Being an informed Bayesian: Assessing prior informativeness and prior–likelihood conflict

BY MATTHEW REIMHERR
Department of Statistics, Pennsylvania State University
mreimherr@psu.edu

XIAO–LI MENG
Department of Statistics, Harvard University
meng@stat.harvard.edu

AND DAN L. NICOLAE
Department of Statistics and Department of Medicine, The University of Chicago
nicolae@galton.uchicago.edu

SUMMARY

Dramatically expanded routine adoption of the Bayesian approach has substantially increased the need to assess both the confirmatory and contradictory information in our prior distribution with regard to the information provided by our likelihood function. We propose a diagnostic approach that starts with the familiar posterior matching method. For a given likelihood model, we identify the difference in information needed to form two likelihood functions that, when combined respectively with a given prior and a baseline prior, will lead to the same posterior uncertainty. In cases with independent, identically distributed samples, sample size is the natural measure of information, and this difference can be viewed as the prior data size $M(k)$, with regard to a likelihood function based on k observations. When there is no detectable prior-likelihood conflict relative to the baseline, $M(k)$ is roughly constant over k, a constant that captures the confirmatory information. Otherwise $M(k)$ tends to decrease with k because the contradictory prior detracts information from the likelihood function. In the case of extreme contradiction, $M(k)/k$ will approach its lower bound -1, representing a complete cancelation of prior and likelihood information due to conflict. We also report an intriguing super-informative phenomenon where the prior effectively gains an extra $(1+r)^{-1}$ percent of prior data size relative to its nominal size when the prior mean coincides with the truth, where r is the percentage of the nominal prior data size relative to the total data size underlying the posterior. We demonstrate our method via several examples, including an application exploring the effect of immunoglobulin levels on lupus nephritis. We also provide a theoretical foundation of our method for virtually all likelihood-prior pairs that possess asymptotic conjugacy.

Some key words: confirmatory information; contradictory information; non-informative prior; prior distribution; prior-likelihood conflict; super-informative prior.

1. The need for Ascertaining Prior Impact

By now, the beauty of Bayesian logic is well appreciated, even by many who do not necessarily adopt it practically. Among all the common reservations about the Bayesian approach, only one appears to have stood the test of time: the difficulty in choosing the prior and in fully understanding its impact. Objective Bayesians face the impossibility of constructing a truly non-informative prior. This is not merely a philosophical quibble, but a problem practitioners routinely face. As is well known, posterior inference can be affected substantially by the scale of the parameter space on which we put a constant prior, even if the constant prior is usually regarded as being "non-informative". Alternatively, subjective Bayesians must translate prior knowledge into a suitable prior distribution. However, perfectly reflective prior specification is only a theoretical desideratum; in practice a prior sensitivity analysis is generally desirable.

As the adoption of Bayesian methods in data-rich applications becomes increasingly routine, it is more important than ever to have methods for quantitatively assessing the impact of priors, permitting at least a check on how weak or strong our prior information is compared to the information in our likelihood function. Putting it differently, even if we have a real (subjective) informative prior, it is still desirable, both scientifically and statistically, to assess how much of our posterior inference is due to our prior knowledge. A posterior inference with 45% prior information contribution may affect our decisions rather differently, even if only psychologically, from one with only 5% prior information contribution.

However, quantifying the impact of a prior has proven to be a rather challenging task, and an all-encompassing approach is seemingly philosophically and mathematically impossible. A key difficulty is that information from the prior may actually be in conflict with that from the likelihood, and hence it can "subtract" rather than "add" to an analysis. Indeed, it is not hard to argue that, to a degree, such conflict is always present, just as all models are wrong. Hence, we explore a strategy that attempts simultaneously to (i) assess if the prior-likelihood conflict is more serious than we expect, and (ii) quantify the impact of the prior on our posterior inference. To circumvent the issue of lacking a truly non-informative prior as the benchmark, we resort to the common approach of assuming a "default" prior as the baseline, as in Evans & Jang (2011). That is, a prior a practitioner would adopt without any real prior information. Furthermore, our procedure accomplishes (i) and (ii) with easily interpretable metrics such as *prior sample size*. At the heart of our method is the idea of determining how many observations it takes, approximately, to match the prior contribution to the reduction in uncertainty of a particular inference. Of special interest is that, for a prior with detectable contradictory information, the "prior data size" tends to decrease with the actual data size underlying the likelihood function.

The idea of matching distributions for equating information is not novel. Among the literature that we are aware of, the approach taken by Morita et al. (2008) seems closest to ours, as they also compute an *effective sample size* of a prior relative to a baseline prior using a distribution matching scheme. Technical differences between the two approaches include the quantity being matched (the curvature of logarithm of a distribution versus a measure of uncertainty) and distributions used for diagnostics (they match the prior to a posterior and we match two posteriors). But most critically, their measure is fully determined by the prior and the *likelihood model*, which does not reflect the specific data at hand. Whereas we fully agree with their emphasis on the usefulness of their measure for prior elicitation before data collection, a data-free measure by definition cannot accomplish the task for assessing either confirmatory or contradictory information in a prior with respect to a specific *likelihood function*.

Other related work include various deviance information criterions for measuring model fittness, such as Spiegelhalter et al. (2002), Watanabe (2010) and Watanabe (2013), and others as reviewed in Gelman et al. (2013). These methods are similar in spirit to our goal for quantifying prior-likelihood conflicts, and they also provide useful information deviances to be employed within our framework, though in this paper we focus on the usual quadratic loss type of measures. Furthermore, the concept of *relative surprise* has been used in Evans (1997), Evans & Moshonov (2006), and Evans & Jang (2011) for a variety of procedures including detecting prior-likelihood conflicts, with the last one also incorporating baseline priors.

The paper is organized as follows. Section 2 provides an overview of our strategy, and then Section 3 implements it for the setting with independent and identically distributed samples. Section 4 provides a proof-of-concept example, a simulation study, and a real-data application. Section 5 establishes theoretical justifications, and Section 6 discusses pros and cons, as well as many open problems. Technical details and proofs are provided in the Appendix.

2. A General Proposal

Let $X \in S$ represent our data set and $f(x|\theta)$ be its posited generative density, with $\theta \in \Theta$ being the model parameter. Define $I = I(\theta)$ to be a *scalar* indicator of information in X, i.e., I is a non-negative real number determined by our model $f(x|\theta)$. We therefore can index X by I, and use the notation X_I and x_I as needed. For example, when X consists of n independent and identically distributed observations, we simply set $I = n$.

Let $\mathcal{P}$ be the set of all distributions over Θ, and $D : \mathcal{P} \to [0, \infty)$ be a measure of dispersion or uncertainty. 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 $\mathcal{P}$. We emphasize that our approach only requires D, not θ, to be univariate.

Assuming θ_0 is the value of θ that generates our data at hand, we define the *average posterior uncertainty with respect to the true model* as

$$U_{\pi,\theta_0}(I) = \int_S D\left[\pi(\theta|x_I)\right] f(x_I|\theta_0) dx_I.$$

We invoke the true density $f(x_I|\theta_0)$ instead of a generic density $f(x_I|\theta)$ as we want to assess information and conflict that reflect the nature of the data at hand. The price we pay for this more specific formulation is the need to estimate θ_0, an issue we will deal with in Section 3.

Here we outline our general strategy, where, for pedagogical simplicity, we treat θ_0 as known. We compare two prior distributions, π_1 and π_2, by matching their corresponding average posterior uncertainty under a given likelihood model. Specifically, for a given θ_0, we define $M_{12}(I)$ as the amount of information that is needed to match the average posterior uncertainty in $\pi_1(\theta|x_I)$ to that in $\pi_2(\theta|x_I)$; we seek an M_{12} that satisfies the identity, exactly or approximately,

$$U_{\pi_1,\theta_0}\left(I + M_{12}(I)\right) = U_{\pi_2,\theta_0}(I). \tag{1}$$

The interpretation of $M_{12}(I)$ is easiest when our data set X consists of independent and identically distributed observations, I taken to be the usual sample size, π_1 is a "baseline" prior, and we are interested in how much information there is in π_2 relative to that in π_1. If the information in the likelihood is proportional to the sample size n, then the information in $\pi_2(\theta)$ can be viewed as proportional to $M_{12}(n)$, relative to the baseline prior $\pi_1(\theta)$.

There is, however, a hidden problem in the above interpretation: there is no guarantee that the solution $M_{12}(I)$ as defined by the equation (1) is non-negative, even if it exists; the only restriction is $M_{12}(I) \geq -I$ because $I + M_{12}(I)$ must be non-negative. Additionally, although we will focus on checking the quality of a prior by assuming the likelihood specification is acceptable, the concept of *prior-likelihood conflict* is applicable even when this assumption is false. We therefore prefer the term *prior-likelihood conflict* over *prior misspecification*. A further subtle point is that even when both the likelihood model and prior distribution are perfectly specified, the true parameter value or the observed data can still happen to be at extreme tails of their respective distribution. Given prior distributions and likelihood functions are largely artificial inferential constructs, it is important to be aware of their conflict because it serves as a warning sign of something amiss, even if just due to bad luck. Therefore, prior-likelihood conflict in this paper should always be understood as the conflict between prior distribution and the *actual likelihood function* based on our data, not the general *likelihood model* specification.

To see how $M(I)$ helps for detecting prior-likelihood conflict, let us assume it is differentiable with respect to I, which, as an index for information, can be treated as a continuous index. Assuming differentiability as needed and taking the derivative with respect to I in identity (1), we see that

$$U'_{\pi_1,\theta_0}\left(I + M_{12}(I)\right)\left[1 + M'_{12}(I)\right] = U'_{\pi_2,\theta_0}(I),$$

and hence, assuming $U'_{\pi_1,\theta_0}\left(I + M_{12}(I)\right) \neq 0$,

$$1 + M'_{12}(I) = \frac{U'_{\pi_2,\theta_0}(I)}{U'_{\pi_1,\theta_0}\left(I + M_{12}(I)\right)}. \tag{2}$$

When D is chosen appropriately, $U(I)$ should be a strictly decreasing function of I, since an appropriate uncertainty measure decreases as information I increases. This implies that the right hand side of (2) will be non-negative, yielding a lower bound of -1 on the derivative of M_{12}. Moreover, a negative $M'_{12}(I)$ implies that the uncertainty decreases slower when using π_2, indicating (when π_1 is a baseline prior) a conflict between the data and the prior π_2. This -1 lower bound has a practical interpretation: the most extreme prior-likelihood conflict *detectable* is when the negative information in the prior wipes out every single piece of information–defined by the information in a single data point–added to the likelihood. Here we emphasize the phrase *detectable* because even when $M_{12}(I)$ reaches its lower bound $-I$ and hence $M'_{12}(I) = -1$, it only reflects the negative information in the prior that is detectable via the matching method. The theoretical limit of the negative information can reach $-\infty$ for example when the supports of the prior and the likelihood function become non-overlapping, in which case the solution to (1) does not exist. Albeit in practice (hopefully) there should be enough warning signs before one needs our method to realize such extreme conflicts, the non-existence of the solution to (1) itself is a diagnosis that the prior data size exceeds that of the likelihood function; see Section 3.

