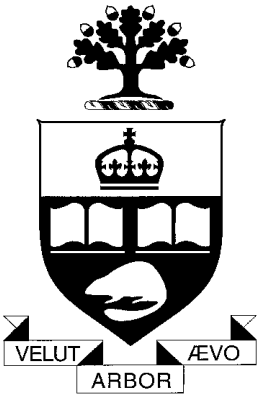

Checking for Prior-Data Conflict

by

Michael Evans
Department of Statistics
University of Toronto

and

Hadas Moshonov
Department of Statistics
University of Toronto

Technical Report No. 0413 December 15, 2004

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Bayesian Analysis (2006) **1**, Number 4, pp. 893–914

Checking for Prior-Data Conflict

Michael Evans* and Hadas Moshonov†

Abstract. Inference proceeds from ingredients chosen by the analyst and data. To validate any inferences drawn it is essential that the inputs chosen be deemed appropriate for the data. In the Bayesian context these inputs consist of both the sampling model and the prior. There are thus two possibilities for failure: the data may not have arisen from the sampling model, or the prior may place most of its mass on parameter values that are not feasible in light of the data (referred to here as prior-data conflict). Failure of the sampling model can only be fixed by modifying the model, while prior-data conflict can be overcome if sufficient data is available. We examine how to assess whether or not a prior-data conflict exists, and how to assess when its effects can be ignored for inferences. The concept of prior-data conflict is seen to lead to a partial characterization of what is meant by a noninformative prior or a noninformative sequence of priors.

Keywords: prior-data conflict, sufficiency, ancillarity, hierarchically specified priors

1 Introduction

Model checking is a necessary component of statistical analyses; this is because statistical analyses are based on assumptions. These assumptions typically take the form of possibilities for how the observed data was generated—the sampling model—and, in Bayesian contexts, a quantification of the investigator's beliefs—the prior—about the actual data generation mechanism, among the possibilities included in the sampling model.

In certain contexts it might be argued that we have specific information that leads to a sampling model or that the investigator has a very clear idea of what an appropriate prior is but, in general, this does not seem to be the case. In fact the sampling model and prior often seem, if not chosen arbitrarily, then at least selected for convenience. Accordingly, it seems quite important that, before carrying out an inferential analysis, we first check to make sure that the choices made make sense in light of the data collected. For, if we can show that the observed data is surprising in light of the sampling model and the prior, then we must be at least suspicious about the validity of the inferences drawn (ignoring any ambiguities about the correct inference process itself).

Our concern here is with the Bayesian context where we have two ingredients—namely, the sampling model and the prior. Several authors have considered model checking in this situation such as Guttman (1967), Box (1980), Rubin (1984), and Gelman, Meng and Stern (1996). We note that all of these authors considered the effect of both the sampling model and the prior simultaneously. There are, however, two possible ways in which the Bayesian model can fail.

First we may have that the sampling model fails. By this we mean that the observed data is surprising for each of the possible distributions in the model. If this is the case, then either we had the misfortune to observe a rare data set or the model is in fact not appropriate. In

*University of Toronto, Toronto, Canada, http://genealogy.math.ndsu.nodak.edu/html/id.phtml?id=15995
†University of Toronto, Toronto, Canada, mailto:hadas@utstat.utoronto.ca

the former situation collecting more data may resolve the issue, but in the latter this will not be the case.

Second, assuming the sampling model is appropriate, the Bayesian model may fail by the prior placing its mass primarily on distributions in the sampling model for which the observed data is surprising. So, in other words, there are distributions in the model for which the data is not surprising, but the prior places little or no mass there. We refer to this as *prior-data conflict*. This conflict again leads to doubts about the validity of inferences drawn from the Bayesian model.

We note that in the second context increasing the amount of data can lead to a resolution of the problem, for as the amount of data increases the effect of the prior decreases, vanishing in the limit. What this implies is that even if there is prior-data conflict, we may have sufficient data so that its effect can be ignored.

The preceding discussion leads us to the conclusion that in Bayesian problems we must check individually for the separate sources of failure. First we must check for the failure of the sampling model. For example, there are many frequentist methods for this and also the Bayesian methods as discussed in Bayarri and Berger (2000). If we have no evidence that the sampling model is in error, then we must check to see whether or not there is any prior-data conflict and, if there is, whether or not this conflict leads to erroneous inferences. Of course, if we find evidence that the sampling model is wrong, then it doesn't make sense to check for prior-data conflict.

We note that, if we do obtain evidence of the sampling model failing, then this failure may not lead to serious problems for our inferences in certain circumstances. This is because the model may only be approximately correct and we have observed a sufficient amount of data to detect a small deviation. If the approximate correctness of the model is appropriate for the application at hand, then we would not want to discard the model. This is a separate issue, however, from what we will discuss in this paper and needs to be addressed by methods other than what we present here. In any case, throughout the remainder of this paper we assume that the sampling model is correct and focus on looking for prior-data conflict.

