Statistical Science
2020, Vol. 35, No. 2, 234–253
https://doi.org/10.1214/19-STS731

Checking for Prior-Data Conflict Using Prior-to-Posterior Divergences

David J. Nott, Xueou Wang, Michael Evans and Berthold-Georg Englert

Abstract. When using complex Bayesian models to combine information, checking consistency of the information contributed by different components of the model for inference is good statistical practice. Here a new method is developed for detecting prior-data conflicts in Bayesian models based on comparing the observed value of a prior-to-posterior divergence to its distribution under the prior predictive distribution for the data. The divergence measure used in our model check is a measure of how much beliefs have changed from prior to posterior, and can be thought of as a measure of the overall size of a relative belief function. It is shown that the proposed method is intuitive, has desirable properties, can be extended to hierarchical settings, and is related asymptotically to Jeffreys' and reference prior distributions. In the case where calculations are difficult, the use of variational approximations as a way of relieving the computational burden is suggested. The methods are compared in a number of examples with an alternative but closely related approach in the literature based on the prior predictive distribution of a minimal sufficient statistic.

Key words and phrases: Bayesian inference, model checking, prior data-conflict, variational Bayes, Bayesian inference.

1. INTRODUCTION

In modern applications, statisticians are often confronted with the task of either combining data and expert knowledge, or of combining information from diverse data sources using hierarchical models. In these settings, Bayesian methods are very useful. However, whenever we combine different sources of information, it is important to check the consistency of the information contributed by different components of the model for inference. This work is concerned with the problem of detecting situations in which information coming from the prior and the data are in conflict in a Bayesian analysis. Such conflicts can highlight a lack of understanding of the information put into the model, and it is only when there is no conflict between prior and data that we can expect Bayesian inference to show robustness to the prior (Al Labadi and Evans, 2017). See Andrade and O'Hagan (2006) for a discussion of Bayesian robustness and the behaviour of Bayesian inference in the case of prior-data conflict.

Here a new and attractive approach to measuring prior-data conflict is introduced based on a prior-to-posterior divergence, and the comparison of the observed value of this statistic with its prior predictive distribution. We show that this method extends easily to hierarchical settings, and has an interesting relationship asymptotically with Jeffreys' and reference prior distributions. For the prior-to-posterior divergence, we consider the class of Rényi divergences (Rényi, 1961), with the Kullback–Leibler divergence as an important special case. In the present context, the Rényi divergence can be thought of as giving an overall measure of the size of a relative belief function, which is a function describing for each possible value of a given parameter of interest how much more or less likely it has become after observing the data. Evans (2015) and

David J. Nott is Associate Professor and the corresponding author, Department of Statistics and Applied Probability, Singapore 117546, and Operations Research and Analytics Cluster, National University of Singapore, Singapore 119077 (e-mail: standj@nus.edu.sg). Xueou Wang is Ph.D Student, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546 (e-mail: a0095911@u.nus.edu). Michael Evans is Professor, Department of Statistical Sciences, University of Toronto, Toronto, Ontario, M5S 3G3, Canada (e-mail: mevans@utstat.utoronto.ca). Berthold-Georg Englert is Professor and Principal Investigator, Center for Quantum Technologies and Department of Physics, National University of Singapore, 117542 and MajuLab, CNRS-UNS-NUS-NTU International Joint Research Unit, UMI 3654, Singapore (e-mail: cqtebg@nus.edu.sg).

Baskurt and Evans (2013) give details of some attractive solutions to many inferential problems based on the notion of relative belief. A large change in beliefs from prior-to-posterior (where this is calibrated by the prior predictive) may be indicative of conflict between prior and likelihood, so that a check with prior-to-posterior Rényi divergence as the checking discrepancy is an intuitive one for prior-data conflict detection.

Checks for prior-data conflict have usually been formulated within the broader framework of Bayesian predictive model checking, although much of this work is concerned with approaches which check the prior and model jointly (see, e.g., Gelman, Meng and Stern (1996) and Bayarri and Castellanos (2007) for entries into this literature). In general the idea is that there is a discrepancy function $D(y)$ of data y (where a large value of this discrepancy might represent an unusual value) and then for some reference predictive density $m(y)$ a p-value is computed as

$$p = P\big(D(Y) \geq D(y_{\text{obs}})\big), \tag{1}$$

where $Y \sim m(y)$ is a draw from the reference predictive distribution and y_{obs} is the observed data. A small p-value indicates that the observed value of the discrepancy is surprising under the assumed model, and that the model formulation might need to be re-examined. The choice of discrepancy will reflect some aspect of the model fit that we wish to check, and this is generally application specific. The reference predictive density $m(y)$ needs to be chosen, and there are many ways that this can be done. For example, $m(y)$ might be the prior predictive density $\int g(\theta)p(y|\theta)\,d\theta$ (Box, 1980), where $g(\theta)$ is the prior density and $p(y|\theta)$ is the density of y given θ. Another common choice of reference distribution is the posterior predictive for a hypothetical replicate (Guttman, 1967, Rubin, 1984, Gelman, Meng and Stern, 1996). More complex kinds of replication can also be considered, particularly in the case of hierarchical models. In some cases, the discrepancy might also be allowed to depend on the parameters, in which case the reference distribution defines a joint distribution on both the parameters and y. When the discrepancy is chosen in a casual way in the posterior predictive approach it may be hard to interpret checks in a similar way across different problems, and a variety of authors have suggested modifications which have better calibration properties (Bayarri and Berger, 2000, Robins, van der Vaart and Ventura, 2000, Hjort, Dahl and Steinbakk, 2006). The choice of a suitable discrepancy and reference distribution in Bayesian predictive model checking often depends on statistical goals, and this is discussed more later.

Checking for prior-data conflict is distinct from the issue of whether the likelihood component of the model is adequately specified. An incorrect likelihood specification means that there are no parameter values which provide a good fit to the data, whereas a prior-data conflict occurs when the prior puts all its mass in the tails of the likelihood. See Chapter 5 of Evans (2015) for a discussion of different kinds of model checks. As a definition of a prior-data conflict check, we can say it is a Bayesian predictive check of the form given in equation (1), where $D(y)$ is a function of a minimal sufficient statistic, and the reference distribution is the prior predictive distribution. To see that this definition captures the statistical intuition of inconsistency between the observed likelihood and the prior, consider the observed likelihood as a kind of functional test statistic, where we want to know whether it is unusual compared to what is expected under the prior. Since we want to compare with what is expected under the prior, we use the prior predictive as a reference distribution, and since we are interested in the likelihood itself, this is determined by the value of a minimal sufficient statistic (i.e., we restrict the discrepancy to be a function of a minimal sufficient statistic since other aspects of the data are irrelevant to the likelihood). So the formulation of a prior-data conflict check as a special kind of Bayesian predictive check with restrictions on the discrepancy and reference distribution is logical.

Although we focus here on prior-data conflict checks, and not on checking the adequacy of the likelihood specification, it should be noted that adequacy of the likelihood specification needs to be checked first before any prior-data conflict check is performed. For if the sampling model is inadequate, this means that there is no value of the model parameter that provides a good fit to the data, and sound inferences cannot result from such a model no matter what prior is chosen. With regard to checking the sampling model, Carota, Parmigiani and Polson (1996) describe one method for this problem related to the current work. They consider checking model adequacy by defining a model expansion and then measuring the utility of the expansion. Their preferred measure of utility is the marginal prior-to-posterior Kullback–Leibler divergence for the expansion parameter, and they consider calibration by comparison of the Kullback–Leibler divergence with its value in some reference situations involving simple distributions. Their use of a prior-to-posterior divergence in a model check is related to our approach and an interesting complement to our method for prior-data conflict checking. The approach is very flexible, but the elements of their construction need to be chosen with care to avoid confounding prior-data conflict checking with assessing the adequacy of the likelihood, and their approach to calibration of the diagnostic measure is also quite different.

Henceforth, we will focus exclusively on model checking with the aim of detecting prior-data conflicts. We postpone a comprehensive survey of the literature on prior-data conflict assessment to the next section, after first describing the basic idea of our own approach. However,

one feature of many existing suggestions for prior-data conflict checking is that they require the definition of a noninformative prior. Among methods that don't require such a choice, our approach is closely related to that of Evans and Moshonov (2006). They modify the approach to model checking given by Box (1980) by considering as the checking statistic the prior predictive density value for a minimal sufficient statistic, and they use the prior predictive distribution as the reference predictive distribution. They show that these choices are logical ones for the specific purpose of checking for prior-data conflict.

In Section 2, we introduce the basic idea of our method and discuss its relationship with other approaches in the literature. In Section 3, a series of simple examples where calculations can be done analytically is described. In Section 4, we consider the asymptotic behaviour of the checks, and some more complex examples are considered in Section 5 where computational implementation using variational approximation methods is considered. Section 6 concludes with some discussion.

2. PRIOR-DATA CONFLICT CHECKING

2.1 The Basic Idea

Let θ be a d-dimensional parameter and y be data to be observed. We will assume henceforth that all distributions such as the joint distribution for (y, θ) can be defined in terms of densities with respect to appropriate support measures and that in the continuous case these densities are defined uniquely in terms of limits (see, e.g., Appendix A of Evans (2015)).

We consider Bayesian inference where the prior density is $g(\theta)$ and $p(y|\theta)$ is the density of y given θ. The posterior density is $g(\theta|y) \propto g(\theta)p(y|\theta)$. We consider checks for prior-data conflict based on a prior-to-posterior Rényi divergence of order α (Rényi, 1961) (sometimes referred to as an α divergence),

$$(2) \quad R_\alpha(y) = \frac{1}{\alpha - 1} \log \int \left\{ \frac{g(\theta|y)}{g(\theta)} \right\}^{\alpha-1} g(\theta|y)\, d\theta,$$

where $\alpha > 0$ and the case $\alpha = 1$ is defined by letting $\alpha \to 1$. This corresponds to the Kullback–Leibler divergence, and we write

$$\mathrm{KL}(y) = \lim_{\alpha \to 1} R_\alpha(y) = \int \log \frac{g(\theta|y)}{g(\theta)} g(\theta|y)\, d\theta.$$

Also of interest is to consider $\alpha \to \infty$, which gives the maximum value of $\log \frac{g(\theta|y)}{g(\theta)}$, and we write

$$\mathrm{MR}(y) = \lim_{\alpha \to \infty} R_\alpha(y)$$

for this maximum relative belief statistic. Our proposed p-value for the prior-data conflict check is

$$(3) \quad p_\alpha = p_\alpha(y_{\mathrm{obs}}) = P\big(R_\alpha(Y) \geq R_\alpha(y_{\mathrm{obs}})\big),$$

where y_{obs} is the observed value of y and $Y \sim p(y) = \int g(\theta)p(y|\theta)\, d\theta$ is a draw from the prior predictive distribution. This is a measure of how surprising the observed value $R_\alpha(y_{\mathrm{obs}})$ is in terms of its prior distribution. For if this is small, then the distance between the prior and posterior is much greater than expected. The use of p-values in Bayesian model checking as measures of surprise is well established, but we emphasize here that these p-values are not measures of evidence, and it may be better to think of the tail probability (3) as a calibration of the observed value of $R_\alpha(y_{\mathrm{obs}})$. However, we will continue to use the well-established p-value terminology in what follows. If $R_\alpha(Y)$, $Y \sim p(y)$, is continuous, and if $y_{\mathrm{obs}} \sim p(y)$, then the tail probability (3) is one minus the transformation of $R_\alpha(y_{\mathrm{obs}})$ by its distribution function, and hence uniformly distributed on $[0, 1]$. So the p-value is a useful measure of suprise in the sense that we know what to expect of it if the data are generated under the prior predictive distribution. Similar to conventional hypothesis testing, if the distribution of the divergence is not continuous the situation is more complex. The case of the hierarchical checks discussed later in Section 2.2 is also more complex, and exact uniformity for finite samples will not usually hold there for the checks of the conditional prior.

We will use the special notation p_{KL} and p_{MR} for the p-values based on the discrepancies $\mathrm{KL}(y)$ and $\mathrm{MR}(y)$ respectively. In the definition (2), it was assumed that we want an overall conflict check for the prior. If interest centers on a particular quantity $\Psi(\theta)$, however, we can look at the marginal prior-to-posterior divergence for Ψ instead of θ in (2). If a predictive perspective is adopted, it is also possible to consider some as yet unobserved data y^* and a prior to posterior divergence involving predictive densities for y^*,

$$\frac{1}{\alpha - 1} \log \int \left\{ \frac{p(y^*|y)}{p(y^*)} \right\}^{\alpha-1} p(y^*|y)\, dy^*,$$

where $p(y^*|y)$ here denotes the posterior predictive and $p(y^*)$ the prior predictive. We consider this later in an example.