On the other hand, when there is no detectable conflict, e.g., when the prior π_2 comes from a well-conducted previous experiment for the same parameter, the information in the prior should stay about the same regardless of the information in the likelihood function. Consequently, $M'_{12}(I)$ will be approximately zero. This interpretation is most obvious when we notice that $\lim_{I\to\infty} M_{12}(I)/I = \lim_{I\to\infty} M'_{12}(I)$ by L'Hôpital's rule when $M_{12} \to \infty$, and that we can write $I + M_{12}(I) = I[1 + R_{12}(I)]$, where $R_{12}(I) = M_{12}(I)/I \geq -1$ (because $M_{12}(I) \geq -I$). Hence $M'_{12}(I)$ for large I is an approximation of the direct measure $R_{12}(I)$, the percentage of information gained or lost, depending on when the prior information adds ($R_{12}(k) > 0$) or subtracts ($R_{12}(k) < 0$) due to prior-likelihood conflict, For example $R(100) = -0.3$ means that although the likelihood function was based on 100 data points, the contradictory information from the prior would cost us about 30% of the data, i.e., our inference result will have approximately the same uncertainty as the posterior inference based on 70 data points and the baseline prior.

For small I's, we find both $M'_{12}(I)$ and $R_{12}(I)$ useful because the former can indicate prior-likelihood conflict even if $R_{12}(I) > 0$, such as in the "too-good-to-be-true" case of super-information to be discussed in Section 4·1. We remark, however, that the discrete counterpart of $M'_{12}(I)$, that is, the finite difference $M_{12}(k+1) - M_{12}(k)$ when $I = k$, is *not* asymptotically equivalent to $R_{12}(k)$ unless $R_{12}(k+1) = [1 + o_p(k^{-1})]R_{12}(k)$. We thus avoid using the finite difference to estimate $\lim_{k\to\infty} R_{12}(k)$, and instead use a regression estimator as in Section 3.

The use of $M_{12}(I)$ to measure I_{prior}, the prior information, leads to a natural information additivity. That is, if we view $I + M_{12}(I)$ as the information measure for the posterior, $I_{posterior}$, then trivially we have

$$I_{\text{posterior}} = I_{\text{likelihood}} + I_{\text{prior}} \tag{3}$$

because the information in the likelihood, $I_{\text{likelihood}}$, is I in our setup. The practical appeal of (3) cannot be overstated. This is, however, a non-standard information decomposition because $I_{\text{prior}} = M_{12}$ can be negative, with negative values pointing to a prior-likelihood conflict. When our likelihood can be trusted, such a negative value reflects a misspecification of the prior. It is also important to emphasize that because we measure information relative to a baseline prior π_1, the prior-likelihood conflict we can detect should be interpreted as the *extra* conflict caused by π_2 in excess of the conflict already existing in π_1.

Before dropping the subscript in $M_{12}(I)$, we mention an obvious link between $M_{12}(k)$ and its "transpose" $M_{21}(k)$:

$$U_{\pi_2,\theta_0}(I) = U_{\pi_1,\theta_0}\left(I + M_{12}(I)\right) = U_{\pi_2,\theta_0}(I + M_{12}(I) + M_{21}(I + M_{12}(I))). \tag{4}$$

Assuming $U_{\pi_2,\theta_0}(I)$ is strictly monotone and all solutions exist, we then arrive at

$$M_{12}(I) = -M_{21}\left(I + M_{12}(I)\right).$$

This is essentially an information preservation identity. It says that if it takes an $M_{12}(I)$ amount of information moving from π_1 to π_2 with I amount of likelihood information to reach $I + M_{12}(I)$, then it will take the same amount back—hence the negative sign—from π_2 to π_1 when the likelihood information is already at $I + M_{12}(I)$.

Furthermore, $M_{ij}(I)$ preserves additivity with multiple priors in the following sense:

$$M_{13}(I) = M_{12}\left(I + M_{23}(I)\right) + M_{23}(I), \tag{5}$$

which is again a consequence of the definition of M, with the additional assumption that $U_{\pi_3,\theta_0}(I)$ is strictly monotone. Here π_3 is a third prior, and identity (5) is the generalization of the intuitive case, $M_{13} = M_{12} + M_{23,}$ which holds when there is no conflict between the likelihood and any of the three priors, and hence all $M_{ij}(I)$ are (approximately) free of I.

3. A Specific Diagnostic Procedure

3·1. *Implementation*

Here we present a specific implementation of the general strategy when our data consist of independent and identically distributed observations, $X_1, \ldots, X_n$, and hence $I = n$, the sample size. As before, we assume that X_i is distributed according to $f(x|\theta)$, with $\pi(\theta)$ being the prior distribution for θ. Under this setup, our general strategy can be realized in the following way:

1. Choose a baseline prior $\pi_b(\theta)$, henceforth called "the baseline", that would be used if no real prior information is available. For example, π_b could be chosen as an "objective" or reference

prior; see Kass & 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 $D(\cdot)$ and then construct an estimator of

$$U_{\pi,\theta_0}(k) = \mathrm{E}[D(\pi(\theta|X_1,\ldots,X_k))|\theta=\theta_0],$$

for $k=1,\ldots,K$, where we choose $K=O(n^{1/2})$ for reasons presented later. In our setup, we can let $\boldsymbol{\omega}_k$ be the $\binom{n}{k}\times k$ matrix enumerating all possible $\binom{n}{k}$ subsamples of $\{1,\ldots,n\}$ of size k. Compute

$$\hat{U}_{\pi,\theta_0}(k) = \frac{1}{\binom{n}{k}}\sum_{j=1}^{\binom{n}{k}} D[\pi(\theta|X_{\boldsymbol{\omega}_k(j,1)},\ldots,X_{\boldsymbol{\omega}_k(j,k)})]. \tag{6}$$

Practically, it is often unnecessary to enumerate completely; a sub-sampling scheme with or without replacement will suffice. That is, we can use a bootstrap estimator.

3. Compute $\hat{U}_{\pi_b,\theta_0}(k)$ analogously to $\hat{U}_{\pi,\theta_0}(k)$, with the baseline π_b in place of the prior π.
4. Plot both $\hat{U}_{\pi,\theta_0}(k)$ and $\hat{U}_{\pi_b,\theta_0}(k)$ against k. If everything behaves properly, there should be a few noticeable characteristics, at least when k is not too small:
   a. The curve for $\hat{U}_{\pi,\theta_0}(k)$ should be lower than that for $\hat{U}_{\pi_b,\theta_0}(k)$. Otherwise the less informative prior outperforms the more informative one, indicating a prior misspecification.
   b. Both curves should decrease monotonically with respect to k. A violation of this monotonicity may indicate a prior-likelihood conflict, which should not happen for the baseline if it is chosen as advertised. In essence, the prior and likelihood would be pointing in such different directions that our uncertainty could actually increase.
5. Next we use $\hat{U}_{\pi,\theta_0}(k)$ and $\hat{U}_{\pi_b,\theta_0}(k)$ to compute the *effective sample sizes* captured by π. We first interpolate the $\hat{U}$ functions so they live on the real line. We use linear interpolation for simplicity, but one can investigate more sophisticated methods. We then define

$$\hat{M}(k) = \arg\min\{m\in\mathbb{R} : \hat{U}_{\pi,\theta_0}(k) = \hat{U}_{\pi_b,\theta_0}(m+k)\}.$$

Note however that in order for $\hat{M}(k)$ to exist, we need to avoid (at least) $\hat{U}_{\pi,\theta_0}(0) < \hat{U}_{\pi_b,\theta_0}(K)$, that is, the prior information in π is so strong that the information contained in the entire likelihood with all K observations plus the baseline prior information still cannot match it. Whereas as a numerical procedure we can try k (and hence K) as large as n, the very fact that we 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 comes from our prior π.

6. Plot the sequence $\hat{M}(k)$ and $\hat{R}(k)=\hat{M}(k)/k$ against k, for $k=1,\ldots,K$, and regress $\hat{M}(k)$ on k for $k=k_0,\ldots,K$ for some suitably chosen k_0 to estimate an *approximate limiting slope* of $\hat{M}(k)$ as a function of k, denoted by S_K. Based on our current theoretical and empirical evidence, we observe the following:
   - when there is no noticeable prior-likelihood conflict, $\hat{M}(k)$ will stay fairly constant, and hence $S_K\approx 0$, and $\hat{R}(k)$ will approach zero rather rapidly as k increases;
   - any serious departure of $\hat{M}(k)$ from being a constant function, especially as a monotone decreasing function, indicates a prior-likelihood conflict;
   - both $R(k)$ and S_K serve as measures of the degree of conflict, where $R(k)$ measures the loss (or gain) due to the prior-likelihood conflict at a finite $k\le K$, and S_K serves an estimator of $R(n)$ for $n >> K$;

- very serious prior-likelihood conflict will cause $\hat{R}(k)$ or S_K to approach -1, that is, in the most extreme cases, the conflict would essentially wipe out all the information in the likelihood function.

We remark here that we use S_K to estimate $R(n)$ instead of $\hat{R}(K)$ because K needs to be chosen such that $n/K = O(n^{1/2}) \to \infty$ and hence $\hat{R}(K)$ is often too far from $R(n)$. However, as long as we are able to choose k_0 such that $\hat{M}(k)$ for $k \geq k_0$ is reasonably linear in k, we can approximate $R(n)$ by the slope of $M(k)$, which can then be estimated via the least-squared estimator from regressing $\hat{M}(k)$ on k for $k \geq k_0$. Furthermore, as we shall demonstrate in Section 4, the signs of S_K and $R(K)$ can be different, with S_K tending to reveal the conflict earlier than any $\{R(k), k = 1, \ldots, K\}$ could. We believe this is largely due to the global nature of S_K as well as the fact that even in the presence of a prior-likelihood conflict, the prior can still help to gain information before the likelihood becomes overwhelming. Indeed, it is also possible to have $S_k < 0$ but $R(k) > 0$ for all $k > k_0$, as in the case of "too-good-to-be-true" super-information phenomenon revealed in Section 4. In other words, S_K does not discriminate between "bad conflict" and "good conflict", though the latter is much less likely in real applications.

