The Information in One Prior Relative to Another

by

Michael Evans
Department of Statistics
University of Toronto

and

Gun Ho Jang
Department of Statistics
University of Toronto

Technical Report No. 0809, November 24, 2008

TECHNICAL REPORT SERIES
UNIVERSITY OF TORONTO
DEPARTMENT OF STATISTICS

The Information in One Prior Relative to Another

Michael Evans and Gun Ho Jang
Department of Statistics
University of Toronto

Abstract

A question of some interest is how to characterize the amount of information that a prior puts into a statistical analysis. Rather than a general characterization of this quantity, we provide here an approach to characterizing the amount of information a prior puts into an analysis, when compared to another base prior. The base prior is considered to be the prior that best reflects the current available information. Our purpose then, is to characterize priors that can be used as conservative inputs to an analysis, relative to the base prior, in the sense that they put less information into the analysis. The characterization that we provide is in terms of *a priori* measures of prior-data conflict.

1 Introduction

Suppose we have two proper priors Π_1 and Π_2 on a parameter space Θ for a statistical model $\{P_\theta : \theta \in \Theta\}$. A natural question to ask is: how do we compare the amount of information each of these priors puts into the problem? While there are natural intuitive ways in which we can express this, such as prior variances, it seems difficult to characterize this precisely in general, e.g., a prior need not have a variance.

We suppose that $P_\theta(A) = \int_A f_\theta(x)\ \mu(dx)$, i.e., each P_θ is absolutely continuous with respect to a support measure μ on the sample space $\mathcal{X}$, with the density denoted by f_θ. With this formulation a prior Π leads to a prior predictive probability measure on $\mathcal{X}$ given by $M(A) = \int_\Theta P_\theta(A)\ \Pi(d\theta) = \int_A m(x)\ \mu(dx)$, where $m(x) = \int_\Theta f_\theta(x)\ \Pi(d\theta)$ is the density of M with respect to μ.

The specification of a model $\{P_\theta : \theta \in \Theta\}$ and a prior Π is equivalent to specifying a joint probability model for (θ, x), namely $P_\theta \times \Pi$. Once we observe x, the principle of conditional probability implies that any probability statements about θ must be computed using the posterior $\Pi(\cdot \mid x)$. If T is a minimal sufficient statistic for $\{P_\theta : \theta \in \Theta\}$, then it is well known that the posterior is the same whether we observe x or $T(x)$. So we will denote the posterior by $\Pi(\cdot \mid T)$ hereafter.

Since T is minimal sufficient we know that the conditional distribution of x given T is independent of θ. We denote this conditional measure by $P(\cdot \mid T)$. We now see that the joint distribution $P_\theta \times \Pi$ can be factored as

$$P_\theta \times \Pi = M \times \Pi(\cdot \mid x) = P(\cdot \mid T) \times M_T \times \Pi(\cdot \mid T) \tag{1}$$

where M_T is the marginal prior predictive distribution of T.

While much of Bayesian analysis focuses on the third factor in (1), there are also roles in a statistical analysis for $P(\cdot \mid T)$ and M_T. In [3] and [4] it is argued that $P(\cdot \mid T)$ is available for checking the sampling model, e.g., if x is a surprising value from this distribution, then we have evidence that the model $\{P_\theta : \theta \in \Theta\}$ is incorrect. Further it is argued that, if we conclude that we have no evidence against the model, then the factor M_T is available for checking whether or not there is any prior-data conflict. So if $T(x)$ is a surprising value from M_T, then we have evidence that the prior Π is placing most of its mass on θ values where the likelihood is relatively low. This is supported by the fact that T is equivalent to the likelihood map. Finally, if we have no evidence against the model, and no evidence of prior-data conflict, then $\Pi(\cdot \mid T)$ is available for probability statements about θ.

Actually the issues involved in model checking and checking for prior-data conflict are more involved than this (see Section 3), but (1) gives the basic idea that the full information, as expressed by the joint distribution of (θ, x), splits into components, each of which is available for a specific purpose in a statistical analysis. In fact, in [1] it is argued that the basic idea of avoiding "double use of the data" can be given precision by saying that each component of (1) is available for one and only one purpose in a statistical analysis. For example, we don't use $P(\cdot \mid T)$ for inference about θ, which is a basic principle of statistics known as the sufficiency principle.