2.2 Motivations for the Check

The prior-data conflict check (3) can be motivated from a number of points of view. First, the choice of discrepancy is intuitive, since $R_\alpha(y)$ is a measure of how much beliefs change from prior to posterior, and comparing this measure for y_{obs} against what is expected under the prior predictive intuitively tells us something about how surprising the observed likelihood is under the prior. This point of view connects with the relative belief framework for inferences summarized in Baskurt and Evans (2013) and Evans (2015). For a parameter of interest $\Psi = \Psi(\theta)$,

the relative belief function is the ratio of the posterior density of Ψ to its prior density,

$$\mathrm{RB}(\Psi|y) = \frac{g(\Psi|y)}{g(\Psi)}.$$

$\mathrm{RB}(\Psi|y)$ measures how much belief in Ψ being the true value has changed after observing data y. If $\mathrm{RB}(\Psi|y)$ is bigger than 1, this says that there is evidence for Ψ being the true value, whereas if it is less than 1 this says that there is evidence against. Use of the Rényi divergence as the discrepancy in (3) is equivalent to the use of the discrepancy

$$\|\mathrm{RB}(\theta|y)\|_s = \{E(\mathrm{RB}(\theta|y)^s|y)\}^{1/s} \tag{4}$$

as a test discrepancy, where $s = \alpha - 1$, since $R_\alpha(y) = \log \|\mathrm{RB}(\theta|y)\|_s$. (4) is a measure of the overall size of the relative belief function. The limit $s \to 0$ gives $\exp(\mathrm{KL}(y))$, $s \to \infty$ gives $\mathrm{RB}(\hat{\theta}|y)$ where $\hat{\theta}$ denotes the maximum relative belief estimate which maximizes the relative belief function, and $s = 1$ is the posterior mean of the relative belief.

In Section 4, we also investigate the asymptotic behaviour of p_α, which under appropriate conditions converges to

$$P\big(g(\theta^*)|I(\theta^*)|^{-1/2} \geq g(\theta)|I(\theta)|^{-1/2}\big) \tag{5}$$

in the large data limit, where $I(\theta)$ is the Fisher information at θ, θ^* is the true value of the parameter that generated the data, and $\theta \sim g(\theta)$. To interpret (5), note that $g(\theta)|I(\theta)|^{-1/2}$ is just the prior density, but written with respect to the Jeffreys' prior as the support measure rather than the Lebesgue measure. So (5) is the probability that a draw from the prior has prior density value less than the prior density value at the true parameter. It is a measure of how far out in the tails of the prior the true value θ^* lies. There is a similar limit result for the check of Evans and Moshonov (2006), but where the densities are with respect to the Lebesgue measure (Evans and Jang, 2011a). Interestingly, (5) might be thought of as giving some kind of heuristic justification for why the Jeffreys' prior could be considered noninformative—if we were to choose $g(\theta)$ as the Jeffreys' prior, $g(\theta) \propto |I(\theta)|^{1/2}$, then the value of the limiting p-value (5) is 1 and hence there can be no conflict asymptotically. Some similar connections with reference priors (Berger, Bernardo and Sun, 2009, Ghosh, 2011) are considered in Section 4 for hierarchical versions of our checks and we discuss these in Section 2.2. While formally inserting the Jeffreys' prior into the limiting p-value (5) leads to a p-value of 1, we note that there is no contradiction here with our earlier observation of uniformity of the p-value from our check under the prior predictive distribution. In fact, a distribution with point mass at 1 is not obtained under continuity by considering the finite sample version of our check using the Jeffreys' prior, considering the p-value as a random variable indexed by sample size n, and then letting n go to infinity. The derivation of the form of the limiting p-value via the arguments of Section 4 fails when $g(\theta)$ is the Jeffreys' prior, as we explain later. Also in the case where the Jeffreys' prior is improper, both our divergence statistic and the prior predictive reference distribution are not well defined. The connection with the Jeffreys' prior obtained by examining the form of the limiting p-value is rather heuristic, and a more rigorous formalization of this connection may be challenging. It would be interesting to investigate to what extent noninformative priors can be characterized through the lens of lack of conflict, but that is not our intention in the present work.

Further motivation for our approach follows from some logical principles that any prior-data conflict check should satisfy. Evans and Moshonov (2006) and Evans and Jang (2011b) consider for a minimal sufficient statistic T a decomposition of the joint model as

$$\begin{aligned} p(\theta, y) &= p(t)g(\theta|t)p(y|\theta, t) \\ &= p(t)g(\theta|t)p(y|t), \end{aligned} \tag{6}$$

where the terms in the decomposition are densities with respect to appropriate support measures, $p(t)$ is the prior predictive density for T, $g(\theta|t)$ is the density of θ given $T = t$ (which is the posterior density since T is sufficient) and $p(y|t)$ is the density of y given $T = t$ (which does not depend on θ because of the sufficiency of T). This decomposition modifies a suggestion of Box (1980) for model checking. In the case where there is no nontrivial minimal sufficient statistic a decomposition (6) can still be contemplated for some asymptotically sufficient T such as the maximum likelihood estimator. The three terms in the decomposition could logically be specified separately in defining a joint model and they perform different roles in an analysis. For example, the posterior distribution $p(\theta|t)$ is used for inference, and $p(y|t)$ is useful for checking the likelihood, since it does not depend on the prior. Ideally a check of adequacy for the likelihood should not depend on the prior since the adequacy of the likelihood has nothing to do with the prior.

For checking for prior-data conflict, Evans and Moshonov (2006) and Evans and Jang (2011b) argue that the relevant part of the decomposition (6) is the prior predictive distribution of T. Since a sufficient statistic determines the likelihood, a comparison between the likelihood and prior can be done by comparing the observed value of a sufficient statistic to its prior predictive distribution. Clearly any variation in y that is not a function of a sufficient statistic does not change the likelihood, and hence is irrelevant to determining whether prior and likelihood are in conflict. Furthermore, a minimal sufficient

statistic will be best for excluding as much irrelevant variation as possible. For a minimal sufficient statistic T, the p-value for the check of Evans and Moshonov (2006) is computed as

$$p_{\mathrm{RB}} = p_{\mathrm{RB}}(y_{\mathrm{obs}}) = P\big(p(T) \le p(t_{\mathrm{obs}})\big), \tag{7}$$

where t_{obs} is the observed value of T and $T \sim p(t)$ is a draw from the prior predictive for T. This approach, however, does not achieve invariance to the choice of the minimal sufficient statistic, which is generally not unique; see, however, Evans and Jang (2010) for an alternative approach which does achieve invariance. They also consider conditioning on maximal ancillary statistics when they are available. Coming back from these general principles to the check (3), we notice that the statistic $R_\alpha(y)$ is automatically a function of any sufficient statistic, since it depends on the data only through the posterior distribution. Furthermore, it is the same function no matter what sufficient statistic is chosen. So our check is a function of any minimal sufficient statistic as Evans and Moshonov (2006) and Evans and Jang (2011b) would require, and is invariant to the particular choice of that statistic. The achievement of invariance is an important attraction of our proposal. Note that if we were to transform the minimal sufficient statistic T in (7) to another minimal sufficient statistic T' by a smooth invertible transformation, then the check (7) would differ for T' through the incorporation of a Jacobian factor. For sufficiently extreme choices of the transformation this Jacobian factor can be made to give any answer at all. While there may sometimes be a natural choice for T, lack of invariance is undesirable. Existing proposals for prior-data conflict checking in the literature lack either invariance to transformation of the test statistic, or invariance to parametrization of the model, with the exception of Evans and Jang (2010). However, this method is very difficult to implement, since it requires a computation of the Jacobian of a mapping of the data onto a minimal sufficient statistic. The method proposed here does not require this, and does not even require the identification of any minimal sufficient statistic, since the check is defined directly in terms of the posterior distribution.

2.3 Hierarchical Versions of the Check

Next, consider implementation of the approach of Section 2.1 in a hierarchical setting. Suppose the parameter θ is partitioned as $\theta = (\theta_1, \theta_2)$, where θ_1 and θ_2 are of dimensions d_1 and d_2 respectively, and that the prior is decomposed as $g(\theta) = g(\theta_1|\theta_2)g(\theta_2)$. Sometimes it is natural to consider the decomposition of the prior into marginal and conditional pieces since it may reflect how the prior is specified (such as in the case of a hierarchical model). We may wish to check the two pieces of the prior separately to identify more precisely the source of any prior-data conflict when it occurs. Mirroring our decomposition of the prior, write $g(\theta|y) = g(\theta_1|\theta_2, y)g(\theta_2|y)$. To define a hierarchically structured check, let

$$R_\alpha(y, \theta_2) = \frac{1}{\alpha - 1} \log \int \left\{ \frac{g(\theta_1|\theta_2, y)}{g(\theta_1|\theta_2)} \right\}^{\alpha-1} \times g(\theta_1|\theta_2, y)\, d\theta_1 \tag{8}$$

denote the conditional prior to conditional posterior Rényi divergence of order α for θ_1 given θ_2, and define

$$R_{\alpha 1}(y) = E_{\theta_2|y_{\mathrm{obs}}}\big(R_\alpha(y, \theta_2)\big). \tag{9}$$

$R_{\alpha 1}(y)$ is a function of both y and y_{obs} although we suppress this in the notation. Also, define

$$R_{\alpha 2}(y) = \frac{1}{\alpha - 1} \log \int \left\{ \frac{g(\theta_2|y)}{g(\theta_2)} \right\}^{\alpha-1} g(\theta_2|y)\, d\theta_2$$

so that $R_{\alpha 2}(y)$ is the marginal prior to posterior divergence for θ_2.

For hierarchical checking of the prior, we consider the p-values

$$p_{\alpha 1} = P\big(R_{\alpha 1}(Y) \ge R_{\alpha 1}(y_{\mathrm{obs}})\big), \tag{10}$$

where

$$Y \sim m(y) = \int p(y|\theta)\, p(\theta_1|\theta_2) g(\theta_2|y_{\mathrm{obs}})\, d\theta \tag{11}$$

and

$$p_{\alpha 2} = P\big(R_{\alpha 2}(Y) \ge R_{\alpha 2}(y_{\mathrm{obs}})\big), \tag{12}$$

where $Y \sim p(y) = \int p(\theta) p(y|\theta)\, d\theta$. The p-value (10) is measuring whether the conditional prior to posterior divergence for θ_1 given θ_2 is unusually large for values of θ_2 and a reference distribution for Y that reflects knowledge of θ_2 under y_{obs}. The p-value (12) is just the nonhierarchical check (3) applied to the marginal posterior and prior for θ_2. We explore the behaviour of these hierarchical checks in examples later, as well as by examining their asymptotic behaviour in Section 4, where we find that these checks are related to two stage reference priors. In the above discussion, we can also consider a partition of the parameters with more than two pieces and the ideas discussed can be extended without difficulty to this more general case. We can also consider functions of θ_1 and θ_2, $\Psi_1(\theta_1)$ and $\Psi_2(\theta_2)$, and prior to posterior divergences involving these quantities in the definition of $R_{\alpha 1}(y)$ and $R_{\alpha 2}(y)$. Later, we will also use the special notation $\mathrm{KL}_1(y)$, $\mathrm{KL}_2(y)$, $p_{\mathrm{KL}1}$ and $p_{\mathrm{KL}2}$ for $\lim_{\alpha\to 1} R_{\alpha 1}(y)$, $\lim_{\alpha\to 1} R_{\alpha 2}(y)$, $\lim_{\alpha\to 1} p_{\alpha 1}$ and $\lim_{\alpha\to 1} p_{\alpha 2}$. As mentioned earlier, the limit $\alpha \to 1$ in the Rényi divergence corresponds to the Kullback–Leibler divergence.

An anonymous referee has asked an intriguing question—is it possible for a sequence of hierarchical checks to all pass when an overall prior check fails, or for one of the hierarchical checks to fail but the overall check

to be passed? We conjecture that the answer is yes, since the hierarchical checks are not just answering the same question as an overall check at a finer level of detail. The overall check is blind to the hierarchy that was used in the specification of the prior, whereas the hierarchical checks are looking for conflict in certain directions making use of the hierarchy. Despite our conjecture that inconsistency between these points of view is possible, we have yet to find examples illustrating this. We feel that the hierarchical checks are the right way to check the prior, if prior elicitation was conducted using a hierarchical approach.

There are a number of ways that the basic approach above can be modified. One possibility is to replace the posterior distribution $g(\theta_2|y_{\text{obs}})$ in the reference distribution (11) with an appropriate partial posterior distribution (Bayarri and Berger, 2000, Bayarri and Castellanos, 2007) $g(\theta_2|y_{\text{obs}}\backslash R_{\alpha 1}(y_{\text{obs}}))$, defined for data y by

$$g\big(\theta_2|y\backslash R_{\alpha 1}(y)\big) \propto g(\theta_2)\frac{p(y|\theta_2)}{p(R_{\alpha 1}(y)|\theta_2)}.$$

The partial posterior removes the information in $R_{\alpha 1}(y)$ about θ_2 from the likelihood $p(y|\theta_2)$ in calculating a reference posterior for θ_2 for use in (11). We would also use the partial posterior in taking the expectation in (9). To get some intuition, imagine receiving the information in y in two pieces where we are told the value of $R_{\alpha 1}(y)$ first, followed by the remainder; if we applied Bayes' rule sequentially, first updating the prior $g(\theta_2)$ by $p(R_{\alpha 1}(y)|\theta_2)$, then the "likelihood" term needed to update the posterior given $R_{\alpha 1}(y)$ to the full posterior $g(\theta_2|y)$ would be $\frac{p(y|\theta_2)}{p(R_{\alpha 1}(y)|\theta_2)}$. So the partial posterior just updates the prior for $g(\theta_2)$ by this second likelihood term that represents the information in the data with that from $R_{\alpha 1}(y)$ removed. This somehow avoids an inappropriate double use of the data where the same information is being used to both construct a reference distribution and assess lack of fit. Use of the partial posterior distribution in (11) makes computation of (10) more complicated, however.