It is also worth remarking that some feel that, when a prior reflects the subjective beliefs of an investigator, then there is no necessity to check for prior-data conflict as the existence of such a conflict would not necessarily lead one to abandon the prior. While this is a possible outcome, we note that it is at least informative to know whether or not a prior-data conflict exists and what its effects on the inferences are. In particular, this seems relevant when reporting the results of a statistical analysis that will be used by others. We are not dictating here what a necessary outcome is, when we have decided a prior-data conflict exists, as this can be determined by the particular context. We do feel, however, that checking for prior-data conflict can be part of good statistical practice.

The question now arises as to how we should check for the existence of a prior-data conflict. In Section 2 we argue that it is appropriate to use the prior predictive distribution of the minimal sufficient statistic to do this. In Section 3 we show how the concept of ancillarity becomes relevant when looking for prior-data conflict. In Section 4 we discuss the implications of our definition of prior-data conflict for the characterization of a noninformative prior. In Section 5 we discuss how to assess whether or not an observed prior-data conflict can be ignored so that the Bayesian model can be used to derive inferences about the unknowns in the sampling model. In Section 6 we discuss a factorization of the joint distribution of the parameter and data that serves as further support for the approach taken here. In Section 7 we discuss the

generalization of the methods developed for checking the whole prior to checking components of a prior.

In preparation for our discussion we specify some notation. We denote the sample space for the data s by S, the parameter space by Ω, and the sampling model by $\{f_\theta : \theta \in \Omega\}$, where each f_θ is a probability density on S with respect to some support measure μ. The prior probability measure on Ω is denoted by Π with density π with respect to a support measure υ on Ω. The joint distribution of (θ, s) leads to the prior-predictive measure for s as given by

$$M(B) = \int_\Omega \int_B f_\theta(s)\,\pi(\theta)\;\mu(ds)\;\upsilon(d\theta) = \int_B m(s)\;\mu(ds),$$

where $m(s) = \int_\Omega f_\theta(s)\,\pi(\theta)\;\upsilon(d\theta)$ is the density of M with respect to μ. We note that we are restricting our attention here to the case where Π is proper so that M is indeed a probability measure. We do, however, make some comments relevant to the improper case.

When we observe the data s_0, then we have the posterior probability measure $\Pi(\cdot \mid s_0)$ for inferences about θ. For a function $T : S \to \mathcal{T}$ we denote the marginal densities by $f_{\theta T}$, with respect to support measure λ on $\mathcal{T}$, and this leads to the marginal prior predictive density for T given by $m_T(t) = \int_\Omega f_{\theta T}(t)\,\pi(\theta)\;\upsilon(d\theta)$ again with respect to λ.

To assess whether or not a prior-data conflict exists, we need to compare an observed value $T(s_0)$ with a fixed distribution P_T to assess whether or not it is surprising. For convenience, and because at this time it is not clear what an alternative approach is in this context, we will use a P-value for this purpose. In particular, we will assess how surprising $T(s_0)$ is, by comparing the value of the density p_T of T at $T(s_0)$ with other possible values—namely, we will compute $P_T(p_T(t) \le p_T(T(s_0)))$. When p_T is unimodal, this is equivalent to computing how far out in the tails the observed value $T(s_0)$ is and, in the multimodal case, this seems to give an appropriate measure when $T(s_0)$ lies between modes. There are various concerns with P-values generally and with this P-value. In practical contexts we recommend also plotting the distribution being used in the comparison, in addition to computing the P-value, to see where the observed value $T(s_0)$ lies with respect to this distribution. Further, the discussion in Section 5 is concerned with diagnostics that can be used to differentiate between statistical significance (i.e., deciding via a P-value that prior-data conflict exists) and practical significance (i.e., whether or not a detected prior-data conflict is really something we need to worry about when constructing inferences).

2 Checking for Prior-data Conflict: Sufficiency

Intuitively, a prior-data conflict exists whenever the data provide little or no support to those values of θ where the prior places its support. So in low dimensions we could plot the posterior and prior distributions of θ to see how different these were. In cases where there is significant prior-data conflict we would expect the effective supports (loosely speaking, a region that contained most of the mass for a distribution) of the prior and the posterior to be quite different. While such plots are useful diagnostics, we would like a more formal measure of the difference and a general methodology for dealing with higher dimensional θ. Further, to completely separate the effect of the prior from the data, we choose instead to compare the effective support of the prior with the region where the likelihood function is relatively high.

This might lead one to compare the MLE $\hat{\theta}(s_0)$ (or some other consistent estimate of θ) to the prior distribution to assess the degree of conflict. If $\hat{\theta}(s_0)$ lay in a region where the prior

placed little support, e.g., out in the tails of π, then we would conclude that a conflict exists. We note, however, that this is not quite correct, for if the spread of the likelihood is sufficiently wide, then there could be a significant overlap between the effective support of the prior and the likelihood, even if $\hat{\theta}(s_0)$ was "out in the tails" of π.

Therefore, a more appropriate measure of prior-data conflict would seem to be one that compared the effective support of the prior with the region where the likelihood was relatively high. If there is very little overlap between these regions, then we have evidence that a prior-data conflict exists. It is not clear, however, how to appropriately measure this degree of overlap.