From this it seems clear that the component relevant for any discussions about the respective merits or features of priors is given by M_T. Since $P(\cdot \mid T)$ is associated with variation that is independent of θ, it can have nothing to do with the choice of a prior. Further, $\Pi(\cdot \mid T)$ is associated with probability statements about θ, after we have observed the data, and so no comparisons among priors should involve this component, at least if we want to avoid using the data to choose a prior. In Section 2 we discuss how one could use M_T to compare priors with respect to the amount of information they bring into an analysis.

One issue that needs to be addressed is how one is to compare the observed data x_0 to $P(\cdot \mid T)$ or compare $t_0 = T(x_0)$ to M_T. In essence we need a measure of surprise. Perhaps the best measure of surprise is the P-value. Effectively, we are in the situation where we have a value from a single fixed distribution and we need to specify the appropriate P-value to use. This issue is addressed in [2]. In [3] and [4] the P-value for checking for prior-data conflict was based on the prior predictive density m_T, namely,

$$M_T(m_T(t) \leq m_T(t_0)). \tag{2}$$

A difficulty with (2) is that it is not invariant under 1-1, smooth transformations on the sample space when the P_θ are continuous. Accordingly, a general invariant P-value is developed in [2] for all such situations. When applied to (2), this leads to using the P-value

$$M_T(m_T^*(t) \leq m_T^*(t_0)) \tag{3}$$

instead, where $m_T^*(t) = \int_\Theta f_{\theta T}^*(t)\, \Pi(d\theta)$, $f_{\theta T}^*(t) = \int_{T^{-1}\{t\}} f_\theta(x)\, \nu_{T^{-1}\{t\}}(dx)$ and $\nu_{T^{-1}\{t\}}$ is geometric measure on $T^{-1}\{t\}$ (the analog of volume measure on a manifold). Note that it can be shown that the marginal density of T is given by $f_{\theta T}(t) = \int_{T^{-1}\{t\}} f_\theta(x) J_T(x)\, \nu_{T^{-1}\{t\}}(dx)$ where $J_T(x) = (\det dT(x) \circ dT'(x))^{-1/2}$ and dT is the differential of T. So $f_{\theta T}^*(t)$ is an adjustment of $f_{\theta T}(t)$ where we do not allow the volume distortions induced by T to affect the density. We will use (3) throughout the remainder of our discussion but note that, for the examples considered in this paper, volume distortion does not play a role and we can take $m_T^*(t) = m_T(t)$.

The motivation for this paper comes from [5] and [6], where the notion of weakly informative priors is introduced as a compromise between informative and noninformative priors. Our purpose here is to give a definition of what it means for one prior to be weakly informative with respect to another. The paper [7] contains some discussion relevant to expressing the absolute information content of a prior in terms of additional sample values, but it seems difficult to adapt this to a comparison of priors.

2 Comparing Priors

There are a variety of measures available for comparing two probability measures. For example, if Π_1 and Π_2 have densities π_1 and π_2 with respect to some support measure υ on Θ, then we could compute the relative entropy $D(\pi_1 \,||\, \pi_2) = \int_\Theta \ln(\pi_1(\theta)/\pi_2(\theta))\, \Pi_1(d\theta)$. A difficulty with any such measure is that it tells us nothing about the how to compare these priors in their role as part of a statistical analysis. For this we must bring the sampling model into play.

Perhaps a natural way to do this is to use $D(m_1 \,||\, m_2)$ where m_i is the prior predictive density of x when using the prior Π_i. Note that $D(m_1 \,||\, m_2) = D(m_{1T} \,||\, m_{2T})$ so, as is appropriate, this measure is not dependent on $P(\cdot \,|\, T)$. Note, however, that $D(m_{1T} \,||\, m_{2T}) \geq 0$ and equals 0 if and only if $M_1 = M_2$. It is not clear, however, how one can interpret $D(m_{1T} \,||\, m_{2T}) > 0$, which arises whenever $M_1 \neq M_2$, in terms of the information one prior puts into the analysis relative to another. For example, suppose M_1 has support at more than one point while M_2 is degenerate at a single point. It seems clear in such a situation that Π_2 is putting more information into the analysis than Π_1 but this is not reflected in the value of $D(m_{1T} \,||\, m_{2T})$. Overall it seems better to interpret $D(m_{1T} \,||\, m_{2T})$ as a measure of distance between these distributions, as opposed to a method of comparison with respect to the amount of relative information the priors put into the analysis.