There are also other ways that the basic hierarchically structured check can be modified in some problems with additional structure. In their discussion of checking hierarchical priors, Evans and Moshonov (2006) consider two situations. The first situation is where the likelihood is a function θ_1 only, $p(y|\theta) = p(y|\theta_1)$. In this case, suppose that T is a minimal sufficient statistic for θ_1 in the model $p(y|\theta_1)$ and that $V = V(T)$ is minimal sufficient for θ_2 in the marginalized model $\int p(y|\theta_1)p(\theta_1|\theta_2)\,d\theta_1$. Writing t_{obs} and v_{obs} for the observed values of T and V, they suggest further decomposing the term $p(t)$ in (6) as $p(v)p(t|v)$ where $p(v)$ denotes the prior predictive density for V and $p(t|v)$ denotes the prior predictive density for T given $V = v$. In this decomposition, it is suggested that $p(t|v)$ should be used for checking $g(\theta_1|\theta_2)$, by comparing $p(t_{\text{obs}}|v_{\text{obs}})$ with $p(T|v_{\text{obs}})$ for draws of T from $p(t|v_{\text{obs}})$, and then if no conflict is found $p(v)$ should then be used for checking $g(\theta_2)$, by comparing $p(v_{\text{obs}})$ with $p(V)$ for $V \sim p(v)$. So checking $g(\theta_2)$ should be based on the prior predictive for V and checking $g(\theta_1|\theta_2)$ should be based on a statistic that is a function of T with reference distribution the conditional for $T|V = v_{\text{obs}}$ induced under the prior predictive for the data. Looking at our hierarchically structured check, if there exists a minimal sufficient statistic V for θ_2, then we see in (12) our checking statistic $R_{\alpha 2}(y)$ is a function of that statistic and it will be invariant to what minimal sufficient statistic is chosen. We are also using the prior predictive for the reference distribution so our approach fits nicely with that of Evans and Moshonov (2006). In the check (10), we can see that the model checking statistic is a function of T and invariant to the choice of T. If we were to change the reference distribution (11) to that of $T|V = v_{\text{obs}}$, then (10) would also fit naturally with the approach of Evans and Moshonov (2006). However, sometimes suitable nontrivial sufficient statistics are not available and the conditional prior predictive of T given $V = v_{\text{obs}}$ might be difficult to work with. Our general approach of using the posterior distribution of θ_2 given v_{obs} to integrate out θ_2 comes close to achieving the ideal considered in Evans and Moshonov (2006) when there are sufficient statistics at different levels of the model. A final observation is that we could consider a cross-validatory version of the check if interest centered on a certain observation specific parameter within the vector θ_1. These cross-validatory checks are also useful when there is a division of the likelihood into pieces representing different data sources. Excluding one of the data sources, we obtain a posterior which we can consider as the prior to be updated with the left out data to get the full posterior. Then the prior to posterior divergence in this sequential updating of one data source can be considered for either parameters or predictive quantities to evaluate the effect on inferences of interest. Cross-validatory checks are considered in a later example.

The other situation considered in Evans and Moshonov (2006) for checking hierarchical priors is the case where $p(y|\theta)$ can depend on both θ_1 and θ_2. Here they suppose there is some minimal sufficient T and a maximal ancillary statistic $U(T)$ for θ, and a maximal ancillary statistic V for θ_1 (ancillary for θ_1 means that the sampling distribution of V given θ depends only on θ_2). Conditioning on ancillaries is relevant since we don't want assessment of prior-data conflict to depend on variation in the data that does not depend on the parameter. They suggest in (6) decomposing $p(t)$ as $p(u)p(v|u)p(t|v,u)$ and using the second term $p(v|u)$ (the conditional distribution of V given U induced under the prior predictive for the data) to check $g(\theta_2)$, with the third term $p(t|v,u)$ (the conditional distribution of T given V and U under the prior predictive for the data) used to check $g(\theta_1|\theta_2)$. Again we

can modify our suggested approach where this additional structure is available. If we change $g(\theta_2|y)$ to $g(\theta_2|v)$ in the definition of $R_{\alpha 2}(y)$, then we are checking $g(\theta_2)$ using a discrepancy which is a function of V. If no maximal ancillary for θ were available, the suggestion of Evans and Moshonov (2006) would use the prior predictive for V for the reference distribution. Because V is ancillary for θ_1 the check does not depend in any way on $g(\theta_1|\theta_2)$, which is desirable because we would like to check for conflict with θ_2 separately from checking for any conflict with $g(\theta_1|\theta_2)$. For the check (10) our discrepancy is a function of T as Evans and Moshonov (2006) would recommend, and if the reference predictive distribution were changed to be that of T given U and V we could use this approach to check for conflict with $g(\theta_1|\theta_2)$. However, in complex situations identifying suitable maximal ancillary statistics may not be possible. Nevertheless consideration of problems like this provides some guidance as an ideal.

2.4 Other Suggestions for Prior-Data Conflict Checking

Now that we have given the basic idea of our method we discuss its connections with other suggestions in the literature. Perhaps the approach to prior-data conflict detection most closely related to the one developed here has been suggested by Bousquet (2008). Similar to us, Bousquet (2008) considers a test statistic based on prior to posterior (Kullback–Leibler) divergences, but uses the ratio of two such divergences. Briefly, a noninformative prior is defined and a reference posterior distribution for this noninformative prior is constructed. Then, the prior to reference posterior divergence for the prior to be examined is computed and divided by the prior to reference posterior divergence for the noninformative prior. When the noninformative prior is improper, some modification of the basic procedure is suggested, and extensions to hierarchical settings are also discussed. The approach we consider here has similar intuitive roots but is simpler to implement because it does not require the existence or precise definition of a noninformative prior. We consider the prior to posterior divergence for the prior under examination, a measure of how much beliefs have changed from prior to posterior, and compare the observed value of this statistic to its distribution under the prior predictive for the data. There is hence no need to define a noninformative prior, although as mentioned earlier there are interesting asymptotic connections between the checks we suggest and Jeffreys' and reference noninformative priors. This will be discussed further in Section 4. Our focus here is not on deriving noninformative prior choices, however, but on detecting conflict for a given proper prior.

A quite general and practically implementable suggestion for measuring prior-data conflict has been given recently by Presanis et al. (2013). Their approach generalizes earlier work by Marshall and Spiegelhalter (2007) and also relates closely to some previous suggestions by Gåsemyr and Natvig (2009) and Dahl, Gåsemyr and Natvig (2007). They give a general conflict diagnostic that can be applied to a node or group of nodes of a model specified as a directed acyclic graph (DAG). The conflict diagnostic is based on formulating two distributions representing independent sources of information about the separator node or nodes which are then compared. Again, in general, there is a need in this approach to specify noninformative priors for the purpose of formulating distributions representing independent sources of information. O'Hagan (2003) is an earlier suggestion for examining conflict at any node of a DAG that was inspirational for much later work in the area, although the specific procedure suggested has been found to suffer from conservatism in some cases. Scheel, Green and Rougier (2011) consider a graphical approach to examining conflict where the location of a marginal posterior distribution with respect to a local prior and lifted likelihood is examined, where the local prior and lifted likelihood are representing different sources of information coming from above and below the node in a chain graph model. Reimherr, Meng and Nicolae (2014) examine prior-data conflict by considering the difference in information in a likelihood function that is needed to obtain the same posterior uncertainty for a given proper prior compared to a baseline prior. Again, some definition of a noninformative prior for the baseline is needed for this approach to be implemented. Finally, the model checking approach considered in Dey et al. (1998) can also be used for checking for prior-data conflict. There is some similarity with our approach in that they use quantities associated with the posterior itself in the test. Specifically, they consider Monte Carlo tests based on vectors of posterior quantiles and the prior predictive with a Euclidean distance measure used to measure similarity between the vectors of quantiles.

2.5 Relationships with Formal Methods for Model Choice and the Role of Explicit Alternatives

Bayesian model checking methods, including the prior-data conflict checking methods discussed in the previous subsection, do not usually make use of any explicit alternative model. Instead, they attempt to reject a current model without having any alternative model to replace it with. Prompted by some comments by an anonymous referee, we reflect here on possible roles for explicit alternative models within Bayesian model checking generally. There are historical debates within statistics surrounding different approaches to significance testing that are pertinent to the discussion. In so-called pure significance testing (Cox and Hinkley, 1974, Chapter 3), the choice of a test statistic is considered more primitive than the construction of an explicit alternative model, and even

though a test statistic in this framework will have certain departures from the assumed model in mind, no alternative model is considered. Gelman and Shalizi (2013) discuss pure significance testing in relation to Bayesian predictive checking. The p-values in this framework are measures of surprise or of lack of consistency of an observed test statistic with the model, and are not used as part of any formal decision making procedure involving an alternative model with associated mathematical optimalities. However, both formal and informal methods of model criticism and choice would seem to have a role to play in choosing good models, depending on the circumstances.

As mentioned, Bayesian model criticism is usually done in a framework without explicit alternative models, but the work of Robins, van der Vaart and Ventura (2000) does do this. They consider model expansions and score-type discrepancies which are locally most powerful. We have some sympathy for the argument that model expansions can be useful even in informal methods for model checking. A difficult practical question in Bayesian model checking is always the choice of discrepancy. We see the use of explicit alternative models as possibly helpful for addressing this issue, with their role being to assist the imagination—a discrepancy derived from a formal model choice procedure and some explicit working model may give a discrepancy with an intuitive form, but one that we might not have thought of without the aid of the working alternative model. The alternative model may not be taken very seriously in itself, however, and concepts such as power might be of limited interest.

The above discussion was concerned with Bayesian model checking generally and not specifically with prior-data conflict checking. We note that our framework provides guidance on the choice of discrepancy—once a function of the parameter of interest is chosen, we suggest using a prior-to-posterior divergence for that function of the parameter as a discrepancy. If an explicit alternative model were to be considered for purposes such as power computations for prior-data conflict checks, what would that involve? We suggest that the logical way to proceed is to embed the original prior used for the analysis into a family of priors. That is, we consider a family of priors $g(\theta|\gamma)$ where γ is some expansion parameter and the original prior is $g(\theta) = g(\theta|\gamma_0)$ for some value γ_0 for γ. Then suppose we were to make a binary decision regarding the existence of a prior-data conflict by thresholding the tail probability (3) at some conventional level like 0.05. We can consider data generated under the prior predictive for the prior with different values of γ, and look at how frequently prior-data conflicts are declared for different values of γ. This is giving something like a notion of power for prior-data conflict checks. This might be of some interest, depending on the context, and we illustrate this idea in the next section in a simple example.

3. FIRST EXAMPLES

To begin exploring the properties of the conflict check (3), we consider a series of simple examples where calculations can be done analytically. Although an analytic form can be obtained for our discrepancies in these examples, their precise form is not always capable of intuitive interpretation. However, as mentioned earlier, an advantage of our framework is that it provides some guidance on the choice of discrepancy as a prior-to-posterior divergence. The p-values (or tail probabilities) in the examples have the usual interpretation in model checks without an explicit alternative of measures of surprise, measuring lack of consistency of the observed discrepancy with the model. The examples considered here were also given in Evans and Moshonov (2006), and we make some comparisons with their check (7) in each case, but leave algebraic details of derivations to the Appendix.

EXAMPLE 1. *Normal location model.* Suppose $y_1, \ldots, y_n \sim N(\mu, \sigma^2)$ where μ is an unknown mean and $\sigma^2 > 0$ is a known variance. In this normal location model, the sample mean is sufficient for μ and normally distributed so without loss of generality, we may consider $n = 1$ and write the observed data point as y_{obs}. The prior density $g(\mu)$ for μ will be assumed to be $N(\mu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are known.

