Bayesian Analysis (2014) **9**, Number 1, pp. 45–52

Rejoinder

Francisco J. Rubio [*] and Mark F. J. Steel [†]

We are very grateful to all discussants of this paper for their thought–provoking and constructive comments and the interesting research directions they indicate. We present our reply to the comments of the discussants in alphabetical order.

[J. M. Bernardo] We agree on the importance of checking the asymptotic normality of the joint posterior distribution induced by the Jeffreys prior, provided it exists. However, we have shown that, in the context of two–piece location–scale models, the Jeffreys–rule prior does not even lead to a proper posterior distribution when the underlying symmetric density f belongs to the family of scale mixture of normals. This result precludes the use of the Jeffreys–rule prior for conducting Bayesian inference altogether.

Although we do not claim optimality of the Jeffreys–rule nor the independence Jeffreys priors in any sense, the optimality of a prior depends, of course, on the optimality criterion. While the Jeffreys–rule prior is not optimal for inferences in certain cases, it represents the optimal choice in some others (see *e.g.* Hedayati and Barlett 2012). Therefore, we think the study of these priors is of interest, especially if this study reveals important pitfalls that occur when such priors are used in apparently simple models. The reality is that Jeffreys priors offer a relatively easily obtained prior candidate in the absence of strong prior information, which has some interesting properties, such as invariance with respect to reparameterizations, and is used quite a lot in practice. In fact, a quick search on Google Scholar (on January 20, 2014) reveals 423 papers since 2013 with the term "Jeffreys prior" (about the same number as containing the phrase "reference prior").

Given the impropriety of the posterior induced by the Jeffreys–rule prior, we present some alternative priors that lead to proper posteriors under certain conditions. The first (standard) alternative studied is the independence Jeffreys prior, which often performs better in practical situations. The Jeffreys–rule and the independence Jeffreys prior coincide for orthogonal parameterizations, which in this case do not seem to be achievable, as discussed by Jones and Anaya-Izquierdo (2011). The second class of alternative priors consists of modifications of the Jeffreys–rule and the independence Jeffreys priors on a more heuristic basis. These priors lead to proper posteriors and, as shown in the Supplementary material (Appendix 2), in a simulation study they also result in posteriors with good frequentist properties.

The study of other types of noninformative priors represents an interesting extension of our work, and the reference prior would indeed be the first alternative we would try.

[*] University of Warwick, Department of Statistics, Coventry, CV4 7AL, UK. Francisco.Rubio@warwick.ac.uk

[†] University of Warwick, Department of Statistics, Coventry, CV4 7AL, UK. M.F.Steel@stats.warwick.ac.uk

DOI:10.1214/14-BA875

For the general location–scale model with any choice of sampling density (skewed or not), Fernández and Steel (1999b) derived that the reference prior is $\pi(\mu, \sigma) \propto \sigma^{-1}$. We conjecture that once we introduce a parameter to capture skewness in the model as in Subsection 2.2, the reference prior will exhibit the structure $\pi(\mu, \sigma, \gamma) \propto \sigma^{-1}\pi(\gamma)$, for some function $\pi(\gamma)$, just like for the skew-normal distribution in Liseo and Loperfido (2006) (note that the structure of the information matrix in their Appendix D is the same as that in the proof of our Theorem 3). This would mean that the reference prior has the same structure as in (23) in Subsection 3.4 with a particular choice of $\pi(\gamma)$ and would thus be covered by Theorem 6. The exact expression for the reference prior and its properties (and especially how these compare with the priors proposed in Subsection 3.4) remain an exciting topic for further research.

[J. G. Scott] The link between two–piece models and finite mixtures models together with the fact that some improper priors lead to proper posteriors for two–piece models may indeed look intriguing at first glance. An intuitive explanation for this seemingly paradoxical situation is that, although two–piece models can be seen as mixture models, the elements of this mixture are restricted in the sense that the weight $\varepsilon = \sigma_1/(\sigma_1 + \sigma_2)$ is tied to the scale parameters (σ_1, σ_2), and the location parameter μ is the same for both components of the mixture. These restrictions produce a mixture model that allows for proper posteriors with certain improper priors. In other words, we have no component–specific parameters in our model, so the "usual" problem with improper priors (due to allocating no observations to one of the components) is avoided.