Another objection that could be raised concerning $D(m_{1T} \,||\, m_{2T})$, is that $D(m_{1T} \,||\, m_{2T}) \neq D(m_{2T} \,||\, m_{1T})$ in general. For our application this is not so serious as we are concerned with the following context. Suppose an analyst has in mind a prior Π_1 that they believe represents the information at hand concerning θ. The analyst, however, prefers to use a prior that is somewhat conservative, with respect to the amount of information put into the analysis, when compared to Π_1. This idea comes from [5] and leads to the notion of weakly informative priors. In such a situation it seems reasonable to consider Π_1 as a base prior and then compare all other priors to it. The question then is, rather than relative entropy, what is an appropriate approach for the comparison of priors?

For a given prior Π_1 and observed value $t_0 = T(x)$ then, from (3), we have that $M_{1T}(m^*_{1T}(t) \leq m^*_{1T}(t_0))$ is the relevant quantity for assessing whether or not there is prior-data conflict with Π_1. Before we observe data, however, we have no way of knowing if we will have a prior-data conflict. Accordingly, since the analyst has determined that Π_1 best reflects the available information, it is reasonable to consider the prior distribution of $M_{1T}(m^*_{1T}(t) \leq m^*_{1T}(t_0))$ when $t_0 \sim M_{1T}$. Of course, this is effectively uniformly distributed (exactly so when $m^*_{1T}(t)$ has a continuous distribution when $t \sim M_{1T}$) and this expresses the fact that all the information about assessing whether or not a prior-data conflict exists, is contained in the P–value, with no need to compare the P-value to its distribution.

Consider now, however, the distribution of $M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0))$. Given that we have identified that *a priori* the appropriate distribution of t_0 is M_{1T}, at least for inferences about an unobserved value, then $M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0))$ is not uniformly distributed. In fact, from this distribution of $M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0))$ we can obtain an intuitively reasonable idea of what it means for a prior distribution Π_2 to be weakly informative with respect to Π_1. For suppose that the prior distribution of $M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0))$ clusters around 1. This implies that, if we were to use Π_2 as the prior when Π_1 is appropriate, then there is a small prior probability that a prior-data conflict would arise. Similarly, if the prior distribution of $M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0))$ clusters around 0, then there is a large prior probability that a prior-data conflict would arise. If one prior distribution results in a larger prior probability of there being a prior-data conflict than another then it seems reasonable to say that the first prior is more informative than the second. In fact, a completely noninformative prior should never produce prior-data conflicts.

So we must compare the distribution of $P_2(t_0) = M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0))$ when $t_0 \sim M_{1T}$, to the distribution of $P_1(t_0) = M_{1T}(m^*_{1T}(t) \leq m^*_{1T}(t_0))$ when $t_0 \sim M_{1T}$, and do this in a way that is relevant to the prior probability of obtaining a prior-data conflict. One such approach is to select a γ-quantile $x_\gamma \in [0, 1]$ of the distribution of $P_1(t_0)$, and then compute the probability

$$M_{1T}(P_2(t_0) \leq x_\gamma). \tag{4}$$

The value γ is presumably some cut-off, dependent on the application, where we will consider that evidence of a prior-data conflict exists whenever $P_1(t_0) \leq \gamma$.

Of course, if $m^*_{1T}(t)$ has a continuous distribution when $t_0 \sim M_{1T}$, then $x_\gamma = \gamma$. If the probability (4) is less than or equal to x_γ, then we call the prior Π_2 *weakly informative relative to* Π_1 *at level* γ, as this indicates the prior distribution of $P_2(t_0)$ is more concentrated about 1 than that of $P_1(t_0)$.

The following example shows that (4) behaves as we think it should in a simple context.

