Bayesian Analysis (2015) **10**, Number 1, pp. 227–231

Comment on Article by Berger, Bernardo, and Sun*

Manuel Mendoza† and Eduardo Gutiérrez-Peña‡

1 Introduction

The search of a prior distribution $p(\boldsymbol{\omega})$ to be used as part of an objective Bayesian analysis of a model $p(x|\boldsymbol{\omega})$ has proved to be a formidable endeavour. This is an area where we do not have a definitive answer yet, and any contribution to the understanding of the subject must be welcome. The authors of this paper are among the most prominent contributors to this field, and reading the manuscript has been very stimulating.

Research on the problem has mainly dealt with three issues: first, a definition of what a *non-informative, reference* or *objective* prior $p(\boldsymbol{\omega})$ must be; second, an operational algorithm to calculate such priors; third, the evaluation of the resulting prior(s) in accordance to certain criteria such as invariance, the avoidance of paradoxes, or desirable frequentist properties.

To us, and this is a subjective judgment, the most convincing approach to produce this sort of priors is reference analysis (Bernardo, 1979; Berger and Bernardo, 1992a,b; Bernardo, 2005; Berger et al., 2009). This procedure: (i) defines the reference prior as the prior maximizing the expected gain of information provided by a sample; (ii) includes a general (although potentially involved) algorithm to calculate the prior; and (iii) avoids a number of paradoxes. Moreover, it generalizes the Jeffreys prior and exhibits its limitations. Among its most remarkable results, it shows that the form of the reference prior $p(\boldsymbol{\omega})$ may depend on the function of the parameters $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ which is considered by the researcher to be of main interest.

Since its inception, the algorithm to obtain reference priors has evolved. This is the case specifically in the multiparameter setting. The most recent version (Berger and Bernardo, 1992a,b) requires all scalar components of the parameter to be *strictly* ordered in terms of their inferential interest. Thus, in principle, the current approach does not offer any solution if the researcher is simultaneously interested in two or more scalar parameters (or functions thereof). Interestingly, the original algorithm of Bernardo (1979) did cover this situation, although the solution was the multivariate Jeffreys prior which leads to unsettling paradoxes in some cases.

In this paper, the authors explore some ideas to extend the reference analysis to this yet unsolved case. They also seem to be considering a more general version of the problem by assuming that the number of scalar parameters (or functions of the parameters) of interest may be greater than the number of parameters in the model.

*Main article DOI: 10.1214/14-BA915.
†Dept. of Statistics, ITAM, Mexico, mendoza@itam.mx
‡Dept. of Probability and Statistics, UNAM, Mexico, eduardo@sigma.iimas.unam.mx

 DOI: 10.1214/14-BA936

So, the question is: What should the objective prior $\pi^R(\boldsymbol{\omega})$ ($\boldsymbol{\omega} \in \mathbb{R}^k$) be if there are m functions $(\theta_1(\boldsymbol{\omega}), \theta_2(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega}))$ which are of simultaneous interest, where m is not constrained to be less than or equal to k? Three methods to produce the required prior distribution are discussed: (i) the common reference prior; (ii) the reference distance approach; and (iii) the hierarchical approach.

2 Common reference prior

This is not really a method. If the reference priors corresponding to $\theta_i(\boldsymbol{\omega})$ as the parameter of interest ($i = 1, \ldots, m$) are the same for any ordering of the remaining parameters, then the posed problem simply vanishes. It is interesting to see some examples illustrating particular cases where the common prior exists, but it is desirable – and would be much more useful – to have general results characterizing sampling models where, for example, Theorem 2.1 applies and hence a common reference prior may be found. In this regard, results such as those in Gutiérrez-Peña and Rueda (2003) and Consonni et al. (2004) could provide a good starting point. These authors find reference priors for wide classes of exponential families that include the family discussed in Section 2.1.3 of the present paper as a particular case.