3·2. *Choosing D*

The choice of D should reflect aspects of the posterior that are most significant to our study, as well as the need for measuring prior *informativeness* and for assessing prior–likelihood *conflict*. In some settings a simple measure such as variance may be sensitive to both. However, even in cases such as the normal data, normal prior, and known variance example, as seen in Section 4·1, variance alone has zero power for detecting prior–likelihood conflict. In such settings the inclusion of a measure of bias is necessary, resulting in a mean squared error measure.

In our theory given in Section 5, we will use a form of the mean squared error:

$$D(\pi(\theta|\vec{X}_k)) = \mathrm{Var}_{\pi}(\theta|\vec{X}_k) + [\mathrm{E}_{\pi}(\theta|\vec{X}_k) - \theta_0]^2, \tag{7}$$

where, for notation simplicity, we assume θ is univariate and denote $\vec{X}_k$ for $\{X_1, \ldots, X_k\}$ for all k. Obviously θ_0 is unknown, so in applications we will estimate it by $\hat{\theta}_0 = \mathrm{E}_{\pi_b}[\theta|X_1, \ldots, X_n]$, the posterior mean under the baseline prior and based on all data. Consequently, we replace (7) by its estimator

$$\hat{D}(\pi(\theta|\vec{X}_k)) = \mathrm{Var}_{\pi}(\theta|\vec{X}_k) + [\mathrm{E}_{\pi}(\theta|\vec{X}_k) - \hat{\theta}_0]^2, \tag{8}$$

which we will use for all the development and examples below.

For additional ideas on choosing D, we reference Morita et al. (2008) for a measure based on curvature of the log likelihood and Gelman et al. (2013) for measures based on deviances. We also emphasize that, as long as it is computationally feasible, there is nothing stopping one from applying our method for multiple Ds. Indeed, it is mathematically impossible to have a single measure assess the impact of a prior on all aspects of our posterior inference, and it is inferentially desirable to assess the impact of prior with a variety of choices of D. In particular, for a multi-dimensional posterior, we can choose the same D or different Ds for different margins, or a D which examines the parameters jointly.

4. Theoretical and Empirical Illustrations

4·1. *Normal with Known Variance*

The normal mean problem, permitting explicit expressions, illustrates nicely our method and the theoretical results in Section 5. Assume that $\{X_1, \ldots, X_n\}$ is a simple random sample from $N(\mu, \sigma^2)$, with μ our estimand θ, and σ^2 known. We adopt the usual conjugate prior on μ, $N(\mu_\pi, \sigma_\pi^2)$, and for the baseline the usual constant prior, which can also be viewed as setting $\sigma_\pi^2 = \infty$. Let $\bar{X}_k$ be the sample mean, $\gamma = \sigma^2/\sigma_\pi^2$ be the variance ratio, and $A_\gamma(k) = (\gamma\ + k)^{-1}$ be the usual shrinkage factor when our data are $\vec{X}_k$ $(k \le n)$. Then it is well-known that

$$\mathrm{E}_\pi[\mu|\vec{X}_k] = A_\gamma(k)\left(\gamma\mu_\pi + k\bar{X}_k\right), \qquad \mathrm{Var}_\pi[\mu|\vec{X}_k] = A_\gamma(k)\sigma^2, \tag{9}$$

$$\mathrm{E}_{\pi_b}[\mu|\vec{X}_k] = \bar{X}_k, \qquad \mathrm{Var}_{\pi_b}[\mu|\vec{X}_k] = \frac{\sigma^2}{k}. \tag{10}$$

Following (6) and (8), under our conjugate prior π, we have

$$\hat{U}_{\pi,\theta_0}(k) = A_\gamma(k)\sigma^2 + \frac{1}{\binom{n}{k}}\sum_{j=1}^{\binom{n}{k}}\left[A_\gamma(k)\left(\gamma\mu_\pi + k\bar{X}_{\boldsymbol{\omega}_k(j)}\right) - \bar{X}_n\right]^2, \tag{11}$$

where $\bar{X}_{\boldsymbol{\omega}_k(j)}$ is the sample mean of $X_{\boldsymbol{\omega}_k(j,1)}, \ldots, X_{\boldsymbol{\omega}_k(j,k)}$. The corresponding expression for $U_{\pi_b,\theta_0}(k)$ is obtained by simply setting $\gamma = 0$ in (11), which renders $A_\gamma(k) = k^{-1}$, and hence

$$\hat{U}_{\pi_b,\theta_0}(k) = \frac{\sigma^2}{k} + \frac{1}{\binom{n}{k}}\sum_{j=1}^{\binom{n}{k}}(\bar{X}_{\boldsymbol{\omega}_k(j)} - \bar{X}_n)^2. \tag{12}$$

To gain theoretical insights, we look at asymptotic cases with large n but with $k \ll n:$

$$\begin{aligned}\hat{U}_{\pi,\theta_0}(k) \approx & A_\gamma(k)\sigma^2 + \mathrm{E}\left[A_\gamma(k)(\gamma\mu_\pi\ + k\bar{X}_k) - \mu_0\right]^2 \\ = & \sigma^2\left[A_\gamma(k) + \gamma\Delta^2\ A_\gamma^2(k) + kA_\gamma^2(k)\right], \qquad (13) \\ \hat{U}_{\pi_b,\theta_0}(k) \approx & k^{-1}\sigma^2 + \mathrm{E}[\bar{X}_k - \mu_0]^2 = 2k^{-1}\sigma^2 \qquad (14)\end{aligned}$$

where μ_0 is the true data generating parameter, and $\Delta = (\mu_\pi - \mu_0)/\sigma_\pi$ is a direct measure of the misspecification of our conjugate prior. Consequently, we can solve explicitly

$$\hat{U}_{\pi,\theta_0}(k) = \hat{U}_{\pi_b,\theta_0}(k + M(k)) \tag{15}$$

for $M(k)$ when we use the limiting expressions given respectively by the most right-hand sides of (13) and (14). We note again that (14) corresponds to (13) with $\gamma = 0$, and hence $A_\gamma(k) = k^{-1}$.

Combining (13)-(15), we can easily arrive at

$$M(k) = \frac{2}{A_\gamma(k) + \gamma\Delta^2 A_\gamma^2(k) + kA_\gamma^2(k)} - k = k\left\{\frac{1}{r[1 + (1-r)(\Delta^2 - 1)/2]} - 1\right\}, \tag{16}$$

where $r = k/(\gamma + k)$. We see immediately that when $\Delta \to \infty$, which indicates extreme prior-likelihood conflict, $M(k)$ approaches $-k$, or equivalently, $M(k)/k$ approaches -1; in Section 5, we will show that this is a general phenomenon.

At the other extreme, if our prior is the posterior from a previous study based on a simple random sample $\vec{Y}_m = \{Y_1, \ldots, Y_m\}$ from the same $N(\mu_0, \sigma^2)$, and we use the same baseline constant prior, then $\mu_\pi = \bar{Y}_m$ and $\sigma_\pi^2 = \sigma^2/m$, and hence $\gamma = m$. Consequently, $\Delta^2 = m(\bar{Y}_m - \mu_0)^2/\sigma^2$ is on average equal to 1 (with respect to the previous data). If we indeed replace Δ^2 by 1, then it is easy to verify that $M(k) = m$, as desired.

We observe that the “other extreme” did not occur at $\Delta = 0$, which seems to eliminate any prior-likelihood conflict. However, we must keep in mind that setting μ_π equal to the true parameter injects more prior information than a typical previous real-data study can offer. This difference is precisely captured by $(\bar{Y}_m - \mu_0)^2/\sigma^2$, which is of order m^{-1}, and hence cannot be ignored. Indeed, if we do set $\Delta = 0$ in (16) but still keep $\gamma = m$, we will arrive at

$$M(k) = \left(\frac{3}{2} + \frac{m}{2(m+2k)}\right) m = \left(2 - \frac{k}{m+2k}\right) m. \tag{17}$$

Because the first factor on the right hand side of (17) is always strictly within the interval $(1.5, 2)$, we know that setting the prior mean to the true mean always increases the prior data/information size m by at least 50%, and often close to 100% when m is large, relative to k.

From (17), we see $M(k)$ is a strictly monotone decreasing function of k, and hence its derivative–treating k as continuous–will be negative. The criterion of negative slope then would suggest that setting the prior mean to be the truth also creates a prior-likelihood conflict. However, this is not illogical because as far as the likelihood function is concerned, the prior is “too good to be true”. In Section 5 we will show that this “too good to be true” or *super-informative* phenomenon is rather general and so is the expression (17).

4·2. *A Simulation Study*

Here we provide numerical illustrations for the normal setting above, as well as for a similar setting but with the exponential distribution. In both settings we explore various parameter values that reflect varying degrees of prior misspecification. The estimated measure of uncertainty, D, we use in all simulations is (8). Throughout, we use a resampling scheme where we take 100,000 sub-samples without replacement, to construct the estimates $\hat{U}_{\pi,\theta}$ and $\hat{U}_{\pi_b,\theta_0}$. In all of our settings the $\hat{U}$ plots were simply monotonically decreasing functions, thus we omit them for brevity.

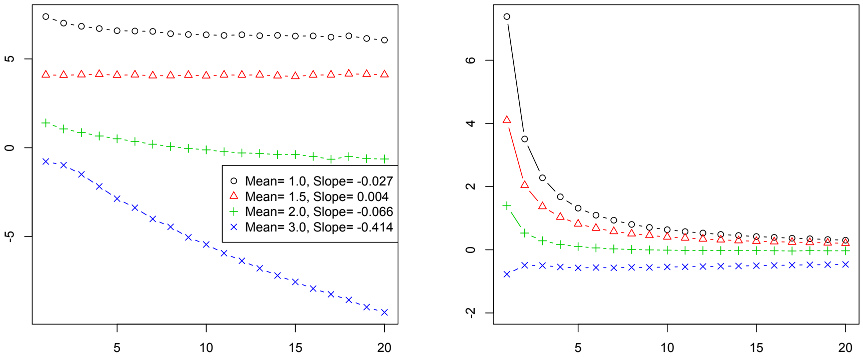

Fig. 1. **Normal with Known Variance** – Plots of the estimated prior sample size $\hat{M}(k)$ (left) and the relative prior sample size $\hat{R}(k) = \hat{M}(k)/k$ (right) versus the likelihood data size k; the four curves correspond to the four scenarios given in Table 1.