Example 1. *Comparing normal priors*

Suppose that $t \sim N(\mu, 1/n)$ with Π_1 on μ a $N(0, \sigma_1^2)$ distribution with σ_1^2 known. Note that t could be considered as the sample average in a sample of n from the $N(\mu, 1)$ distribution, as this statistic is minimal sufficient. We then have that M_{1T} is the $N(0, 1/n + \sigma_1^2)$ distribution. Now suppose that Π_2 is a $N(0, \sigma_2^2)$ distribution with σ_2^2 known. Then M_{2T} is the $N(0, 1/n + \sigma_2^2)$ distribution and

$$\begin{aligned} P_2(t_0) &= M_{2T}(m^*_{2T}(t) \leq m^*_{2T}(t_0)) = M_{2T}(m_{2T}(t) \leq m_{2T}(t_0)) \\ &= M_{2T}(t^2 \geq t_0^2) = 1 - G_1(t_0^2/(1/n + \sigma_2^2)), \end{aligned}$$

where G_k denotes the Chi-squared(k) distribution function. Now under M_{1T} we have that $t_0^2/(1/n + \sigma_1^2) \sim$ Chi-squared(1). Therefore,

$$\begin{aligned} M_{1T}(P_2(t_0) \leq \gamma) &= M_{1T}(1 - G_1(t_0^2/(1/n + \sigma_2^2)) \leq \gamma) \\ &= M_{1T}\left(\frac{t_0^2}{1 + \sigma_1^2} \geq \frac{1/n + \sigma_2^2}{1/n + \sigma_1^2} G_1^{-1}(1 - \gamma)\right) \\ &= 1 - G_1\left(\frac{1/n + \sigma_2^2}{1/n + \sigma_1^2} G_1^{-1}(1 - \gamma)\right). \end{aligned} \tag{5}$$

We see immediately that (5) will be less than γ whenever $\sigma_2 > \sigma_1$. In other words Π_2 will be strictly weakly informative relative to Π_1 whenever Π_2 is more diffuse than Π_1 and not otherwise. Note that $M_{1T}(P_2(t_0) \leq \gamma)$ converges to 0 as $\sigma_2^2 \to \infty$ to reflect noninformativity. Also, as $n \to \infty$, then (5) increases to $1 - G_1((\sigma_2^2/\sigma_1^2) G_1^{-1}(1 - \gamma))$.

We can generalize this to $t \sim N_k(\mu, n^{-1}I)$ with Π_1 given by $\mu \sim N_k(0, \Sigma_1)$ so that M_{1T} is the $N_k(0, n^{-1}I + \Sigma_1)$ distribution. If Π_2 is given by $\mu \sim N_k(0, \Sigma_2)$, then M_{1T} is the $N_k(0, n^{-1}I + \Sigma_1)$ distribution. It is then easy to see that $P_2(t_0) = 1 - G_k(t_0'(n^{-1}I + \Sigma_2)^{-1}t_0)$ and $M_{1T}(P_2(t_0) \leq \gamma) = M_{1T}(t_0'(n^{-1}I + \Sigma_2)^{-1}t_0 \geq G_k^{-1}(1-\gamma))$. Now, using the ordering on positive definite matrices, we have that $n^{-1}I + \Sigma_1 < n^{-1}I + \Sigma_2$ whenever $\Sigma_1 < \Sigma_2$ and so $(n^{-1}I + \Sigma_2)^{-1} > (n^{-1}I + \Sigma_1)^{-1}$. This implies that $M_{1T}(t_0'(n^{-1}I + \Sigma_2)^{-1}t_0 \geq G_k^{-1}(1 - \gamma))$ is greater than γ when $\Sigma_1 < \Sigma_2$ and so the $N_k(0, \Sigma_2)$ prior is stricly weakly informative with respect to the $N_k(0, \Sigma_1)$ prior. In this case (4) converges to the probability that $t_0'{\Sigma_2}^{-1}t_0 \geq G_k^{-1}(1 - \gamma)$ where $t_0 \sim N_k(0, \Sigma_2)$ as $n \to \infty$, and this probability can be easily computed via simulation.

An interesting feature of Example 1 is that the M_{iT} converge as $n \to \infty$. This will hold in many examples and, in such a case, we can expect that (4) will converge to some value as well. This limit can then be taken as a measure

of the amount of information in Π_2 relative to Π_1, when we will find evidence against no prior-data conflict at level γ, that is independent of sample size.

We also notice in Example 1 that, when $\Sigma_1 \leq \Sigma_2$, then $\Pi_2 = N_k(0, \Sigma_2)$ is weakly informative relative to $\Pi_1 = N_k(0, \Sigma_1)$ at every level γ, i.e., Π_2 is *uniformly weakly informative relative to* Π_1. While being uniformly weakly informative seems like a more desirable property, we still have to choose a weakly informative prior in a particular context. This seems to require selecting a γ and computing (4), as a measure of how much less informative Π_2 is, for a particular application. The following examples show that a prior may be weakly informative but not uniformly weakly informative, with respect to another prior.