Here and in later examples, we use the notation $A(y) \doteq B(y)$ to mean that $A(y)$ and $B(y)$ are related (as a function of y) by a monotone transformation. When conducting a Bayesian model check with discrepancies $D_1(y)$ and $D_2(y)$ then they will result in the same predictive p-values if $D_1(y) \doteq D_2(y)$ (although care must be taken to compute the appropriate left or right tail area, since in our definition of the $\doteq$ notation the relationship between $A(y)$ and $B(y)$ can be either monotone increasing or decreasing). Consider the prior-data conflict check based on the Rényi divergence statistic. The posterior density for μ is $N(\tau^2\gamma, \tau^2)$ where $\tau^2 = (1/\sigma_0^2 + 1/\sigma^2)^{-1}$ and $\gamma = (\mu_0/\sigma_0^2 + y/\sigma^2)$ and the prior to posterior Rényi divergence of order α is (using, e.g., the formula in Gil, Alajaji and Linder (2013)),

$$R_\alpha(y) = \log\frac{\sigma_0}{\tau} + \frac{1}{2(\alpha-1)}\log\frac{\sigma_0^2}{\sigma_\alpha^2} + \frac{1}{2}\frac{\alpha(\tau^2\gamma-\mu_0)^2}{\sigma_\alpha^2},$$

where $\sigma_\alpha^2 = \alpha\sigma_0^2 + (1-\alpha)\tau^2$. Here only γ depends on y, so that

$$\begin{aligned} R_\alpha(y) &\doteq (\tau^2\gamma - \mu_0)^2 \\ &\doteq (\gamma - \mu_0/\tau^2)^2 \\ &= (y-\mu_0)^2/\sigma^2 \doteq (y-\mu_0)^2, \end{aligned}$$

The divergence based check turns out to be equivalent to the Evans and Moshonov (2006) check in this example for every value of α. To implement the conflict check of Evans and Moshonov (2006), we need $p(y)$ which is the $N(\mu_0, \sigma^2 + \sigma_0^2)$ density (the sufficient statistic in this case of a single observation is just y). We can write $\log p(y) \doteq (y - \mu_0)^2$ and just like the divergence based check the check of Evans and Moshonov (2006) compares $(y_{\text{obs}} - \mu_0)^2$ to the distribution of $(Y - \mu_0)^2$ for $Y \sim p(y)$. Following the similar example of Evans and Moshonov (2006), page 897, the p-value is

$$p_{\text{RB}} = 2\left(1 - \Phi\left(\frac{|y_{\text{obs}} - \mu_0|}{\sqrt{\sigma^2 + \sigma_0^2}}\right)\right).$$

It is interesting to examine also the predictive version of our check where $R_\alpha(y)$ is defined in terms of the prior-to-posterior divergence for a predictive replicate y^* of y. The prior predictive density for y^* is the $N(\mu_0, \sigma^2 + \sigma_0^2)$ density, and the posterior predictive density for y^* given y is the $N(\tau^2\gamma, \tau^2 + \sigma_0^2)$ density. Writing down the Rényi divergence we see once again that $R_\alpha(y) \doteq (\tau^2\gamma - \mu_0)^2 \doteq (y - \mu_0)^2$ so the predictive perspective leads to the same check here as the divergence based check on the parameters.

Following the discussion of Section 2.4, we consider how a power calculation for an alternative prior could proceed here if that were of interest. Consider a prior family $g(\theta|\mu') = N(\mu', \sigma^2)$, where μ' is a prior hyperparameter that is allowed to vary. Choosing $\mu' = \mu_0$ gives the original prior $g(\theta)$. The prior predictive for $g(\theta|\mu')$ is normal, $g(y|\mu') = N(\mu', \sigma^2 + \sigma_0^2)$. For data y generated under $g(y|\mu')$, we can study how frequently the p-value (3) is less than some cutoff, which we choose here as 0.05, as μ' varies. The probability of a p-value less than 0.05 in the conflict check for $y \sim g(y|\mu')$ is

$$P(\mu') = P\left(2\left(1 - \Phi\left(\frac{|y - \mu_0|}{\sqrt{\sigma^2 + \sigma_0^2}}\right)\right) < 0.05\right),$$

which after some simple algebra leads to

$$P(\mu') = \Phi\left(\frac{|\mu_0 - \mu'|}{\sqrt{\sigma^2 + \sigma_0^2}} - \Phi^{-1}(0.975)\right) + \Phi\left(-\frac{|\mu_0 - \mu'|}{\sqrt{\sigma^2 + \sigma_0^2}} - \Phi^{-1}(0.975)\right).$$

Figure 1 shows plots of this power curve $P(\mu')$ versus μ' with $\mu_0 = 0$, $\sigma_0^2 = 1$ and $\sigma^2 = 1$ and 0.1. The case of $\sigma^2 = 0.1$ can equivalently be thought of as corresponding to conflict checking based on the sample mean (a normally distributed minimal sufficient statistic) for a sample of size 10 from a distribution with variance 1. In this example, other expansions of the original prior could have been used, such as varying the variance hyperparameter.

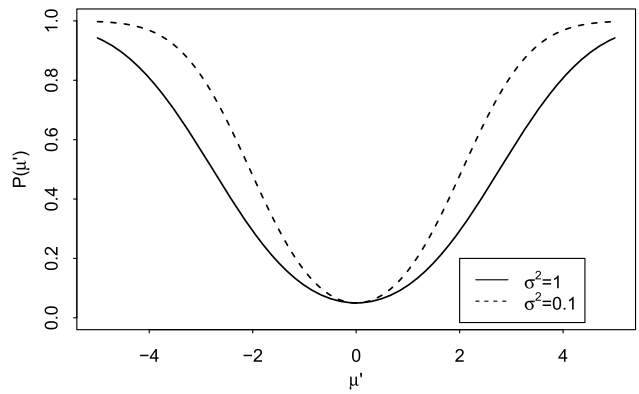

FIG. 1. *Plots of $P(\mu')$ versus μ' in the location normal example with $\mu_0 = 0$, $\sigma_0^2 = 1$, and $\sigma^2 = 1$ (solid) and 0.1 (dashed).*

EXAMPLE 2. *Binomial model.* Suppose that $y \sim \text{Binomial}(n, \theta)$ and write y_{obs} for the observed value. The prior density $g(\theta)$ of θ is Beta(a, b), which for data y results in the posterior density $g(\theta|y)$ being Beta$(a + y, b + n - y)$. Using the expression for the Rényi divergence between two beta distributions (Gil, Alajaji and Linder, 2013)

$$R_\alpha(y) = \log\frac{B(a, b)}{B(a + y, b + n - y)} + \frac{1}{\alpha - 1}\log\frac{B(a + \alpha y, b + \alpha(n - y))}{B(a + y, b + n - y)},$$

where $B(\cdot, \cdot)$ denotes the beta function.

Calculations in the Appendix show that

$$\begin{aligned} R_\alpha(y) &\doteq \log B(a, b) - \frac{1}{2}\log\hat{\theta}_n \\ &\quad - \frac{1}{2}\log(1 - \hat{\theta}_n) \\ &\quad - (a - 1)\log\hat{\theta}_n \\ &\quad - (b - 1)\log(1 - \hat{\theta}_n) + O(1/n) \\ &\doteq -\log g(\hat{\theta}_n) + \frac{1}{2}\log|I(\hat{\theta}_n)| \\ &\quad + O(1/n), \end{aligned} \tag{13}$$

where $\hat{\theta}_n = (a + y)/(a + b + n)$ is the posterior mean, $I(\theta) = n/(\theta(1 - \theta))$ is the Fisher information, and $g(\hat{\theta}_n)$ is the prior density evaluated at $\hat{\theta}_n$. The posterior mean can be replaced by any other estimator differing from it by $O(1/n)$ such as the maximum likelihood estimator. We explain in Section 4 why the form of the result above is expected much more generally.

As a comparison, for the check of Evans and Moshonov (2006), we have

$$\log p(y) \doteq \log g(\hat{\theta}_n) + O(1/n),$$

where as before $\hat{\theta}_n$ is the posterior mean for θ. See the Appendix for further details. A general result about the

check of Evans and Moshonov (2006) explaining the limiting form of the check above is given in Evans and Jang (2011a). So the two checks differ asymptotically owing to the presence of the term $-0.5\log I(\hat{\theta}_n(y))$. See the next section for further discussion.

It is helpful to consider finite sample behaviour in some particular cases. We see that for $R_\alpha(y)$ if we consider $\alpha \to \infty$, we obtain

$$\mathrm{MR}(y) = \log \frac{B(a,b)}{B(a+y, b+n-y)} + \frac{y}{n}\log\frac{y}{n} + \left(1-\frac{y}{n}\right)\log(n-y).$$

If $a = b = 1$ so that the prior is uniform, we see that

$$p_{\mathrm{MR}} = \left(\#\left\{y : \binom{n}{y}\left(\frac{y}{n}\right)^y\left(1-\frac{y}{n}\right)^{n-y} \geq \binom{n}{y_{\mathrm{obs}}}\left(\frac{y_{\mathrm{obs}}}{n}\right)^{y_{\mathrm{obs}}}\left(1-\frac{y_{\mathrm{obs}}}{n}\right)^{n-y_{\mathrm{obs}}}\right\}\right)/(n+1)$$

and plotting $\binom{n}{y}(\frac{y}{n})^y(1-\frac{y}{n})^{n-y}$ reveals that it is symmetric with an antimode at $n/2$ when n is even and at $\{(n+1)/2, 1+(n+1)/2\}$ when n is odd. So prior-data conflict is detected whenever y_{obs} is near 0 or n. This does seem strange when the prior is uniform but is perhaps not surprising given the asymptotic connection between our checks and the Jeffreys' prior, which is also not uniform in this example. On the other hand note that, letting $p(m)$ denote the prior predictive density of $\mathrm{MR}(y)$, then $p(m) = 2/(n+1)$ when n is even for all m except when m is the antimode and when n is odd then $p(m) = 1/(n+1)$ for all m. So if we were to check the prior using $p(m)$ as the discrepancy rather than $\mathrm{MR}(y)$, the p-value would never be small and any conflict would be avoided.

EXAMPLE 3. *Normal location-scale model, hierarchically structured check.* Extending our previous location normal example, suppose $y_1, \ldots, y_n$ are independent $N(\mu, \sigma^2)$ where now both μ and σ^2 are unknown. Write $y = (y_1, \ldots, y_n)$. We consider a normal inverse gamma prior for $\theta = (\mu, \sigma^2)$, $\mathrm{NIG}(\mu_0, \lambda_0, a, b)$ say, having density of the form

$$g(\theta) = \frac{\sqrt{\lambda_0}}{\sigma\sqrt{2\pi}}\frac{b^a}{\Gamma(a)}\left(\frac{1}{\sigma^2}\right)^{a+1} \times \exp\left(-\frac{2b+\lambda_0(\mu-\mu_0)^2}{2\sigma^2}\right).$$

This prior is equivalent to $g(\theta) = g(\theta_2)g(\theta_1|\theta_2) = g(\sigma^2)g(\mu|\sigma^2)$ with $g(\sigma^2)$ inverse gamma, $\mathrm{IG}(a,b)$, and $g(\mu|\sigma^2)$ normal, $N(\mu_0, \sigma^2/\lambda_0)$. In this model a sufficient statistic is $T = (\bar{y}, s^2)$ where $\bar{y}$ denotes the sample mean and s^2 the sample variance and we write $t_{\mathrm{obs}} = (\bar{y}_{\mathrm{obs}}, s^2_{\mathrm{obs}})$ for its observed value. The normal inverse gamma prior is conjugate, and the posterior is $\mathrm{NIG}(\mu_0'(y), \lambda_0', a', b'(y))$ where $\mu_0'(y) = (n+\lambda_0)^{-1}(\mu_0\lambda_0 + n\bar{y})$, $\lambda_0' = n+\lambda_0$, $a' = (a+n/2)$ and $b' = b'(y) = b + (n-1)s^2/2 + n(\bar{y}-\mu_0)^2/(2(n/\lambda_0+1))$. It is natural to consider the hierarchical checks we discussed earlier for testing the two components of $g(\theta)$. First, let us consider the check for conflict with $g(\mu|\sigma^2)$. Using the expression for the Rényi divergence between normal densities, we get

$$R_\alpha(y, \sigma^2) = \log\frac{\lambda_0'}{\lambda_0} + \frac{1}{2(\alpha-1)}\log\frac{{\lambda_0'}^2}{\lambda_0^2} + \frac{1}{2}\frac{\alpha(\mu_0'(y)-\mu_0)^2}{\sigma_\alpha^2},$$

where $\sigma_\alpha^2 = \alpha\sigma_0^2/\lambda_0 + (1-\alpha)\sigma^2/{\lambda_0'}^2$ and we note that

$$R_{\alpha 1}(y) \doteq (\mu_0'(y)-\mu_0)^2 \doteq (\bar{y}-\mu_0)^2.$$

Our suggested hierarchical check compares $R_{\alpha 1}(y_{\mathrm{obs}})$ to a reference distribution based on $Y \sim m(y) = \int p(\sigma^2|y_{\mathrm{obs}}) \int p(y|\mu,\sigma^2)p(\mu|\sigma^2)\,d\mu\,d\sigma^2$. Noting that the distribution of $\bar{y}$ under $m(y)$ is $t_{2a'}(\mu_0, \sqrt{\frac{b'(y_{\mathrm{obs}})}{a'}(\frac{1}{\lambda_0}+\frac{1}{n})})$ we see that the divergence based check just computes whether

$$\frac{\bar{y}_{\mathrm{obs}}-\mu_0}{\sigma^*} = \frac{\bar{y}_{\mathrm{obs}}-\mu_0}{\sqrt{b'(y_{\mathrm{obs}})/a'(1/\lambda_0+1/n)}}$$

is in the tails of a $t_{2a'}(0,1)$ distribution. The hierarchical check of Evans and Moshonov (2006), page 909, on the other hand calculates the probability that $(\bar{y}_{\mathrm{obs}}-\mu_0)/\tilde{\sigma}$ is in the tails of a $t_{2a'-1}(0,1)$ distribution, where $\tilde{\sigma}^2 = (1/\lambda_0(n/\lambda_0+1)(2b+(n-1)s^2_{\mathrm{obs}}))/(n/\lambda_0(n+2a-1))$. Clearly these checks are very similar, since both σ^* and $\tilde{\sigma}$ are approximately $s/\sqrt{\lambda_0}$ for large n and there is only one degree of freedom difference in the reference t-distribution. We also note that in our check if we change the reference distribution to be that of y given s^2 (noting that s^2 is ancillary for μ and following the discussion of Section 2.2) then our check would coincide with that of Evans and Moshonov (2006).