Normal with Known Variance

We simulate our procedure in the setting given in Section 4·1, with $n = 1000$ and under the parameter scenarios given in Table 1. The true mean is $\mu_0 = 1$ and the variance is $\sigma^2 = 1$. The baseline is taken to be constant on the real line. The resulting $\hat{M}(k)$ and $\hat{R}(k)$ plots are given in Figure 1. The estimated slopes for $\hat{M}(k)$ in Table 1 are from least squares by using $k_0 = 6$, because the plots reveal that $\hat{M}(k)$ is closer to being linear once we ignore the segments with $k \leq 5$. We also give $\hat{R}(k)$ at the largest k used, $K = 20$. From the right panel in Figure 1, it is clear that $\hat{R}(k)$ will have an asymptote, and as we discussed in Section 2, its limiting value will be the same as $\lim_{k\to\infty} \hat{M}(k)$. However, Table 1 shows that $K = 20$ is too small for the asymptotic equivalence to kick in (which would also require both k_0 and K to approach ∞). In such cases, our current belief is that our slope estimator is a better estimator for $R(n)$, the actual relative gain or loss of information due to our prior for the entire data set at hand.

μ_π	σ_π^2	Slope: S_{20}	$\hat{R}(20)$
1.0	0.25	-0.027	0.303
1.5	0.25	0.004	0.206
2.0	0.25	-0.066	-0.032
3.0	0.25	-0.414	-0.463

Table 1. *Prior parameter values for Figure 1, and the estimated slope S_{20} and $\hat{R}(20)$.*

As we can see from the plots, the procedure behaves as expected from the asymptotic calculations given in the previous section. The case of $\mu_\pi = 1 = \mu_0$ represents the super-informative case, with the prior worth approximately 6 data points, 50% more than the nominal prior size $\gamma = \sigma^2/\sigma_\pi^2 = 4$. When $\mu_\pi = 1.5$, $\Delta^2 = 1$ using the notation in Section 4·1, and hence $\hat{M}(k)$ recovers the nominal size $\gamma = 4$ for all k, as the line of triangles demonstrates. When $\mu_\pi = 2$, the values of $\hat{M}(k)$ become negative, though only slightly, indicating a mild prior-likelihood conflict. This is not surprising in view of the fact that the ideal case recovering the prior size being exactly 4, is given by $\mu_\pi = 1.5$, which is only 0.5 units away from $\mu_\pi = 2$.

However, when $\mu_\pi = 3$, there is a rather serious prior-likelihood conflict, both visually and as measured by S_{20} or by $\hat{R}(20)$, which amounts to subtracting almost half of a sample per data point added. That is, for a likelihood based on 20 data points, the misspecified $N(3, 0.25)$ prior would cost about 9 data points worth of information. Concretely, the Bayesian estimator—posterior mean or mode—for μ will have a mean squared error that is about double that of the maximum likelihood estimator, the sample mean. In practice of course we are unlikely to be able to attribute the conflict solely to the misspecification of the prior. But being able to detect such conflicts can help us to re-examine our assumptions, conduct more model checking, contemplate using multiple estimates (e.g., considering both the maximum likelihood estimator and Bayesian estimator), etc. Minimally it would help to prevent us from blindly trusting our posterior inference and letting misspecifications manifest into potentially consequential damages.

Exponential under Two Parameterizations

We now assume that $X_1, \ldots, X_n$ are exponential random variables with mean $\mu = \lambda^{-1}$ and variance $\mu^2 = \lambda^{-2}$. By comparing the case of $\theta = \mu$ with $\theta = \lambda$, we reveal the rather sensitive nature of the prior-likelihood conflict to parameterizations when our Bayesian estimators or the uncertainty measure are not invariant to parameterization. This is the case when we use the common posterior mean as our estimator, or mean squared error as our D. We emphasize that such sensitivities are *not* due to artifacts of our procedure, but rather because it honestly captures how a chosen uncertainty measure of a chosen estimator behaves as a functional of the likelihood, the prior, and their interplay.

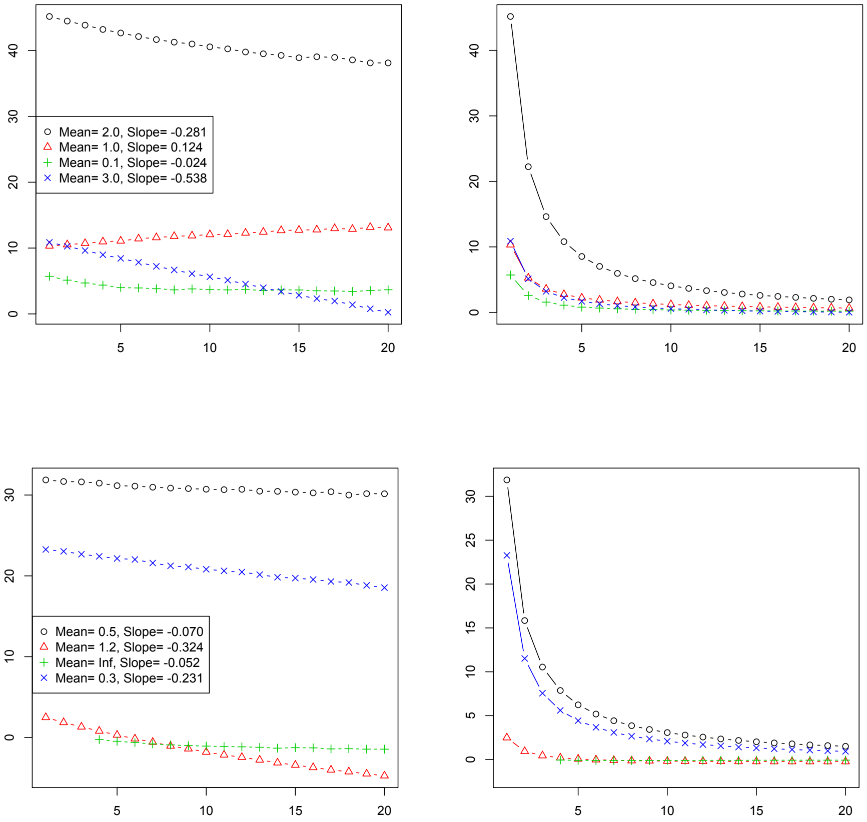

Fig. 2. **Exponential** – Plots of the estimated prior sample size $\hat{M}(k)$ (left) and the relative prior sample size $\hat{R}(k) = \hat{M}(k)/k$ (right) versus the likelihood data size k; the four curves correspond to the four scenarios given respectively in the upper and bottom panels in Table 2, where the top panel/plots refer to the prior placed on the rate λ while the bottom ones refer to the mean $\mu = \lambda^{-1}$.

As is well-known, the conjugate prior on λ is gamma, $\Gamma(\alpha, \beta)$, with $\mathrm{E}(\lambda) = \alpha/\beta$, and $\mathrm{Var}(\lambda) = \mathrm{E}^2(\lambda)/m$, where $m = \alpha$ can be viewed as the nominal prior size. The conjugate prior on μ is then the inverse gamma $\Gamma^{-1}(\alpha, \beta)$ where $\mathrm{E}(\mu) = \beta(\alpha - 1)^{-1}$ and $\mathrm{Var}(\mu) = (\alpha - 2)^{-1}\,\mathrm{E}^2(\mu)$. Notice that $\mathrm{E}(\mu) = \infty$ when $\alpha \leq 1$ and $\mathrm{Var}(\mu) = \infty$ when $\alpha \leq 2$. Our baseline is given by taking $(\alpha, \beta) \to 0$, which is equivalent to taking $\pi_b(\theta) \sim \theta^{-1}$, regardless of whether $\theta = \lambda$ or $\theta = \mu$. The posterior of λ under the $\Gamma(\alpha, \beta)$ prior is also a gamma distribution, with mean and variance given by

$$\mathrm{E}_\pi[\lambda|\vec{X}_k] = \frac{\alpha + k}{\beta + k\bar{X}_k}, \qquad \mathrm{Var}_\pi[\lambda|\vec{X}_k] = \frac{\alpha + k}{(\beta + k\bar{X}_k)^2}. \tag{18}$$

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

$$\mathrm{E}_\pi[\mu|\vec{X}_k] = \frac{\beta + k\bar{X}_k}{\alpha + k - 1}, \qquad \mathrm{Var}_\pi[\mu|\vec{X}_k] = \frac{(\beta + k\bar{X}_k)^2}{(\alpha + k - 1)^2(\alpha + k - 2)}. \tag{19}$$

We apply our procedure to simulated data with $n = 1000$ and under the parameter scenarios given in Table 2 for the rate λ and the mean μ. We attempt to control prior misspecification by fixing the prior variance and adjusting the prior mean (for λ) and then matching the scenarios for μ to the ones for λ. The true parameter value is $\lambda = 2$ or $\mu = 1/2$. The resulting plots are given in Figure 2, and the estimated slope and $\hat{R}(20)$ are in Table 2. Again we see the asymptote tendencies in the right panels of Figure 2, yet the K, the largest value of k we used, is still too small to deliver reliable numerical limiting value, as seen from the rather large discrepancies between the columns of $\hat{S}_{20}$ and $\hat{R}(20)$ revealed in Table 2. As discussed before, we will trust more the slope estimator than $\hat{R}(20)$ as an approximation to $R(n)$.

