WILEY

Author Query Form

Journal: RSSB

Article: 12414

Dear Author,

During the copyediting of your manuscript, the following queries arose.

Please refer to the query reference callout numbers in the page proofs and respond.

Please remember illegible or unclear comments and corrections may delay publication.

Many thanks for your assistance.

AUTHOR: Please note that missing content in references have been updated where we have been able to match the missing elements without ambiguity against a standard citation database, to meet the reference style requirements of the journal. It is your responsibility to check and ensure that all listed references are complete and accurate.

Query reference	Query	Remarks
1	**AUTHOR: Please supply date of submission and acceptance.**	
2	**AUTHOR: Please confirm that given names (blue) and surnames/family names (vermilion) have been identified correctly.**	
3	**AUTHOR: Table 1 was not cited in the text. An attempt has been made to insert the table into a relevant point in the text - please check that this is OK. If not, please provide clear guidance on where it should be cited in the text.**	
4	**AUTHOR: Brown, Cai, DasGupta (2001) not cited. Please cite reference in text or delete from the list.**	
5	**AUTHOR: Lee (1990) not cited. Please cite reference in text or delete from the list.**	
6	**AUTHOR: Please provide the volume number and page range for reference Li and Meng (2021).**	

DOI: 10.1111/rssb.12414

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

Prior sample size extensions for assessing prior impact and prior-likelihood discordance

Matthew Reimherr[1] | Xiao-Li Meng[2] | Dan Nicolae[3]

[1]The Pennsylvania State University, University Park, PA, USA

[2]Harvard University, Cambridge, MA, USA

[3]The University of Chicago, Chicago, IL, USA

Correspondence
Matthew Reimherr, Department of Statistics, The Pennsylvania State University, 411 Thomas Building, University Park, PA 16802, USA.
Email: mreimherr@psu.edu

Abstract

This paper outlines a framework for quantifying the prior's contribution to posterior inference in the presence of prior-likelihood discordance, a broader concept than the usual notion of prior-likelihood conflict. We achieve this dual purpose by extending the classic notion of *prior sample size*, M, in three directions: (I) estimating M beyond conjugate families; (II) formulating M as a relative notion that is as a function of the likelihood sample size k, $M(k)$, which also leads naturally to a graphical diagnosis; and (III) permitting negative M, as a measure of prior-likelihood conflict that is harmful discordance. Our asymptotic regime permits the prior sample size to grow with the likelihood data size, hence making asymptotic arguments meaningful for investigating the impact of the prior relative to that of likelihood. It leads to a simple asymptotic formula for quantifying the impact of a proper prior that only involves computing a centrality and a spread measure of the prior and the posterior. We use simulated and real data to illustrate the potential of the proposed framework, including quantifying how weak is a 'weakly informative' prior adopted in a study of lupus nephritis. Whereas we take a pragmatic perspective in assessing the impact of a prior on a given inference problem under a specific evaluative metric, we also touch upon conceptual and theoretical issues such as using improper priors and permitting priors with asymptotically non-vanishing influence.

KEYWORDS

conjugate prior, prior-likelihood conflict, weakly informative prior

Journal Name RSSB | Manuscript No. 12414 | WILEY | No. of pages: 25 | Dispatch: 12-2-2021 | CE: Wiley | PE: Bhuvaneswari R.

1 | MOTIVATION AND ILLUSTRATION

The difficulty in choosing priors and fully understanding their impact on statistical analyses has been a primary concern of Bayesian methods since their inception. The common approach to alleviate such concerns is to conduct a sensitivity analysis, investigating how the results are affected by perturbations of the prior. However, such an approach does not typically reveal how a chosen prior has actually contributed to the analysis in comparison to the information from the data, as captured by the posited likelihood. Although many have asked questions along the line 'How much of your conclusion is actually due to your prior assumptions?', to the best of our knowledge, there are no well-recognized approaches to quantitatively address such legitimate inquires.

As the adoption of Bayesian tools continues to grow, we need quantitative assessments on the impact of priors, permitting at least a check on their impact compared to the likelihood. This is desirable, both scientifically and statistically. A posterior inference with 45% prior information contribution may affect our decisions rather differently, at least psychologically, from one with only 5% prior contribution. However, quantifying the impact of a prior has been a very challenging task, partially explaining the lack of routinely adopted methods. A key difficulty is that the information from the prior may be in conflict with that from the likelihood to a point that it can actually 'subtract' rather than 'add' to an analysis. Recently, Efron (2015) explored frequentist properties of Bayesian estimates, illustrating that, in many ways, researchers are often still unable to understand/quantify the impact of their priors on their inference. Our paper aims to make a substantive contribution in this direction, though we by no means declare that we have found *the* solution as this is an area that requires much more research.

1.1 | Building upon the classic notion of prior sample size

This paper presents a strategy for simultaneously assessing the degree of conflict between the prior and likelihood, and quantifying the information contribution of the prior to the posterior. We accomplish this by extending the easily interpretable information metric *prior sample size* (PSS), a common notion in the literature of conjugate priors (e.g. Diaconis et al., 1979; Gutiérrez-Peña et al., 1997; Meng & Zaslavsky, 2002). As is well known, a conjugate prior can be equated to the posterior from a prior study with, say, M i.i.d. (hypothetical) observations and a baseline 'non-informative' prior. This equivalence provides a concrete practical guideline. If our likelihood is based on N i.i.d. observations from the same conjugate family, then we can consider that the conjugate prior has contributed $M/(N+M)\times 100\%$ of the information to our posterior inference.

Given its practical appeal, multiple efforts have been made to extend the concept of PSS beyond conjugate families. The approach by Clarke (1996) is particularly significant, taking advantage of reference priors (Bernardo, 1979), which are equivalent to Jeffreys priors in univariate cases (George & McCulloch, 1993). Specifically, by minimizing their relative entropy (i.e. Kullback–Leibler divergence), Clarke's approach matches a target prior—typically considered to be informative—with the posterior based on a likelihood from a given family and a given reference prior. The resulting likelihood function is then interpreted as representing implicit data information in the target prior, relative to the reference prior. The PSS of the target prior is then approximated by this likelihood data size, though Clarke (1996) also identifies the actual data values used by the likelihood (typically not unique). Subsequently, Clarke and Yuan (2006) developed closed-form expressions for the prior sample size, and Lin et al. (2007) examined how to quantify the information content with non i.i.d. data. Ginebra (2007) discussed, more generally, how to quantify the information content in an experiment, and Berger et al. (2014) investigated how to quantify effective sample sizes of various linear models.

Recently Wiesenfarth and Calderazzo (2019) explored the usage of historical data for quantifying PSS in the context of clinical trials.

In addition, Morita et al. (2008) took a similar approach but with a different baseline and divergence. Their baseline prior was constructed by keeping the prior mean and correlations (for multivariate cases) the same as the target prior, but with the prior variance greatly inflated to render ϵ-amount of information. Their divergence is based on the trace of the (expected) curvature of the log density, which they reported was best after extensive trial and error. This is intuitive as the curvature of the log density is to the prior variance what Fisher information is to the posterior variance. This pairing between the divergence and the baseline prior was also emphasized by Clarke (1996), because the reference prior is the minimizer of an expected K–L divergence, and hence his pairing also guarantees his PSS to be non-negative when it exists. However, permitting a negative PSS turns out to be the key to resolve the thorny issue of prior-likelihood conflict when measuring the impact of prior information, as we will discuss in Section 2. The methods of Morita et al. (2008) were further illustrated on biomedical applications in Morita et al. (2010), and extended to hierarchical models in Morita et al. (2012).

A commonality between the settings in Clarke (1996) and Morita et al. (2008) (and their subsequent extensions) is that both treat the likelihood model as a device for measuring the information in the target prior, with the hypothetical observations optimized over or averaged over, respectively. Hence these are *pre-data* measures, most useful for design purposes and theoretical investigations. In this paper, however, we address a harder and more common *post-data* question: how many observations are required, approximately, to match the prior's contribution—in terms of some statistical efficiency—to a posterior inference based on a particular *likelihood function* from a set of observed data? That is, we intend our PSS to be *inherently data dependent*.

In theory, we all hope that, at the very least, our prior does no harm. However, in practice, typically there are some degrees of prior-likelihood conflict. This can lead to, for example, a 95% posterior interval that is wider than an analogous 95% confidence interval, or which has deficient coverage. This can occur regardless of the correctness of the prior or likelihood, because a particular data set from a known model can still exhibit 'tail' behaviour. Any measure of the prior contribution to a particular posterior inference must then allow for the possibility of negative prior contribution. But how should we formulate negative 'sample size'? We report a practical way to circumvent this problem by matching two posteriors corresponding to two priors (target and baseline), instead of the prior-posterior matching as in Clarke (1996) or in Morita et al. (2008).

The more fundamental question is the meaning of measuring the statistical efficiency in a particular study. All statistical inference paradigms require a specification of *reference replications* (see, e.g. Liu & Meng, 2016), because otherwise there is no variation, and hence no information, to speak of. The reference replications in Clarke (1996) and Morita et al. (2008) are pre-data hypothetical observations as respectively specified, designed as frequentist measures, averaging over many hypothetic data sets that we will never observe. But if we insist on using the Bayesian replication, that is, all data sets that are exactly the same as the observed ones, then our entire information measure will be driven by the prior, the only source of variation. To avoid either extreme (see Liu & Meng, 2016, for reasons for this avoidance), we adopt a compromise: we measure information/efficiency with respect to all data sets that are *exchangeable* with the observed data, that is, they are not identical to the observed data but they share the same generating model with the same parameter value for the latter. We do not know this parameter value, but when we have adequate internal replications (e.g. with i.i.d. data), we will be able to estimate our measure via common methods such as bootstrap.

Before we proceed to illustrate the key ingredients of our proposal, we emphasize that the need for choosing a baseline prior, as above and in other similar works such as Evans and Jang (2011), is

unavoidable because it is mathematically impossible to represent ignorance via a probability distribution (see Martin & Liu, 2016, Proposition 2.1). In responding to an insightful question raised by a reviewer, we will also report a 'prior size paradox' caused by this impossibility (see Section 5.2). Our preference therefore is to choose a prior that represents what a practitioner would adopt without real prior information, such as those documented in Kass and Wasserman (1996). We also note that various developments on deviance information criteria such as Spiegelhalter et al. (2002), Watanabe (2010) and Watanabe (2013), and others as reviewed in Gelman et al. (2014), are similar in spirit to our goal of quantifying prior-likelihood conflicts. They also provide information deviances to be examined within our framework, in addition to the quadratic loss measures, a focus of this paper. Furthermore, the concept of *surprise* has been used in Evans (1997), Evans and Moshonov (2006), Bousquet (2008), and Evans and Jang (2011) for a variety of procedures including detecting prior-likelihood conflicts. In particular, Evans and Moshonov (2006) provided methods to check for conflict in proper priors, an assumption we avoid making due to the heavy reliance on improper priors in practice, whereas Bousquet (2008) expanded on this work, giving a binary decision rule for determining if there is conflict.