Accordingly, we take a different approach. We note that the prior induces a probability distribution on the set of possible likelihood functions via the prior predictive probability measure M. If the observed likelihood $L(\cdot \mid s_0) = f_{\cdot}(s_0)$ is a surprising value from this distribution, then this would seem to indicate that a prior-data conflict exists.

We can simplify this somewhat by noting that the likelihood is in effect equivalent to a minimal sufficient statistic T—namely, if we know the value $T(s_0)$ then we can obtain the map $L(\cdot \mid s_0)$, and conversely. So, instead of comparing the observed likelihood to its marginal model, we can compare $T(s_0)$ to its prior distribution M_T. We can then choose T so that this comparison can be made relatively simply.

We might ask if anything in the observed data s_0, beyond the value of $T(s_0)$, has any relevance to checking for prior-data conflict. The following result, proved in the appendix, suggests that this is not the case.

Theorem 1. Suppose T is a sufficient statistic for the model $\{f_\theta : \theta \in \Omega\}$ for data s. Then the conditional prior predictive distribution of the data s given T is independent of the prior π.

So the conditional prior distribution of any aspect of the data, beyond the value of $T(s_0)$, doesn't depend on the prior. Therefore such information can tell us nothing about whether or not a prior-data conflict exists, and so can be ignored for this purpose. Now note the following result, also proved in the appendix.

Theorem 2. If $L(\cdot \mid s_0)$ is nonzero only on a θ-region where π places no mass then $T(s_0)$ is an impossible outcome for M_T.

So we see that, at least in this extreme case of nonoverlapping support for the likelihood and prior, comparing $T(s_0)$ to M_T leads to definite evidence of a prior-data conflict.

These results support our assertion that comparing the observed value $T(s_0)$, of a minimal sufficient statistic, to its prior distribution M_T is an appropriate method for checking for prior-data conflict. We note that Box (1980) proposed comparing the observed data s_0 to M as a method for model checking. So what we are suggesting here is really just a restriction of Box's approach to minimal sufficient statistics, and we are further suggesting that this is appropriate only for checking for prior-data conflict and not model checking generally. This, together with the discussion in Section 6, would seem to lead to a resolution of some anomalous behavior of Box's approach in particular examples. We note also that the discussion in Section 3 leads to a further modification of Box's approach.

The following examples show that basing the check for prior-data conflict by comparing $T(s_0)$ to M_T produces acceptable results.

Example 1. *Location normal model*

Suppose $s = (s_1, ..., s_n)$ is a sample from a $N(\theta, 1)$ distribution with $\theta \in R^1$ and $\theta \sim N(\theta_0, \sigma_0^2)$. A minimal sufficient statistic is $T(s) = \bar{s} \sim N(\theta, 1/n)$.

For the prior predictive distribution of $\bar{s}$ we can write $\bar{s} = \theta + Z$ where $Z \sim N(0, 1/n)$ independent of θ. So the prior predictive distribution is $N(\theta_0, \sigma_0^2 + 1/n)$. Accordingly, it makes sense to see if $\bar{s}_0$ lies in the tails of this distribution; we assess this via the P-value

$$M_T(m_T(\bar{s}) \leq m_T(\bar{s}_0)) = 2(1 - \Phi(|\bar{s}_0 - \theta_0|/(\sigma_0^2 + 1/n)^{1/2})). \tag{1}$$

This shows that when $\bar{s}_0$ lies in the tails of its prior distribution, we have evidence of a prior-data conflict existing. Note that using the prior-predictive distribution of $\bar{s}$ to assess this, instead of the prior for θ, results in the standardization by $(\sigma_0^2 + 1/n)^{1/2}$ rather than σ_0. So $\bar{s}_0$ has to be somewhat further out in the tail than if we simply compared the MLE to the prior of θ.

Now suppose that the true value of θ is θ_*. Then as $n \to \infty$ we have that (1) converges almost surely to $2(1 - \Phi(|\theta_* - \theta_0|/\sigma_0))$. Therefore, if the true value of θ_* lies in the tails of the prior, then asymptotically (1) will detect this and lead to evidence that a prior-data conflict exists.

Also note that when $\sigma_0 \to \infty$, (1) converges to 1 and no evidence will be found of a prior-data conflict. This is not surprising; recall that we are proceeding as if the sampling model is correct and a diffuse prior simply indicates that all values are equally likely, so we should definitely not find any conflict. Further as $\sigma_0 \to 0$, so we have a very precise prior, then (1) will definitely find evidence of a prior-data conflict for large n, unless θ_0 is indeed the true value.

Example 2. *Bernoulli model*

Suppose $s = (s_1, ..., s_n)$ is a sample from a Bernoulli(θ) model with $\theta \sim \text{Beta}(\alpha, \beta)$. A minimal sufficient statistic is $T(s) = \sum s_i \sim \text{Binomial}(n, \theta)$.