Example 2. *Comparing a Student prior with a normal prior*

A conventional belief is that a Student prior is less informative than a normal prior. Our results here show that this isn't quite true.

Suppose that $t \sim N(\mu, 1/n)$ with Π_1 on μ a $N(0, \sigma^2)$ distribution and we take Π_2 to be a $t(\lambda)$ distribution, with $\lambda > 2$, scaled so that the variance of Π_2 is σ^2. So when we compare Π_2 to Π_1 we are only comparing the Student quality of the prior with normality. Numerical experience suggests that the value of σ^2 plays almost no role in this comparison. In Figure 1 we have plotted the prior predictive densities of t, for various choices of λ, when $n = 20$ and $\sigma^2 = 1$.

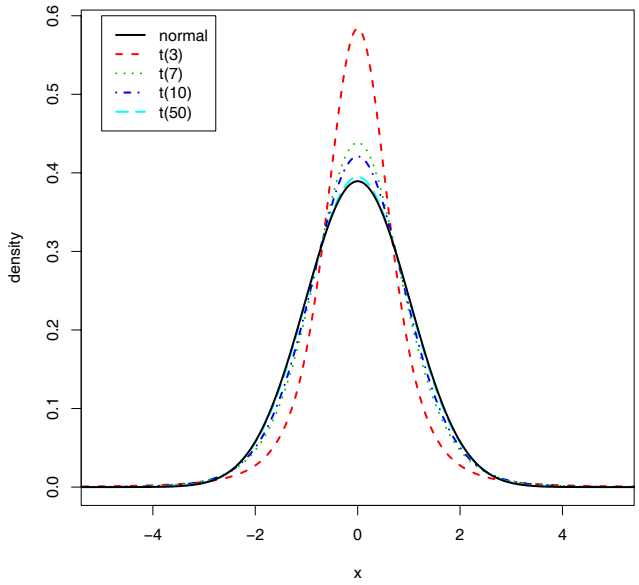

Figure 1: Plot of m_T densities, arising from standardized Student densities, when $n = 20$ and $\sigma^2 = 1$, in Example 2.

In Figure 2 we have plotted the value of (4), that arises with various standardized Student priors relative to the normal prior, against γ. We see immediately that none of these Student priors is uniformly weakly informative with

respect to the normal prior. A standardized $t(\lambda)$ prior is strictly weakly informative at level γ for all values of γ less than some cut-off value that depends on λ. For example, a standardized $t(3)$ distribution is strictly weakly informative with respect to a $N(0,1)$ prior whenever γ is less than .03573 and not otherwise, e.g. when $\gamma = .05$. Clearly this has something to do with the peakedness of the Student priors. While these results do depend on n, numerical evidence suggests that this dependence is not very strong.

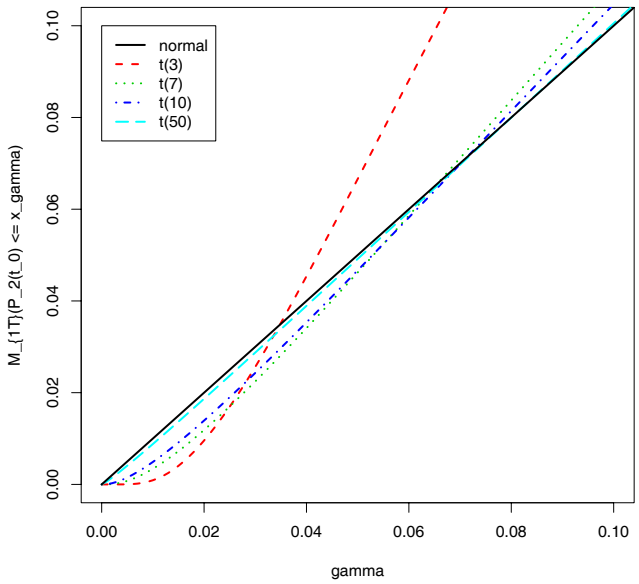

Figure 2: Plot of (4) versus γ for various standardized Student priors relative to a $N(0,1)$ prior when $n = 20$ in Example 2.