1.2 | The normal enlightenment

As usual, a normal distribution example sheds much light on what lies ahead, and permits exact analytic results. Our main theoretical results, given in Section 2, show that under an asymptotic regime that permits the influence of the prior to grow with the likelihood data size, the exact normal results are special cases of the asymptotic results for a large class of likelihood-prior models. We will also use this example (and others) in Section 4 to check the implementation and computational procedures outlined in Section 3.

We begin by assuming $\vec{X}_n = \{X_1, \ldots, X_n\}$ to be an i.i.d. sample from $N(\mu, \sigma_0^2)$, where for simplicity of illustration, we assume σ_0^2 is known. We adopt the usual conjugate prior on μ, $\pi = N(\mu_\pi, \sigma_0^2/m)$, but we parameterize the prior variance as σ_0^2/m since if our prior was set according to a previous data set also from the model $N(\mu, \sigma_0^2)$, say $\{Y_1, \ldots, Y_m\}$, then the variance of the previous MLE of μ would be σ_0^2/m. For the baseline prior, π_b, we take $m = 0$ that is a constant prior. With a slight abuse of notation, we let $\vec{X}_k$ be a generic notation for any subset of $\vec{X}_n$ with size $k \le n$, and $\overline{X}_k$ denotes the corresponding sample average. The posterior of μ given $\vec{X}_k$ under either prior is normal with, respectively,

$$\mathrm{E}_\pi[\mu \mid \vec{X}_k] = \frac{m\mu_\pi + k\overline{X}_k}{m+k}, \quad \mathrm{Var}_\pi[\mu \mid \vec{X}_k] = \frac{\sigma_0^2}{m+k}; \tag{1}$$

$$\mathrm{E}_{\pi_b}[\mu \mid \vec{X}_k] = \overline{X}_k, \quad \mathrm{Var}_{\pi_b}[\mu \mid \vec{X}_k] = \frac{\sigma_0^2}{k}. \tag{2}$$

It is natural to ask, how π has changed our posterior inference for μ compared to the baseline? To be specific, let us examine the posterior mean-squared error (MSE), averaged over all data sets that are generated (under the normal model) by the same μ_0 that generated our observed data $\vec{X}_n$. Therefore, denoting by U this expected MSE, we would like to compare

$$U_\pi(k) = \frac{\sigma_0^2}{m+k} + \mathrm{E}\left[\frac{m\mu_\pi + k\overline{X}_k}{m+k} - \mu_0\right]^2 \quad \text{with} \quad U_{\pi_b}(k) = \frac{\sigma_0^2}{k} + \mathrm{E}\left[\overline{X}_k - \mu_0\right]^2.$$

Let $\Delta = \sqrt{m}(\mu_0 - \mu_\pi)/\sigma_0$, which encapsulates the degree of *discordance* between the prior and the likelihood. It is then straight forward to show that

$$U_\pi(k) = \frac{[2k + m(1+\Delta^2)]\sigma_0^2}{(m+k)^2} \text{ and } U_{\pi_b}(k) = \frac{2\sigma_0^2}{k}. \tag{3}$$

Our approach is to find $M(k)$ such that $U_{\pi_b}(k + M(k)) = U_\pi(k)$, so $M(k)$ can be viewed as the PSS of π relative to π_b.

After some algebra, we can express

$$M(k) = m\left[1 - \frac{(r+1)(\Delta^2 - 1)}{2(r+1) + r(\Delta^2 - 1)}\right], \tag{4}$$

where $r = m/k$ is the nominal prior size relative to the likelihood data size. Expression (4) reveals something unexpected: $M(k) = m$, the *perceived* PSS, if and only if $\Delta^2 = 1$. This may surprise those who expect that $M(k) = m$ when Δ=0. However, if our prior was specified according to a prior data set $\{Y_1, \ldots, Y_m\}$, then we would have set $\mu_\pi = \overline{Y}_m$, and hence $\Delta^2 = m(\overline{Y}_m - \mu_0)^2/\sigma_0^2$, which is distributed as χ_1^2 when the prior data set is indeed from the same population. That is, on average we should expect Δ^2 to be 1, not 0. Therefore, $\Delta^2 < 1$ means we have a 'fortuitous' prior (as compared to a no-conflict prior), and hence $M(k) > m$ because of the additional 'lucky' information brought in by μ_π. When $1 < \Delta^2 < 1 + 2(1+r)$, we have $0 < M(k) < m$, meaning that, although the prior is not as informative as its nominal size m advertises, it is still helpful in the sense of reducing the MSE over using the baseline. However, when $\Delta^2 \geq 1 + 2(1+r)$, the prior has zero or negative impact, because $M(k) \leq 0$. 1

In summary, (4) tells us that,*with respect to the impact on MSE for estimating* μ,

$$\begin{array}{lll} \text{When } \Delta^2 < 1, & M(k) > m; & \text{lucky prior} \\ \text{When } \Delta^2 = 1, & M(k) = m; & \text{"advertised" prior} \\ \text{When } 1 < \Delta^2 < 1 + 2(1+r), & 0 < M(k) < m; & \text{unlucky but helpful prior} \\ \text{When } \Delta^2 = 1 + 2(1+r), & M(k) = 0; & \text{zero-impact prior} \\ \text{When } \Delta^2 > 1 + 2(1+r), & -k \leq M(k) < 0; & \text{harmful prior} \end{array} \tag{5}$$

This calls for a more general concept of *prior-likelihood discordance* than prior-likelihood conflict to describe the lack of harmony between our *likelihood model* and *prior model*, because not all such discordance is harmful, which the phrase 'conflict' would suggest. Indeed, $\Delta^2 = 0$ represents our most fortunate case, with the prior mean being exactly the true parameter value and hence $M(k)$ reaches its maximal value $m(3+2r)/(2+r)$. We observe that this maximal value is always between $1.5m$ and $2m$. We suggest that the term *prior-likelihood conflict* is reserved for cases when our prior becomes harmful, that is, when $M(k) < 0$. We must emphasize that we take a pragmatic perspective in suggesting these terms, by considering primarily the impact of the target prior on the chosen inference with respect to a specified evaluative metric (and a baseline prior). Hence a zero-impact prior does not mean a zero-information prior (which is a self-contradictory phrase in the Bayesian framework), nor does a helpful prior imply no harmful consequences, such as lack of robustness; see Al-Labadi and Evans (2017).

Regardless of the value of Δ^2, we see from (4) that $M(k)$ is a strictly decreasing function of k, unless $\Delta^2 = 1$ when it is a constant function. Its decreasing rate is controlled by $\Delta^2 - 1$, with the most rapid decreasing occurring when $\Delta^2 = \infty$, in which case $M(k) = -k$. We see that, whenever there is a

prior-likelihood discordance (regardless of being lucky or unlucky), the slope of $M(k)$ will be negative. Its extreme value, -1, is also achieved if and only if $\Delta = \infty$, which means that the prior-likelihood conflict is so extreme that it wipes out the entire likelihood. When we treat k as a continuous index, the derivative of $M(k)$ will be always bounded below by -1.

Therefore, this normal example leads to (at least) five observations:

1. PSS is a relative concept, relative to the size of the likelihood sample size (LSS);
2. The dependence of PSS on LSS is governed by the prior-likelihood discordance;
3. The prior-likelihood discordance can be both beneficial and harmful;
4. PSS can take negative values, when the prior-likelihood discordance is severe;
5. The PSS as a function of LSS, $M(k)$, has a slope that is bounded below by -1.

The main contribution of this paper is to show, theoretically and empirically, that these observations hold rather generally. Theoretically, we show that the normal formula (4), not surprisingly, holds asymptotically for a rather general class of distributions and hence (5) holds as well. This asymptotic approximation provides a quick (and not too dirty) assessment of the prior impact almost as a by-product of the original posterior computation. But for those who are willing and able to do more, we also describe a finite-sample bootstrap-like method to estimate $M(k)$, especially its slope, as a function of k, which provides a diagnostic tool for detecting the discordance.

A reviewer's comment also reminded us to stress that the notion of *prior-likelihood discordance* is a qualitative and absolute concept, intended to indicate any kind of incompatibility between the prior and likelihood (function), harmful or not. In contrast, the classic notion of *prior sample size* (PSS) is a quantitative and relative concept, designed to provide a practically appealing measure to numerically index the strength or weakness of an adopted prior with respect to our likelihood function. Whereas both concepts are needed, quantitative measures can do more harm than qualitative ones because of their seductive nature of being precise, regardless of their validity. It is therefore critical for those of us who develop such measures to be explicit about their limitations and potential misuse, a practice we follow whenever appropriate.

The remainder of the paper is organized as follows. Section 2 provides our general framework, and implements it asymptotically, as well as for finite-sample i.i.d. data. Section 3 establishes theoretical results for a large class of distributions to justify the implementations in Section 2. Section 4 gives a simulation study, and a real-data application. Section 5 concludes with a discussion of complications, limitation and open problems. Some secondary proofs and technical verifications are in the online supplemental material. All computations were done using `R` and the accompanying code can be found at the corresponding authors website.

2 | A GENERAL FORMULATION OF PRIOR SAMPLE SIZE

Let $Y \in S$ represent a data set and $f(y|\theta)$ its density, with $\theta \in \Theta$ being the model parameter, and θ_0 the value that generated Y. Let $I = I(\theta)$ be a user defined *scalar* indicator of information content in Y about θ, that is, I is a non-negative real number determined by $f(y|\theta)$. We can index Y by I, and use the notation Y_I and y_I as needed. When Y consists of n i.i.d. observations, we typically set $I=n$.

Let $\mathcal{P}$ be the set of all distributions over Θ, and $D: \mathcal{P} \to [0, \infty)$ be a user defined measure quantifying the amount of uncertainty in a particular distribution or a loss function when one invokes a decision-theoretic perspective. For example, when θ is univariate, D can be the variance, the mean absolute deviation, the mean squared error to a specified value of the parameter, etc.

The range of D implies that it exists and is finite, a condition which may require us to restrict its domain to a subset of $\mathscr{P}$. We emphasize that our approach only requires D be real valued, not that θ be univariate.

In general, the choice of D should reflect aspects of a posterior that are most relevant to what we want to learn. A common choice is the posterior MSE as in Section 1.2:

$$D_{\theta_0}(\pi(\cdot\,|Y)) = \mathrm{Var}_\pi(\theta|Y) + [\mathrm{E}_\pi(\theta|Y) - \theta_0]^2, \tag{6}$$

where, for notation simplicity, we assume θ is univariate, and we use the subscript θ_0 to highlight the dependence of D on the true value. Recall that expected measures such as MSE are typically not invariant even to one-to-one transformations, regardless of whether they are for estimation uncertainty or prediction error. We stress that as a pragmatic measure to capture the impact of a prior on the actual values of these expected measures, the proposed PSS can vary with the scale of what we want to estimate or predict; see Section 5.3. We can also consider other L_p measures, such as L_1 distance (see Section E of the appendix). We discuss several possible directions in Section 5.4 on choices for D as future work. For additional ideas on choosing D, see Morita et al. (2008) for a measure based on curvature of the log likelihood and Gelman et al. (2014) for measures based on deviances. One can also use multiple Ds to serve for different purposes.