The prior-predictive probability function of $T(s)$ is then given by

$$m_T(t) = \binom{n}{t} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(t + \alpha)\Gamma(n - t + \beta)}{\Gamma(n + \alpha + \beta)}.$$

To assess whether or not $T(s) = t_0$ is surprising, we must compare t_0 to m_T; we do this by computing the tail probability $M_T(m_T(t) \leq m_T(t_0))$ which, in this case, must be done numerically. In essence, this is the prior probability of obtaining a value of the minimal sufficient statistic with probability of occurrence no greater than that for the observed value. Because of cancellations we can write the P-value as

$$M_T\left(\frac{\Gamma(t + \alpha)\Gamma(n - t + \beta)}{\Gamma(t + 1)\Gamma(n - t + 1)} \leq \frac{\Gamma(t_0 + \alpha)\Gamma(n - t_0 + \beta)}{\Gamma(t_0 + 1)\Gamma(n - t_0 + 1)}\right). \tag{2}$$

Notice that when $\alpha = \beta = 1$, then (2) equals 1 and we never have any evidence of prior-data conflict. This makes sense, as when the sampling model is correct, there can be no conflict between the data and a noninformative prior.

To illustrate the use of the P-value given by (2), we consider a numerical example. For this suppose that $n = 10, \alpha = 5$ and $\beta = 20$. In this case the prior distribution puts most of its mass on values of θ in $(0, 1/2)$. Figure 1 is a plot of the prior predictive probability function of the minimal sufficient statistic.

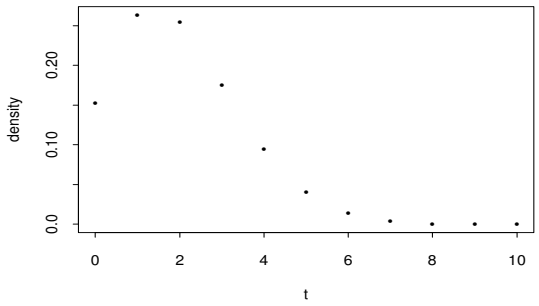

Figure 1: Plot of m_T for a sample of $n = 10$ from the Bernoulli(θ) distribution when $\theta \sim$ Beta(5,20) in Example 2.

Generating a sample of size 10 from the Bernoulli(0.9) distribution, we obtained $t_0 = 9$. Note that $\theta = 0.9$ is outside the effective support of the prior distribution. The prior predictive P-value, is given by $M_T\left(m_T(y) \leq m_T(9)\right) = m_T(9) + m_T(10) = 0.000117$. As expected, this is an indication of a prior-data conflict. Generating a sample of size 10 from the Bernoulli(0.25) distribution we obtained $t_0 = 1$. The P-value is given by $M_T\left(m_T(y) \leq m_T(1)\right) = 0.736635$. As expected, this does not indicate any prior-data conflict.

The following example illustrates that the notion of a noninformative prior is somewhat subtle. We discuss this further in Section 4.

Example 3. *Negative-binomial sampling*

Suppose that s is distributed Negative-binomial(k, θ); then s is minimal sufficient. Note that the likelihood function is a positive multiple of the likelihood function in Example 2 (Bernoulli) when we obtain s_0 zeros in a sample of size $n = s_0 + k$. It is then clear that posterior inferences about θ are the same from the two models. We will see, however, that the checks for prior-data conflict are somewhat different.

Suppose that we place a uniform prior on θ. In Example 2 (Bernoulli), the prior predictive for the minimal sufficient statistic is uniform and we never have evidence of a prior-data conflict existing. The prior predictive under negative binomial sampling, however, is given by

$$m(s) = \binom{s+k-1}{k-1} \int_0^1 \theta^k (1-\theta)^s \, d\theta = \frac{k}{(s+k)\,(s+k+1)} \tag{3}$$

and this is not uniform. Indeed, it cannot be uniform because the support for this distribution is the nonnegative integers. Since (3) is decreasing in s,

$$M\left(m(s) \leq m(s_0)\right) = \sum_{s=s_0}^{\infty} \frac{k}{(s+k)\,(s+k+1)} = \frac{k}{s_0+k}. \tag{4}$$

We see that this P-value is small only when s_0 is large compared to k.

While this P-value may seem unusual, when compared to the Binomial case, there is a subtle difference between the two situations. In particular, in the Binomial(n, θ) case it is permissible to have $\theta = 0$ and so, for example, observe n consecutive tails. In the Negative-binomial(k, θ) case, however, it is not possible to have $\theta = 0$ as there is no such thing as the Negative-binomial$(k, 0)$ distribution. The uniform prior on the unit interval places positive mass on every interval about 0 and so is not really a sensible prior for this model, as the positive mass indicates a prior belief that $\theta = 0$ is a possible value. Recall our restriction that, when looking for prior-data conflict, we assume that the sampling model is correct.

If we place a uniform prior on $[\epsilon, 1]$ with $\epsilon > 0$, as this places no mass at 0, and denote the prior predictive by m_ϵ, then the dominated convergence theorem establishes that $\lim_{\epsilon \to 0} M_\epsilon \left(m_\epsilon(s) \leq m_\epsilon(s_0) \right)$ is equal to (4). In fact, when $k = 1$, we have that $M_\epsilon \left(m_\epsilon(s) \leq m_\epsilon(s_0) \right) = (1 - \epsilon)^{s_0}/(s_0 + 1)$. Accordingly we can view (4) as an approximation to the P-value we would obtain for a prior that is 0 in a small neighborhood of 0.