Consider next the check on $p(\sigma^2)$. For two inverse gamma distributions, $p_1(\sigma^2)$ and $p_2(\sigma^2)$, being $\mathrm{IG}(a', b')$ and $\mathrm{IG}(a,b)$ respectively, the Rényi divergence between them is

$$\log\left\{\frac{\Gamma(a){b'}^{a'}}{\Gamma(a')b^a}\right\} + \frac{1}{\alpha-1}\log\left\{\frac{\Gamma(a_\alpha)}{\Gamma(a')}\frac{{b'}^{a'}}{{b_\alpha}^{a_\alpha}}\right\},$$

where $a_\alpha = a'\alpha + (1-\alpha)a$ and $b_\alpha = \alpha b' + (1-\alpha)b$. Since a, b and a' don't depend on the data, this gives

$$R_{\alpha 2}(y) \doteq a'\log b' + \frac{1}{\alpha-1}a'\log b' - \frac{1}{\alpha-1}a_\alpha\log b_\alpha.$$

Using $\log b_\alpha = \log(\alpha b' + (1-\alpha)b) = \log \alpha b' + (1-\alpha)b/(\alpha b') + O(1/n)$ and collecting terms

$$R_{\alpha 2}(y) \doteq \frac{a}{a'}\log b' + \frac{a_\alpha}{a'\alpha}\frac{b}{b'} + O\left(\frac{1}{n}\right) \doteq \log\frac{b'/a'}{b/a} + \frac{b/a}{b'/a'} + O\left(\frac{1}{n}\right).$$

Note also that $s^2 \approx b'/a'$ for large n, so that for large n using $R_{\alpha 2}(y)$ as discrepancy is approximately the same as using

$$\log\frac{s^2}{b/a} + \frac{b/a}{s^2}. \tag{14}$$

As a comparison, for the check in Evans and Moshonov (2006) it is shown in the Appendix that we have approximately for large n

$$\mathrm{RB}(y) \doteq \frac{a-1}{2a}\log\frac{s^2}{b/a} + \frac{b/a}{s^2},$$

which, comparing with (14), clarifies the relationship to the divergence based check.

EXAMPLE 4. *A nonregular example*. The following example is adapted from Jaynes (1976) and Li et al. (2016). Suppose we observe $y_1, \ldots, y_n \sim f(y|\theta)$ where $f(y|\theta) = r\exp(-r(y-\theta))I(y>\theta)$ where r is a known parameter, $\theta > 0$ is unknown and $I(\cdot)$ denotes the indicator function. We consider an exponential prior on θ, $g(\theta) = \kappa\exp(-\kappa\theta)I(\theta>0)$. Note that this is a nonregular example when inference about θ is considered, due to the way that the support of the density for the data depends on θ. This means, for example, that the MLE as well as the posterior distribution are not asymptotically normal. Writing $t = (nr-\kappa)y_{\min}$ (where $y_{\min}$ is the minimum of $y_1, \ldots, y_n$), $\nu = nr/\kappa$ and t_{obs} for the observed value of t, it can be shown (see the Appendix) that the p-value p_α is

$$p_\alpha = p_\alpha(y) = 1 - \int_{t_1}^{t_2} \frac{\nu}{(\nu-1)^2}\left[\exp\left(-\frac{t}{\nu+1}\right) - \exp\left(-\frac{\nu t}{\nu-1}\right)\right]dt,$$

where t_1 and t_2 are such that $R_\alpha(t_1) = R_\alpha(t_2) = R_\alpha(t_{\mathrm{obs}})$ with $t_1 < t_0 < t_2$ and t_0 is the value of t at which $R_\alpha(y) = R_\alpha(t)$ is minimal. There is a single global minimum with $R_\alpha(t)$ decreasing for $t < t_0$ and increasing for $t > t_0$. Either t_1 or t_2 will be equal to t_{obs}. We can easily see that if $t_{\mathrm{obs}} = t_0$ then $p_\alpha = 1$, and if $t_{\mathrm{obs}} \to \infty$ then $p_\alpha \to 0$. Figure 2 considers the special case of the KL divergence and shows some plots of how p_{KL} varies with t_{obs} for a few different values of $\nu = nr/\kappa$.

4. LIMITING BEHAVIOUR OF THE CHECKS

We now give derivations of some of the limit results stated in Section 2. We will consider the special case of the Kullback–Leibler divergence first. Let $y_1, \ldots, y_n$ be independent and identically distributed from $p(y|\theta)$ and denote the true value of θ by θ^*. Write $nI(\theta)$ for the Fisher information and $n\hat{I}_n$ for the observed information. Then under suitable regularity conditions (see, e.g., Theorem 1 of Ghosh, 2011, which summarizes the discussion in Ghosh, Delampady and Samanta, 2006; see also Johnson, 1970) an asymptotic expansion of the posterior distribution gives

$$\log g(\theta|y) + \frac{d}{2}\log\frac{2\pi}{n} - \frac{1}{2}\log|\hat{I}_n| + \frac{n(\theta-\hat{\theta}_n)^T\hat{I}_n(\theta-\hat{\theta}_n)}{2} = O_p\left(\frac{1}{\sqrt{n}}\right)$$

almost surely P_{θ^*}. Adding and subtracting $\log g(\theta)$ from the left-hand side and taking expectation with respect to $g(\theta|y)$ gives

$$\mathrm{KL}(y) + \int (\log g(\theta))g(\theta|y)\,d\theta + \frac{d}{2}\log\frac{2\pi}{n} - \frac{1}{2}\log|\hat{I}_n| + \int \frac{n(\theta-\hat{\theta}_n)^T\hat{I}_n(\theta-\hat{\theta}_n)}{2}g(\theta|y)\,d\theta$$

is $o_p(1)$ and using the asymptotic normality of the posterior and noting that $\hat{I}_n - I(\theta)$ converges to zero almost surely, and $\hat{\theta}_n$ converges to θ^* almost surely under the assumed regularity conditions, gives

$$\mathrm{KL}(y) + \log g(\theta^*) + \frac{d}{2}\log\frac{2\pi e}{n} - \frac{1}{2}\log|I(\theta^*)| = o_p(1).$$

Hence provided that $\log g(\theta^*) - 1/2\log|I(\theta^*)|$ and $\log g(\theta) - 1/2\log|I(\theta)|$ for $\theta \sim g(\theta)$ are not equal with positive probability (which excludes the case where $g(\theta)$ is the Jeffreys' prior), the p-value (3) converges as $n \to \infty$ to

$$P\left(\frac{1}{2}\log|I(\theta)| - \log g(\theta) \geq \frac{1}{2}\log|I(\theta^*)| - \log g(\theta^*)\right),$$

where $\theta \sim g(\theta)$, and this can be written as $P(g(\theta^*) \times |I(\theta^*)|^{-1/2} \geq g(\theta)|I(\theta)|^{-1/2})$.

Next, consider our hierarchical checks and the conflict p-values (10) and (12). The check (12) is really just the same check as in the nonhierarchical case, but ap-

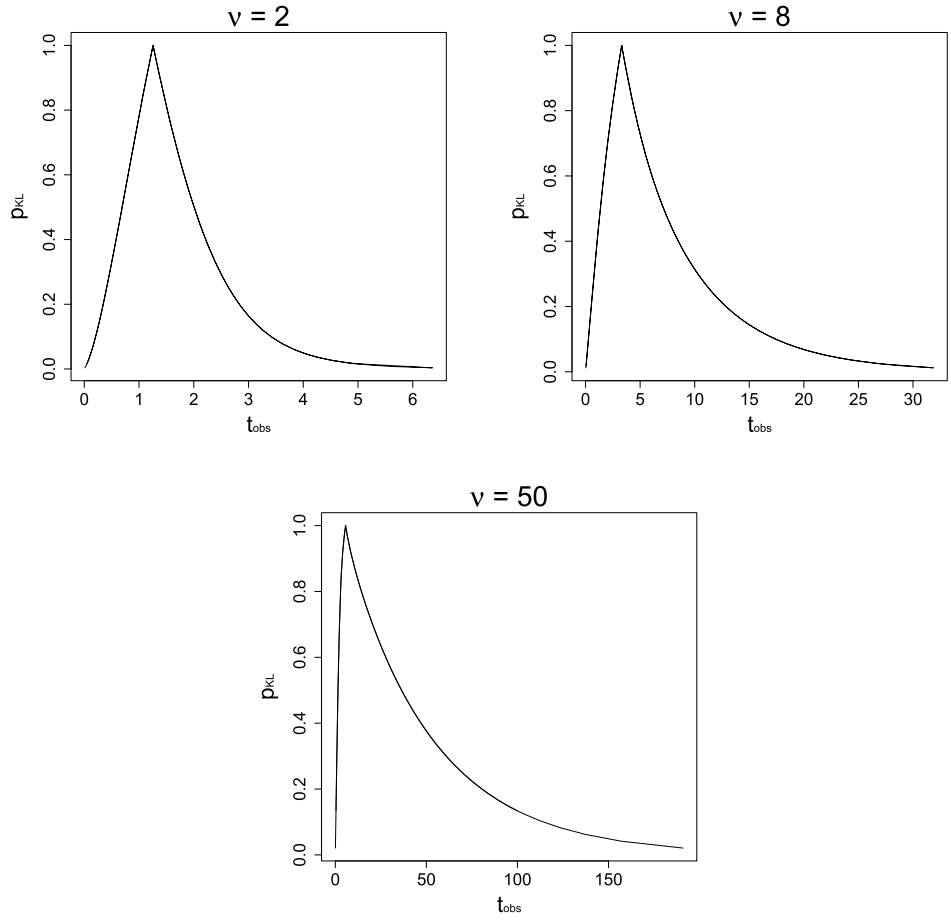

FIG. 2. *Plots of p_{KL} versus t_{obs} for $\nu = 2, 8$ and 50 in Example 4. Note the different scale on the x-axes for the three plots.*

plied to the model and prior with θ_1 integrated out so the limit is the same as in the nonhierarchical case with the Fisher information being that for the marginalized model $p(y|\theta_1) = \int p(y|\theta)p(\theta_2|\theta_1)\,d\theta$, provided that an appropriate asymptotic expansion of the marginal posterior is available. For the check (10), the reference predictive distribution $m(y)$ converges as $n \to \infty$ to $p(y|\theta_2^*) = \int p(y|\theta)p(\theta_1|\theta_2^*)\,d\theta_1$ and in this model with $\theta_2 = \theta_2^*$ fixed we will get the limiting p-value

$$P\big(g(\theta_1^*|\theta_2^*)|I_{11}(\theta_1^*,\theta_2^*)|^{-1/2} \geq g(\theta_1|\theta_2^*)|I_{11}(\theta_1,\theta_2^*)|^{-1/2}\big),$$

where $I_{11}(\theta)$ denotes the submatrix of $I(\theta)$ formed by the first d_1 rows and d_1 columns and $\theta_1 \sim g(\theta_1|\theta_2^*)$. Just as the choice of $g(\theta)$ as the Jeffreys' prior results in a limiting p-value of 1 in the nonhierarchical case, choosing $g(\theta)$ according to the two stage reference prior (Berger, Bernardo and Sun, 2009, Ghosh, 2011) results in the limiting p-values corresponding to (10) and (12) being 1. This provides at least some heuristic reason why, from the point of view of avoidance of conflict, a reference prior might be considered desirable. It is not our intention here, however, to develop methodology for default nonsubjective prior choice or even to justify existing choices, but rather to develop methods for checking for conflict with given proper priors.

Regarding the extension of the above ideas to the more general case of the Rényi divergence, using a Laplace approximation to the integral

$$\int \left\{\frac{g(\theta|y)}{g(\theta)}\right\}^{\alpha-1} g(\theta|y)\,d\theta = \int g(\theta)^{-(\alpha-1)} g(\theta|y)^{\alpha}\,d\theta,$$

expanding about the mode $\hat{\theta}$ of $g(\theta|y)$ and replacing the Hessian of $\log g(\theta|y)$ at the mode with $n\hat{I}_n$, gives

$$(2\pi)^{d/2} g(\hat{\theta}|y)^{\alpha} g(\hat{\theta})^{-(\alpha-1)} |\alpha n \hat{I}_n|^{-1/2}, \tag{15}$$

and using the asymptotic normal approximation to $g(\theta|y)$, $N(\hat{\theta}, n^{-1}\hat{I}_n^{-1})$, so that

$$g(\hat{\theta}|y) \approx (2\pi)^{-d/2} |n\hat{I}_n|^{1/2}, \tag{16}$$

and combining (15) and (16), gives

$$R_\alpha(y) \approx \frac{1}{\alpha-1}\left(\frac{d}{2}\log 2\pi - \frac{\alpha d}{2}\log 2\pi + \frac{\alpha d}{2}\log n + \frac{\alpha}{2}\log|\hat{I}_n|\right.$$

$$- (\alpha - 1)\log g(\hat{\theta}) - \frac{\alpha n d}{2} - \frac{1}{2}\log|\hat{I}_n|\Big)$$

$$\doteq -\log g(\hat{\theta}) + \frac{1}{2}\log|\hat{I}_n|,$$

which converges to $-\log g(\theta) + \frac{1}{2}\log|I(\theta)|$. Hence, we expect a similar limit will hold for the p-value as for the Kullback–Leibler case, under suitable conditions.