Notice that, as $n \to \infty$, then M_{1T} converges to a $N(0, \sigma^2)$ distribution, while M_{2T} converges to a $t(\lambda)$ distribution with variance equal to σ^2. Therefore, $P_2(t_0)$ converges to $1-H_{1,\lambda}(\lambda t_0^2/(\lambda-2)\sigma^2)$ where $H_{1,\lambda}$ is the distribution function of an $F_{1,\lambda}$ distribution. This implies that (4) converges to $1-G_1((\lambda-2)H_{1,\lambda}^{-1}(1-\gamma)/\lambda)$. Note that this quantity does not depend on σ^2 and it converges to γ, as $\lambda \to \infty$.

Example 3. *Comparing beta priors*

Suppose that $T \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$. This implies that $m_T(t) = \binom{n}{t}\Gamma(\alpha+\beta)\Gamma(t+\alpha)\Gamma(n-t+\beta)/\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)$ and from this we can compute (4) for various choices of (α, β).

As a specific example, suppose that $n = 20$, the base prior is given by $(\alpha, \beta) = (7, 7)$, and we take $\gamma = .05$ so that $x_{.05} = .0479$. As alternatives to this base prior, we consider $\text{Beta}(\alpha, \alpha)$ priors. In Figure 3 we have plotted the value of (4) as a function of α, for those values of α such that the $\text{Beta}(\alpha, \alpha)$ prior is weakly informative with respect to the $\text{Beta}(7, 7)$ prior. We see that a $\text{Beta}(\alpha, \alpha)$ prior is strictly weakly informative whenever $1 \leq \alpha < 7$.

In Figure 4 we have plotted all the (α, β) corresponding to Beta(α, β) distributions that are weakly informative with respect to the Beta$(7, 7)$ distribution at level .05, together with the subset of all (α, β) corresponding to Beta(α, β) distributions that are uniformly weakly informative with respect to the Beta$(7, 7)$ distribution. The graph on the left corresponds to $n = 20$ while the one on the right corresponds to $n = 100$. The plot for $n = 20$ shows some anomalous effects due to the discreteness. For example, a Beta(α, α) prior for α satisfying $7 < \alpha < 15.46486$ is weakly informative relative to the Beta$(7, 7)$ prior although Figure 3 shows that such a prior is not strictly weakly informative. When we increase n these cases are eliminated as shown in the plot for $n = 100$. The limiting regions are difficult to determine in this example.

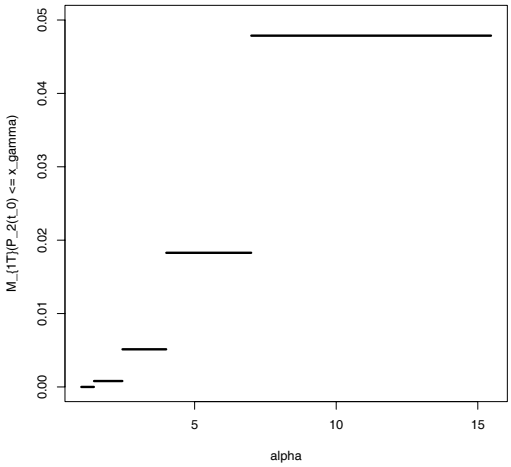

Figure 3: Plot of (4) versus α in Example 3.

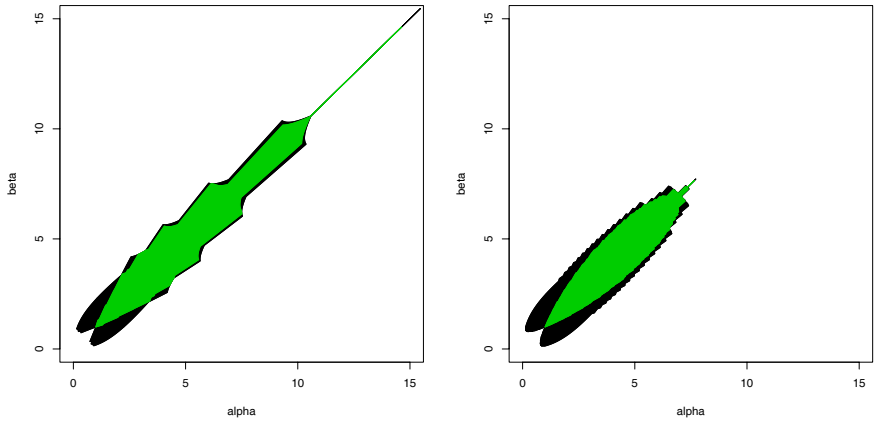

Figure 4: Plot of all (α, β) corresponding to weakly informative priors at level $\gamma = .05$ (light and dark shading) and all (α, β) corresponding to uniformly weakly informative priors (light shading) for $n = 20$ (on the left) and $n = 100$ (on the right) in Example 3.