Example 4. *Location-scale normal model*

Suppose that $x = (x_1, ..., x_n)$ is a sample from a $N(\mu, \sigma^2)$ distribution where $\mu \in R^1$ and $\sigma > 0$ are unknown. With $s^2 = (n-1)^{-1} \sum (x_i - \bar{x})^2$, then $(\bar{x}, s^2)$ is a minimal sufficient statistic for (μ, σ^2) with $\bar{x} \sim N(\mu, \sigma^2/n)$ independent of $s^2 \sim (\sigma^2/(n-1))\chi^2_{(n-1)}$. Suppose the prior on (μ, σ^2) is given by

$$\mu \,|\, \sigma^2 \sim N(\mu_0, \tau_0^2 \sigma^2), \quad \frac{1}{\sigma^2} \sim \text{Gamma}(\alpha_0, \beta_0). \tag{5}$$

The posterior distribution of (μ, σ^2) is then given by

$$\begin{aligned} \mu \,|\, \sigma^2, x_1, \ldots, x_n &\sim N\left(\mu_x, \left(n + 1/\tau_0^2\right)^{-1} \sigma^2\right) \\ \frac{1}{\sigma^2} \,|\, x_1, \ldots, x_n &\sim \text{Gamma}\left(\alpha_0 + \frac{n}{2}, \beta_x\right) \end{aligned}$$

where $\mu_x = (n + 1/\tau_0^2)^{-1}(\mu_0/\tau_0^2 + n\bar{x})$ and

$$\beta_x = \beta_0 + (n-1)s^2/2 + n(\bar{x} - \mu_0)^2/2(n\tau_0^2 + 1).$$

The joint prior predictive density $m(\bar{x}, s^2)$ of $(\bar{x}, s^2)$ is proportional to

$$(s^2)^{(n-1)/2-1} \beta_x^{-n/2-\alpha_0}. \tag{6}$$

We then assess whether or not observed $(\bar{x}_0, s_0^2)$ is a reasonable value by computing the P-value $M\left(m(\bar{x}, s^2) \leq m(\bar{x}_0, s_0^2)\right)$; note that we don't need the norming constant of $m(\bar{x}, s^2)$ for this.

To compute this, for specified values of the hyperparameters α_0, β_0, μ_0 and τ_0^2, we generate (μ, σ^2) using (5), then generate $(\bar{x}, s^2)$ from their joint distribution given (μ, σ^2) and evaluate $m(\bar{x}, s^2)$ using (6). Repeating this many times, and recording the proportion of values of $m(\bar{x}, s^2)$ that are less then or equal to $m(\bar{x}_0, s_0^2)$, gives us a Monte Carlo estimate of the P-value.

For example, suppose that $(\mu, \sigma) = (0, 1)$ and that for a sample of size $n = 20$ from this distribution we obtained $\bar{x}_0 = 0.0358324$ and $s_0^2 = 0.836563$. Then for the prior specified by $\tau_0^2 = 1$, $\mu_0 = 50$, $\alpha_0 = 1$ and $\beta_0 = 5$, based on a Monte Carlo sample of size $N = 10^3$, the P-value is estimated as 0.000 and so clear evidence of a prior-data conflict is obtained.

Rather than computing a P-value based on the full prior predictive distribution of T, it is also possible to use the prior predictive distribution of some function of the minimal sufficient statistic. For example, we could instead use the marginal prior predictive distributions of $\bar{x}$ and s^2.

We can write $\bar{x} = \mu + (\sigma^2/n)z$ where $z \sim N(0,1)$ and $\mu \,|\, \sigma^2 \sim N(\mu_0, \tau_0^2\sigma^2)$. This implies that $\bar{x} \,|\, \sigma^2 \sim N(\mu_0, \sigma^2(\tau_0^2 + 1/n))$. Then, after multiplying the $N(\mu_0, \sigma^2(\tau_0^2 + 1/n))$ density by the prior for $1/\sigma^2$ and integrating out σ^2, we have that

$$\bar{x} \sim t_{2\alpha_0}\left(\mu_0, (\beta_0\left(\tau_0^2 + \frac{1}{n}\right)\alpha_0^{-1})^{1/2}\right),$$

where $t_\lambda(0,1)$ denotes the t distribution with λ degrees of freedom and $t_\lambda(\mu,\sigma) = \mu + \sigma t_\lambda(0,1)$, is the marginal prior predictive distribution of $\bar{x}$. Therefore the P-value based on $\bar{x}$ alone is $2(1 - G_{2\alpha_0}(|\bar{x}_0 - \mu_0|/(\beta_0\left(\tau_0^2 + 1/n\right)/\alpha_0)^{1/2}))$, where $G_{2\alpha_0}$ is the cdf of the t distribution with $2\alpha_0$ degrees of freedom. For the above generated data, the P-value equals to .0021, which also indicates the existence of a prior-data conflict.