5. MORE COMPLEX EXAMPLES AND VARIATIONAL APPROXIMATIONS

To calculate the check (3) or its hierarchical extensions can be difficult. Computation of $R_\alpha(y)$ involves an integral which is usually intractable, and an expensive Monte Carlo procedure may be needed to approximate it. Furthermore, the integrand involves the posterior distribution. Even worse, as well as computing $R_\alpha(y_{\text{obs}})$, we need to compute a reference distribution for it, and this may involve calculating $R_\alpha(y^{(i)})$ for $y^{(i)}$, $i = 1, \ldots, m$, independently drawn from the prior predictive distribution. So a straightforward Monte Carlo computation of p_α may involve calculating $R_\alpha(y)$ for $m+1$ different datasets where m might be large and with each of these calculations itself being expensive. Here we suggest a way to make the computations easier using variational approximation methods. Tan and Nott (2014) also considered the use of variational approximations for computation of conflict diagnostics in hierarchical models and they show a relationship between the diagnostics they consider and the mixed predictive checks of Marshall and Spiegelhalter (2007). Their use of variational approximations for conflict detection is very different to that considered here, however.

In the variational approximation literature there are quite general methods for learning approximations to the posterior that are in the exponential family (Attias, 1999, Jordan et al., 1999, Winn and Bishop, 2005, Rohde and Wand, 2016). If the prior distribution for a certain block of parameters is also in the same exponential family as its variational approximation, it is possible to compute the Rényi divergence in closed form (Liese and Vajda, 1987). Furthermore, because variational approximations are fast to compute, they are ideally suited to the repeated posterior computations for samples under a reference predictive distribution that we need to compute p_α.

More generally there are also useful methods for learning approximations which are mixtures of Gaussians (Salimans and Knowles, 2013, Gershman, Hoffman and Blei, 2012) and if the prior can also be approximated by a mixture of Gaussians then useful closed form approximations to Kullback–Leibler divergences are available (Hershey and Olsen, 2007). We illustrate the use of variational methods for computing approximations of our conflict p-values in two examples. In these examples, we use the Kullback–Leibler divergence as the divergence measure. In the first example, we use a variational mixture approximation, and in the second a Gaussian approximation in a hierarchically structured check for a logistic random effects model. In both cases, there is a parametric family of approximating densities, and the variational approximation involves finding the distribution in the parametric family closest to the true posterior distribution in the Kullback–Leibler sense. Note that because the variational approximation procedure involves an optimization that depends only on the data through the true posterior distribution, the variational approximation is a function of the data only through the posterior distribution, and our prior-data conflict checks making use of the approximation still satisfy the defining property of such checks of the discrepancy being only a function of a minimal sufficient statistic.

EXAMPLE 5. *Beta-binomial example*. We consider the example in Albert (2009), Section 5.4. This example estimates the rates of death from stomach cancer for males at risk aged 45–64 for the 20 largest cities in Missouri. The data set cancer mortality is available in the `R` package `LearnBayes` (Albert, 2009). It contains 20 observations denoted by (n_i, y_i), $i = 1, \ldots, 20$, where n_i is the number of people at risk and y_i is the number of deaths in the ith city. An interesting model for these data is a beta-binomial model with mean η and precision K, where the probability function for the ith observation is

$$p(y_i|\eta, K) = \binom{n_i}{y_i} \frac{B(K\eta + y_i, K(1-\eta) + n_i - y_i)}{B(K\eta, K(1-\eta))}.$$

Albert (2009) considers the prior $g(\eta, K) \propto \frac{1}{\eta(1-\eta)}\frac{1}{(1+K)^2}$ and reparametrizes to $\theta = (\theta_1, \theta_2)$ where

$$\theta_1 = \text{logit}(\eta) = \log\left(\frac{\eta}{1-\eta}\right), \quad \theta_2 = \log(K).$$

We use this parametrization, but since Albert's prior on (η, K) is improper we consider a Gaussian prior for θ, $g(\theta) = N(\mu_0, \Sigma_0)$, where μ_0 is the mean and Σ_0 the covariance matrix. The posterior distribution $g(\theta|y)$ has a nonstandard form, and we approximate it using a Gaussian mixture model (GMM). Variational computations are done using the algorithm in Salimans and Knowles (2013), Section 7.2, where the same dataset was also considered but with Albert's original prior. We consider a two-component mixture approximation,

$$g(\theta|y) \approx q(\theta) = \omega_1 q_1(\theta) + \omega_2 q_2(\theta),$$

where $q(\theta)$ denotes the variational approximation, ω_1 and ω_2 are mixing weights with $\omega_1 + \omega_2 = 1$, and $q_1(\theta)$ and $q_2(\theta)$ are the normal mixture component densities with means and covariance matrices μ_1, Σ_1 and μ_2, Σ_2, respectively. In our check, we replace

$$\text{KL}(y) = \int \log\frac{g(\theta|y)}{g(\theta)} g(\theta|y)\, d\theta$$

with

$$\widetilde{\mathrm{KL}}(y) = \int \log \frac{q(\theta)}{g(\theta)} q(\theta)\, d\theta. \tag{17}$$

$\widetilde{\mathrm{KL}}(y)$ replaces the true posterior $g(\theta|y)$ with its variational approximation. Then we replace the exact computation of (17) with the closed form approximation of Hershey and Olsen (2007), Section 7, which here takes the form

$$\omega_1 \cdot \log \frac{\omega_1 + \omega_2 \cdot \exp(-D(q_1\|q_2))}{\exp(-D(q_1\|g))} + \omega_2 \cdot \log \frac{\omega_1 \cdot \exp(-D(q_2\|q_1)) + \omega_2}{\exp(-D(q_2\|g))},$$

where $D(q_1\|q_2)$, $D(q_1\|g)$, $D(q_2\|g)$ are the Kullback–Leibler divergences between q_1 and q_2, q_1 and g and q_2 and g respectively where g is the prior. There are closed form expressions for these Kullback–Leibler divergences since they are between pairs of multivariate Gaussian densities. After application of the Hershey–Olsen bound, we have an approximating statistic $\mathrm{KL}^*(y)$ to $\mathrm{KL}(y)$. Then we can approximate p_{KL} by simulating datasets $y^{(i)}$, $i = 1, \ldots, M$ under the prior predictive, computing $\mathrm{KL}^*(y^{(i)})$ and $\mathrm{KL}^*(y_{\mathrm{obs}})$ and then

$$p_{\mathrm{KL}} \approx \frac{1}{M} \sum_{i=1}^{M} I\big(\mathrm{KL}^*\big(y^{(i)}\big) \geq \mathrm{KL}^*(y_{\mathrm{obs}})\big).$$

For illustration, consider three different normal priors, all with prior covariance matrix Σ_0 diagonal with diagonal entries 0.25, but with prior means representing a lack of conflict, moderate conflict and a clear conflict ($\mu_0 = (-7.1, 7.9)$, $\mu_0 = (-7.4, 7.9)$ and $\mu_0 = (-7.7, 7.9)$ respectively). Figure 3 shows for the three cases contour plots of the prior and likelihood (left column) and the true posterior together with its two component variational posterior approximation computed using the algorithm of Salimans and Knowles (2013). The three rows from top to bottom show the cases of lack of conflict, moderate conflict and a clear conflict. The p-values approximated by the variational method and Hershey–Olsen bound with $M = 1000$ are 0.58, 0.25 and 0.03 for the three cases. We can see that the variational posterior approximation is excellent even with just two mixture components and the p-values behave as we would expect.

EXAMPLE 6. *Bristol Royal Infirmary Inquiry data.* We illustrate the computation of our conflict checks in a hierarchical setting using a logistic random effects model. Here the data are part of that presented to a public enquiry into excess mortality at the Bristol Royal Infirmary in complex paediatric surgeries prior to 1995. The data are given in Marshall and Spiegelhalter (2007), Table 1, and a comprehensive discussion is given in Spiegelhalter et al. (2002). The data consists of pairs (y_i, n_i), $i = 1, \ldots, 12$ where i indexes different hospitals, y_i is the number of deaths in hospital i and n_i is the number of operations. The first hospital ($i = 1$) is the Bristol Royal Infirmary. Marshall and Spiegelhalter (2007) consider a random effects model of the form $y_i \sim \mathrm{Binomial}(n_i, p_i)$ where $\log(p_i/(1 - p_i)) = \beta + u_i$ and $u_i \sim N(0, D)$ so that u_i are hospital specific random effects, and they consider formal measures of conflict involving the prior for u_i given D. Particular interest is in whether there is a prior data conflict for $i = 1$ (Bristol) which would indicate that this hospital is unusual compared to the others. In our analysis here, we consider priors on β and D where $\beta \sim N(0, 1000)$ and $\log D \sim N(-3.5, 1)$ which were chosen to be roughly similar to priors chosen in Tan and Nott (2014) for this example. So we have a hierarchical prior, $g(\theta) = g(u, \beta, D) = g(u|D)g(\beta, D)$ and we can use our methods for checking hierarchical priors to check for conflict involving each of the u_i.

We will use a multivariate normal variational approximation to $g(\theta|y)$ (but with D transformed by taking logs) and computed using the method described in Kucukelbir et al. (2017), which is implemented in the software package Stan (Carpenter et al., 2017). The conditional prior $g(u|D)$ is normal, and in the variational posterior the conditional for u given β, D is also normal, so that conditional prior to (variational) posterior divergences can be computed in closed form. For checking for conflict for the u_i's, we will use the statistic $\mathrm{KL}_1(y) = \lim_{\alpha \to 1} R_{\alpha 1}(y)$, except that we replace the conditional posterior and prior for u given β, D in the definition (8) with that of u_i given β, D when checking u_i. This is because we are interested in checking for conflicts for individual hospital specific effects. We will approximate $\mathrm{KL}_1(y)$ by $\mathrm{KL}_1^*(y)$ obtained by replacing all computations involving the true posterior with the equivalent calculations for the variational Gaussian posterior.

Figure 4 shows for the observed data the variational posterior distribution, together with the true posterior approximated by MCMC. Table 1 also shows our conflict p-values for the different hospitals. Also listed are cross-validated mixed predictive p-values obtained by the method of Marshall and Spiegelhalter (2007) by MCMC and given in Tan and Nott (2014), Table 1, as well as a cross-validated version of our divergence based p-values. The cross-validated divergence based p-values use the posterior distribution for (β, D) obtained when leaving out the ith observation, $g(\theta_2|y_{\mathrm{obs},-i})$, instead of $g(\theta_2|y_{\mathrm{obs}})$ in the definition of the reference distribution (11) and in taking the expectation in (9). We can see that the p-values are similar although the priors on the parameters (β, D) were not exactly the same in Tan and Nott's analysis. For comparison with previous analyses of the data, we have computed a one-sided version of our conflict p-value here, which makes

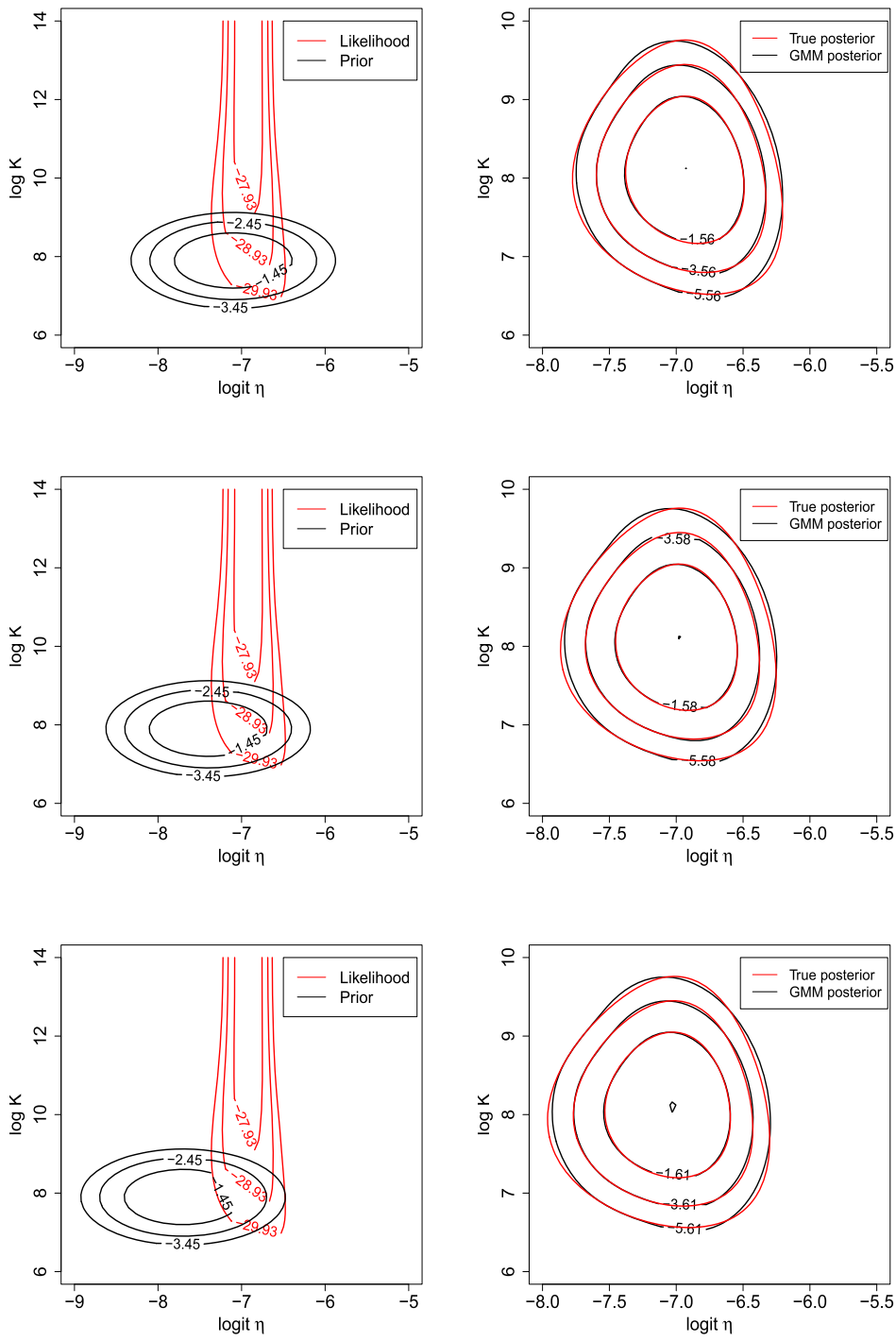

FIG. 3. *Contour plots of log-likelihood and prior* (*left*) *and true posterior together with Gaussian mixture approximation* (*right*) *for priors centered at* $(-7.1, 7.9)$, $(-7.4, 7.9)$ *and* $(-7.7, 7.9)$ (*from top to bottom*).

