mathcal{R}^d$, where

$$Y(x_i) = \sum_{j=1}^{p} \theta_j f_j(x_i) + \epsilon(x_i) \, .$$

Here the θ_j are unknown, the f_j are known regression functions, and $\epsilon(x)$ is a mean-zero isotropic Gaussian process with

$$\mathrm{cov}\{\epsilon(s), \epsilon(t)\} = \tau^2 K\left(\|s-t\|; \psi\right) \, .$$

In this case, a wide class of objective priors are of the form

$$\pi(\theta_1, \ldots, \theta_p, \sigma^2, \psi) \propto \frac{\pi(\psi)}{(\sigma^2)^c}$$

for different choices of c; see Handcock and Stein (1993). Interestingly, the ordinary Jeffreys-rule prior does result in a proper posterior, but the independence Jeffreys-rule prior does not—just the opposite of the current situation.

Fourth, if I were to encounter a data set tomorrow where skewness played an important role, my initial reaction would have been to use a half-Cauchy prior for the scale parameter in the two-piece model, instead of the independence Jeffreys prior favored by the authors. Although Rubio and Steel have shown that the independence prior leads to a proper posterior, I do not find this enough to recommend the prior in practice. By contrast, the half-Cauchy prior cannot be justified by formal rules, as far as I am aware. But it has proven itself to be a sensible default prior for a scale parameter across a wide range of situations (Gelman 2006; Gelman et al. 2008; Carvalho et al. 2010; Polson and Scott 2012), including model-selection problems where improper priors cannot be used (e.g. Scott and Berger 2006). I have never encountered a case where it leads to obviously stupid results, unlike every formal rule I know except the reference prior.

Finally, there's an old saying in American football that a team with two starting quarterbacks is a team with none. Judging by the sheer number of formal systems out there for defining objective priors—uniform, Jeffreys-rule, reference, right-Haar, left-Haar, matching, fiducial, maximum-entropy, minimum-description length, and so forth—it could be said with some justice that we don't have one at all. In the face of such embarrassing riches, I would be the last to suggest that the Jeffreys-rule prior is obviously wrong, and some other rule obviously right, and that this question could be decided on first principles. Far better to set aside the philosophizing and judge the operational success of a particular prior in a particular problem, just as Rubio and Steel have done here.

Nonetheless, a good clue that the Jeffreys-rule prior may be a poor choice in general is that Harold Jeffreys himself often recommended priors that were different than the one suggest by his rule. This was especially true in multiparameter problems. Thus a natural question is whether one of those other formal systems for choosing a prior would lead to a similarly poor result in this case. One particular method that has enjoyed a lot of practical success—and that "magically" seems to avoid the pathologies that befall other formal rules for constructing objective priors—is the reference-prior approach. Have the authors looked into the reference prior for this problem, perhaps using the newer techniques described in Berger et al. (2009)? It would be especially depressing if the reference prior gave an improper posterior, too!

References

Berger, J., Pericchi, L., and Varshavsky, J. (1998). "Bayes factors and marginal distributions in invariant situations." *Sankhya, Series A*, 60: 307–321. 26

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). "The formal definition of reference priors." *The Annals of Statistics*, 37(2): 905–38. 28

Bernardo, J. M. and Girón, F. (1988). "A Bayesian analysis of simple mixture problems." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics 3*. Oxford University Press. 26

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–80. 27

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3): 515–33. 27

Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008). "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics*, 2(4): 1360–83. 27

Handcock, M. S. and Stein, M. (1993). "A Bayesian analysis of kriging." *Technometrics*, 35: 403–10. 27

Neyman, J. and Scott, E. (1948). "Consistent estimates based on partially consistent observations." *Econometrica*, 16. 25

Polson, N. G. and Scott, J. G. (2012). "On the half-Cauchy prior for a global scale parameter." *Bayesian Analysis*, 7(4): 887–902. 27

— (2013). "Data augmentation for non-Gaussian regression models using variance-mean mixtures." *Biometrika*, 100(2): 459–71. 26

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136(7): 2144–2162. 27

Wasserman, L. (2000). "Asymptotic inference for mixture models by using data-dependent priors." *Journal of the Royal Statistical Society (Series B)*, 62: 159–80. 26