2.1 | Define the prior information function

For a given D, the *expected posterior loss (i.e. risk) with respect to the true model* is

$$U_\pi(I) = \int_S D_{\theta_0}\left[\pi(\cdot\,|y_I)\right] f(y_I|\theta_0)\; d\mu(y_I).$$

Given a baseline prior π_b, we define $M(I)$ as the amount of information needed to match the risk in $\pi(\theta|y_I)$ to that in $\pi_b(\theta|y_{I+M(I)})$, that is, we seek M such that

$$U_{\pi_b}(I + M(I)) = U_\pi(I), \tag{7}$$

just as in the normal example, where $I = k$.

To see how $M(I)$, as a function of I, is useful for detecting prior-likelihood discordance, let us assume it is differentiable with respect to I, which, as an index for information, can be treated as continuous. Assuming differentiability as needed, and taking the derivative with respect to I in (7),

$$U'_{\pi_b}(I + M(I))\left[1 + M'(I)\right] = U'_\pi(I),$$

we arrive at, assuming $U'_{\pi_b}(I + M(I)) \neq 0$,

$$1 + M'(I) = \frac{U'_\pi(I)}{U'_{\pi_b}(I + M(I))}. \tag{8}$$

When D is chosen appropriately, $U(I)$ should be a strictly decreasing function of I, since an appropriate uncertainty measure should decrease as the information I increases *in expectation* (see the on-line supplement for why we need to emphasize this issue, as well as Meng and Xie (2014) on how variance measures

violate this monotonicity for inefficient procedures). This implies that the right hand side of (8) will be non-negative, yielding $M'(I) \geq -1$, confirming observation (V) from the normal example. Moreover, a negative $M'(I)$ implies that the uncertainty decreases slower when using π because the left hand side of (8) is less than one, indicating a discordance between the likelihood function $L(\theta|Y) \propto f(Y|\theta)$ and the prior π. The -1 lower bound has a practical interpretation: the most extreme prior-likelihood conflict *detectable* by $M(I)$ is when the negative information in the prior erases every single piece of information (defined by the information in a single data point) added to the likelihood.

On the other hand, when there is no detectable discordance, for example, when the prior π comes from a (exchangeable) previous experiment on the same θ_0, the information in the prior should stay about the same regardless of the information in the likelihood function. Hence $M'(I)$ will be approximately zero. This interpretation is most obvious when we notice that $\lim_{I\to\infty} M(I)/I = \lim_{I\to\infty} M'(I)$ by L'Hôpital's rule if $M(I)\to\infty$, and that $I+M(I) = I[1+R(I)]$, where $R(I) = M(I)/I$. Hence $M'(I)$, for large I, approximates the direct measure $R(I)$, *the information gained or lost due to the prior relative to that in the likelihood*. Therefore, when the information in the likelihood grows but the prior information stays about the same, $M'(I) \approx R(I) \approx 0$ for large I, that is, the prior information is negligible asymptotically.

In contrast, if say $M'(I)$ or $R(I) \approx -0.5$, then the prior-likelihood conflict has caused a reduction of 50% information, for example, our posterior mean/mode based on 1000 i.i.d. observations and our prior π behaves like the posterior mean/mode based on 500 i.i.d. observations and the baseline prior π_b, which typically behaves like the MLE based on 500 i.i.d. observations. Clearly it is helpful for users of Bayesian methods to be aware of such loss of efficiency, just as they should be aware of the uncertainty in their estimators.

An appealing property of using $M(I)$ to measure I_{prior}, the prior information, is that if we view $I+M(I)$ as the information measure for the posterior, $I_{posterior}$, then trivially

$$I_{\text{posterior}} = I_{\text{likelihood}} + I_{\text{prior}} \tag{9}$$

because the information in the likelihood, $I_{\text{likelihood}}$, is I in our setup. Whereas (9) is practically appealing, it is a non-standard information decomposition because $I_{\text{prior}} = M$ can be negative, pointing to a prior-likelihood conflict.

2.2 | Implementing the asymptotic formula

As we will demonstrate in Section 3, the normal formula (4) holds asymptotically for a general class of distributions. Our theoretical and empirical investigations provide us sufficient confidence to suggest that, *in the absence of other more reliable methods*, it can be adopted to be a rule-of-thumb for a quick assessment of the impact of the prior. Specifically, formula (4) implies that, when $k=n$ (which is our target case),

$$R(n) = \frac{M(n)}{n} = r\left[1 - \left(\frac{2}{\Delta^2 - 1} + \frac{r}{1+r}\right)^{-1}\right]. \tag{10}$$

Therefore, to compute $R(n)$, we only need to compute Δ^2 and r. Here r and Δ can take on a number of asymptotically equivalent forms and hence they can be estimated in a number of different ways.

For computational simplicity, in general, we recommend using the estimates

$$\widehat{r} = \frac{1}{d}\text{trace}\left(\widehat{\Sigma}_{\pi_b}\Sigma_{\pi}^{-1}\right), \text{ and } \widehat{\Delta}^2 = (\widehat{\mu}_{\pi_b} - \mu_{\pi})^{\top}\Sigma_{\pi}{}^{-1}(\widehat{\mu}_{\pi_b} - \mu_{\pi}); \tag{11}$$

where d is the dimension for multivariate θ. Note here we have deviated from our assumption of d=1 in order to provide explicit general formula, which might not be immediate for general practitioners if we only give the univariate version $\widehat{r} = \widehat{\sigma}^2_{\pi_b}/\sigma^2_{\pi}$ and $\widehat{\Delta}^2 = (\widehat{\mu}_{\pi_b} - \mu_{\pi})^2/\sigma^2_{\pi}$. Here μ_{π} and Σ_{π} (or σ^2_{π}) are, respectively, the *prior* mean and variance of θ from the target prior π, and $\widehat{\mu}_{\pi_b}$ and $\widehat{\Sigma}_{\pi_b}$ (or $\widehat{\sigma}^2_{\pi_b}$) the *posterior* mean and variance of θ under the baseline prior π_b (and use all the data). These four quantities are readily available for the vast majority of Bayesian analyses where a proper prior is used; note the need of assessing the prior impact relative to a baseline prior (often improper) arise typically only when the prior is proper. Just as a sanity check, for the example in Section 1.2, $\sigma^2_{\pi} = \sigma^2_0/m$, $\widehat{\sigma}^2_{\pi_b} = \sigma^2_0/n$, and hence $\widehat{r} = m/n = r$ (when k=n). Furthermore, $\widehat{\mu}_{\pi_b} = \overline{X}_n$, hence $\widehat{\Delta}^2 = m(\overline{X}_n - \mu_{\pi})^2/\sigma^2_0$, which consistently estimates $\Delta^2 = m(\mu_0 - \mu_{\pi})^2/\sigma^2_0$ (recall here σ_0 is a known constant).

There are cases, however, where a proper prior does not have variance or even mean, such as the Cauchy prior in Section 4.3. Our theory actually does not require them to exist, but rather the existence of a prior estimate of θ and the associated uncertainty measure, denoted by $\tilde{\mu}_{\pi}$ and $\tilde{\Sigma}_{\pi}$ (or $\tilde{\sigma}^2_{\pi}$) respectively. For the Cauchy prior, for example, we can use its median for $\tilde{\mu}_{\pi}$ and its scale parameter for $\tilde{\sigma}_{\pi}$.

In those cases where the prior estimate μ_{π} and its associate uncertainty Σ_{π} are not readily available, one can use the same routine for computing the posterior mean and variance to approximate them by applying the routine to a random selected subsample of size n_0. Ideally we want to set $n_0 = 0$, but if that is not permissible (e.g. resulting in a non-convergent MCMC), we can use n_0 the smallest possible one that still rends a well-defined output from the posterior routine. That is, we are willing to move a very small part of the likelihood into the prior in order to gain computational simplification, and then assessing the contribution of this enhanced prior, as an approximation to the actual prior contribution. This might cause a slightly over-estimation or underestimation of our prior contribution, but as long as n_0 is a few percentages of the total n, the resulting estimate $\widehat{r}_{\Delta}$ should serve the same purpose as that from the actual $R(n)$. Again, our practical interest is to gain a reasonably quantified feeling of the impact of the prior (e.g. whether it is 5% or over 30%), not to pinpoint the exact prior contribution, which will not be a fruitful pursuit even if it is theoretically possible.

2.3 | A finite-sample procedure for i.i.d. data

Even as a quick-and-not-so-dirty assessment metric, the accuracy of (10) will depend on how soon the asymptotic kicks in. For data consist of $X_1, \ldots, X_n \overset{i.i.d}{\sim} f(x|\theta)$, with $\pi(\theta)$ being the target prior, we can also implement $M(r)$ empirically and numerically, as long as we are willing to perform some non-trivial computation. Specifically,

1. Choose a baseline prior $\pi_b(\theta)$, such as an objective or reference prior; see Kass and Wasserman (1996) and Berger et al. (2009). A flexibility of our strategy is the allowance of atypical baselines (e.g. Protassov et al., 2002).
2. Choose $D_{\theta_0}(\cdot)$ and then construct an estimator of

$$U_{\pi,\theta_0}(k) = \mathrm{E}[D_{\theta_0}(\pi(\cdot|X_1, \ldots, X_k))|\theta = \theta_0], \quad k = 1, \ldots, K,$$

where we choose $K = O(n^{1/2})$ for reasons given in Section 3. Letting $\boldsymbol{\omega}_k$ be the $\binom{n}{k} \times k$ matrix enumerating all possible $\binom{n}{k}$ subsamples of $\{1, \ldots, n\}$ of size k, we can then estimate $U_{\pi, \theta_0}(k)$ by $\widehat{U}_{\pi,\widehat{\theta}_n}(k)$, where $\widehat{\theta}_n$ is an efficient estimator of θ_0 based on all data $\{X_1, \ldots, X_n\}$, and

$$\widehat{U}_{\pi,\widehat{\theta}_n}(k) = \frac{1}{\binom{n}{k}} \sum_{j=1}^{\binom{n}{k}} D_{\widehat{\theta}_n}[\pi(\cdot \mid X_{\boldsymbol{\omega}_k(j,1)}, \ldots, X_{\boldsymbol{\omega}_k(j,k)})]. \tag{12}$$

In practice, a sub-sampling strategy, that is, bootstrapping, will typically suffice. Obtain $\widehat{U}_{\pi_b,\widehat{\theta}_n}(k)$ analogously to $\widehat{U}_{\pi,\widehat{\theta}_n}(k)$, with the baseline π_b in place of π.