sense because excess mortality is of interest. We have modified our p-value measuring surprise to $p_{\mathrm{KL1}} = P(\mathrm{KL}_1(Y) \geq \mathrm{KL}_1(y_{\mathrm{obs}})$ and $E_q(u_i|Y) > 0)$ for clusters i with $E_q(u_i|y_{\mathrm{obs}}) > 0$, and to $p_{\mathrm{KL1}} = P(\mathrm{KL}_1(Y) \leq \mathrm{KL}_1(y_{\mathrm{obs}})) + P(\mathrm{KL}_1(Y) \geq \mathrm{KL}_1(y_{\mathrm{obs}})$ and $E_q(u_i|Y) > 0)$ for clusters i with $E(u_i|y_{\mathrm{obs}}) < 0$, where in these expressions $E_q(\cdot)$ denotes expectation with respect to the appropriate variational posterior distribution. Although it is not expected that these conflict p-values should be exactly the same, it is seen that they give a similar picture about the degree of consistency of the data for each hospital with the hierarchical prior.

6. DISCUSSION

We have proposed a new approach for prior-data conflict assessment based on comparing the prior to posterior Rényi divergence to its distribution under the prior pre-

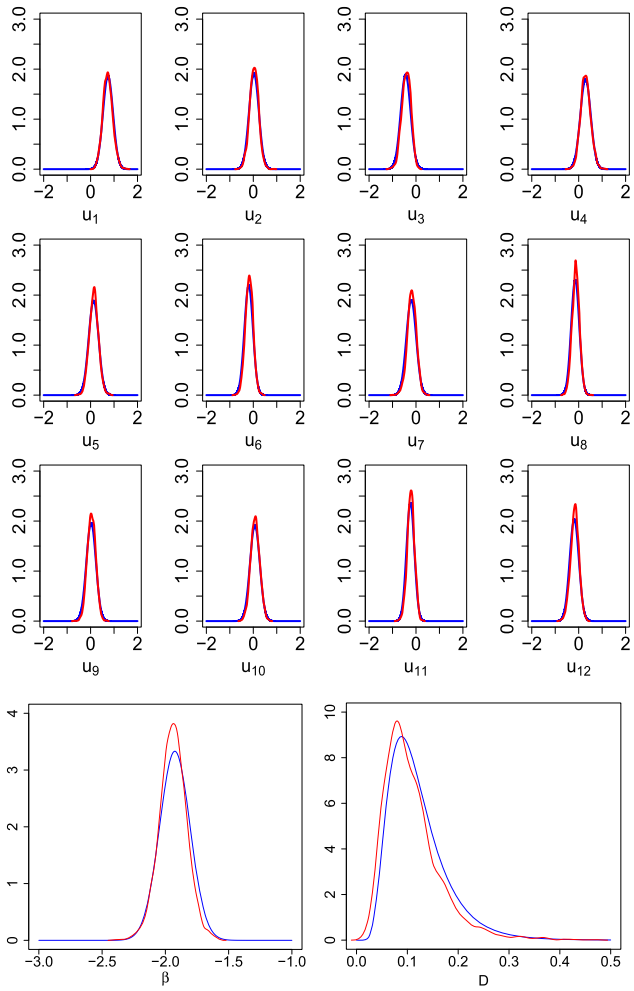

FIG. 4. *Marginal posterior distributions computed by MCMC (red) and Gaussian variational posteriors (blue) for u (top) and (β, D) (bottom).*

dictive for the data. The method can be extended to hierarchical settings where it is desired to check different components of a prior distribution, and has some interesting connections with the methodology of Evans and Moshonov (2006) and with Jeffreys' and reference prior distributions. Similar to Evans and Moshonov (2006), the discrepancy will be a function of any minimal sufficient statistic, but the new approach achieves invariance to the choice of that statistic which is not unique. The method works well in the examples we have examined, and we have suggested the use of variational approximations for making the methodology implementable in complex settings.

There are a number of ways that this work could be further developed. One line of future development concerns the computational approximations developed in Section 5, which can no doubt be improved. On the more statistical side, Evans and Jang (2011b) define a notion of weak informativity of a prior with respect to a given base prior, inspired by ideas of Gelman (2006), and their particular formulation of this concept makes use of the notion of prior-data conflict checks. It will be interesting to examine how the prior-data conflict checks we have developed here perform in relation to this application.

TABLE 1

Cross-validatory conflict p-values using the method of Marshall and Spiegelhalter ($p_{MS,CV}$), KL divergence conflict p-values (p_{KL}), and cross-validated KL divergence p-values ($p_{KL,CV}$) for hospital specific random effects

Hospital	$p_{MS,CV}$	p_{KL}	$p_{KL,CV}$
Bristol	0.001	0.010	0.002
Leicester	0.436	0.527	0.516
Leeds	0.935	0.912	0.947
Oxford	0.125	0.173	0.123
Guys	0.298	0.398	0.383
Liverpool	0.720	0.690	0.745
Southampton	0.737	0.680	0.715
Great Ormond St	0.661	0.595	0.628
Newcastle	0.440	0.455	0.430
Harefield	0.380	0.474	0.452
Birmingham	0.763	0.761	0.787
Brompton	0.721	0.591	0.631

APPENDIX

Details of Example 2

Consider the check of Evans and Moshonov (2006). y is minimal sufficient and the prior predictive for y is beta-binomial,

$$p(y) = \binom{n}{y} \frac{B(a+y, b+n-y)}{B(a,b)}, \quad y = 0, \ldots, n.$$

Hence a suitable discrepancy for the check of Evans and Moshonov (2006), which we denote by $\mathrm{RB}(y)$, is

$$\begin{aligned} \mathrm{RB}(y) &= \log p(y) \\ &= \log \binom{n}{y} + \log \frac{B(a+y, b+n-y)}{B(a,b)} \\ &\doteq \log \Gamma(a+y) + \log \Gamma(b+n-y) \\ &\quad - \log \Gamma(y+1) - \log \Gamma(n-y+1). \end{aligned} \tag{18}$$

The check of Evans and Moshonov (2006) and the divergence based check are not equivalent in this example. However, they can be related to each other when y and $n-y$ are both large.

The form of the check with the Rényi divergence is

$$\begin{aligned} R_\alpha(y) &= \log \frac{B(a,b)}{B(a+y, b+n-y)} \\ &\quad + \frac{1}{\alpha - 1} \log \frac{B(a+\alpha y, b+\alpha(n-y))}{B(a+y, b+n-y)} \\ &= T_1 + T_2 \end{aligned} \tag{19}$$

where $B(\cdot,\cdot)$ denotes the beta function. Using Stirling's approximation for the beta function

$$B(x,z) \approx \sqrt{2\pi}\frac{x^{x-\frac{1}{2}}z^{z-\frac{1}{2}}}{(x+z)^{x+z-\frac{1}{2}}},$$

for x and z large, we obtain

$$\begin{aligned} T_1 \doteq{}& \log B(a,b) - (a+b+n)\hat{\theta}_n \log\hat{\theta}_n \\ &+ \frac{1}{2}\log\hat{\theta}_n \\ &- (a+b+n)(1-\hat{\theta}_n)\log(1-\hat{\theta}_n) \\ &+ \frac{1}{2}\log(1-\hat{\theta}_n) + O\left(\frac{1}{n}\right), \end{aligned} \tag{20}$$

where some constants not depending on y have been ignored on the right hand side and $\hat{\theta}_n = (a+y)/(a+b+n)$ is the posterior mean of θ. Another application of Stirling's approximation to T_2 in (19) gives

$$\begin{aligned} T_2 &= \frac{1}{\alpha-1}\log\frac{B(a+\alpha y, b+\alpha(n-y))}{B(a+y,b+n-y)} \\ &= \frac{1}{\alpha-1}\{(a+b+\alpha n)\tilde{\theta}_n\log\tilde{\theta}_n \\ &\quad + (a+b+\alpha n)(1-\tilde{\theta}_n)\log(1-\tilde{\theta}_n) \\ &\quad - (a+b+n)\hat{\theta}_n\log\hat{\theta}_n \\ &\quad - (a+b+n)(1-\hat{\theta}_n)\log(1-\hat{\theta}_n)\} \\ &\quad + O\left(\frac{1}{n}\right), \end{aligned}$$

where $\tilde{\theta}_n = (a+\alpha y)/(a+b+\alpha n)$. Making the Taylor series approximations

$$\begin{aligned} \tilde{\theta}_n\log\tilde{\theta}_n &= \hat{\theta}_n\log\hat{\theta}_n \\ &\quad + (\tilde{\theta}_n-\hat{\theta}_n)(1+\log\hat{\theta}_n) \\ &\quad + O\left(\frac{1}{n^2}\right), \\ (1-\tilde{\theta}_n)\log(1-\tilde{\theta}_n) &= (1-\hat{\theta}_n)\log(1-\hat{\theta}_n) \\ &\quad - (\tilde{\theta}_n-\hat{\theta}_n)(1+\log(1-\hat{\theta}_n)) \\ &\quad + O\left(\frac{1}{n^2}\right) \end{aligned}$$

and also observing that $n(\tilde{\theta}_n-\hat{\theta}_n) = \frac{\alpha-1}{\alpha}\{(a+b)\hat{\theta}_n - a\} + O(\frac{1}{n})$ gives

$$\begin{aligned} T_2 = {}& n\hat{\theta}_n\log\hat{\theta}_n \\ &+ n(1-\hat{\theta}_n)\log(1-\hat{\theta}_n) \\ &+ \big((a+b)\hat{\theta}_n - a\big)\log\hat{\theta}_n \\ &+ \big((a+b)\hat{\theta}_n - b\big)\log(1-\hat{\theta}_n) \\ &+ O\left(\frac{1}{n}\right). \end{aligned} \tag{21}$$

Combining (20) and (21) gives the expression (13).

Turning now to the check of Evans and Moshonov (2006), and writing $\psi(\cdot)$ for the digamma function, appropriate Taylor expansions in (18) gives

$$\begin{aligned} &\log\Gamma(a+y) \\ &\quad = \log\Gamma(y+1) + (a-1)\psi(a+y) \\ &\quad = \log\Gamma(y+1) + (a-1)\log(a+y) \\ &\qquad + O(1/n), \\ &\log\Gamma(b+n-y) \\ &\quad = \log\Gamma(n-y+1) \\ &\qquad + (b-1)\psi(b+n-y) \\ &\quad = \log\Gamma(n-y+1) \\ &\qquad + (b-1)\log(b+n-y) + O(1/n) \end{aligned}$$

which gives

$$\begin{aligned} \log p(y) \doteq{}& \log\Gamma(y+1) \\ &+ (a-1)\log(a+y) \\ &+ \log\Gamma(n-y+1) \\ &+ (b-1)\log(b+n-y) \\ &- \log\Gamma(y+1) \\ &- \log\Gamma(n-y+1) + O(1/n) \\ \doteq{}& (a-1)\log(a+y) \\ &+ (b-1)\log(b+n-y) + O(1/n) \\ \doteq{}& \log g(\hat{\theta}_n) + O(1/n), \end{aligned}$$

where as before $\hat{\theta}_n$ is the posterior mean for θ.