Similarly, we find that $s^2 \sim (\beta_0/\alpha_0)F_{(n-1,2\alpha_0)}$ is the marginal prior predictive of s^2. For the above example we compare $s^2/5 = 0.1673126$ with the $F(19,2)$ distribution. Computing the probability of obtaining a value from the $F(19,2)$ distribution with density smaller than that obtained at the observed value, leads to the P-value $\Pr(F(19,2) \leq 0.1673126) + \Pr(F(19,2) > 1.5295) = .47832$, which does not indicate any prior-data conflict. We see from these two tests that the conflict arises from the location of that data and not its spread.

3 Checking for Prior-data Conflict: Ancillarity

In principle we could look for prior-data conflict by comparing the observed value of $U(T(s))$ with its marginal prior-predictive distribution for any function U—e.g., consider Example 4 (Location-scale normal). We note, however, that for certain choices of U this will not be appropriate. For, if U is ancillary, then the prior-predictive distribution of U is the same as its sampling distribution—i.e., the marginal prior-predictive distribution does not depend on the prior. So, if $U(T(s_0))$ is surprising this cannot be evidence of a prior-data conflict, but rather indicates a problem with the sampling model. Since we are assuming here that the sampling model has passed its checks, we want to avoid this possibility.

A simple answer would be to simply not choose any ancillary U. But note that the prior-predictive distribution of T will also be affected by aspects that are ancillary. In other words, $T(s_0)$ may be a surprising value because $U(T(s_0))$ is surprising for some ancillary U and we want to avoid this. Accordingly, it makes sense to compare the observed value $T(s_0)$ with its prior-predictive distribution given the value $U(T(s_0))$, i.e., we remove the variation in the prior-predictive distribution of T, that is associated with U, by conditioning on U.

Note that we avoid the necessity of conditioning on an ancillary whenever we have a complete minimal sufficient statistic T. In this situation Basu's theorem implies that every ancillary U is independent of T and so the prior predictive distribution of T does not exhibit variation that can be ascribed to U. Therefore, for the examples discussed in Section 2 we do not need to condition on ancillaries.

If $U_1 \circ T$ and $U_2 \circ T$ are ancillary and $U_1 \circ T = h \circ U_2 \circ T$ for some h then we only need to do the check based on the prior-predictive distribution for T given $U_2(T(s_0))$ since this is

conditioning on more information and so removing more variation. When an ancillary U_1 is such that there is no function h and ancillary U_2 (other than a choice where h is one-to-one) such that $U_1 \circ T = h \circ U_2 \circ T$ then U_1 is a *maximal ancillary*. So we want to condition on maximal ancillaries.

We note, however, as discussed in Lehmann and Scholz (1992), the lack of a general method for obtaining maximal ancillaries. Still if we have an ancillary U available it seems better to condition on U, and so remove the variation due to U, even if we cannot determine whether or not U is maximal ancillary or determine a maximal ancillary based on U.

Furthermore, it can happen that there is more than one maximal ancillary in a problem, e.g., see Lehmann and Scholz (1992). The lack of a unique maximal ancillary can cause problems in frequentist approaches to inference because inference about θ can be different, depending on which ancillary we choose to condition on. This does not cause a problem here, however, for if the check based on conditioning on one maximal ancillary indicates that a prior-data conflict exists, then a conflict exists irrespective of what happens when we condition on other maximal ancillaries. In essence we are considering different aspects of the prior-predictive distribution when we condition on different maximal ancillaries.

The following example is presented in Cox and Hinkley (1974) as a compelling argument for the use of ancillary statistics in inference.

Example 5. *Mixture*

Suppose we have that a response x is either from a $N(\theta, \sigma_0^2)$ or $N(\theta, \sigma_1^2)$ distribution where $\theta \in R^1$ is unknown and σ_0^2, σ_1^2 are both known and unequal. Suppose these two distributions correspond to the variation exhibited by two measuring instruments and the particular instrument used is chosen according to $c \sim \text{Bernoulli}(p_0)$ where p_0 is known. Then we see that (c, x) is minimal sufficient and c is ancillary. We place a $N(\theta_0, 1)$ prior on θ. Therefore, when $c = 0$ we would use the prior predictive based on the prior and the $N(\theta, \sigma_0^2)$ family—namely $x \sim N(\theta_0, \sigma_0^2 + 1)$—and when $c = 1$ we would use the prior predictive based on the prior and the $N(\theta, \sigma_1^2)$ family—namely $x \sim N(\theta_0, \sigma_1^2 + 1)$—to check for prior-data conflict.

This example nicely illustrates the use of ancillary statistics; there are two "components" to the prior predictive distribution of the minimal sufficient statistic and the ancillary statistic selects one. The data x may lead to a prior-data conflict for both components, one component or neither component. This will depend on the prior and the values of σ_0^2 and σ_1^2. Also, if p_0 is very small and we observed $c = 1$, then we consider this as evidence of a problem with the model, not an indication that a prior-data conflict exists (see Section 6).

The following example is presented in Cox and Hinkley (1974) as a case where the use of ancillary statistics leads to ambiguity for inferences about the model parameter.