It must be pointed out that this section relies on the analysis of the information matrix $\boldsymbol{I}(\boldsymbol{\theta})$, so all reviewed scenarios assume $m \leq k$. Also, a somewhat disquieting result is that of Section 2.2.2, where the authors show that $\pi^R(\psi_1, \psi_2, \psi_3, \mu_1, \mu_2) \propto (\psi_1\psi_2)^{-1}$ is the one-at-a-time reference prior for any of these parameters and any possible ordering. In particular, it is the reference prior for the case where μ_2 is the parameter of main interest. It so happens, however, that this prior is equivalent to the right-Haar prior which leads to a problematic posterior precisely for μ_2. This result would imply that, in general, reference analysis might produce inadequate posteriors *for the parameter of interest*, depending on the specific accompanying parameters.

3 Reference distance method

In order to introduce this method, the authors explicitly assume that $\boldsymbol{\theta} = \boldsymbol{\omega}$, hence $m = k$. The idea is to find an overall prior $\pi(\boldsymbol{\theta})$ such that each of its marginal posteriors $\pi(\theta_i|\boldsymbol{x})$ is close to the corresponding marginal posterior $\pi_i(\theta_i|\boldsymbol{x})$ obtained when θ_i is the parameter of interest ($i = 1, \ldots, m$). As a measure of approximation the authors propose a weighted average of expected logarithmic divergences, although other measures could in principle be used. Also, the search for the overall prior is restricted to a specific parametric family $\mathcal{F} = \{\pi(\boldsymbol{\theta}|\boldsymbol{a}), \boldsymbol{a} \in \mathcal{A}\}$. Apart from the fact (acknowledged by the authors) that the existence of an optimal $\boldsymbol{a}$ is not guaranteed, a rather unappealing feature of this proposal is its dependence on the family $\mathcal{F}$. The authors offer no guidance on how to choose $\mathcal{F}$ *in general*. If the aim is to produce an objective approach, it seems desirable that $\mathcal{F}$ be somehow *intrinsic* to the sampling model. The examples in the paper suggest that perhaps this could be achieved through some kind of conjugacy.

Incidentally, the reference distance method bears some resemblance to the mean-field approach to variational inference, which is relatively straightforward in the case of

exponential families with conjugate priors; see, for example, Bishop (2006, Chapter 10). What is the authors' take on this?

We would like now to comment on Example 3.2.4. There, the normal model $N(x|\mu, \sigma)$ is considered, and the parameters of interest are μ, σ and $\phi = \mu/\sigma$. (Note that, despite the authors' remark at the beginning of Section 3, here $\boldsymbol{\theta} \neq \boldsymbol{\omega}$ and $m > k$.) In any case, the authors remind us that the reference prior when μ or σ is the parameter of interest is $\pi(\mu, \sigma) = \sigma^{-1}$, whereas the reference prior for $\phi = \mu/\sigma$ is given by $\pi_\phi(\mu, \sigma) = (2\sigma^2 + \mu^2)^{-1/2}\sigma^{-1}$. They then propose, as a "natural" choice, the class of relatively invariant priors $\mathcal{F} = \{\pi(\mu, \sigma) = \sigma^{-a}; a > 0\}$. For this family, they show that the overall prior for (μ, σ, ϕ) can be approximated by $\pi^o(\mu, \sigma) = \sigma^{-1}$, so that inclusion of ϕ as an additional parameter of interest makes no difference. We find this rather disappointing. From an algorithmic point of view, this outcome is not surprising given the choice of $\mathcal{F}$ and the form of the reference priors for μ and σ. Only a large weight on the divergence corresponding to ϕ could lead to a different result. An idea that springs to mind is to try another (arguably more "natural" family) such as $\mathcal{G} = \{\pi(\mu, \sigma|a_1, a_2) = (2\sigma^2 + \mu^2)^{-a_1}\sigma^{-a_2}; a_1 > 0, a_2 > 0\}$, which includes all three reference priors for μ, σ and ϕ. On the other hand, since $\pi_\mu(\mu, \sigma)$ and $\pi_\sigma(\mu, \sigma)$ are equal in this case, the authors could alternatively have minimized the sum of the two divergences corresponding to the marginal posterior of ϕ and the *joint* posterior of (μ, σ)). We wonder how these alternative ideas compare with that proposed in the paper for this example.