α	β	$\mathrm{E}(\lambda\vert\alpha,\beta)$	$\mathrm{Var}(\lambda\vert\alpha,\beta)$	S_{20}: Slope	$\hat{R}(20)$
20.0	10.0	2.0	0.2	-0.281	1.906
5.0	5.0	1.0	0.2	0.124	0.655
0.1	0.5	0.1	0.2	-0.024	0.184
45.0	15.0	3.0	0.2	-0.538	0.012

α	β	$\mathrm{E}(\mu\vert\alpha,\beta)$	$\mathrm{Var}(\mu\vert\alpha,\beta)$	S_{20}: Slope	$\hat{R}(20)$
20.0	10.0	0.526	0.015	-0.070	1.508
5.0	5.0	1.250	0.521	-0.324	-0.236
0.1	0.5	Inf	Inf	-0.052	-0.072
45.0	15.0	0.341	0.003	-0.231	0.927

Table 2. *Prior parameter values for Figure 2, and the estimated slopes for $M(k)$ and $\hat{R}(20)$. The prior placed on λ is $\Gamma(\alpha,\beta)$, and on $\mu = \lambda^{-1}$ is $\Gamma^{-1}(\alpha,\beta)$.*

In this setting we see some characteristics resembling that of the normal setting, as well as some new phenomena. In the first scenario ($\alpha = 20, \beta = 10$), the prior mean is set to be the truth, the prior sample size is around 40 for the gamma and 30 for the inverse gamma, which is 100% and 50% higher, respectively, than $\alpha = 20$, the nominal prior sample size in this setting. However, there is a larger negative slope associated with gamma setting which means that, for larger k, the two settings will be closer. This demonstrates that the super-informative phenomenon is not an artifact of the normal setting, and indeed we will verify this theoretically in Section 5. However, the approximations on which we base the theory are obviously sensitive to the chosen parameterization (at least for small m/k values).

The second and fourth scenarios tell an interesting and important story concerning the impact of parameterization on inference. We see that the second scenario ($\alpha = 5, \beta = 5$) shows a positive slope for gamma, but a negative one for the inverse gamma, indicating stronger prior-likelihood conflict. The fourth scenario ($\alpha = 45, \beta = 15$) appears to be the opposite, being mildly misspecified for the inverse gamma but strongly so for the gamma.

In the third scenario ($\alpha = 0.1, \beta = 0.5$) , one might expect to see a prior-likelihood conflict for the gamma setting, given how far the prior mean is from the truth. But instead we see that the prior size is essentially constant around 3-4, higher than $\alpha = 0.1$, indicating essentially no prior-likelihood conflict, but rather a super-information phenomenon. Turning to the inverse gamma, we also see little conflict and that $\hat{M}$ is very close to zero. This apparent contradiction can be explained by examining the mean and variance of the prior and the baseline. As mentioned earlier, the baseline is obtained by taking α, β to zero, which corresponds to the Jeffreys prior. However, in terms of the mean and variance (of the gamma setting), the baseline can be approached in two seemingly different ways: keep the mean constant and send the variance to infinity, or keep

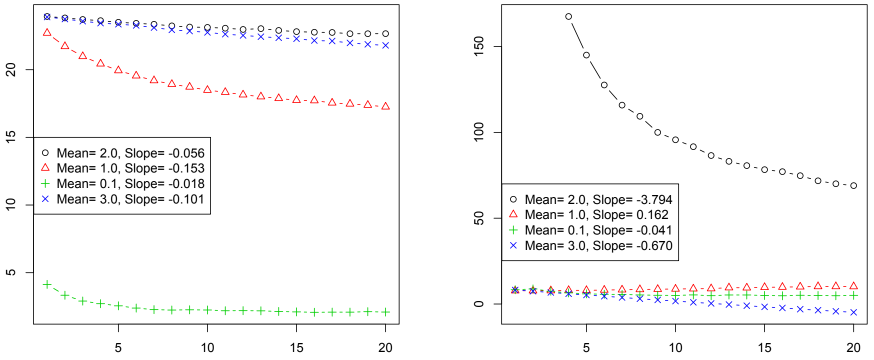

Fig. 3. **Exponential** – Plots of estimated prior size $\hat{M}(k)$ when our uncertainty measure D is taken to be the variance (left) and bias (right) only, under the four scenarios given in Table 2 for the rate (λ) only.

the variance constant and send the mean to zero. In both cases α and β will tend to zero and the prior will approach the baseline. Thus, the prior in the third scenario is actually closer to the baseline than one might initially expect. As we emphasized before, our approach does not detect conflict already existing in the baseline, but only the extra conflict induced by our prior. When we fix the prior variance but send the prior mean to the other direction, that is, making it large, as in the fourth scenario, we see a much clearer prior-likelihood conflict, subtracting about 60% information per datum for the gamma case and about 20% for the inverse gamma case.

Lastly we explore the seemingly peculiar behavior of the second scenario with the gamma prior on the rate parameter, where the slope of $\hat{M}(k)$ is slightly positive. Such a result seems counterintuitive and is not seen in the normal or inverse gamma settings. To better understand this behavior we recompute $\hat{M}$, but for the variance and bias separately. That is, in one case we use a D function which consists only of the variance term in (8), and in another only of the squared bias term. This is done to examine which part of the mean squared error is responsible for the monotone increasing behavior, not to suggest that the $\hat{M}(k)$ function can be decomposed into two parts corresponding respectively to variance and bias. The results are given in Figure 3. There we see that the variance plot behaves as expected, flat or slightly decreasing, but the bias plot exhibits a positive slope for the second scenario.

As an uncertainty measure, variance works well for determining prior impact, but poorly for detecting prior-likelihood conflict. The bias has the opposite problem, great for conflict detection, but poor for impact determination. The super information phenomenon is especially prominent with the bias measure, because it will take a large sample to bring down the bias in the posterior mean under the baseline prior to the level that is enjoyed by the posterior mean derived from a prior whose mean is set at the true value of the estimand. Furthermore, with no data the prior alone cannot provide evidence for bias, so adding data can only make things worse, hence the large negative slope for the bias in the first scenario. Regarding the positive slope phenomenon in the second scenario, we suspect it is possible because the bias term is a *nonlinear* function of the sample mean, in contrast to the normal and inverse gamma setting, where the bias is a linear function of the mean. Taking k large enough will alleviate this behavior as the curve will level off, but the positive slope serves as a good reminder of the complex and erratic nonlinear behavior with small k.

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 & 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 and G4, which are IgG3 and IgG4 respectively. 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.

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 informative such a prior really is and whether there is any misspecification using such a prior in the present setting. We compare a Cauchy prior with scales 2.5, 5, and 10 against a Cauchy with scale 10000, which in essence "flattens" out the prior. The results are insensitive to the choice of scale for the baseline as we see the same patterns when the baseline scale parameter is 1000. We use the metric (6) to compare the two priors, however instead of taking the mean of this metric over subsamples we take the median. This helps with the stability of the procedure over smaller subsamples which can be a significant problem in a logistic regression, but it poses an open theoretical question on how sensitive our information assessment is to the choice of estimator for the uncertainty measure. Furthermore, we examine $\hat{M}(k)$ only for $k > 10$, because logistic regressions are notoriously unstable for small sample sizes. To reduce the impact of the non-linear part of $\hat{M}(k)$ on the estimation of $R(55)$, we can further take $k_0 = 20$, that is, to estimate $R(55)$ 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 smaller sample size: 55 versus 1000. The prior suggested by Gelman et al. (2008), that is, with scale=2.5, seems to indeed depict a *weakly informative prior*, as it does not add a substantial amount of information, but only about 2-6 data points, that is, no more than 10% of the information provided by the likelihood function. There might be some small amount of prior-likelihood conflict. By taking the scale up to 5 or 10, the conflict is reduced, 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 weekly informative prior, see Gelman (2006) and Polson & Scott (2012).

5. Theoretical Results

In this section, we provide theoretical justifications for the method presented in Section 3. These results are by no means exhaustive, but they are applicable to essentially all posterior-prior families that possess conjugacy, exactly or asymptotically. This (approximate) conjugacy permits us to index prior information via the intuitive notion of "prior data size", by equating our prior to the outcome from a previous study of similar nature to the data forming the likelihood. More importantly, the similarity, or rather the lack of it, allows us to model the prior-likelihood conflict. Specifically, the following assumption plays a critical role in our theoretical formulation.

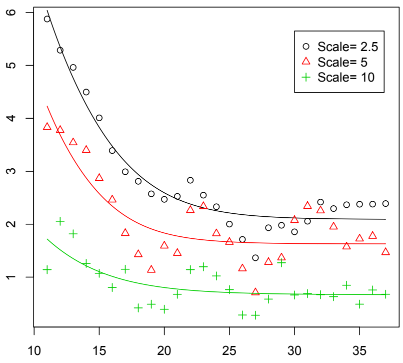

Fig. 4. Plot of the estimated prior size $\hat{M}(k)$ for the application in Section 4·3. Estimated slopes are -0.1173, -0.0646, and -0.0283 using $k > 10$, for scales $2.5, 5$, and 10 respectively. The estimated slopes become 0.0032, -0.0024, and -0.0079 respectively when using $k > 20$.

We remark that for simplicity of both derivation and presentation, we will restrict the parameter θ to be univariate, but the results hold for multivariate θ with necessary extensions of notation (e.g., replace σ^2 by Σ, and absolute value by Euclidean norm).

Assumption 1. *Assume that $X_1, X_2, \ldots, X_n$ are independent and identically distributed according to density $f(x|\theta)$ with respect to a measure on $\mathbb{R}$, where $\theta \in \mathbb{R}$. Assume that our prior $\pi(\theta)$ is such that there exists $m > 0$ and $\mu_m \in \mathbb{R}$ such that for any $\vec{X}_k = \{X_1, \ldots, X_k\}$, we have the following expansions for the corresponding posterior mean and variance:*

$$\mathrm{E}_\pi[\theta|\vec{X}_k] = u(T_{k,m}) + O_p(l^{-1}) \quad \textit{and} \quad \mathrm{Var}_\pi[\theta|\vec{X}_k] = \frac{v(T_{k,m})}{l} + O_p(l^{-2}), \tag{20}$$

where u is a twice differentiable function and $v > 0$ is a differentiable function on $\Omega_u \equiv \{a \in \mathbb{R} : |u(a)| < \infty\}$, $l = k + m$,

$$T_{k,m} = \frac{m\mu_m + k\bar{T}_k}{m + k}, \tag{21}$$

and $\bar{T}_k$ is the average of some $T_i = T(X_i)$ over $i = 1, \ldots, k$, 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,

$$\mathrm{E}_{\pi_b}[\theta|\vec{X}_k] = u(\bar{T}_k) + O_p(k^{-1}) \quad \textit{and} \quad \mathrm{Var}_{\pi_b}[\theta|\vec{X}_k] = \frac{v(\bar{T}_k)}{k} + O_p(k^{-2}). \tag{22}$$

Assumption 1 is satisfied by many common conjugate prior distributions including the six natural exponential families with quadratic variance functions (Morris, 1982); see the Appendix. Perhaps the easiest way to gain insight is to consider the normal case in Section 4·1, by which it is particularly easy to understand the $T_{k.m}$ expression in (21). This is because the normal case should remind us of expressing the posterior mean as a weighted average of the sample mean and the prior mean, with weights proportional to their respective precisions. In particular, by comparing (9) to (20), we see that $T(x) = x$ and hence $\mu_T = \mu$ and $\sigma_T^2 = \sigma^2$; and $u(t) = t, v(t) = \sigma^2$, $m = \sigma^2/\sigma_\pi^2$, $\mu_m = \mu_\pi$, and without the high-order error terms in (20). That is, m

here is the total prior precision, $1/\sigma_\pi^2$, where σ_π^2 is the prior variance, relative to the data precision *per sample* (in terms of $T_i = X_i$): $1/\sigma_T^2$.