3. Interpolate the $\widehat{U}$ functions so they live on the real line. We use linear interpolation for simplicity, but one can investigate more sophisticated methods. We then define

$$\widehat{M}(k) = \arg\min\{m \in \mathbb{R}: \widehat{U}_{\pi,\widehat{\theta}_n}(k) = \widehat{U}_{\pi_b,\widehat{\theta}_n}(m+k)\}.$$

For $\widehat{M}(k)$ to exist, we need to avoid (at least) $\widehat{U}_{\pi,\widehat{\theta}_n}(0) < \widehat{U}_{\pi_b,\widehat{\theta}_n}(K)$, that is, the information in π is so strong that it exceeds the combined information from the entire likelihood with all K observations and from the baseline prior. Whereas we can try k (and hence K) as large as n, the very need to do so should serve as a warning that the prior is very informative. Indeed, if the solution still does not exist when k=n, then it suggests that at least 50% of our posterior information will come from our prior π.

4. Plot the sequence $\widehat{M}(k)$ and $\widehat{R}(k) = \widehat{M}(k)/k$ against k, for k=1, …, K, and regress $\widehat{M}(k)$ on k for $k = k_0, \ldots, K$ for some suitably chosen k_0 to estimate an *approximate limiting slope* of $\widehat{M}(k)$ as a function of k, denoted by S_K. Based on our current theoretical and empirical evidence, we observe the following:
5. When there is no noticeable prior-likelihood discordance, $\widehat{M}(k)$ stays fairly constant, and hence $S_K \approx 0$, and $\widehat{R}(k)$ will approach zero rapidly as k increases;
6. Any serious departure of $\widehat{M}(k)$ from being a constant function, especially as a monotone decreasing function, indicates a prior-likelihood discordance;
7. Both $R(k)$ and S_K serve as measures of the degree of discordance, where $R(k)$ measures the loss (or gain) due to the prior-likelihood discordance at a finite $k \leq K$, and S_K serves an estimator of $R(n)$, the object of our central interest, for $n >> K$; Very serious prior-likelihood conflict will cause $\widehat{R}(k)$ or S_K to approach -1, that is, the conflict would essentially wipe out all the information in the likelihood.

We use S_K instead of $\widehat{R}(K)$ to estimate $R(n)$ because K needs to be chosen such that $n/K = O(n^{1/2}) \to \infty$ and hence $\widehat{R}(K)$ is often too far from $R(n)$. However, as long as we are able to choose k_0 such that $\widehat{M}(k)$, for $k \geq k_0$, is reasonably linear in k, we can approximate $R(n)$ by the slope from regressing $\widehat{M}(k)$ on k for $k \geq k_0$. The theoretical and empirical evidence provided below indicates that this approximation is of practical value. Nevertheless, we do not have any evidence, nor intuition, to suggest that it cannot be improved; we hence invite readers to search for improvements.

3 | THEORETICAL UNDERPINNING

This section establishes an asymptotic result to provide some theoretical insight about the procedure given in Section 2.3, with the D being the estimated posterior MSE given by, for k=1, …, $K(\ll n)$,

$$D_{\widehat{\theta}_n}(\pi(\cdot\,|\vec{X}_k)) = \mathrm{Var}_\pi(\theta|\vec{X}_k) + [\mathrm{E}_\pi(\theta|\vec{X}_k) - \widehat{\theta}_n]^2, \text{ with } \widehat{\theta}_n = \mathrm{E}_{\pi_b}[\theta|\overrightarrow{X_n}]. \tag{13}$$

Any suitable asymptotic regime here must permit the prior influence to grow in some suitable way with the likelihood data size. Otherwise the prior contribution would become negligible by design, as with the standard asymptotic framework for the large-sample equivalence between Bayesian and likelihood inferences. We emphasize that the standard asymptotic framework is statistical, meaning that its limiting process is a statistically feasible one, at least conceptually. The non-standard asymptotics strategy we adopt is mathematical, invoked purely for obtaining a tractable mathematical expression to approximate a target quantity. This is in the same spirit as the popular large-p-small-n asymptotics, where the number of parameter p is assumed to grow with the sample size n, a process typically with no scientific or statistical reality, because nature and humans do not collaborate with each other in choosing the number of variables (p) relative to the sample size (n). See Li and Meng (2021) for a discussion about the importance of distinguishing between *mathematical asymptotics* and *statistical asymptotics*.

Our non-standard regime shows that the key identity for the normal case, (4), holds asymptotically, essentially for all posterior-prior families that satisfy the following 'functional shrinkage' assumption. For simplicity, we restrict θ to be univariate, but the results hold generally with necessary extensions of notation, as illustrated by (11).

Assumption 1 *Assume* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta)$ *with respect to a measure on* $\mathbb{R}$*, where* $\theta \in \mathbb{R}$*. Assume that the prior,* $\pi(\theta)$*, is such that there exists* $m>0$ *and* $\mu_m \in \mathbb{R}$ *such that for any* $\vec{X}_k = \{X_1, \ldots, X_k\}$*, where* $k \geq k^*$ *for some fixed* k^**, the following hold*

$$\mathrm{E}_\pi[\theta|\vec{X}_k] = u(T_{k,m}) + O_p((m+k)^{-1}) \text{ and } \mathrm{Var}_\pi[\theta|\vec{X}_k] = \frac{v(T_{k,m})}{m+k} + O_p((m+k)^{-2}), \tag{14}$$

where u *is a twice differentiable and* $v>0$ *is a differentiable,*

$$T_{k,m} = \frac{m\mu_m + k\overline{T}_k}{m+k}, \tag{15}$$

and $\overline{T}_k$ *is the average of some* $T_i = T(X_i)$ *over* i=1, …, k*, whose mean* $\mu_T = \mathrm{E}[T_i|\theta]$ *and variance* $\sigma_T^2 = \mathrm{Var}[T_i|\theta]$ *are assumed to exist. Furthermore, assume that our baseline prior* π_b *corresponds to the limiting case of* π *when* m *is set to zero. That is,*

$$\mathrm{E}_{\pi_b}[\theta|\vec{X}_k] = u(\overline{T}_k) + O_p(k^{-1}) \text{ and } \mathrm{Var}_{\pi_b}[\theta|\vec{X}_k] = \frac{v(\overline{T}_k)}{k} + O_p(k^{-2}). \tag{16}$$

Assumption 1 is satisfied by many common conjugate prior distributions including the six natural exponential families (NEFs) with quadratic variance functions (Morris, 1982); see the online Supplement. More broadly, under standard regularity conditions, a log-likelihood function resulting from i.i.d. data is known to be asymptotically quadratic, and hence we can expect Assumption 1 to

hold at least asymptotically. Perhaps the easiest way to gain insight is to consider the parallel to the normal case in Section 1.2, where is particularly easy to understand the $T_{k,m}$ expression in (15), as a weighted average of the sample mean and the prior mean, with weights proportional to their respective precisions. As is well known, this weighted average is the backbone of the much celebrated shrinkage estimation from a Bayesian perspective (e.g. Efron & Morris, 1973). Hence we view Assumption 1 as an assumption of *functional shrinkage* because it requires both the posterior mean and variance as functions of the standard linear shrinkage estimator as in the normal example of Section 1.2, where $T(x) = x$.

The comparison with the normal example also gives us the insight that m can be interpreted in general as the *nominal* PSS measured on the same unit scale as the likelihood data size. We say m is *nominal* because the real PSS must take into account the potential prior-likelihood discordance, as emphasized previously. Furthermore, Assumption 1 does not require the existence of the prior mean, but only the existence of m and μ_m (see the exponential example in Section 4.2). Therefore, in general, μ_m should be regarded as a measure of prior centrality, and is not necessarily the prior mean for μ_T.

We use the notation μ_m to indicate that the prior centrality can depend on m. This is obvious when our prior information actually comes from a previous study based on a data set $\{\tilde{X}_1, \ldots, \tilde{X}_m\}$, which are i.i.d. samples from $f(x|\theta_1)$, where θ_1 may differ from θ_0, the generating value for our data $\{X_1, \ldots, X_n\}$. Assuming the previous Bayesian analysis used the same baseline prior π_b, we know from (16) that the prior mean for μ_T will be approximately $\tilde{T}_m$, the average of $\{T(\tilde{X}_i), i = 1, \ldots, m\}$.

The simple concept that we can approximate μ_m by $\tilde{T}_m$ turns out to provide rather useful insights for forming an appropriate asymptotic regime, a regime that permits m to grow with k such that $k/(k+m)$ stays within the interval (0,1). Specifically, if we let $\Delta = \sqrt{m}(\mu_m - \mu_T)/\sigma_T$, then the fact that $\Delta \approx \sqrt{m}(\tilde{T}_m - \mu_T)/\sigma_T$ means that even under the assumption $\theta_1 = \theta_0$, Δ^2 will not approach zero, because Δ^2 is a test statistic—based on data $\tilde{T}_m$—of the null hypothesis $H_0: \theta_1 = \theta_0$; its asymptotic null distribution, as $m \to \infty$, is the chi-squared distribution χ^2_1, which is exact in the normal example of Section 1.2. It is therefore meaningful in our asymptotic regime to consider Δ as fixed while permitting m to grow, because Δ provides a probabilistic yardstick for assessing how the prior data set, as a proxy for the prior information, differs from the current data set used for the likelihood function. Consequently, we build our asymptotic regime under the following assumption:

Assumption 2 *For the μ_m given in Assumption 1, we assume that it can be expressed as*

$$\mu_m = \mu_T + \Delta \frac{\sigma_T}{\sqrt{m}} + O_p(m^{-1}) \tag{17}$$

for some fixed constant $\Delta \in \mathbb{R}$.

As shown shortly, the two assumptions above play a critical role in establishing an asymptotic expression for $R(k)=M(k)/k$. The next assumption is of a technical nature to ensure that our asymptotic expression is unique, and it holds trivially in virtually all applications. Nevertheless, it is needed for eliminating pathological cases where properties that hold in probability, as in (14), fail to hold almost surely, as required by Assumption 3; for practical purposes, this difference is almost immaterial.

Assumption 3 *We assume (i) both $\hat{U}_{\pi,\hat{\theta}_n}(I)$ and $\hat{U}_{\pi_b,\hat{\theta}_n}(I)$ converge almost surely to zero as $I \to \infty$, and (ii) for any finite stopping time $\hat{I}$, $\hat{U}_{\pi_b,\hat{\theta}_n}(\hat{I}) > 0$, almost surely.*

We are now ready to state our main theoretical results; see Appendix A for proof.