4 Hierarchical approach

The idea of this approach is, first, to find a "natural" parametric family of proper priors $\pi(\boldsymbol{\theta}|a)$ such that $a \in \mathbb{R}$ and the integrated likelihood results in a proper density $p(x|a)$. Then, the univariate reference prior for a, $\pi^R(a)$, is obtained for this latter model. Finally, the overall prior $\pi^o(\boldsymbol{\theta})$ is defined as the expectation of $\pi(\boldsymbol{\theta}|a)$ with respect to $\pi^R(a)$. This is an intuitive and seemingly reasonable idea. However, it is not clear how to make explicit that $\boldsymbol{\theta}$ is the parameter of interest even though the model is originally indexed by $\boldsymbol{\omega}$, especially when the dimension of $\boldsymbol{\theta}$ is larger than that of $\boldsymbol{\omega}$. (See the comment below concerning the multi-normal means example.) We wonder if the authors can provide some advice on how this could be achieved in general. On the other hand, as in the reference distance case, dependence upon a specific family of priors introduces no small amount of arbitrariness in the method. Here, again, a proper objective method would use an *intrinsic* family entirely determined by the sampling model. One possibility, particularly suitable for the case of hierarchical models, would be to elaborate on the idea of conjugate likelihood distributions (George et al., 1993), although a suitable restriction should be imposed on the corresponding conjugate family in order to get a one-dimensional hyperparameter. Concerning the implementation of the method, the authors suggest that integration to get the overall prior can be avoided by using $\pi^o(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\hat{a})$ instead, where $\hat{a}$ is the mode of the posterior $p(a|\boldsymbol{x})$. This proposal may be efficient from a computational point of view, but it is both surprising and disappointing since it essentially reduces the hierarchical approach to a standard empirical Bayes procedure and leads to a data-dependent prior.

The example in Section 4.2 concerning the multivariate hypergeometric model is confusing and does not quite illustrate the method described above. First, the parameters of the sampling model are given a multinomial prior (which does not depend on a single scalar parameter a, but on a vector of probabilities $\boldsymbol{p}_k$); then, the likelihood is integrated and shown to yield a multinomial distribution. In the process, the k original parameters $R_1, R_2, \ldots, R_k$ are replaced by the parameters $p_1, p_2, \ldots, p_k$, so the idea of reducing the problem to the determination of the reference prior for a scalar parameter is abandoned. Next, in the multinomial model, the approximate overall prior obtained *using the reference distance method* is adopted for the hyperparameters $\boldsymbol{p}_k$. Finally, the corresponding integrated Multinomial–Dirichlet distribution is declared as the overall prior for $R_1, R_2, \ldots, R_k$. We find this *ad hoc* combination of methods difficult to justify as a general procedure.

An alternative formulation could be based on the idea of super-populations (quite common in the field of survey sampling) as follows. Let us assume that a random sample of size N is obtained from a multinomial distribution $Mu_k(\boldsymbol{Y}_k|1, \boldsymbol{p}_k)$. As a result we get a vector $R_1, R_2, \ldots, R_k$ describing the number of sampled units in each category. Now imagine that we then get a subsample of size n, *without replacement*, from the sample of size N. In this setting, the multinomial distribution describes an infinite super-population, the sample of size N is the finite population of interest and the subsample of size n is the actual sample we observe. Given the sample, the likelihood based on the subsample corresponds to that of a hypergeometric distribution. However, with respect to the super-population, the subsample is just a sample of the original multinomial population whose parameters are given by the vector $\boldsymbol{p}_k$. Within this framework, $R_1, R_2, \ldots, R_k$ are observables and any inference regarding these quantities must be produced through the corresponding posterior predictive distribution. This argument shows that the hypergeometric problem can be viewed as a multinomial one where the interest is not really on the parameters but on observables, and the relevant overall prior is that for $\boldsymbol{p}_k$, no matter which method we use.

