Bayesian Analysis (2014) **9**, Number 1, pp. 39–44

Comment on Article by Rubio and Steel

Xinyi Xu *

Prior elicitation is an important and challenging problem in Bayesian analysis. When little prior knowledge is available for the model parameters (which is commonly the case when the model is high dimensional), a standard approach is to use "noninformative" or "weakly-informative" prior distributions. When the posterior distribution is proper and the sample size is reasonably large, the Bayesian estimates under these priors are usually close to those obtained by frequentist methods, and thus are viewed as "objective" estimates. However, this approach falls apart in some situations.

The authors of this paper show an interesting case where even for quite simple models such as the two-piece location-scale models, the widely used "noninformative" Jeffreys prior leads to improper posteriors and thus prevents valid Bayesian inference for the models. They cleverly propose two alternative classes of priors for the two-piece location-scale models and particularly recommend one of them, which focuses on the Arnold-Groeneveld (AG) measure of skewness. These AG priors have nice interpretations, lead to proper posteriors for all practically interesting subclasses of these models, and can be easily implemented by practitioners in many scientific and industrial fields. This work provides significant methodological and practical contributions to the literature. I'd like to discuss two aspects of prior elicitation that are reflected in this work.

1 The impact of model parametrization on prior elicitation

Although some priors such as the Jeffreys priors are invariant to model parametrization, many common priors are not, so different model parameterizations can lead to different prior choices. In this paper, the authors consider two different parameterizations of the two-piece location-scale models in Sections 2.1 and 2.2. In both model specifications, the model parameters are not directly interpretable, nor can people easily collect information on them. Therefore, improper "noninformative" priors are placed on the model parameters in pursuit of "objective" analysis. This is an all too common practice in Bayesian inference. However, when the models are parameterized with non-interpretable parameters, the "noninformative" priors on convenient model specifications are not necessarily noninformative; instead, they could implicitly contain strong undesirable information on important model features.

In the work of Rubio and Steel, for the Inverse Scale Factors (ISF) model, the Jeffreys

*The Ohio State University xinyi@stat.osu.edu

 DOI:10.1214/13-BA864

prior and the independent Jeffreys prior are provided in (13) and (14) as

$$\begin{aligned}\pi_J(\mu,\sigma,\gamma) &\propto \frac{1}{\sigma^2(1+\gamma^2)}\\ \pi_I(\mu,\sigma,\gamma) &\propto \frac{1}{\sigma}\sqrt{\frac{\alpha_2}{\gamma^2}+\frac{4}{(\gamma^2+1)^2}},\end{aligned}$$

where α_2 is a constant determined by the symmetric density f. The AG measure of skewness is a function of γ and can be represented for the ISF model as $AG = (\gamma^2-1)(\gamma^2+1)$. Therefore, it is easy to derive the implied priors on AG under the Jeffreys and the independent Jeffreys priors as

$$\begin{aligned}\pi_J(AG) &\propto \frac{1}{2}\sqrt{\frac{1}{1-AG^2}}\\ \pi_I(AG) &\propto \sqrt{\frac{\alpha_2}{(1-AG^2)^2}+\frac{1}{1-AG^2}},\end{aligned}$$

respectively. As shown in Figure 1, these priors have infinite peaks at $AG = 1$ or -1, which correspond to models that are either extremely left-skewed or extremely right-skewed; and reach their minima when $AG = 0$, which corresponds to the models that are symmetric. This mass assignment is counter-intuitive in most practical situations. Similarly, for the ε-skew model, the implied priors on AG under the Jeffreys and the independent Jeffreys prior in (16) and (17) are

$$\begin{aligned}\pi_J(AG) &\propto \frac{1}{1-AG^2}\\ \pi_I(AG) &\propto \sqrt{\frac{1}{1-AG^2}},\end{aligned}$$

which also strongly favor extremely skewed models and place little mass around symmetric models. This undesirable assignment of mass across the model space is hidden when the models are parameterized with (μ,σ,γ), and could be dangerous in practice if not examined carefully. On the other hand, the AG beta prior in (25) recommended by the authors is constructed directly on the rescaled AG measure, and thus allows us to incorporate information on this meaningful parameter when we have prior knowledge and to avoid unintentionally incorporating strong prior information when we do not.