Theorem 1 *Assume $\widehat{D}$ as defined in (13) and that both* k *and* m *increase to infinity with* n*, with the restriction $k = O(n^{1/2})$ and that* r=m/k *is strictly bounded away from zero and infinity even at its limit. Letting $c = [u'(\mu_T)]^2\sigma_T^2/\{[u'(\mu_T)]^2\sigma_T^2 + v(\mu_T)\} \leq 1$, we then have the following results.*

1. Under Assumptions 1 and 2, any $M(k) = kR(k)$, where

$$R(k) = R_r(\Delta^2) + O_p(k^{-1/2}), \text{ with } R_r(\Delta^2) = r\left(1 - \frac{c(1+r)(\Delta^2-1)}{(1+r)+cr(\Delta^2-1)}\right), \tag{18}$$

gives the PSS to the order of $O_p(k^{-1/2})$, in the sense that it satisfies

$$\frac{\widehat{U}_{\pi,\hat{\theta}_n}(k)}{\widehat{U}_{\pi_b,\hat{\theta}_n}(k+M(k))} = 1 + O_p(k^{-1/2}). \tag{19}$$

2. Furthermore, under Assumption 3, (19) holds if and only if (18) holds.

Expression (18) illustrates the role of Δ^2 in determining the behavior of $R(k)$=$M(k)/k$. As in the normal example, when $\Delta^2 = 1$, $R_r(1) = m/k = r$, implying that $M(k)$ will recover the nominal PSS m asymptotically. When $\Delta^2 \to \infty$, representing the extreme prior-likelihood conflict, $R_r(\Delta^2)$ goes to its lower limit -1; clearly $R_r(\Delta^2)$ decreases strictly monotonically to -1 as Δ^2 increases to ∞.

At the other extreme, that is, when Δ=0, we see that because we can write

$$R_r(\Delta^2) = rA_r(\Delta^2), \quad \text{where} \quad A_r(\Delta^2) = 1 - \frac{c(1+r)(\Delta^2-1)}{1+r+rc(\Delta^2-1)}, \tag{20}$$

we have $R_r(0) = rA_r$, with (recall $0 \leq c \leq 1$)

$$A_r = 1 + \frac{c(1+r)}{1+r(1-c)} \geq 1.$$

Therefore, asymptotically, $M(k)$ is larger than the nominal size m by the factor A_r. This is the beneficial discordance phenomenon seen in the normal example, where $c = 1/2$.

Intriguingly, $c = 1/2$ holds for a wide range of models. Specifically, let us assume the usual large-sample equivalence between the likelihood inference and the Bayesian inference under our baseline prior π_b, that is, as $k \to \infty$, the posterior variance of θ, $\mathrm{Var}_{\pi_b}[\theta|\overline{X}_k]$ is almost surely the same as the sampling variance of the posterior mean $\mathrm{E}_{\pi_b}[\theta|\vec{X}_k]$. Then we have from (16), by the δ-method, that

$$1 = \lim_{k\to\infty} \frac{V\left[\mathrm{E}_{\pi_b}(\theta|\vec{X}_k)|\theta\right]}{V_{\pi_b}(\theta|\vec{X}_k)} = \lim_{k\to\infty} \frac{[u'(\mu_T)]^2\sigma_T^2/k}{v(\mu_T)/k} = \frac{[u'(\mu_T)]^2\sigma_T^2}{v(\mu_T)}, \tag{21}$$

and hence the c as specified in Theorem 1 is 1/2. In Appendix B, we will verify that (21) holds for all models examined there. Furthermore, when $c = 1/2$, the increase in PSS due to beneficial discordance is

always an additional $(1+r)/(2+r)$ percent of information, and hence it is between 50% and 100%, exactly the same as in the normal example. This is an unexpected finding, especially because of its simple and general nature. The assumption (21) holds rather generally because of the standard asymptotic equivalence between the likelihood and Bayesian inferences. For some convolution families under the single observation unbiased prior (SOUP; under which posterior mean is unbiased as a point estimator; see Meng & Zaslavsky, 2002), (21) holds exactly for any k.

More generally, we see from (20) that the beneficial information kicks in as soon as $\Delta^2 < 1$, and the amount of increased information is monotone in $1 - \Delta^2$. Similarly, when $\Delta^2 > 1$, the amount of the information lost is a monotone increasing function of $\Delta^2 - 1$.

This result also says that when $|\Delta^2 - 1|$ is too small, our method will not be able to detect the prior-likelihood discordance, especially considering we can only detect the type of discordance with the likelihood that is not already presented in the baseline prior, as demonstrated below.

4 | EMPIRICAL ILLUSTRATIONS

4.1 | Computational considerations

To empirically demonstrate the performance of our procedure, we first address its computational requirements. The brute-force implementation of the procedure outlined in Section 2.3 can impose a substantial burden because of the need to recalculate posterior summaries (e.g. means and variances) for many different subsamples. For some models, this may simply be infeasible. However, when working with conjugate families, means and variances of the posterior can be immediately calculated, and hence our methods can be carried out very efficiently. For example, the simulations presented below were carried out in R making heavy use of vectorization and parallelization, and each simulation study (e.g. the entire normal example) took 5–30 min, depending on the sample size, on a laptop running an Intel i7 processor.

When one is not working with conjugate families, usually posterior calculations are carried out using MCMC. Having to obtain thousands of separate MCMC samples is often impractical. However, this can be sidestepped by a careful use of importance weights, when they are easy to compute. This can be very effective as the posteriors for two different subsamples are usually quite close. Therefore, only one or a few large MCMC samples need to be generated, which can then be used for importance sampling. To illustrate, let $(X_1, \ldots, X_k)$ and $(X'_1, \ldots, X'_{k'})$ be two separate subsamples. If the MCMC samples, $\theta_1, \ldots, \theta_m$, were generated using $\pi(\theta|X_1, \ldots, X_k)$ but we wish to calculate the posterior mean $\theta'_{post} = \mathrm{E}[\theta|X'_1, \ldots, X'_{k'}]$, then we can estimate it by

$$\widehat{\theta}'_{post} = \frac{\sum_{i=1}^{m} w_i \theta_i}{\sum_{i=1}^{m} w_i} \quad \text{where} \quad w_i \propto \frac{\pi(\theta_i|X'_1, \ldots, X'_{k'})}{\pi(\theta_i|X_1, \ldots, X_k)}.$$

This approach works well as long as we choose $(X_1, \ldots, X_k)$ reasonably wisely so the (sample) variance of the importance weight w is not too large, because the effective Monte Carlo sample size for $\widehat{\theta}'_{post}$ is largely *bounded below*[‡] by $m_{\text{eff}} = m/(1 + \mathrm{Var}(w))$ (see Kong, 1992; Liu, 1996), assuming the underly-

[‡]We thank the AE for pointing out that this often quoted formula for m_{eff} can be very misleading in cases where the importance sampling weights are designed to *improve* the Monte Carlo efficiency, for which cases a term neglected in deriving m_{eff} is in fact not negligible.

ing MCMC chain has mixed well (and without taking into account the comparison of CPU time). Using this approach, we were able to recreate the normal example, employing parallelization and the Rcpp package (to execute the necessary loops in C++), with a computation time less than 2.5 h (with the same equipment). There we took an MCMC sample of size $m = 100{,}000$, 10,000 subsamples, and k up to 100, with m_{eff} varying from 7000 to 60,000.

4.2 | A simulation study

Here we provide numerical illustrations for the normal and exponential settings. Throughout, the estimated measure of uncertainty, D, is from (13), and we take 100,000 subsamples with replacement, to construct the estimates $\widehat{U}_{\pi,\hat{\theta}_n}$ and $\widehat{U}_{\pi_b,\hat{\theta}_n}$.

The goal of this section is to highlight the empirical performance of our procedure and validate the approximation formula given in (10). To that end, we will choose the hyperparameters to reflect a desired level of r and Δ; in all examples below, we take 100 replicates, n=1000, $r = 0.05$, and consider $\Delta = 0{,}0.5, 1, 1.25$ to reflect differing levels of conflict between the prior and the likelihood. To produce the 'truth' in each setting, we also use a Monte Carlo approach with 100,000 draws to approximate the population values for the U, which can then be used to approximate the population level values for $M(k)$ and $R(k)$. These values will be marked as dashed lines throughout the plots.

Example 1: Normal with known variance

The true mean is $\mu_0 = 1$ and the variance is $\sigma^2 = 1$; the latter will be treated as known. The baseline is taken to be normal with mean zero and infinite variance (i.e. a 'flat prior'). The first row of Figure 1 depicts plots of $\widehat{M}(k)$ for 100 replications given in grey. Starting from the left column, the plots correspond to Δ=0, 0.5, 1, 1.25. A solid black line is included as the cross-sectional average of the grey lines and a dashed red line is given as the population level $M(k)$, computed via Monte Carlo. We see the solid and dash lines are practically the same. The second row of Figure 1 provides boxplots comparing $\widehat{R}(n)$ when using the plots to approximate $\widehat{M}(n)$ (*RPlot*) versus using the asymptotic formula (*QuickEst*). When using the plots, we average the last 10 values of $\widehat{M}(k)$ and then divide by n to approximate $R(n)$. That is

$$\widehat{R}_{plot}(n) = \frac{\frac{1}{10}\sum_{k=K-10}^{K}\widehat{M}(k)}{n}, \tag{22}$$

where $K = n/2 = 500$ in our simulation studies. For the normal case, the asymptotic formula is exact and thus we see that the two procedures agree as we would expect.

Lastly, we note that the slope of the $\widehat{M}(k)$ plots reflect the level of discordance between the likelihood and prior. The case of 'no conflict' corresponds to Δ=1 which is the third plot and has basically a slope of 0. There the effective sample size of the prior is about 50, which corresponds to $r = 0.05$ with $n = 1000$. As Δ moves away from 1, we see negative slopes reflecting the discordance. However, for the two Δ smaller than 1, the discordance is beneficial because it leads to larger effective sample sizes, about 70–80, demonstrating the super information phenomena.