This comparison gives us the insight that m can be interpreted in general as the *nominal* "prior data size" measured on the same unit scale as the data for the likelihood, and similarly that μ_m can be viewed in general as the prior mean for μ_T, the mean of T under $f(x|\theta)$. We say m is *nominal* because the real prior data size, as an appropriate indicator (which therefore does not need to be an integer) of the amount of information introduced by the prior, must take into account the potential conflict between μ_T and μ_m, a key issue as we emphasized previously. Furthermore, Assumption 1 does not require the existence of a mean, but only the existence of m and μ_m. For example, in the inverse gamma case in Section 4·2, although $\mathrm{E}(\mu) = \beta/(\alpha-1)$ does not exist when $\alpha \leq 1$, we can still take $m = \alpha$ and $\mu_m = \beta/\alpha$, which satisfies Assumption 1 because of (19). Therefore, in general, it would be more accurate to consider μ_m a measure of prior centrality then necessarily the prior mean.

We use the notation μ_m to denote the prior centrality to remind ourselves that it can, and often does, depend on the nominal prior data size m, even though this issue has been largely overlooked in the literature. This is seen most clearly when our prior information actually comes from a previous study based on a data set $\{\tilde{X}_1, \ldots, \tilde{X}_m\}$, which were also independent and identically distributed according to our model $f(x|\theta)$ but possibly with a different parameter value of θ, say θ_1, from the one generating the current data $\{X_1, \ldots, X_n\}$, say θ_0. Assuming the previous Bayesian analysis used the same baseline prior π_b, we know from (22) that the prior mean for μ_T will be approximately $\tilde{T}_m$, where $\tilde{T}_m$ is the average of $\{T(\tilde{X}_i), i = 1, \ldots, m\}$.

Of course, the fact that $\mu_m \approx \tilde{T}_m$ is needed much more than merely to justify the notation μ_m. It guides us to carry out an asymptotic analysis that can render meaningful statistical insights for our purposes. Specifically, the usual asymptotic regime where the data size k going to infinity but with the prior specifications considered fixed, e.g., the prior data size m as a fixed constant, is inapplicable here because then the impact of the prior is asymptotically negligible, providing no insight whatsoever on any prior-likelihood conflict. Indeed, any asymptotic regime where the prior impact becomes negligible would run into the same problem.

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 $r = m/(k+m)$ 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 $H_0 : \theta_1 = \theta_0$, and its asymptotic null distribution, as $m \to \infty$, is the chi-squared distribution χ_1^2. It is therefore meaningful in our asymptotic regime to consider Δ^2 as fixed while allowing m to grow, especially because Δ^2 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. Perhaps an even clearer justification of Δ is to write $m = \sigma_T^2/\sigma_\pi^2$, as we did before. Then, $\Delta = (\mu_m - \mu_T)/\sigma_\pi$, which can be viewed as the relative difference between the prior mean (or centrality) and the true mean with respect to the prior standard deviation, is clearly a good measure of how our prior specification deviates from the actual data forming our likelihood. Consequently, we make the following assumption for our asymptotic regime.

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

$$\mu_m = \mu_T + \Delta \frac{\sigma_T}{\sqrt{m}} + O_p(m^{-1}) \tag{23}$$

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

As we shall see in the Appendix, the two assumptions above play a critical role in establishing an asymptotic expression for the relative prior size $M(k)/k$. The next assumption is of a technical nature to ensure that our asymptotic expression is unique. This assumption holds trivially in virtually all applications, but nevertheless we need it to eliminate pathological cases where properties hold in probability, as assumed in (20), fail to hold almost surely, as required by Assumption 3, though for practical purposes, this difference is almost immaterial.

Assumption 3. *We assume (i) both $\hat{U}_{\pi,\theta_0}(I)$ and $\hat{U}_{\pi_b,\theta_0}(I)$ converge almost surely to zero as $I \to \infty$, and (ii) for any finite stopping time $\hat{I}$, $\hat{U}_{\pi_b,\theta_0}(\hat{I}) > 0$, almost surely.*

THEOREM 1. *Assume $\hat{D}$ as defined in* (8)*, and that both k and m increase to infinity with n, with the restriction $k = O(n^{1/2})$ and $r = k/(k+m)$ is strictly between zero and one even at its limit. We then have the following results, where $c = [u'(\mu_T)]^2\sigma_T^2/\{[u'(\mu_T)]^2\sigma_T^2 + v(\mu_T)\} \leq 1$.*

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

$$R(k) = R_r(\Delta^2) + O_p(k^{-1/2}), \text{ with } R_r(\Delta^2) = \frac{1}{r[1 + c(1-r)(\Delta^2 - 1)]} - 1, \tag{24}$$

*is an asymptotic solution of (*15*) to the order of $O_p(k^{-1/2})$, that is,*

$$\frac{\hat{U}_{\pi,\theta_0}(k)}{\hat{U}_{\pi_b,\theta_0}(k + M(k))} = 1 + O_p(k^{-1/2}). \tag{25}$$

(B) *Under further Assumption* 3*, (*25*) holds if and only if (*24*) holds.*

Expression (24) makes clear the role Δ^2 plays in determining the limiting behavior of the relative size ratio $R(k) = M(k)/k$. In particular, as in the normal example, when $\Delta^2 = 1$, $R_r(1) = m/k$, implying that $M(k)$ will recover the nominal prior size m asymptotically. When $\Delta^2 \to \infty$, representing 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 $\Delta = 0$, we see that because we can write

$$R_r(\Delta^2) = A_r(\Delta^2)\frac{m}{k}, \quad \text{where} \quad A_r(\Delta^2) = 1 - \frac{c(\Delta^2 - 1)}{1 + c(1-r)(\Delta^2 - 1)}, \tag{26}$$

we have $R_r(0) = (m/k)A_r$, with

$$A_r = 1 + \frac{c}{1 - (1-r)c} \geq 1. \tag{27}$$

Therefore, asymptotically, the actual prior size $M(k)$ is larger than the nominal size m by the factor A_r. This is the same super-information phenomenon we saw in the normal example, where $c = 1/2$, in which case (27) is the same as (17). Intriguingly, the fact $c = 1/2$ holds much broadly than in the normal case, though it is not surprising since normality holds asymptotically under broad regularity conditions.

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, $\mathrm{Var}_{\pi_b}[\theta|\bar{X}_k]$ 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 (22), by the δ-method, that

$$1 = \lim_{k\to\infty} \frac{V\left[\mathrm{E}_\pi(\theta|\vec{X}_k)|\theta\right]}{V_\pi(\theta|\vec{X}_k)} = \lim_{k\to\infty} \frac{[u'(\mu_T)]^2\sigma_T^2/k}{v(\mu_T)/k} = \frac{[u'(\mu_T)]^2\sigma_T^2}{v(\mu_T)}, \tag{28}$$

and hence the c as specified in Theorem 1 is $1/2$. In Appendix, we will see that (28) holds for all examples examined there. Consequently, we see the phenomenon that the super-information is always between 150% and 200% is general because, whenever $c = 1/2$,

$$\frac{3}{2} \leq A_r = 1 + \frac{1}{1+r} \leq 2 \tag{29}$$

and A_r is close to the lower bound 1.5 when r is closer to 1, that is, when the nominal prior size m is small compared to k. This explains the phenomena we observed in our simulation studies, especially the exponential example. This is a rather unexpected finding, especially because of its simple and general nature, precisely quantifying the super-information as an additional $(1+r)^{-1}$ percent of information. The assumption (28) is rather mild because it holds whenever the large-sample variance approximation via the Fisher information is appropriate for both likelihood and Bayesian inferences. As a matter of the fact, for some convolution families under the single observation unbiased prior (Meng & Zaslavsky, 2002), (28) holds exactly for finite samples as well, that is, without the need to take $k \to \infty$.

More generally, we see from (26) that the super-information phenomenon kicks in as soon as $\Delta^2 < 1$, and the amount of increased information is monotone in $|\Delta^2 - 1|$. 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 conflict even if it exits because our method can only detect the additional conflict not already present in the baseline prior, a fact we have demonstrated empirically in Section 4.

6. LIMITATIONS AND FUTURE WORK

The methods we proposed have many limitations and therefore additional work is needed. Maybe 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 obviously also need to establish theoretical results for scenarios that go beyond those covered in Section 5, and more critically to cases where the likelihood itself is misspecified in consequential ways.

In applying our methods, we encountered two practical problems. The first is the computational demand. Our procedure involves computing some posterior quantities many times, and hence the overall computational load depends critically on how the posterior calculations are performed. For example, using conjugate priors, we were able to carry out most of our simulation studies in less than a minute each on a 2.6 GHz Intel i7 laptop. In contrast, for the application in Section 4·3, we used the *bayesglm* package in R, which took closer to an hour. For complex models, the computational load could become impractical with brute force execution. Therefore seeking effective computational strategies is an area of much needed research.

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 L_2 norm. Extending our theoretical results to cover other norms, especially L_1 norm, as well as more general choices of the discrepancy or uncertainty measure D is another direction for future research.

Finally, we can explore other methodological applications using the idea of assessing conflict via monitoring the changes in $M(k)$. For example, we can compare two subjective priors constructed by two different investigators, and determine whether one is in more serious conflict with a likelihood function than the other. Going even further, it is possible to 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 the misspecification of the other, unless we just have very bad luck. Of course, whether we assess prior-likelihood conflict or misspecification of a likelihood, our general message is the same, that is, to be an informed Bayesian, or more generally, an informed statistical analyst.