Example 6. *Special Multinomial Distribution*

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 cell 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 $(f_1, f_2, f_3, f_4) \mid U_1$ is given by $f_1 \mid U_1 \sim \text{Binomial}(f_1 + f_2, (1-\theta)/2)$ independent of

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

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

This distribution is 2-dimensional and can be used as described in Example 4 (Location-scale normal) for prior-data conflict checking. We can also use the 1-dimensional distributions $f_1 \mid U_1$ and $f_3 \mid U_1$.

For example, suppose the prior π is the distribution given by $\theta = 1 - 2U$ where $U \sim \text{Beta}(\alpha, \beta)$. When $\alpha = \beta = 20$, so the prior concentrates about 0, and $f_1 + f_2 = 10$, Figure 2 shows the conditional prior predictive probability function for $f_1 \mid U_1$. We see that we will find evidence of a prior-data conflict whenever $f_1 \in \{0, 1, 9, 10\}$. On the other hand when $\alpha = 1, \beta = 20$, so the prior concentrates about 1, and $f_1 + f_2 = 10$, Figure 3 shows the conditional prior predictive probability function for $f_1 \mid U_1$. In this case we will find evidence of a prior-data conflict whenever $f_1 \in \{3, \ldots, 10\}$.

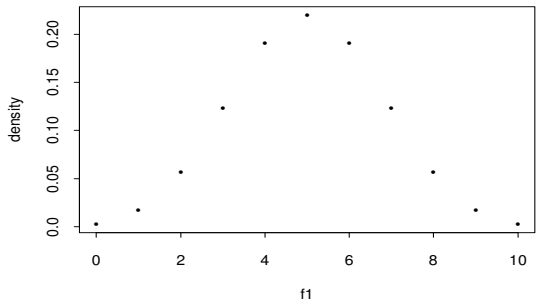

Figure 2: Conditional prior predictive of f_1 given U_1 when $\alpha = \beta = 20$ in Example 6.

Alternatively, we could use $f_1 \mid U_2 \sim \text{Binomial}(f_1 + f_4, (1-\theta)/3)$. In this case, when $\alpha = \beta = 20$ and $f_1 + f_4 = 10$, a similar calculation shows that we will find evidence of a prior-data conflict whenever $f_1 \in \{0, 8, 9, 10\}$. This differs from the previous check involving f_1, but there is no conflict between them. If either check indicates that f_1 is a surprising value, then we have evidence of a prior-data conflict existing. We also have available the checks using $f_3 \mid U_1$ and $f_2 \mid U_2$.

The following example can be considered as an archetype for the situation where an ancillary is determined as the maximal invariant under a transformation group acting on a sample space.

Example 7. *Location Cauchy*

Suppose we have a sample $s = (s_1, \ldots, s_n)$ from a distribution with density proportional to $1/(1+(x-\theta)^2)$ where $\theta \in R^1$. Then it is known that $T(s) = \left(s_{(1)}, \ldots, s_{(n)}\right)$, the order statistic,

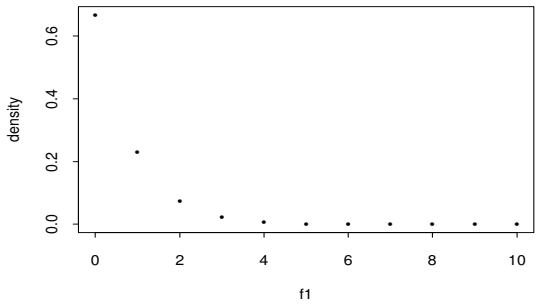

Figure 3: Conditional prior predictive of f_1 given U_1 when $\alpha = 1, \beta = 20$ in Example 6.

is a minimal sufficient statistic. Further we have that $U(T(s)) = (s_{(2)} - s_{(1)}, \ldots, s_{(n)} - s_{(1)}) = (u_1, \ldots, u_{n-1})$ is ancillary. Clearly the conditional distribution of T given U can be expressed as the conditional distribution of $s_{(1)}$ given U and this has conditional density proportional to

$$1/\{(1 + \left(s_{(1)} - \theta\right)^2)\Pi_{i=1}^{n-1}(1 + \left(s_{(1)} + u_i - \theta\right)^2)\}.$$

Note that the normalizing constant will not involve θ and so the conditional prior predictive density of $s_{(1)}$ is

$$\begin{aligned} & m_T(s_{(1)} \mid u_1, \ldots, u_{n-1}) \\ \propto & \int_{-\infty}^{\infty} \frac{1}{\left(1 + \left(s_{(1)} - \theta\right)^2\right)} \prod_{i=1}^{n-1} \frac{1}{\left(1 + \left(s_{(1)} + u_i - \theta\right)^2\right)} \pi(\theta)\, d\theta \\ = & \int_{-\infty}^{\infty} \frac{1}{(1+v^2)} \prod_{i=1}^{n-1} \frac{1}{\left(1 + (v + u_i)^2\right)} \pi\left(s_{(1)} - v\right)\, dv. \end{aligned}$$

Integrating $s_{(1)}$ out of the above expression shows that the normalizing constant is given by $\int_{-\infty}^{\infty}\{(1+v^2)\Pi_{i=1}^{n-1}(1+(v+u_i)^2)\}^{-1}\, dv$. We then check for any prior-data conflict by comparing the observed value of $s_{(1)}$ with the distribution given by $m(\cdot \mid u_1, \ldots, u_{n-1})$.