Analogous results for the existence of the posterior distribution in skew–symmetric scale mixtures of normals using the independence Jeffreys prior (which has a similar product structure) are provided in Rubio and Liseo (2014).

The comparison with variance-mean mixture models provides an interesting parallel: indeed, if $s = 0$ then skewness is maximized in these models (just as when one of the scales tends to zero in our two–piece model). Of course, the amount of skewness generated by the variance-mean mixture model will depend not only on s but also on the mixing distribution. Barndorff-Nielsen (1977) pointed out that the hyperbolic distribution can be represented as a normal-mean mixture with a generalized inverse Gaussian mixing distribution. This is one of the two choices for the mixing distribution that Polson and Scott (2013) focus on. The fact that it is not trivial to conduct formal "objective" Bayesian analyses in this context is illustrated by Fonseca et al. (2012) who derive the Jeffreys prior for the hyperbolic model and conclude that it does not exist in closed form and use numerical integration. They also report some numerical experiments using this prior but do not present a proof of the propriety of the corresponding posterior.

The use of the half–Cauchy prior (or half–Student–t prior, more generally) indeed represents another appealing choice for priors on scale parameters. In Rubio and Steel (2013a) we construct a "vague proper prior" for skew–symmetric and two–piece sampling distributions, in the context of stress–strength models, using this sort of prior for the scale parameters. Recently, other types of vague proper priors for scale parameters have been proposed. A common feature of these priors is that they are heavy–tailed, which can be interpreted as a translation of "vague prior beliefs". The impact on the inference of this kind of "vague" priors depends, however, on the units of measurement of the data.

Inspired by this comment, we investigated the use of half–Cauchy priors on the scales in the context of our two–piece model in (2). Given the intuition in Subsection 3.3 and the fact that these priors integrate close to zero (or anywhere else, for that matter), we would expect the posterior to exist. Indeed, we can show the following:

Theorem R.10. *Let* $\mathbf{y} = (y_1, \dots, y_n)$ *be an independent sample from the model in (2), where* f *is a scale mixture of normals. Define the prior*

$$\pi(\mu, \sigma_1, \sigma_2) \propto \frac{1}{1+\sigma_1^2}\frac{1}{1+\sigma_2^2}. \tag{R.1}$$

Then, the posterior distribution of $(\mu, \sigma_1, \sigma_2)$ *is proper under the following conditions:*

(i) If $n \geq 2$ *and all the observations are different.*

(ii) Suppose that the sample $\mathbf{y}$ *contains repeated observations and* k *is the largest number of observations with the same value in* $\mathbf{y}$*. If* $1 < k < n$*, then the posterior is proper if the mixing distribution of* f *satisfies* (18)*. In the case of the two-piece normal sampling model (i.e. normal* f*), it suffices to have two different observations.*

Proof. See Appendix

We have investigated the empirical coverage of the two-piece model in (2) with the prior in (R.1) using the same simulation setup as in Appendix 2 of the Supplementary material. Results for the coverage of the 95% posterior credible intervals with sample size $n = 100$ are summarized in Table R.1, which indicates that the coverage with this prior based on half–Cauchy distributions for the scales is quite good (about the same as with the independence Jeffreys prior in one case, and even slightly better in the other), in line with the conjecture of the discussant.

Sample size	$n = 100$	
Parameters	$\sigma_1 = 2.0$ $\sigma_2 = 0.5$	$\sigma_1 = 0.66$ $\sigma_2 = 1.50$
μ	0.957	0.952
σ_1	0.961	0.959
σ_2	0.952	0.954

Table R.1: Coverage proportions. Two–piece model in (2) with half–Cauchy priors on the scales.

We agree that the reference prior would be another interesting candidate to examine in this context. On this issue, please see our replies to the comments of the previous discussant.

[**R. E. Weiss and M. A. Suchard**] In line with most of the literature, we merely used "valid" as a shorthand for inference that can be conducted in a fundamental probabilistic sense, leaving aside properties of possible estimators and similarity to hypothetical perfect analyses. Of course, we agree that sensible Bayesian inference has to go beyond propriety of the posterior distribution; propriety is just the least we should check when using improper priors. In our case, this motivated the simulation study presented in Appendix 2 of the Supplementary material where we checked the coverage of some posterior credible intervals in order to assess the frequentist properties of the proposed priors. A good posterior coverage is typically considered to be a desirable property of a posterior induced by an improper/benchmark prior. However, it is impossible to propose a criterion without hurting someone's feelings.