7. Appendix

7·1. *Verifying Theoretical Assumptions and Results*

This section presents several prior-likelihood examples that satisfy Assumptions 1-3, and verifies the conclusions given in Section 5. All our examples form conjugate prior-likelihood pairs with the following exponential forms: X_i has a density of the form

$$f(x|\theta) = \exp\{T(x)\eta(\theta) + \xi(\theta) + B(x)\}, \tag{30}$$

and the prior is a two-parameter conjugate family

$$g(\theta; a, d) = \exp\left\{ad\eta(\theta) + d\xi(\theta) + \zeta(\theta) + C(a, d)\right\}. \tag{31}$$

As before, letting $\vec{X}_n = \{X_1, \ldots, X_n\}$ denote an independent and identically distributed sample from (30), we then have that the posterior is proportional to

$$\begin{aligned} p(\theta|\vec{X}_n) &\sim \exp\{(ad + n\bar{T})\eta(\theta) + (d+n)\xi(\theta) + \zeta(\theta)\} \\ &= \exp\left\{\left(\frac{ad + n\bar{T}}{d+n}\right)(d+n)\eta(\theta) + (d+n)\xi(\theta) + \zeta(\theta)\right\}, \end{aligned}$$

where $\bar{T} = (1/n)\sum T(X_i)$. Therefore

$$p(\theta|\vec{X}_n) = g\left(\theta; \frac{ad + n\bar{T}}{d+n}, d+n\right).$$

This means (20) and (22) hold with $m = d$ and $\mu_m = a$ if for the $g(\theta; a, d)$ family we have

$$\mathrm{E}(\theta) = u(a) + O(d^{-1}) \quad \text{and} \quad \mathrm{Var}(\theta) = \frac{v(a)}{d} + O(d^{-2}), \tag{32}$$

where $u(a)$ and $v(a)$ satisfy the properties given in Assumption 1. Below we show this is the case for four common applications, where the expressions of posterior means and variances will also make it transparent that Assumptions 3(i) is a consequence of the strong law of large numbers. We therefore only need to verify Assumption 3(ii). Note Assumption 2 is a restriction on the hyper-parameters in our asymptotic regime, and hence it is satisfied whenever we treat the value $\Delta = m(\mu_m - \mu_T)^2/\sigma_T^2$ as fixed as we let m vary.

Exponential

Assume that $X_1, \ldots, X_n$ are exponential random variables, and hence $\mu_X = \lambda^{-1}$ and $\sigma_X^2 = \lambda^{-2}$. The conjugate prior on λ is the gamma distribution with parameters α and β, $\Gamma(\alpha, \beta)$. The baseline is given by taking $(\alpha, \beta) \to 0$, yielding $\pi_b(\lambda) \sim \lambda^{-1}$, the Jeffreys prior. The corresponding posteriors are respectively gamma distributions with

$$\mathrm{E}_{\pi}[\lambda|\vec{X}_n] = \frac{\alpha + n}{\beta + n\bar{X}_n}, \qquad \mathrm{Var}_{\pi}[\lambda|\vec{X}_n] = \frac{\alpha + n}{(\beta + n\bar{X}_n)^2}; \tag{33}$$

$$\mathrm{E}_{\pi_b}[\lambda|\vec{X}_n] = \frac{1}{\bar{X}_n}, \qquad \mathrm{Var}_{\pi_b}[\lambda|\vec{X}_n] = \frac{1}{n\bar{X}_n^2}. \tag{34}$$

It is easy to see from the first expression of (33) that for $\theta = \lambda$, we should take $a = \beta\alpha^{-1}$, $d = \alpha$, and $T = X$. Condition (32) then is satisfied by $u(a) = a^{-1}$ and $v(a) = a^{-2}$ exactly without the O terms because $\Gamma(\alpha, \beta)$ has mean and variance $\alpha\beta^{-1}$ and $\alpha\beta^{-2}$, respectively. Assumption 3(ii) follows trivially from (34) because it shows that $\mathrm{Var}_{\pi_b}[\lambda|\vec{X}_{\hat{I}}] > 0$ for any finite $\hat{I}$. Condition (28) can also be verified directly from $v(\mu_X) = \mu_X^{-2} = \lambda^2$, and $[u'(\mu_X)]^2\sigma_X^2 = [-\mu_X^{-2}]^2\sigma_X^2 = \lambda^2$.

Using the alternative parameterization of the exponential, we can also put a prior directly on μ_X. The conjugate family then becomes the inverse gamma, also with parameters α and β. Denote the baseline parameters as α_b and β_b. While the parameters update in the same way, the mean and variance functions are different:

$$\mathrm{E}_{\pi}[\mu_X|\vec{X}_n] = \frac{\beta + n\bar{X}_n}{\alpha + n - 1}, \qquad \mathrm{Var}_{\pi}[\mu_X|\vec{X}_n] = \frac{(\beta + n\bar{X}_n)^2}{(\alpha + n - 1)^2(\alpha + n - 2)}; \tag{35}$$

$$\mathrm{E}_{\pi_b}[\mu_X|\vec{X}_n] = \frac{\beta_b + n\bar{X}_n}{\alpha_b + n - 1}, \qquad \mathrm{Var}_{\pi_b}[\mu_X|\vec{X}_n] = \frac{(\beta_b + n\bar{X}_n)^2}{(\alpha_b + n - 1)^2(\alpha_b + n - 2)}. \tag{36}$$

We are now left with a more interesting choice for the baseline than in other examples. Taking $\beta_b \to 0$ seems natural given the other examples, but there is a minor concern with taking $\alpha_b \to 0$ because the baseline variance does not exist for $\alpha_b \leq 2$ and the baseline mean does not exist for $\alpha_b \leq 1$. (But as we mentioned in Section 5, this is not a requirement for Assumption 1 to hold.) Thus, taking $\alpha_b \to 2$ might also be a reasonable baseline. However, in that case the prior sample size is not α, but $\alpha - \alpha_b$ or $\alpha - 2$. Taking $\alpha_b \to 2$ and fixing β_b at some finite value corresponds to a prior on μ_X which has a relatively small mean and a very large variance. Taking the same parametrization for a and d, we have that the mean function is given by $ad(d-1)^{-1} = a + O(d^{-1})$, and the variance function is given by $a^2d^2[(d-1)^2(d-2)]^{-1} = a^2d^{-1} + O(d^{-2})$. Condition (32) is therefore satisfied. Assumption 3(ii) still follows by the same reasoning, while Condition (28) can be verified from $v(\mu_X) = \mu_X^2$, and $[u'(\mu_X)]^2\sigma_X^2 = \sigma_X^2 = \mu_X^2$.

Bernoulli

Assume that $X_1, \ldots, X_n$ are Bernoulli random variables, and hence $\mu_X = p$ and $\sigma_X^2 = p(1 - p)$. The conjugate prior on p is the beta distribution with parameters α and β, $B(\alpha, \beta)$. By taking α and β to zero, our baseline is $\pi(p) \propto p^{-1}(1-p)^{-1}$. The posteriors are beta distributions with means and variances

$$\mathrm{E}_{\pi}[p|\vec{X}_n] = \frac{\alpha + n\bar{X}_n}{\beta + \alpha + n}, \qquad \mathrm{Var}_{\pi}[p|\vec{X}_n] = \frac{(\alpha + n\bar{X}_n)(\beta + n - n\bar{X}_n)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}; \tag{37}$$

$$\mathrm{E}_{\pi_b}[p|\vec{X}_n] = \bar{X}_n, \qquad \mathrm{Var}_{\pi_b}[p|\vec{X}_n] = \frac{\bar{X}_n(1 - \bar{X}_n)}{n + 1}. \tag{38}$$

As before, (37) implies that we can take $a = \alpha(\alpha + \beta)^{-1}$, $d = \alpha + \beta$, and $T = X$. We then see that (32) holds for $u(a) = a$ and $v(a) = a(1-a)$ because $B(\alpha, \beta)$ has mean a and variance $a(1-a)(d+1)^{-1} = v(a)d^{-1} - v(a)[d(d+1)]^{-1} = v(a)d^{-1} + O(d^{-2})$. Again Assumption 3(ii) follows from the second expression in (38), excluding the trivial case where all Xs are equal. Condition (28) is verified because $[u'(\mu_X)]^2\sigma_X^2 = p(1-p) \equiv v(\mu_X)$.

Poisson

Assume that $X_1, \dots X_n$ are Poisson random variables, and hence $\mu_X = \sigma_X^2 = \lambda$. The conjugate prior on λ is the gamma distribution $\Gamma(\alpha, \beta)$. The prior and the baseline are therefore the same as for the exponential, but the posterior means and variances become

$$\mathrm{E}_\pi[\vec{X}_n] = \frac{\alpha + n\bar{X}_n}{\beta + n}, \qquad \mathrm{Var}_\pi[\lambda|\vec{X}_n] = \frac{\alpha + n\bar{X}_n}{(\beta+n)^2}; \tag{39}$$

$$\mathrm{E}_{\pi_b}[\lambda|\vec{X}_n] = \bar{X}_n, \qquad \mathrm{Var}_{\pi_b}[\lambda|\vec{X}_n] = \frac{\bar{X}_n}{n}. \tag{40}$$

The first expression of (39) tells us to take $a = \alpha\beta^{-1}$, $d = \beta$ and $T = X$. Condition (32) then holds with $u(a) = a$ and $v(a) = a$ because $\Gamma(\alpha, \beta)$ has mean $\alpha\beta^{-1}$ and $\alpha\beta^{-2}$. Assumption 3(ii) follows from the second expression in (40), excluding the pathological case where all Xs are zero. Condition (28) is verified because $[u'(\mu_X)]^2\sigma_X^2 = \lambda = v(\mu_x)$.