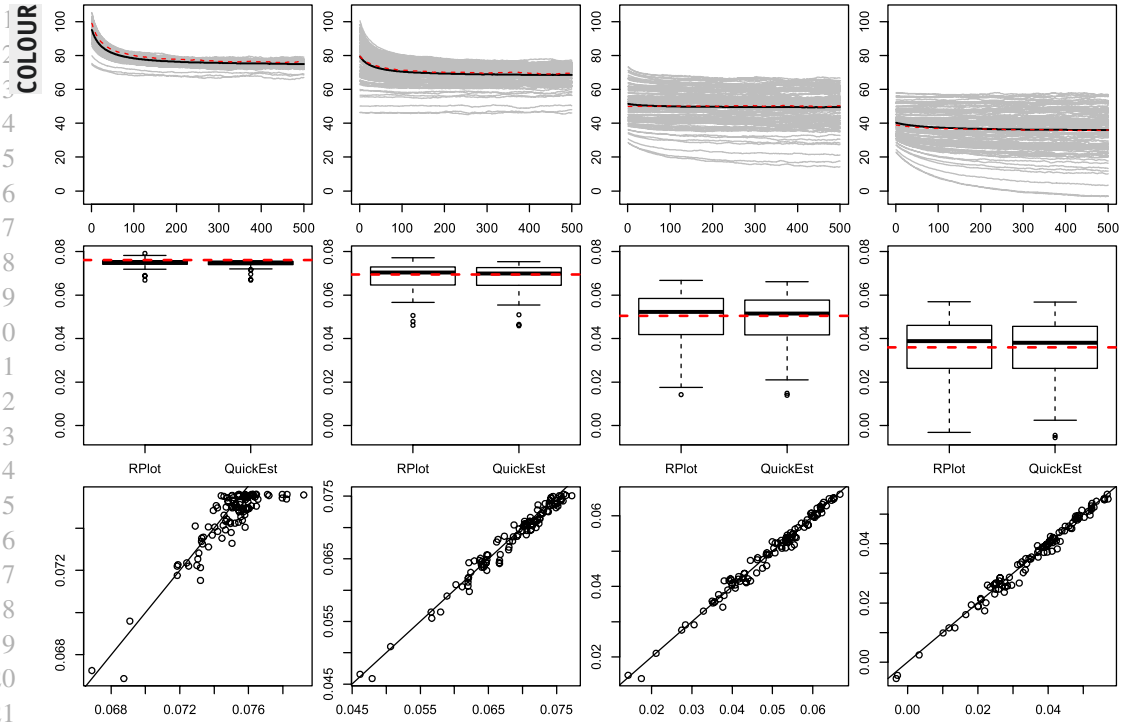

FIGURE 1 Normal Distribution. From the left, the four columns correspond to $\Delta = 0, 0.5, 1, 1.25$, respectively. The first row displays 100 replications of the $\widehat{M}(k)$ curve, with their point-wise average indicated as the solid line. The second and third rows compare the two estimates of $R(k)$, (4.1) and (10), via boxplots and scatter plots (with the QuickEst method on the y-axis). In all plots, the red dash-line represent the estimand, which itself is approximated by using 100,000 Monte Carlo samples

Example 2: Exponential under two parameterizations

We now assume that $X_1, \ldots, X_n$ are i.i.d. exponential random variables with mean $\mu = \lambda^{-1} = 1/2$ and variance $\sigma^2 = \lambda^{-2} = 1/4$. By comparing the case of $\theta = \mu$ with $\theta = \lambda$, we explore the nature of prior-likelihood discordance with respect to parameterizations. As we will see, while the two parametrizations do differ, the differences are not drastic.

The conjugate prior on λ is gamma, $\Gamma(\alpha,\beta)$, while the conjugate prior on μ is then the inverse gamma $\Gamma^{-1}(\alpha, \beta)$. Our baseline is given by taking $(\alpha, \beta)\to 0$, which yields $\pi_b(\theta) \sim \theta^{-1}$, regardless of whether $\theta = \lambda$ or $\theta = \mu$, and corresponds to the Jeffreys prior for this likelihood. The posterior of λ under the $\Gamma(\alpha,\beta)$ prior is also a gamma distribution with mean and variance

$$\mathrm{E}_\pi[\lambda|\vec{X}_k] = \frac{\alpha + k}{\beta + k\overline{X}_k}, \quad \mathrm{Var}_\pi[\lambda|\vec{X}_k] = \frac{\alpha + k}{(\beta + k\overline{X}_k)^2}. \tag{23}$$

Similarly, the posterior for μ under the $\Gamma^{-1}(\alpha, \beta)$ prior is the inverse gamma with mean and variance

$$\mathrm{E}_\pi[\mu|\vec{X}_k] = \frac{\beta + k\overline{X}_k}{\alpha + k - 1}, \quad \mathrm{Var}_\pi[\mu|\vec{X}_k] = \frac{(\beta + k\overline{X}_k)^2}{(\alpha + k - 1)^2(\alpha + k - 2)}. \tag{24}$$

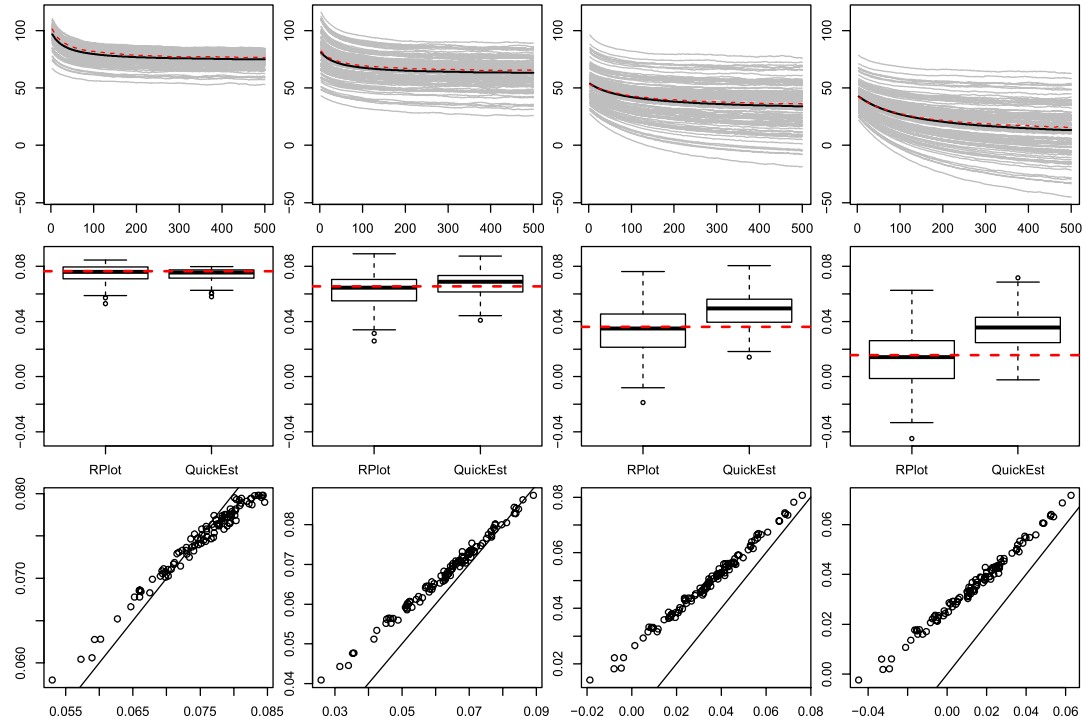

FIGURE 2 Exponential distribution with $\theta = \lambda$. The same caption for Figure 1 applies

The results for the rate and mean parametrizations are given in Figures 2 and 3, respectively. The plots are analogous to the normal plots given in Figure 1. We once again see the plots for $\widehat{M}(k)$ are centered around the population quantity, as are the boxplots for $\widehat{R}_{plot}(n)$. However, unlike the normal case, the asymptotic formula is less accurate. While the distortion is still not extreme and hence the asymptotic formula can still be used as a quick approximation, it highlights the benefits of our more computationally involved algorithm.

In terms of the parametrization, there are slight differences. In particular, the variability of our procedure is higher for the rate parameter λ than for the mean parameter μ and our asymptotic approximation seems to work better for the mean as well. However, the broader message is essentially the same. The asymptotic formula provides a computationally cost-effective approximation, but the full algorithm is useful, especially as we move away from normality and linear estimators.

4.3 | Application: Logistic regression for predicting lupus

We apply our methods on a data set provided by Dr. Haas, a client at the University of Chicago's consulting program, as reported in van Dyk and Meng (2001). The data set consists of 55 patients, 18 of which have membranous lupus nephritis also known as stage V lupus. We also have measurements on the difference between immunoglobulin G3 (IgG3) and G4 (IgG4). Haas (1994) was interested in the relationship between this difference and the presence of stage V lupus. To that end, a logistic regression model on disease status was used where a covariate representing the difference between IgG3 and IgG4 was included. A summary of the data (in counts) is reported below (Table 1).

COLOUR

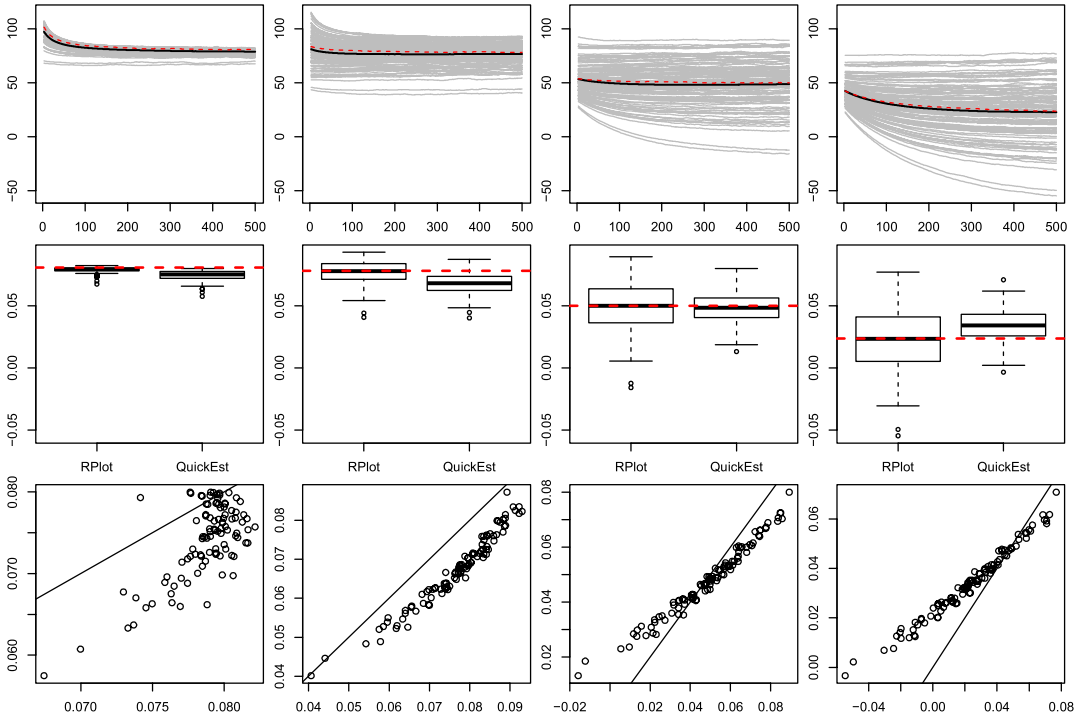

FIGURE 3 Exponential distribution with $\theta = \mu$. The same caption for Figure 1 applies

TABLE 1 Counts of patients with and without stage V lupus against IgG3-IgG4, the difference between immunoglobulin G3 and G4 levels

	IgG3 - IgG4				
Lupus	**0**	**0.5**	**1**	**1.5**	**2**
0	31	2	2	0	2
1	5	0	2	6	5

Gelman et al. (2008) investigated the idea of a *weakly informative prior*, and for logistic regression suggested, after standardizing appropriately, that one use a Cauchy prior with a scale of 2.5 on the slope parameter. We use our methodology to explore how weak or strong such a prior really is. We evaluate candidate priors from the Cauchy distribution with scales 2.5, 5 and 10 against a baseline prior, taken to be Cauchy with a scale of 10,000. The results end up being fairly robust against the choice of baseline, as we also tried Cauchy with scales 100 and 1000, as well as a normal baseline, as reported in Section E of the online supplement. We use the metric (12) to compare the two priors. However, instead of taking the mean of this metric over subsamples, we take the median to combat the well-known small-sample instability of logistic regressions, and for the same reason, we examine $\widehat{M}(k)$ only for k>10. (To be more consistent with the median approach, in the online supplemental we also explore using the mean absolute deviation for D instead of the MSE, however the results are nearly identical.) To reduce the impact of the non-linear part of $\widehat{M}(k)$ on the estimation of $R(55)$, we take $k_0 = 20$ to approximate $R(55)$ by the least squared estimator based on $k = 20, \ldots, 35$.