The danger of unintentionally including strong prior information by using "noninformative" priors on conventional model parameters is not restricted to inference in two-piece location-scale models. In fact, it exists in a wide range of problems. For example, Hans et al. (2012) points out that in normal linear models, the standard normal priors on regression coefficients might contain strong information on the regression relationship (measured by R^2), which depends critically on the model dimensions. As a remedy, they construct a class of priors that focuses directly on the regression relationship, such that the same prior on R^2 is maintained over models of difference sizes. Another example in the literature is associated with the portfolio choice problem, in which the excess

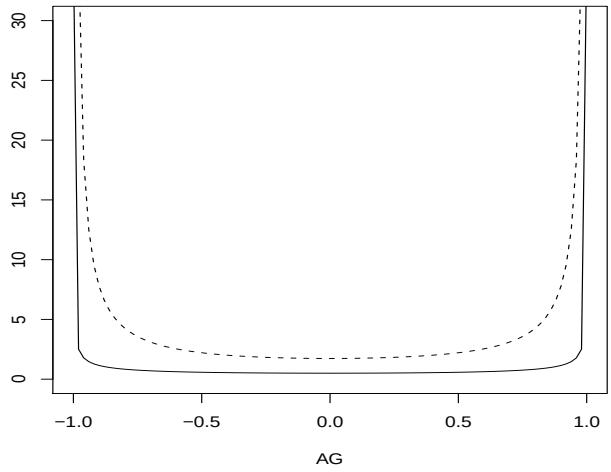

Figure 1: The solid line represents the density $\pi_J(AG)$ implied by the Jeffreys prior on (μ, σ, γ), and the dashed line represents the density $\pi_I(AG)$ implied by the independent Jeffreys prior on (μ, σ, γ). The constant α_2 in π_I is computed under the standard normal distribution.

returns of risky assets are usually assumed to follow multivariate normal distributions with unknown mean and unknown covariance matrix. The seemingly innocuous diffuse priors on these parameters can actually imply rather strong prior information in various applications. For testing portfolio efficiency, Kandel et al. (1995) shows that a diffuse prior on the model parameters implies strong information on inefficiency of a given portfolio; and for predicting portfolio returns, Lamoureux and Zhou (1996) shows that the diffuse prior implicitly assigns most prior mass on either high or low degrees of return predictability. To fix these issues, several approaches (e.g. Chevrier and McCulloch 2008; Tu and Zhou 2010) have been recently proposed to construct priors on the tangency portfolio weights, which is a function of the unknown mean and covariance matrix, or to construct priors based on economic theories.

As shown by the above results, careful examinations of the implications of "objective" priors on model features can help us to understand the behaviors of the corresponding Bayes estimates and should be an important step in Bayesian inference. Focusing directly on important model features can often help us to construct better priors and to avoid pitfalls of implicitly placing mass at undesirable regions of the model space.

2 The impact of information content in priors on Bayesian inference

In the work of Rubio and Steel, the major criticism of using Jeffreys or independent Jeffreys priors is that they lead to improper posteriors for some interesting two-piece location-scale models. What if the posterior is proper? Then even if a prior distribution does not reflect our prior belief, won't the prior information be washed out by the data information? And if so, why do we need to carefully examine the information contained in the prior? It is true that in many estimation problems, when the model is low to medium dimensional and the sample size is relatively large, the prior information is dominated by the likelihood. However, this is not always the case under increasingly prevalent high-dimensional or even infinite-dimensional models and hierarchical models with complex structures. Moreover, in Bayesian hypothesis testing and model comparison, the prior distributions on the model parameters can have huge impacts even with a large amount of data (see Kass and Wasserman (1995) and the references therein).

Xu et al. (2011) show that for Bayesian model selection between two models with different dimensions, the use of arbitrarily vague proper priors can yield misleading Bayes factors. They illustrate an example where the data is generated from a skew-normal distribution and a Gaussian parametric model is compared with a Mixture of Dirichlet Process (MDP) model. Improper noninformative priors are not amenable to Bayes factor calculations when the models differ in dimension, because they are determined only up to an arbitrary constant. Additionally, for infinite-dimensional models (such as the MDP model), a proper prior distribution is required to produce a proper posterior. Therefore, standard diffuse but proper priors are placed on the model parameters, and thus the Bayes factor is computable. However, at a fixed sample size, the value of the Bayes factor can be vastly different under priors with different levels of diffuseness, which is arbitrarily selected by the analyst. This poses serious questions to the robustness of the model preference. A key observation in Xu et al. (2011) is that to obtain robust and reliable results in Bayesian hypothesis tests, priors on model parameters must be proper and not have too big a spread (a similar comment was made by Jeffreys about Lindley's paradox). This leads to the issue of how to measure the "information level" in a prior distribution.