3 Refinements Based Upon Ancillarity

The approach in Section 2 works whenever T is a complete minimal sufficient statistic. This is a consequence of Basu's Theorem as, in such a case, any ancillary is statistically independent of T and so conditioning on such an ancillary is irrelevant. When $U(T)$ is a meaningful ancillary, however, then the variation due to $U(T)$ is independent of θ and so should be removed from the P-value (3) when checking for prior-data conflict. Removing this variation is equivalent to conditioning on $U(T)$ and so we replace (3) by

$$M_T(m_T^*(t) \leq m_T^*(t_0) \,|\, U(T)), \tag{6}$$

i.e., we use the conditional prior predictive given the ancillary $U(T)$. To remove the maximal amount of ancillary variation we must have that $U(T)$ is a maximal ancillary. Therefore (4) becomes

$$M_{1T}(P_2(t_0 \,|\, U(T)) \leq x_\gamma \,|\, U(T)), \tag{7}$$

i.e., we have replaced $P_2(t_0)$ by $P_2(t_0 \,|\, U(T)) = M_{2T}(m_{2T}^*(t) \leq m_{2T}^*(t_0) \,|\, U(T))$ and M_{1T} by $M_{1T}(\cdot \,|\, U(T))$.

When ancillary $U(T)$ exists, then we can also check the model by comparing the observed value $U(t_0)$ against $P_{U(T)}$, the marginal distribution of $U(T)$ induced by the model, as this distribution is independent of θ. This leads to a more refined factorization than (1) given by

$$P_\theta \times \Pi = P(\cdot \,|\, T) \times P_{U(T)} \times M_T(\cdot \,|\, U(T)) \times \Pi(\cdot \,|\, T)$$

and each component is available for a specific purpose in a statistical analysis. In models where T is not a complete minimal sufficient statistic, the factor $P_{U(T)}$ typically plays the more important role in model checking.

One problem with ancillaries is that multiple maximal ancillaries may exist. When ancillaries are used for frequentist inferences about θ via conditioning, this poses a problem because it is not clear which multiple ancillary to use and confidence regions depend on the maximal ancillary chosen. For checking for prior-data conflict via (6), however, this does not pose a problem. This is because we simply get different checks depending on which maximal ancillary we condition on. For example, if conditioning on maximal ancillary $U_1(T)$ does not lead to prior-data conflict, but conditioning on maximal ancillary $U_2(T)$ does, then we have evidence against no prior-data conflict existing.

Similarly, when we go to use (7), we can also simply look at the effect of each maximal ancillary on the analysis and make our assessment about Π_2 based on this. For example, we can use the maximum value of (7) over all maximal ancillaries to assess whether or not Π_2 is weakly informative relative to Π_1. When this maximum is small, we conclude that we have a small prior probability of finding evidence against the null hypothesis of no prior-data conflict when using Π_2.

Example 4. *Nonunique maximal ancillaries*

Suppose that we have a sample of n from the

$$\text{Multinomial}\left(1, (1-\theta)/6, (1+\theta)/6, (2-\theta)/6, (2+\theta)/6\right)$$

distribution where $\theta \in [-1, 1]$ is unknown. Then the counts (f_1, f_2, f_3, f_4) constitute a minimal sufficient statistic and $U_1 = (f_1 + f_2, f_3 + f_4)$ is ancillary as is $U_2 = (f_1 + f_4, f_2 + f_3)$.