An equally loaded term is, of course, "benchmark", and the discussants astutely point out that this is by no means a clear-cut concept. We have in mind (again, like most of the literature) a prior that can be used (perhaps only once for a particular practitioner) relatively safely (without having any unexpected impact on the results) in a situation where genuine prior information is either lacking or not used (in an attempt to make the analyses as palatable as possible to a wide set of readers), and that can be formulated without taking "time and data". This is something we feel would be a useful addition to the toolbox of applied users of relatively standard models.

The example using the AIS data set raises an important question in the context of modelling data using flexible distributions: are location, scale, and skewness parameters enough to model asymmetric, seemingly unimodal, data? The answer to this question in general is "no". Departures from normality are usually studied in terms of skewness and tail behaviour, and the latter cannot always be controlled by only one shape parameter. In order to properly model these two aspects, several four or five–parameter distributions have been studied (see Rubio and Steel 2013b and the references therein). So, although three–parameter two–piece models provide extra flexibility (with respect to the original symmetric model), this may not be sufficient for modelling all sorts of data, and additional shape parameters may be necessary. We believe the study of benchmark priors obtained by formal rules for other types of flexible distributions is an interesting research line. We entirely agree on the possible influence of outliers on the inference on the skewness parameter (γ).

Bayesian nonparametric methods might indeed be able to produce a better fit, provided that the sample size is large enough, given their intrinsic higher flexibility. However, it is more difficult to learn about the asymmetry and the tails of the data using this approach, in contrast to the use of parametric distributions containing skewness and kurtosis parameters, such as the two–piece Student–t, which has parameters that are readily interpretable. It is also more complicated to conduct for the applied practitioner than the parametric analysis outlined in our paper. In addition, it does not free the analyst from having to elicit a prior, which is actually a harder task in the nonparametric case. Thus, although the DPM model could certainly be better at fitting (large amounts of) data, we actually do not feel the use of such a model would be "easier". The issues mentioned in the discussion for the implementation of DPM models illustrate the level of complexity and we agree they deserve serious consideration. The use of skewed components in DPM models has been studied, *e.g.* in Kottas and Gelfand (2001), who use median zero components obtained as in (1) but with $\epsilon = 1/2$ so they are generally discontinuous at the origin, and in Kalli et al. (2013), who use uniforms with skewness induced through inverse scale factors. Rodríguez and Walker (2014) adopt infinite mixtures of such uniforms, which amounts to using flexible unimodal components, in the context of modelling the number of clusters in the data.

[**X. Xu**] We fully agree that looking at implications of a prior in terms of an interpretable quantity can help us understand much better the (often hidden) implications of prior choices. The discussion of these Jeffreys–type priors in terms of the interpretable skewness measure AG is quite illuminating and clearly shows that these "objective" priors (especially the Jeffreys prior) have quite extreme implications. However, the reason we can do this so easily in the context of our models is that skewness (as measured by AG) is a function of one single parameter (γ), which is not the case in most competing models for dealing with skewness (for example, the skew–normal model of Azzalini 1985 and its extensions, and the variance–mean mixture models mentioned by the second discussant). This emphasises the importance of using models with interpretable parameters, especially for prior elicitation.

This discussion also focuses on Bayesian model selection with improper priors and in high-dimensional problems. The problems arising with the use of Bayes factors (BF) as a model selection tool, in particular their dependence on the prior, have been widely discussed. The solution proposed in Xu et al. (2011) through the use of the information level is quite interesting and seems to be widely applicable. One other model selection criterion that is not overly sensitive to the impropriety of the prior or the dimension of the problem consists of comparing the predictive performance of the various models considered, measured, for example, through the log predictive scores. A related general question is how to select a model from the growing catalogue of flexible/skew distributions. Given that several proposed families of skewed distributions produce similar degrees of flexibility, the answer to this question is not straightforward. However, we believe that for a sensible model selection we need to consider both formal model selection tools as well as an *ad hoc* evaluation of other intrinsic properties of the competing models, such as interpretability of the parameters, ease of use and inferential properties.