Suppose that $Y_1, Y_2, \ldots, Y_n \mid \theta \overset{iid}{\sim} f_\theta$ and $\theta \sim \pi$, where π is a prior distribution. The Fisher information is often used to measure the amount of information for a parametric model, however, it is not applicable when π is nonparametric. Therefore, for a general π, Xu et al. (2011) propose to measure the information by the proximity of the distributions $f_{\theta^{(1)}}$ and $f_{\theta^{(2)}}$, where $\theta^{(1)}$ and $\theta^{(2)}$ are two random draws from π. The intuition is that when π is highly concentrated (high information), $\theta^{(1)}$ and $\theta^{(2)}$ tend to be close to each other, so $f_{\theta^{(1)}}$ and $f_{\theta^{(2)}}$ are also close; when π is diffuse (low information), $\theta^{(1)}$ and $\theta^{(2)}$ tend to be far away, so $f_{\theta^{(1)}}$ and $f_{\theta^{(2)}}$ are also very different. The proximity of $f_{\theta^{(1)}}$ and $f_{\theta^{(2)}}$ is measured by the symmetrized Kullback-Leibler (SKL) divergence

$$\mathrm{SKL}(f_{\theta^{(1)}}, f_{\theta^{(2)}}) = \frac{1}{2}\int f_{\theta^{(1)}}(y)\log\frac{f_{\theta^{(1)}}(y)}{f_{\theta^{(2)}}(y)}dy + \frac{1}{2}\int f_{\theta^{(2)}}(y)\log\frac{f_{\theta^{(2)}}(y)}{f_{\theta^{(1)}}(y)}dy.$$

The randomness of $(\theta^{(1)}, \theta^{(2)})$ induces a distribution on SKL. Then the information contained in π is evaluated by the percentiles of this distribution of SKL. The advantages of the above information measurement are that it is well defined for both parametric and nonparametric priors and for both proper and improper diffuse priors. Under this information metric, Xu et al. (2011) show that we can mimic the performance of the Bayes factor under a reasonable default prior by calibrating overdispersed prior distributions using part of the data as training samples, such that they achieve a sensible level of "information", and then compute the Bayes factor based on the calibrated priors and the remaining data.

Recently, another approach for evaluating prior informativeness has been proposed in Meng et al. (2013) using the posterior matching method, but they focus on measuring the prior-likelihood conflict instead of the prior concentration.

In summary, during the process of prior elicitation in Bayesian inference, statisticians should pay close attention to: 1) the model parametrization and the hidden messages implied by the priors; and 2) the information level contained in the priors and its impact on the analysis results. When the model dimension is high or the model structure is complex, this could be challenging because it is not always obvious which model features we should focus on, and it is not always clear how to represent the model features using (potentially a large number of) model parameters. Additionally, it could be difficult to calibrate the amount of prior information when the data has special features such as being censored and being spatial/temporal dependent. Careful examinations of prior information in these situations pose important tasks for future development of prior elicitation.

References

Chevrier, T. and McCulloch, R. (2008). "Using Economic Theory to Build Optimal Portfolios." Working papers, University of Chicago. 41

Hans, C., MacEachern, S., and Som, A. (2012). "Structuring Dependence in Regression with R-prior Distributions." Presented in Special topic session at Eleventh ISBA World Meeting, Kyoto, Japan. 40

Kandel, S., McCulloch, R., and Stambaugh, R. F. (1995). "Bayesian inference and portfolio efficiency." *Review of Financial Studies*, 9: 1–53. 41

Kass, R. E. and Wasserman, L. (1995). "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association*, 90: 928–934. 42

Lamoureux, C. and Zhou, G. (1996). "Temporary components of stock returns: What do the data tell us?" *Review of Financial Studies*, 9: 1033–1059. 41

Meng, X. L., Reimherr, M., and Nicolae, D. (2013). "What Your Prior Isn't Telling You: Assessing Prior Informativeness and Prior-Likelihood Conflict." Submitted papers. 43

Tu, J. and Zhou, G. (2010). "Incorporating economic objectives into Bayesian priors: portfolio choice under parameter uncertainty." *Journal of Financial and Quantitative Analysis*, 45: 959–986. 41

Xu, X., Lu, P., MacEachern, S. N., and Xu, R. (2011). "Calibrated Bayes Factors for Model Comparison." Technical Report 855, Department of Statistics, The Ohio State University. 42, 43