The results are plotted in Figure 4 with the slopes given in the caption, using both $k_0 = 10$ and $k_0 = 20$. The plots are a bit more chaotic than in our simulations due, likely, to the aforementioned

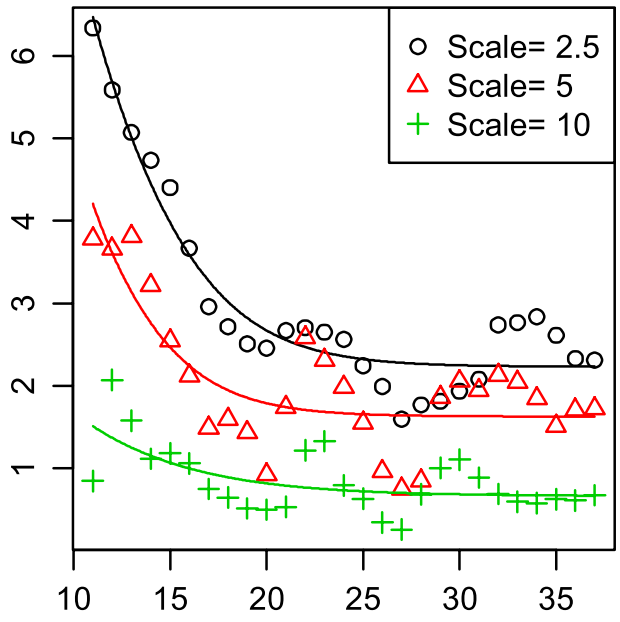

FIGURE 4 Plot of the estimated PSS $\widehat{M}(k)$ for the application in Section 4.3. Estimated slopes are −0.1173, −0.0646, and −0.0283 using k>10, for scales 2.5,5, and 10 respectively. The estimated slopes become 0.0032, −0.0024, and −0.0079 respectively when using k>20

instability of logistic regression with small sample sizes, but none of them suggests more than 6 PSS. The prior suggested by Gelman et al. (2008), that is, with scale = 2.5, seems to indeed depict a *weakly informative prior*, equivalent to between 2 and 6 data points, which is no more than 10% of the information provided by the likelihood. There might be some small amount of prior-likelihood discordance. By taking the scale up to 5 or 10, the discordance is reduced, and so is the prior impact. Indeed, the slope estimators based on $k > k_0 = 20$ are essentially zero regardless of the scale, indicating essentially negligible prior impact with $n = 55$. Such practical, quantifiable, and interpretable assessments can help greatly to strengthen our inferential conclusions and to communicate them convincingly, by reducing both the impact and the appearance of ad hoc choices made during our inference process. Evidently, it is more scientific to numerically demonstrate that the impact of a prior is no more than adding 10% of data than to simply declare that it is weakly informative. For more studies on weakly informative prior, see Gelman (2006) and Polson and Scott (2012).

We remark that logistic regression is a telling example about why it is important to formulate PSS as a measure relative to the likelihood function, and hence it is data dependent. It is well-known that when the observed data exhibit a (nearly) perfect separation pattern, that is, when a predictor (nearly) perfectly separates those with positive outcome from those with negative outcome, we will run into a (nearly) non-identifiability issue. This issue occurs rather frequently in practice, and Bayesian methods have been suggested as an effective way to address the problem, as detailed in Gelman et al. (2008), and more recently in Rainey (2016), who concluded that 'When facing separation, researchers must *carefully* choose a prior distribution ....' (emphasize is original). But since separation is a data-dependent phenomenon, and the prior is brought in to combat the issue that the likelihood function is too flat (e.g. MLE does not exist) to form a proper posterior without a suitably informative prior, the relative contribution of the prior information must be data dependent. Furthermore, to be meaningfully 'careful' in choosing such a prior, we need to be able to quantify the amount of contribution of

various choices and hence one can judicate based on quantifiable evidence, which should also help to communicate—and hence to ensure the reproducibility of—our findings.

5 | RESPONDING TO REVIEWERS: OLD CONTEMPLATION AND NEW EXPLORATIONS

5.1 | Why do we still work on such an old problem?

One reviewer expressed a strong disbelief that the Bayesian paradigm had been promoted for so long without having had settled on such a basic assessment about prior influence. We were sceptical as well when we started this project, but the progress we made (since 2012) has shown us why quantifying prior information is fundamentally as problematic as the use of an improper prior, a matter of ongoing debate. Indeed, another reviewer questioned our use of an improper prior as the base prior. They asked if one does not have a proper prior to start with, then should one question the use Bayesian methods in the first place? Philosophically, we agree with the reviewer, because it is a mathematical fact that the Bayesian paradigm cannot handle complete ignorance (e.g. Martin & Liu, 2016). Practically, we also agree that whenever meaningful, we should use a proper prior as the base, which is not restricted by our proposed framework in any way.

There are inferential paradigms that can quantify complete ignorance in fully logical and mathematical ways, such as belief functions (Dempster, 1968; Shafer, 1976). However, a recent investigation (Gong & Meng, 2021) reveals that there is currently no known inferential paradigm that can do so without having to pay a 'leap-of-faith' price somewhere else (e.g. trading indeterminacy in updating rules for that in prior specifications). The implication is that we have to live with imperfection or even paradoxes one way or another. Using improper priors seems to be the least problematic for now, given its wide-spread practice, which does not justify its use, but it does suggest that we know much more about its pros and cons, or at least know where to look for source of troubles. The following is such an example.

5.2 | A prior sample size paradox?

We are grateful to yet another reviewer for raising an intriguing question regarding what should be viewed as nominal prior size in Example 2 discussed above. The reviewer pointed out that, inspecting the density function based on an i.i.d. sample with size n, we would see a term λ^n in the exponential likelihood. When this is compared to the term $\lambda^{\alpha-1}$ in the conjugate Gamma prior, it seems natural to define the 'nominal prior size' as $n_0 = \alpha - 1$, instead of what we defined as $m{=}\alpha$. Furthermore, since $n_0 = m - 1$, with our definition, we seem to run into a negative nominal prior size when $m{=}0$. This is undesirable because as a *nominal* prior size, the worst it can be should be zero, representing no prior information. A negative nominal prior size would suggest that we know *a priori* the degree of prior-likelihood conflict before seeing any data, which would be an odd position to take, if not illogical.

However, if we follow the reviewer's definition, then our prior mean would be $\mathrm{E}(\lambda) = (n_0 + 1)/\beta$. This would imply that even when the prior size is zero, meaning there is no prior information whatsoever, we would still have a finite 'prior mean' $\mathrm{E}(\lambda){=}1/\beta$. This seems at least as illogical as having a negative prior sample size. This dilemma is rooted exactly in the indeterminacy of an 'ignorant prior' discussed in Section 5.1: if the prior is the posterior obtained from a previous study of sample size n_0, what was the prior that led to that posterior? If that was the constant prior on λ, then that

posterior would be $Gamma(n_0 + 1, \beta)$, and hence $\alpha = n_0 + 1$, consistent with the reviewer's suggestion. However, if that prior was Jeffreys prior $\pi(\lambda) \propto \lambda^{-1}$, which the reviewer also suggested to be a natural choice, then that posterior would be $Gamma(n_0, \beta)$, and hence $\alpha = n_0$, as we formulated. Hence 'looking directly at the likelihood form' and 'using Jefferys prior' cannot be 'natural' simultaneously.

Although the meaning of 'natural' is debatable in this context, we see a more compelling reason to adopt the latter, since there are various theoretical justifications for using the Jeffreys prior (Gutiérrez-Peña et al., 1997). The former runs into the danger of mixing the form of λ arising from normalizing constant for the *sampling distribution* with the form of λ arising from modelling it directly. As far as we are aware of, there is no theoretical justification on why the meanings of the powers in these two different usages of λ should be the same. Of course, the central difficulty here lies in the fundamental impossibility of using a probabilistic distribution to represent ignorance, and hence it further illustrates the necessity of choosing a baseline when we assess prior contribution.

5.3 | Should we always check likelihood and prior, regardless of prior contribution?

The answer to this reviewer's question of course is a resounding YES. We should always worry about the inadequacy of any part of our model, even or especially when we cannot check it. As many would argue, rightly, checking the likelihood is more important than assessing the prior and it should be done first, because it is the likelihood that permits us to make (Bayesian) inference from data to parameters. The prior serves typically an important but nevertheless supplemental role, except when we have very little information from our likelihood. But it is exactly because of the perceived supplemental role of our prior that we need to have some reasonable ways to assess the actual impact of a posited prior relative to the likelihood contribution, even if just for the purposes of calibrating with our original expectations. For example, in the logistic model in Section 4.3, we surely should check the adequacy of the logistic model first. But once it is adopted for whatever reasons (e.g. convenience), then if our intention is to use weak priors for its parameters, we should at least to check whether these 'weak' priors actually have weak impact; see the mortality example in Gelman et al. (1996), where a seemingly innocent uniform prior on convex curves turned out to be very influential.

This reviewer also emphasized the need to check data-prior conflict as a *falsification of the prior*, an important modelling step regardless of the need to assess the prior contribution. We also agree, without getting into the debate about checking 'subjective priors' versus 'objective priors'. Indeed, the whole industry of prior predictive and posterior checks (e.g. Box, 1980; Gelman et al., 1996; Meng, 1994; Rubin, 1984) were designed for such purposes, though as the reviewer noted there are multiple complications.

First, essentially all checks are local in the sense that they can detect only some model defects in the likelihood, prior, or both. This is due to the necessarily limited capacity of the checking/testing statistics or more generally 'realized discrepancy' (Gelman et al., 1996). We view this locality a feature rather than a deficiency, because an almighty test would or at least should reject essentially all models, because 'all models are wrong'. George Box's mantra 'but some are useful' reminds us that our job—through judicious choices of assessments—is to ensure the relevant parts are usable.

Second, not all model defects are consequential for the substantive questions at hand. Again, we avoid this issue by choosing the measure of uncertainty directly reflective of the analysis of interest, as emphasized in Section 2. We also agree with the reviewer that even when a defect is inconsequential for one study, it may still be useful to understand it since it can be very consequential for another study using the same model.