Geometric

Assume that $X_1, \dots, X_n$ are geometric random variables, and hence $\mu_X = p^{-1}$ and $\sigma_x^2 = p^{-2}(1-p)$. The conjugate prior for p is the beta distribution as given in the Bernoulli example, but with the posterior means and variances given by

$$\mathrm{E}_\pi[p|\vec{X}_n] = \frac{\alpha + n}{\alpha + \beta + n\bar{X}_n}, \quad \mathrm{Var}_\pi[p|\vec{X}_n] = \frac{(\alpha+n)(\beta + n\bar{X}_n - n)}{(\alpha+\beta+n\bar{X}_n)^2(\alpha+\beta+n\bar{X}_n+1)}; \tag{41}$$

$$\mathrm{E}_{\pi_b}[p|\vec{X}_n] = \frac{1}{\bar{X}_n}, \qquad \mathrm{Var}_{\pi_b}[p|\vec{X}_n] = \frac{\bar{X}_n - 1}{\bar{X}_n^2(n\bar{X}_n + 1)}. \tag{42}$$

The first expression of (41) then tells us to take $a = (\alpha+\beta)\alpha^{-1}$, $d = \alpha$, and $T = X$. Condition (32) then holds with $u(a) = a^{-1}$ and $v(a) = (a-1)a^{-3}$ because $B(\alpha, \beta)$ has mean a^{-1} and variance $a^{-1}(1-a^{-1})(\alpha+\beta+1)^{-1} = (a-1)a^{-3}(d+a^{-1})^{-1} = v(a)d^{-1} - v(a)[ad(ad+1)]^{-1} = v(a)d^{-1} + O(d^{-2})$. The second expression of (42) shows that Assumption 3(ii) is trivially satisfied other than the pathological case where all Xs are one. Furthermore, because $[u'(\mu_X)]^2\sigma_X^2 = [-p^2]^2p^{-2}(1-p) = p^2(1-p) = v(\mu_X)$, condition (28) also holds.

7·2. *Proofs*

We first establish the following lemma needed for proving Theorem 1. For simplicity of notation, we will abbreviate $\hat{U}_{\pi,\theta_0}(k)$ and $\hat{U}_{\pi_b,\theta_0}(k)$ as $\hat{U}(k)$ and $\hat{U}_b(k)$ respectively.

LEMMA 1. *Suppose $k = O(n^{1/2})$ and $r = k/(m+k)$ is strictly between zero and one even at its limit as $n \to \infty$. Then under the Assumptions* 1 *and* 2, *we have the following expansion:*

$$\hat{U}(k) = \frac{\alpha}{k} + O_p(k^{-3/2}), \text{ with } \alpha = r\left\{v(\mu_T) + \sigma_T^2[u'(\mu_T)]^2(r + (1-r)\Delta^2)]\right\}, \tag{43}$$

and

$$\hat{U}_b(k) = \frac{\beta}{k} + O_p(k^{-3/2}), \quad \text{where} \quad \beta = v(\mu_T) + [u'(\mu_T)]^2\sigma^2. \tag{44}$$

Proof. Because $k/(m+k)$ is strictly between zero and one, k, m and $l = k+m$ are of the same order, hence we can use them exchangeably when using the O notation. Let $\delta_k = \sqrt{k}(\bar{T}_k - \mu_T)$ and $d_m = \sqrt{m}(\mu_m - \mu_T)$, then δ_k is $O_p(1)$ by the central limit theorem and $d_m = \sigma_T\Delta + O_p(m^{-1/2})$ by Assumption 2, and hence

$$\delta_{k,m} \equiv T_{k,m} - \mu_T = l^{-1/2}[\sqrt{r}\delta_k + \sqrt{1-r}d_m] = O_p(k^{-1/2}). \tag{45}$$

Consequently $v(T_{k,m}) - v(\mu_T) = O_p(k^{-1/2})$ by a one-term Taylor expansion. Assumption 1 then allows us to write

$$\mathrm{Var}_\pi[\theta|\vec{X}_k] = \frac{v(\mu_T)}{k} r + O_p(k^{-3/2}). \tag{46}$$

For the bias term $B = \mathrm{E}_\pi(\theta|\vec{X}_k) - \mathrm{E}_{\pi_b}(\theta|\vec{X}_n)$, we expand $u(T_{k,m})$ in (20) around μ_T to obtain

$$\mathrm{E}_\pi(\theta|\vec{X}_k) = u(\mu_T) + u'(\mu_T)\delta_{k,m} + O_p(k^{-1}); \quad \mathrm{E}_{\pi_b}(\theta|\vec{X}_n) = u(\mu_T) + O_p(n^{-1/2}). \tag{47}$$

Only one term expansion of $\mathrm{E}_{\pi_b}(\theta|\vec{X}_n)$ is needed because $O_P(n^{-1/2}) = O_p(k^{-1})$ under our assumption. Consequently, we have

$$B^2 = \left[u'(\mu_T)\delta_{k,m} + O_p(k^{-1})\right]^2 = [u'(\mu_T)]^2\delta_{k,m}^2 + O_p(k^{-3/2}). \tag{48}$$

But

$$\delta_{k,m}^2 = l^{-1}[\sqrt{r}\delta_k + \sqrt{1-r}d_m]^2 = l^{-1}[r\delta_k^2 + 2\sqrt{r(1-r)}\delta_k d_m + (1-r)d_m^2]. \tag{49}$$

From (46) and (48), we see that when we take a bootstrap sample of $\hat{D}$ of (8) to obtain (6), to an error order of $O_p(k^{-3/2})$, it amounts to replacing the $\delta_k^i \equiv \delta_k^i(\vec{X}_k)$ term in (49) by its bootstrap average $\hat{\delta}_k^i$ $(i = 1, 2)$, which is defined similarly as in (6). Because $\hat{\delta}_k = \sqrt{k}(\bar{T}_n - \mu_T)$, it differs from its mean, that is, zero, by an order of $\sqrt{k}O_p(n^{-1/2}) = O_p(k^{-1/2})$. Hence the middle term on the most right hand side of (49) can be dropped without introducing more than an error of order $l^{-1}O_p(k^{-1/2}) = O_p(k^{-3/2})$, which is of the same order as the error term in (46) or in (48).

For the $\hat{\delta}_k^2$ term, we will need to use some standard results for U-statistics (e.g., see Ch. 3 of Lee (1990)). Let $h(X_1, \ldots, X_k) = (X_1 + \ldots + X_k)^2/k$. Then $\hat{\delta}_k^2$ is exactly the U-statistics generated by the kernel h, with $X_i = T_i - \mu_T$. Therefore it is known that

$$\mathrm{Var}(\hat{\delta}_k^2) \leq \frac{k}{n}\,\mathrm{Var}[h(X_1, \ldots, X_k)] = \frac{k}{n}\sigma_T^4(2 + \frac{\kappa_T}{k}), \tag{50}$$

where κ_T is the kurtosis of T_i. This implies that asymptotically the $\hat{\delta}_k^2 - \mathrm{E}[\hat{\delta}_k^2]$ is controlled by the order $O_p((k/n)^{1/2}) = O_p(k^{-1/2})$. Therefore, as before, replacing $\hat{\delta}_k^2$ by $\mathrm{E}[\hat{\delta}_k^2] = \sigma_T^2$ in (49) introduces an error of order controlled by $l^{-1}O_p(k^{-1/2}) = O_p(k^{-3/2})$, no more than what is already permitted by (46) or (48). Expansion (43) then follows because from Assumption 2, $l^{-1}d_m^2 = l^{-1}[\sigma_T^2\Delta^2 + O(m^{-1/2})] = l^{-1}\sigma_T^2\Delta^2 + O(k^{-3/2})$.

The derivation above clearly is valid when we start it by setting $m = 0$, and hence $r = 1$ and $\delta_{k,m} \equiv \delta_k$ (and then Δ is immaterial), but this is exactly the proof needed for establishing (44).□

We are now ready to prove Theorem 1. Let $R(k) = M(k)/k$, then by Lemma 1, expression (25) is equivalent to

$$[1 + R(k)][\alpha + O_p(k^{-1/2}))] = [1 + O_p(k^{-1/2})][\beta + [1 + R(k)]^{-1/2}O_p(k^{-1/2})]. \tag{51}$$

We can then verify directly that (51) is equivalent to (24) as long as $R(k) + 1$ stays away from zero almost surely in its limit, that is, $\liminf_{k\to\infty}[R(k) + 1] > 0$ almost surely, since otherwise we cannot write $[1 + R(k)]^{-1/2}O_p(k^{-1/2}) = O_p(k^{-1/2})$, which is needed for the equivalence.

Now it is easy to see that when $R(k)$ is given by (24) itself, then $\liminf_{k\to\infty}[R(k) + 1] > 0$ holds almost surely. This is because otherwise with positive probability, say $p > 0$, there exists a subsequence $\{k_i, i \geq 1\}$ such that $k_i \to \infty$ and $1 + R(k_i) \to 0$. But $1 + R(k_i) = 1 + R_{r_i}(\Delta^2) + \epsilon_i$, where $r_i = k_i/(k_i + m)$ and $\sqrt{k_i}\epsilon_i = O_p(1)$. Consequently we know with probability $p > 0$, ϵ_i converges to $-(1 + R_{r_\infty}(\Delta^2)) < 0$, where $r_\infty = \lim_{i\to\infty} r_i$. Therefore, with probability p, $|\epsilon_i|$ will be bounded away from zero when i is large enough, hence it is impossible for $\sqrt{k_i}|\epsilon_i|$ to be bounded away from infinity as k_i goes to infinity. This contradicts the fact that $\sqrt{k_i}\epsilon_i = O_p(1)$. This proves that (24) is the solution to (25).

To prove that any solution to (25) must take the form (24), we will need the additional regularity condition Assumption 3. Again we prove this by contradiction, by assuming with probability $p > 0$, the subsequence $\{k_i, i \geq 1\}$ defined above exists. Then for such subsequences the left hand side of (51) goes to zero. But the right hand side can have the zero limit only if $\sqrt{k_i + M(k_i)} = -\epsilon_i/\beta$, where $\epsilon_i = O_p(1)$. This means with positive probability (possibly smaller than p), $\hat{I} = \limsup_{i\to\infty}[k_i + M(k_i)]$ is finite. Hence with a positive probability $\liminf_{k\to\infty} \hat{U}_b(k + M(k)) > 0$ under Assumption 3(ii) because $\Pr(U_b(\hat{I}) > 0) \geq \Pr(\hat{I} < \infty) > 0$. But this contradicts (25) because its left hand side then will go to zero with a positive probability, yet its right hand side will go to 1 with probability one for the same reason as in the previous paragraph.

Acknowledgements

We thank Andrew Gelman for insightful comments and for help regarding the application in Section 4·3, Davis Jones and Steven Finch for very helpful comments and proofreading, and the U.S. National Science Foundation for partial financial support.

References

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

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

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., Hwang, J. & Vehtari, A. (2013). Understanding predictive information criteria for bayesian models. *Statistics and Computing* , 1–20.

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.

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.

Lee, A. J. (1990). *U Statistics: Theory and Practice*. New York: Marcel Dekker, Inc.

Meng, X. & Zaslavsky, A. (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.

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

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.

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.

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

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.