References

Azzalini, A. (1985). "A class of distributions which includes the normal ones." *Scandinavian Journal of Statistics*, 12: 171–178. 49

Barndorff-Nielsen, O. (1977). "Exponentially decreasing distributions for the logarithm of particles size." *Proceedings of the Royal Society of London, A*, 353: 401–419. 46

Fonseca, T. C. O., Migon, H. S., and Ferreira, M. A. R. (2012). "Bayesian analysis based on the Jeffreys prior for the hyperbolic distribution." *Brazilian Journal of Probability and Statistics*, 26: 327–343. 46

Hedayati, F. and Barlett, P. L. (2012). "The optimality of Jeffreys prior for online density estimation and the asymptotic normality of Maximum Likelihood Estimators." *JMLR: Workshop and Conference Proceedings*, 23: 7.1–7.13. 45

Jones, M. C. and Anaya-Izquierdo, K. (2011). "On parameter orthogonality in symmetric and skew models." *Journal of Statistical Planning and Inference*, 141: 758–770. 45

Kalli, M., Walker, S. G., and Damien, P. (2013). "Modeling the conditional distribution of daily stock index returns: An alternative Bayesian semiparametric model." *Journal of Business and Economic Statistics*, 31: 371–383. 49

Kottas, A. and Gelfand, A. E. (2001). "Bayesian Semiparametric Median Regression Modeling." *Journal of the American Statistical Association*, 96: 1458–1468. 49

Liseo, B. and Loperfido, N. (2006). "A note on reference priors for the scalar skew-normal distribution." *Journal of Statistical Planning and Inference*, 136: 373–389. 46

Polson, N. G. and Scott, J. G. (2013). "Data augmentation for non-Gaussian regression models using variance-mean mixtures." *Biometrika*, 100: 459–471. 46

Rodríguez, C. E. and Walker, S. G. (2014). "Univariate Bayesian nonparametric mixture modeling with unimodal kernels." *Statistics and Computing*, 24: 34–49. 49

Rubio, F. J. and Liseo, B. (2014). "On the independence Jeffreys prior for skew-symmetric models." *Statistics and Probability Letters*, 85: 91–97. 46

Rubio, F. J. and Steel, M. F. J. (2013a). "Bayesian Inference for $P(X < Y)$ Using Asymmetric Dependent Distributions." *Bayesian Analysis*, 8: 43–62. 46

— (2013b). "Bayesian modelling of skewness and kurtosis with two-piece scale and shape transformations." Technical Report CRiSM working paper 13-10, University of Warwick. 48

Xu, X., Lu, P., MacEachern, S. N., and Xu, R. (2011). "Calibrated Bayes factors for model comparison." Technical Report Technical Report 855, Department of Statistics, The Ohio State University. 49

Appendix: Proof of Theorem R.1

Consider the change of variable given by $\sigma_1 = \sigma(1+\gamma)$ and $\sigma_2 = \sigma(1-\gamma)$, with $\gamma \in (-1,1)$. The determinant of the Jacobian of this transformation is given by $|\mathcal{J}| = 2\sigma$. Then, the prior (R.1) becomes

$$\pi(\mu,\sigma,\gamma) \propto \sigma \frac{1}{1+\sigma^2(1+\gamma)^2} \frac{1}{1+\sigma^2(1-\gamma)^2}.$$

For $-1 < \gamma \leq 0$ we have that there exists a constant k_1 such that

$$\pi(\mu, \sigma, \gamma) \leq \frac{k_1 \sigma}{1 + \sigma^2 (1-\gamma)^2} \leq \frac{k_1 \sigma}{1+\sigma^2} < \frac{k_1}{\sigma}.$$

Analogously, for $0 < \gamma < 1$ we have that there exists a constant k_2 such that

$$\pi(\mu, \sigma, \gamma) \leq \frac{k_2 \sigma}{1 + \sigma^2 (1+\gamma)^2} \leq \frac{k_2 \sigma}{1+\sigma^2} < \frac{k_2}{\sigma}.$$

Therefore, we have that $\pi(\mu, \sigma, \gamma) \leq \dfrac{1}{\sigma}$, up to a proportionality constant. This upper bound corresponds to the *AG*–Beta prior with hyperparameters $\alpha_0 = \beta_0 = 1$, and, consequently, the result follows from Theorem 6. □