Then $T = (f_1, f_2, f_3, f_4) \mid U_1$ is given by

$$f_1 \mid U_1 \sim \text{Binomial}\left(f_1 + f_2, (1-\theta)/2\right)$$

independent of

$$f_3 \mid U_1 \sim \text{Binomial}\left(f_3 + f_4, (2-\theta)/4\right)$$

giving

$$m_T(f_1, f_2, f_3, f_4 \mid U_1) = \binom{f_1 + f_2}{f_1}\binom{f_3 + f_4}{f_3} \times$$
$$\int_{-1}^{1} \left(\frac{1-\theta}{2}\right)^{f_1} \left(\frac{1+\theta}{2}\right)^{f_2} \left(\frac{2-\theta}{4}\right)^{f_3} \left(\frac{2+\theta}{4}\right)^{f_4} \pi(\theta)\, d\theta.$$

We then have two 1-dimensional distributions $f_1 \mid U_1$ and $f_3 \mid U_1$ to use for checking for prior-data conflict. A similar result holds for the conditional distribution given U_2.

For example, suppose π is a Beta$(20, 20)$ distribution on $[-1, 1]$, so the prior concentrates about 0, and for a sample of $n = 18$ we have that $U_1 = f_1 + f_2 = 10$ and $U_2 = f_1 + f_4 = 8$. In Figure 5 we have plotted all the values of (α, β) that correspond to a Beta(α, β) prior that is weakly informative relative to the Beta$(20, 20)$ prior at level $\gamma = .05$. So for each such (α, β) we have that (7) is less than or equal to .05 for both $U = U_1$ and $U = U_2$, i.e., (7) is less than or equal to .05 for these values of (α, β) when $U = U_1$ and when $U = U_2$.

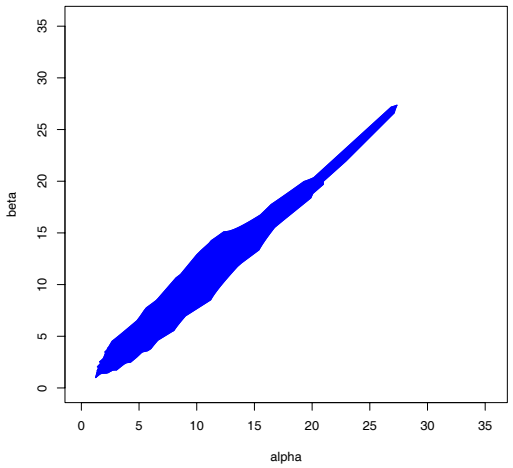

Figure 5: Plot of all (α, β) corresponding to weakly informative priors at level $\gamma = .05$ in Example 4.

4 Conclusions

We have developed an approach to measuring the amount of information a prior puts into an a statistical analysis relative to another base prior. This base prior can be considered as the prior that best reflects current information and our goal is to determine a prior that is weakly informative with respect to the base prior. Our measure is in terms of the prior predictive probability measure, using the base prior, of obtaining a prior-data conflict. This was applied in several examples where the approach is seen to work quite well. More involved contexts require some additional computational complexities, but these are not insurmountable.

As noted in Example 3, we need to be careful when we conceive of a prior being weakly informative relative to another. This entails choosing a γ and computing (4). The difference between these quantities indicates to what extent the prior is less informative than the base prior in terms of the prior probability of observing a prior data conflict. While this is a design consideration, i.e., (4) is used to choose a prior Π_2 before we observe the data, we still should check for prior-data conflict with Π_2 after observing the data.

References

[1] Evans, M. (2007) Comment on "Bayesian checking of the second levels of hierarchical models" by Bayarri and Castellanos. Statistical Science, 22, 3, p. 344-348.

[2] Evans, M. and Jang, G. (2008) Invariant P-values for model checking and checking for prior-data conflict. Technical Report No. 0803 June, 2008, Dept. of Statistics, University of Toronto.

[3] Evans, M. and Moshonov, H. (2006) Checking for prior-data conflict. Bayesian Analysis, 1, 4, p. 893-914.

[4] Evans, M. and Moshonov, H. (2007) Checking for prior-data conflict with hierarchically specified priors. Bayesian Statistics and its Applications, eds. A.K. Upadhyay, U. Singh, D. Dey, Anamaya Publishers, New Delhi, p. 145-159.

[5] Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1, 3, p. 515-533.

[6] Gelman, A., Jakulin, A., Pittau, M.G. and Su, Y-S. (2008) A weakly informative default prior distribution for logistic and other regression models. To appear in Annals of Applied Statistics.

[7] Kass R., and Wasserman L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association, 90, 431, p. 928-934. 376.