The example on the multi-normal means (Section 4.3) deserves a few words as well. Here, the parameters of interest are, using the same notation as the authors, μ_i; $i = 1, \ldots, m$ and $|\boldsymbol{\mu}|^2 = \mu_1^2 + \cdots + \mu_m^2$. First, we note that throughout the paper k refers to the dimension of $\boldsymbol{\omega}$ and m is the number of parameters of interest (the dimension of $\boldsymbol{\theta}$), so in this example we have $m = k + 1$. It must be pointed out, however, that the hierarchical method, as defined, cannot be applied when $m > k$ since the distribution $\pi(\boldsymbol{\theta}|a)$ would then be defined over a space of functionally related components of $\boldsymbol{\theta}$ and would be singular. This fact is implicitly recognized by the authors when they propose a prior for $(\mu_1, \ldots, \mu_m)$ only, ignoring the last parameter of interest, $|\boldsymbol{\mu}|^2$. They then argue that the resulting overall prior is reasonable not only for each mean μ_i but also for $|\boldsymbol{\mu}|^2$. The key issue here is the convenient choice of $\pi(\boldsymbol{\mu}|a)$ as the product of the normals $N(\mu_i|0, a)$. So, strictly speaking, this problem is not actually solved by using the hierarchical approach as proposed in the paper but by an *ad hoc* choice of $\pi(\boldsymbol{\mu}|a)$.

5 Final remarks

This paper contains many interesting ideas and examples. However, it offers more of a brainstorming than a systematic treatment and a general solution to the problem. It is somewhat disappointing that the methods proposed in the paper bear little resemblance with the original reference prior approach, where the problem is clearly stated, the criterion used is sensible, and one can typically obtain *unique* and reasonable solutions. The approaches proposed here are still far from becoming operational algorithms since they require a number of arbitrary inputs. Hopefully, at least one of these methods will evolve into an *overall objective procedure* to find overall objective priors. We believe the reference distance method to be the most promising in this regard.

References

Berger, J. O. and Bernardo, J. M. (1992a). "On the development of reference priors." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, 35–60. Oxford: University Press (with discussion). MR1380269. 227

Berger, J. O. and Bernardo, J. M. (1992b). "Ordered group reference priors, with applications to multinomial problems." *Biometrika*, 79: 25–37. MR1158515. doi: http://dx.doi.org/10.1093/biomet/79.1.25. 227

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). "The formal definition of reference priors." *Annals of Statistics*, 37: 905–938. MR2502655. doi: http://dx.doi.org/10.1214/07-AOS587. 227

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society, Series B*, 41: 113–147 (with discussion). MR0547240. 227

Bernardo, J. M. (2005). "Reference analysis." In Dey, D. K. and Rao, C. R. (eds.), *Bayesian Thinking: Modeling and Computation*, *Handbook of Statistics 25*, 17–90. Amsterdam: Elsevier. MR2490522. doi: http://dx.doi.org/10.1016/S0169-7161(05)25002-2. 227

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer. MR2247587. doi: http://dx.doi.org/10.1007/978-0-387-45528-0. 229

Consonni, G., Veronese, P., and Gutiérrez-Peña, E. (2004). "Order-invariant group reference priors for natural exponential families having a simple quadratic variance function." *Journal of Multivariate Analysis*, 88: 335–364. MR2025617. doi: http://dx.doi.org/10.1016/S0047-259X(03)00095-2. 228

George, E. I., Makov, U. E., and Smith, A. F. M. (1993). "Conjugate Likelihood Distributions." *Scandinavian Journal of Statistics*, 20: 147–156. MR1229290. 229

Gutiérrez-Peña, E. and Rueda, R. (2003). "Reference priors for exponential families." *Journal of Statistical Planning and Inference*, 110: 35–54. MR1944632. doi: http://dx.doi.org/10.1016/S0378-3758(01)00281-6. 228