Details of Example 3

The check described in Evans and Moshonov (2006), page 910, compares $s^2/(b/a)$ to an $F_{n-1,2a}$ density. Plugging in $s^2/(b/a)$ to the expression for the log of the F density, we have the statistic

$$\mathrm{RB}(y) \doteq \frac{n-3}{2}\log\frac{s^2}{b/a}$$

$$-\frac{n+2a-1}{2}\log\left(1+\frac{n-1}{2a}\frac{s^2}{b/a}\right),$$

and then using the approximation $\log(1+x)\approx\log x+1/x$ for large x gives approximately

$$\begin{aligned}\mathrm{RB}(y)&\doteq\frac{n-3}{2}\log\frac{s^2}{b/a}\\&\quad-\frac{n+2a-1}{2}\log\left(\frac{s^2}{b/a}\right)\\&\quad-\frac{n+2a-1}{2}\frac{2a}{n-1}\frac{b/a}{s^2}+O\left(\frac{1}{n}\right)\\&\doteq-\frac{a-1}{2}\log\frac{s^2}{b/a}-\frac{n+2a-1}{n-1}\frac{b}{s^2}\\&\quad+O\left(\frac{1}{n}\right).\end{aligned}$$

So for large n, we have approximately

$$\mathrm{RB}(y)\doteq\frac{a-1}{2a}\log\frac{s^2}{b/a}+\frac{b/a}{s^2},$$

which, comparing with (14), clarifies the relationship to the divergence based check.

Details of Example 4

The likelihood function is

$$p(y|\theta)=c(y)\exp\bigl(-nr(y_{\min}-\theta)\bigr)I(0<\theta<y_{\min}),$$

where $y_{\min}$ denotes the minimum of $y_1,\ldots,y_n$ and $c(y)=r^n\exp(-nr(\bar{y}-y_{\min}))$ where $\bar{y}$ denotes the sample mean. A sufficient statistic is $y_{\min}$, and its sampling distribution has density

$$\begin{aligned}&p(y_{\min}|\theta)\\&\quad=nr\exp\bigl(-nr(y_{\min}-\theta)\bigr)I(0<\theta<y_{\min}).\end{aligned}$$

The prior predictive of $y_{\min}$ is

$$\begin{aligned}p(y_{\min})&=nr\kappa\exp(-nry_{\min})\\&\quad\times\int_0^{y_{\min}}\exp\bigl((nr-\kappa)\theta\bigr)\,d\theta\\&=\frac{nr\kappa}{nr-\kappa}\bigl(\exp(-\kappa y_{\min})\\&\quad-\exp(-nry_{\min})\bigr),\end{aligned}\tag{22}$$

and this is the discrepancy for the test of Evans and Moshonov (2006). Consider now the statistic $R_\alpha(y)$. We have $g(\theta|y)\propto\exp((nr-\kappa)\theta)I(0<\theta<y_{\min})$ so that

$$\begin{aligned}g(\theta|y)&=\frac{(nr-\kappa)}{\exp(t)-1}\\&\quad\times\exp\bigl((nr-\kappa)\theta\bigr)I(0<\theta<y_{\min}),\end{aligned}$$

where $t=(nr-\kappa)y_{\min}$. Then

$$\begin{aligned}&\int_0^{y_{\min}}\left(\frac{g(\theta|y)}{g(\theta)}\right)^{\alpha-1}g(\theta|y)\,d\theta\\&\quad=\frac{\kappa}{\alpha nr-\kappa}\left(\frac{(nr-\kappa)}{\kappa(\exp(t)-1)}\right)^{\alpha}\\&\qquad\times\bigl[\exp\bigl((\alpha nr-\kappa)y_{\min}\bigr)-1\bigr],\end{aligned}$$

and so

$$\begin{aligned}R_\alpha(y)&=\frac{1}{\alpha-1}\log\frac{\kappa}{\alpha nr-\kappa}\\&\quad+\frac{\alpha}{\alpha-1}\log\left(\frac{(nr-\kappa)}{\kappa(\exp(t)-1)}\right)\\&\quad+\frac{1}{\alpha-1}\log\bigl(\exp\bigl((\alpha nr-\kappa)y_{\min}\bigr)-1\bigr).\end{aligned}$$

ACKNOWLEDGEMENTS

David Nott was supported by a Singapore Ministry of Education Academic Research Fund Tier 2 grant (R-155-000-143-112). Berthold-Georg Englert's work is funded by the Singapore Ministry of Education (partly through the Academic Research Fund Tier 3 MOE2012-T3-1-009) and the National Research Foundation of Singapore. Michael Evans' work was supported by a Natural Sciences and Engineering Research Council of Canada Grant Number 10671. We thank the Editor, Associate Editor and referees for their help in improving the paper.

REFERENCES

AL LABADI, L. and EVANS, M. (2017). Optimal robustness results for relative belief inferences and the relationship to prior-data conflict. *Bayesian Anal.* **12** 705–728. MR3655873 https://doi.org/10.1214/16-BA1024

ALBERT, J. (2009). *Bayesian Computation with R*, 2nd ed. *Use R!* Springer, Dordrecht. MR2839312 https://doi.org/10.1007/978-0-387-92298-0

ANDRADE, J. A. A. and O'HAGAN, A. (2006). Bayesian robustness modeling using regularly varying distributions. *Bayesian Anal.* **1** 169–188. MR2227369 https://doi.org/10.1214/06-BA106

ATTIAS, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the* 15*th Conference on Uncertainty in Artificial Intelligence* (K. Laskey and H. Prade, eds.) 21–30. Morgan Kaufmann, San Francisco, CA.

BASKURT, Z. and EVANS, M. (2013). Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Anal.* **8** 569–590. MR3102226 https://doi.org/10.1214/13-BA824

BAYARRI, M. J. and BERGER, J. O. (2000). p values for composite null models. *J. Amer. Statist. Assoc.* **95** 1127–1142, 1157–1170. MR1804239 https://doi.org/10.2307/2669749

BAYARRI, M. J. and CASTELLANOS, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statist. Sci.* **22** 322–343. MR2416808 https://doi.org/10.1214/07-STS235

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37** 905–938. MR2502655 https://doi.org/10.1214/07-AOS587

BOUSQUET, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis. *J. Appl. Stat.* **35** 1011–1029. MR2522125 https://doi.org/10.1080/02664760802192981

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745 https://doi.org/10.2307/2982063

CAROTA, C., PARMIGIANI, G. and POLSON, N. G. (1996). Diagnostic measures for model criticism. *J. Amer. Statist. Assoc.* **91** 753–762. MR1395742 https://doi.org/10.2307/2291670

CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. CRC Press, London. MR0370837

DAHL, F. A., GÅSEMYR, J. and NATVIG, B. (2007). A robust conflict measure of inconsistencies in Bayesian hierarchical models. *Scand. J. Stat.* **34** 816–828. MR2396940 https://doi.org/10.1111/j.1467-9469.2007.00560.x

DEY, D. K., GELFAND, A. E., SWARTZ, T. B. and VLACHOS, P. K. (1998). A simulation-intensive approach for checking hierarchical models. *TEST* **7** 325–346.

EVANS, M. (2015). *Measuring Statistical Evidence Using Relative Belief. Monographs on Statistics and Applied Probability* **144**. CRC Press, Boca Raton, FL. MR3616661

EVANS, M. and JANG, G. H. (2010). Invariant P-values for model checking. *Ann. Statist.* **38** 512–525. MR2589329 https://doi.org/10.1214/09-AOS727

EVANS, M. and JANG, G. H. (2011a). A limit result for the prior predictive applied to checking for prior-data conflict. *Statist. Probab. Lett.* **81** 1034–1038. MR2803740 https://doi.org/10.1016/j.spl.2011.02.025

EVANS, M. and JANG, G. H. (2011b). Weak informativity and the information in one prior relative to another. *Statist. Sci.* **26** 423–439. MR2917964 https://doi.org/10.1214/11-STS357

EVANS, M. and MOSHONOV, H. (2006). Checking for prior-data conflict. *Bayesian Anal.* **1** 893–914. MR2282210 https://doi.org/10.1016/j.spl.2011.02.025

GÅSEMYR, J. and NATVIG, B. (2009). Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models. *Scand. J. Stat.* **36** 822–838. MR2573310 https://doi.org/10.1111/j.1467-9469.2009.00659.x

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A. and SHALIZI, C. R. (2013). Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* **66** 8–38. MR3044854 https://doi.org/10.1111/j.2044-8317.2011.02037.x

GERSHMAN, S., HOFFMAN, M. D. and BLEI, D. M. (2012). Nonparametric variational inference. In *Proceedings of the* 29*th International Conference on Machine Learning*, *ICML* 2012.

GHOSH, M. (2011). Objective priors: An introduction for frequentists. *Statist. Sci.* **26** 187–202. MR2858380 https://doi.org/10.1214/10-STS338

GHOSH, J. K., DELAMPADY, M. and SAMANTA, T. (2006). *An Introduction to Bayesian Analysis*: *Theory and Methods*. *Springer Texts in Statistics*. Springer, New York. MR2247439

GIL, M., ALAJAJI, F. and LINDER, T. (2013). Rényi divergence measures for commonly used univariate continuous distributions. *Inform. Sci.* **249** 124–131. MR3105467 https://doi.org/10.1016/j.ins.2013.06.018

GUTTMAN, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. Roy. Statist. Soc. Ser. B* **29** 83–100. MR0216699

HERSHEY, J. R. and OLSEN, P. A. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. In 2007 *IEEE International Conference on Acoustics*, *Speech and Signal Processing—ICASSP* '07 **4** IV-317–IV-320.

HJORT, N. L., DAHL, F. A. and STEINBAKK, G. H. (2006). Post-processing posterior predictive p-values. *J. Amer. Statist. Assoc.* **101** 1157–1174. MR2324154 https://doi.org/10.1198/016214505000001393

JAYNES, E. T. (1976). Confidence intervals vs. Bayesian intervals (1976). In *Foundations of Probability Theory*, *Statistical Inference*, *and Statistical Theories of Science*, *Vol. II* 175–257. Reidel, Dordrecht.

JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Stat.* **41** 851–864. MR0263198 https://doi.org/10.1214/aoms/1177696963

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KUCUKELBIR, A., TRAN, D., RANGANATH, R., GELMAN, A. and BLEI, D. M. (2017). Automatic differentiation variational inference. *J. Mach. Learn. Res.* **18** 14. MR3634881

LI, X., SHANG, J., NG, H. K. and ENGLERT, B.-G. (2016). Optimal error intervals for properties of the quantum state. *Phys. Rev. A* **94** 062112.

LIESE, F. and VAJDA, I. (1987). *Convex Statistical Distances*. *Teubner-Texte zur Mathematik* [*Teubner Texts in Mathematics*] **95**. BSB B. G. Teubner Verlagsgesellschaft, Leipzig. With German, French and Russian summaries. MR0926905

MARSHALL, E. C. and SPIEGELHALTER, D. J. (2007). Identifying outliers in Bayesian hierarchical models: A simulation-based approach. *Bayesian Anal.* **2** 409–444. MR2312289 https://doi.org/10.1214/07-BA218

O'HAGAN, A. (2003). HSSS model criticism. In *Highly Structured Stochastic Systems*. *Oxford Statist. Sci. Ser.* **27** 423–453. Oxford Univ. Press, Oxford. MR2082418

PRESANIS, A. M., OHLSSEN, D., SPIEGELHALTER, D. J. and DE ANGELIS, D. (2013). Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statist. Sci.* **28** 376–397. MR3135538 https://doi.org/10.1214/13-STS426

REIMHERR, M., MENG, X.-L. and NICOLAE, D. L. (2014). Being an informed Bayesian: Assessing prior informativeness and prior likelihood conflict. Available at arXiv:1406.5958.

RÉNYI, A. (1961). On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.*, *Vol. I* 547–561. Univ. California Press, Berkeley, CA. MR0132570

ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of p values in composite null models. *J. Amer. Statist. Assoc.* **95** 1143–1167, 1171–1172. MR1804240 https://doi.org/10.2307/2669750

ROHDE, D. and WAND, M. P. (2016). Semiparametric mean field variational Bayes: General principles and numerical issues. *J. Mach. Learn. Res.* **17** 172. MR3567440

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681 https://doi.org/10.1214/aos/1176346785

SALIMANS, T. and KNOWLES, D. A. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Anal.* **8** 837–881. MR3150471 https://doi.org/10.1214/13-BA858

SCHEEL, I., GREEN, P. J. and ROUGIER, J. C. (2011). A graphical diagnostic for identifying influential model choices in Bayesian

hierarchical models. *Scand. J. Stat.* **38** 529–550. MR2833845 https://doi.org/10.1111/j.1467-9469.2010.00717.x

SPIEGELHALTER, D. J., AYLIN, P., BEST, N. G., EVANS, S. J. W. and MURRAY, G. D. (2002). Commissioned analysis of surgical performance using routine data: Lessons from the Bristol inquiry. *J. Roy. Statist. Soc. Ser. A* **165** 191–231. MR1869173 https://doi.org/10.1111/1467-985X.02021

TAN, L. S. L. and NOTT, D. J. (2014). A stochastic variational framework for fitting and diagnosing generalized linear mixed models. *Bayesian Anal.* **9** 963–1004. MR3293964 https://doi.org/10.1214/14-BA885

WINN, J. and BISHOP, C. M. (2005). Variational message passing. *J. Mach. Learn. Res.* **6** 661–694. MR2249835