Such choices, however, lead to a third issue. Our uncertainty measure is not invariant to reparametrization, which is the case for the quadratic measure. Whereas we agree with the reviewer that an invariant measure has some general appeal, our proposed methods would not have much practical impact if we do not consider measures such as MSE, which are well understood and most commonly adopted for good reasons.

5.4 | Limitations and future work

Although our method has a number of appealing properties, much more needs to be done. Perhaps the most important extension is for problems where sample size is not a good indicator of information, as is typically the case with time series and spatially dependent data. We also need to establish theoretical results for scenarios that go beyond those covered in Section 3, and more critically to cases where the likelihood itself is misspecified in consequential ways. A reviewer also reminded us to study the issue of assessing likelihood-prior combination that could lead to substantial bias, in the sense of creating regions of parameter space that are highly probable a priori; see Baskurt et al. (2013); Evans and Guo (2019). Applications to high-dimensional and/or non-parametric problems are another important direction to explore, and the growing literature on the relationship between prior and posterior concentrations (see for example van der Pas et al. (2014) and Strawn et al. (2014) and references therein) may provide some theoretical insight on this exploration.

In applying our method, we also encountered three practical problems. The first is the computational demand. Seeking effective computational strategies is an area of much needed research, and the importance sampling approach presented in Section 4.1 merely is a starting point. The second issue involves instability with small k. We did not encounter any problem for our simulation studies, where conjugate priors were used. However, for the lupus nephritis application, we had to avoid small k because logistic regressions can be very unstable for small sample sizes. Any model which has stability problems for small samples can generate similar issues. We found switching the means to medians in our resampling scheme helped, but obviously this creates a discrepancy between the application and the current theoretical results, which are mean-based, that is, using the L_2 norm. Extending our theoretical results to cover other norms, especially the L_1 norm, as well as more general choices of the discrepancy or uncertainty measure D is another direction for future research. Third, we need to search for more reliable estimate of $R(n)$, the relative gain or loss corresponding to the actual data size, as our current extrapolation via the slope of the PSS curve $M(k)$ is more of exploratory nature.

A reviewer reminded us that a particularly interesting direction for choosing D involves moving to a prediction based uncertainty measure. This can help, to a degree, with the parametrization problem as it fixes the scale of the outcome as default. However, there are at least a few options as to how to construct a prediction based measure. One possibility is to take a similar approach to the MSE measure we introduced, which involves conditioning on the true underlying parameter. Another option could be based on the posterior predictive distribution and not conditioning on the true parameters, while yet another option would be similar to cross-validation, where observations not included in the $\vec{Y}_k$ could be used for evaluating prediction. However, preliminary explorations using cross-validation idea, as in the on-line supplement, are not very encouraging. In particular, any newly proposed measure for D has to reasonably quantify the bias of the estimates. As demonstrated in Section D of the Appendix, if this is not done, then results are quite unreliable.

Finally, we can explore other methodological applications using the idea of assessing discordance via monitoring $M(I)$. For example, we can compare two subjective priors constructed by two different

investigators, and determine whether one has more serious discordance with a likelihood function than the other. Or perhaps we can convert this diagnostic tool into something helpful in selecting a prior via tuning the measure we proposed, as a *functional* of a candidate prior, according to some sensible criterion. For instance, we may want our prior to be weakly informative in the sense that the PSS should not exceed, say, 10% of the likelihood sample size, for a chosen purpose.

Going even further, we can extend the idea of comparing two priors to comparing two likelihood functions, by using a common baseline prior. If one of the likelihood models is saturated, then the conflict between them can be viewed as a misspecification of the other, unless we just have very bad luck. Of course, whether we assess prior-likelihood discordance or misspecification of a likelihood, our general goal is the same: to be an informed Bayesian, or more generally, an informed statistical analyst.

ACKNOWLEDGEMENTS

We thank Andrew Gelman for insight comments and help regarding Section 4.3, Murali Haran for his help in Section 4.1, David Jones and Steven Finch for very helpful proofreading, and a good number of reviewers for their comments that have led to a much improved paper. We also thank the U.S. National Science Foundation, National Institutes of Health, as well as the John Templeton Foundation for partial financial support.

REFERENCES

Al-Labadi, L. & Evans, M. (2017) Optimal robustness results for some Bayesian procedures and the relationship to prior-data conflict. *Bayesian Analysis*, 12, 702–728.

Baskurt, Z. & Evans, M. (2013) Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Analysis*, 8, 569–590.

Berger, J.O., Bernardo, J.M. & Sun, D. (2009) The formal definition of reference priors. *Annals of Statistics*, 37, 905–938.

Berger, J., Bayarri, M. & Pericchi, L. (2014) The effective sample size. *Econometric Reviews*, 33, 197–217.

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

Bousquet, N. (2008) Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics*, 35, 1011–1029.

Box, G.E. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A*, 143, 383–430.

Brown, L.D., Cai, T.T. & DasGupta, A. (2001) Interval estimation for a binomial proportion. *Statistical Science*, 16, 101–117.

Clarke, B. (1996) Implications of reference priors for prior information and for sample size. *Journal of the American Statistical Association*, 91, 173–184.

Clarke, B. & Yuan, A. (2006) Closed form expressions for Bayesian sample size. *The Annals of Statistics*, 34, 1293–1330.

Dempster, A.P. (1968) A generalization of Bayesian inference. *Journal of the Royal Statistical Society, B*, 30, 205–247.

Diaconis, P. & Ylvisaker, D. (1979) Conjugate priors for exponential families. *The Annals of Statistics*, 7, 269–281.

van Dyk, D. & Meng, X.-L. (2001) The art of data augmentation (with discussions). *Journal of Computational and Graphical Statistics*, 10, 1–50.

Efron, B. (2015) Frequentist accuracy of Bayesian estimates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 617–646.

Efron, B. & Morris, C. (1973) Stein's estimation rule and its competitors—An empirical Bayes approach. *Journal of the American Statistical Association*, 68, 117–130.

Evans, M. (1997) Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics*, 26, 1125–1143.

Evans, M. & Guo, Y. (2019) Measuring and controlling bias for some Bayesian inferences and the relation to frequentist criteria. *arXiv preprint arXiv:1903.01696*.

Evans, M. & Jang, G.H. (2011) Weak informativity and the information in one prior relative to another. *Statistical Science*, 26, 423–439.

Evans, M. & Moshonov, H. (2006) Checking for prior–Data conflict. *Bayesian Analysis*, 1, 893–914.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1, 515–534.

Gelman, A., Meng, X.-L. & Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–760.

Gelman, A., Jakulin, A., Pittau, M. & Su, Y. (2008) A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2, 1360–1383.

Gelman, A., Hwang, J. & Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016.

George, E.I. & McCulloch, R. (1993) On obtaining invariant prior distributions. *Journal of Statistical Planning and Inference*, 37, 169–179.

Ginebra, J. (2007) On the measure of the information in a statistical experiment. *Bayesian Analysis*, 2, 167–211.

Gong, R. & Meng, X.-L. (2021) Judicious judgment meets unsettling updating: dilation, sure loss, and simpson's paradox (with discussions). *Statistical Science*. To appear.

Gutiérrez-Peña, E., Smith, A.F., Bernardo, J.M., Consonni, G., Veronese, P., George, E., et al. (1997) Exponential and Bayesian conjugate families: Review and extensions. *Test*, 6, 1–90.

Haas, M. (1994) IgG subclass deposits in glomeruli of lupus and nonlupus membranous nephropathies. *American Journal of Kidney Disease*, 23, 358–364.

Kass, R.E. & Wasserman, L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343–1370.

Kong, A. (1992) A note on importance sampling using standardized weights. *University of Chicago, Department of Statistics*, Technical Report 348.

5 Lee, A.J. (1990) *U statistics: Theory and practice*. New York: Marcel Dekker, Inc.

Li, X. & Meng, X.-L. (2021) A multi-resolution theory for approximating infinite-*p*-zero-*n*: Transitional inference, individualized predictions, and a world without bias-variance trade-off. *Journal of the American Statistical*
6 *Association*. To appear.

Lin, X., Pittman, J. & Clarke, B. (2007) Information conversion, effective samples, and parameter size. *IEEE Transactions on Information Theory*, 53, 4438–4456.

Liu, J.S. (1996) Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6, 113–119.

Liu, K. & Meng, X.-L. (2016) There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and Its Application*, 3, 79–111.

Martin, R. & Liu, C. (2016) *Inferential models: Reasoning with uncertainty*. Boca Raton: CRC Press.

Meng, X.-L. (1994) Posterior predictive *p*-values. *The Annals of Statistics*, 22, 1142–1160.

Meng, X.-L. and Xie, X. (2014) I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Reviews*, 33, 218–250.

Meng, X.-L. & Zaslavsky, A.M. (2002) Single observation unbiased priors. *Annals of Statistics*, 30, 1345–1375.

Morita, S., Thall, P.F. & Müller, P. (2008) Determining the effective sample size of a parametric prior. *Biometrics*, 64, 595–602.

Morita, S., Thall, P.F. & Müller, P. (2010) Evaluating the impact of prior assumptions in Bayesian biostatistics. *Statistics in Biosciences*, 2, 1–17.

Morita, S., Thall, P.F. & Müller, P. (2012) Prior effective sample size in conditionally independent hierarchical models. *Bayesian Analysis*, 7, 591–614

Morris, C.N. (1982) Natural exponential families with quadratic variance functions. *Annals of Statistics*, 10, 65–80.

van der Pas, S., Kleijn, B. & van der Vaart, A. (2014) The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8, 2585–2618.

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

Protassov, R., van Dyk, D., Connors, A., Kashyap, V. & Siemiginowska, A. (2002) Statistics, handle with care: Detecting multiple model components with the likelihood ratio test. *The Astrophysical Journal*, 571, 545–559.

Rainey, C. (2016) Dealing with separation in logistic regression models. *Political Analysis*, 24, 339–355.

Rubin, D.B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12, 1151–1172.

Shafer, G. (1976) *A mathematical theory of evidence*. Princeton: Princeton University Press.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64, 583–639.

Strawn, N., Armagan, A., Saab, R., Carin, L. & Dunson, D. (2014) Finite sample posterior concentration in high-dimensional regression. *Information and Inference*, 3, 103–133.

Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11, 3571–3594.

Watanabe, S. (2013) A widely applicable Bayesian information criterion. *The Journal of Machine Learning Research*, 14, 867–897.

Wiesenfarth, M. & Calderazzo, S. (2019) Quantification of prior impact in terms of effective current sample size. *Biometrics*, 76, 326–336.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Reimherr M, Meng X, Nicolae D. Prior sample size extensions for assessing prior impact and prior-likelihood discordance. *J R Stat Soc Series B*. 2021;00:1–25. https://doi.org/10.1111/rssb.12414