To evaluate a P-value associated with this conditional prior predictive we must integrate numerically. For example, consider the following ordered sample of $n = 10$ from the Cauchy distribution with $\theta = 0$. Further suppose that $\theta \sim N(0, 1)$.

-4.4829	-2.9692	-0.8915	-0.7164	-0.5501
-0.2805	0.0474	2.1665	4.1467	18.7272

The conditional density $m_T(\cdot \mid u_1, \ldots, u_{n-1})$ is shown in Figure 4. From this we can see, as we would expect, that the observed value $s_{(1)} = -4.4829$ provides no evidence of prior-data conflict.

Suppose, however, the prior distribution is $N(\theta_0, 1)$. In this case $m_T(\cdot \mid u_1, \ldots, u_{n-1})$ looks just like that plotted in Figure 4 but is translated by θ_0 units. So we see that if $|\theta_0|$ is more than 3 then the observed value $s_{(1)} = -4.4829$ will provide evidence of a prior-data conflict.

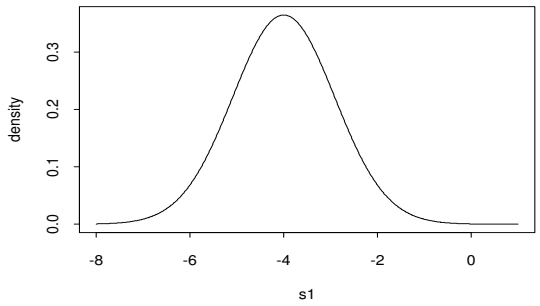

Figure 4: The density function $m_T(\cdot \mid u_1, \ldots, u_{n-1})$ with a $N(0,1)$ prior in Example 7.

As noted in Lehmann and Scholz (1992), there can be a conflict between ancillarity and sufficiency in frequentist inference, as it is not entirely clear which should be imposed first. In other words there are situations where ancillary statistics exist that are not functions of the minimal sufficient statistic. In the context of checking for a prior-data conflict, however, this does not pose a difficulty, as it is clear from Theorem 1 that we only consider functions of the minimal sufficient statistic for this. Of course, for Bayesian inferences about θ the concept of ancillarity has no relevance.

4 Noninformative Priors

Various definitions are available for expressing what it means for a prior to be noninformative. For example, see Kass and Wasserman (1996) for a discussion of the various possibilities and difficulties inherent in this, and also Bernardo (1979) and Berger and Bernardo (1992). A somewhat different requirement for noninformativity arises from considerations about prior-data conflict.

For if a prior is such that we would never conclude that a prior-data conflict exists, no matter what data is obtained, then it seems reasonable to say that such a prior is at least a candidate for being called noninformative. Example 2 (Bernoulli) gives an example where the uniform prior on a compact parameter space satisfies this requirement. We saw in Example 3 (Negative-binomial) that, while the parameter space is contained in a compact set, it is not the case that the uniform distribution is always noninformative.

Strictly interpreted, our requirement only applies to proper priors because M_T is not a probability measure unless Π is proper. We can, however, extend this to sequences of priors that converge in some sense to an improper prior. As in Example 1 (Location normal), we can say that a sequence of priors satisfies this requirement for noninformativity if the P-value we choose to compute converges to 1 for every possible set of data.

In general it is not clear whether a proper prior (or sequence of proper priors) that satisfies this requirement will exist for a given problem. Even if we can show that such objects do exist, it is likely that further requirements must be satisfied for a prior to be called noninformative.

If it is possible, however, for a so-called noninformative prior to conflict with the data in some sense, then it does seem that it is putting some information into the analysis. So we consider the absence of the possibility of any prior-data conflict as a necessary characteristic of noninformativity rather than a characterization.

The following example demonstrates that more than one sequence can satisfy the requirement.

Example 8. *Location normal model*

Consider the context of Example 1 (Location normal) but suppose we take the prior Π_w for θ to be the uniform distribution on $(-w, w)$ for some $w > 0$. The prior predictive density of $T(s) = \bar{s}$ is given by $m_{T,w}(\bar{s}) = (\Phi(\sqrt{n}(w - \bar{s})) - \Phi(\sqrt{n}(-w - \bar{s})))/2w$. It is clear that $m_{T,w}$ is unimodal, with mode at $\bar{s} = 0$, and is symmetric about 0. Then the relevant P-value is given by

$$M_{T,w}(m_{T,w}(\bar{s}) < m_{T,w}(\bar{s}_0)) = 1 - M_{T,w}((-\bar{s}_0, \bar{s}_0))$$
$$= 1 - \int_0^{\bar{s}_0} \frac{1}{w}\left(\Phi\left(\sqrt{n}(w - \bar{s})\right) - \Phi\left(\sqrt{n}(-w - \bar{s})\right)\right) d\bar{s}$$

which converges to 1, as $w \to \infty$, by the dominated convergence theorem.

So in this case we see that a natural sequence of priors converging to the uniform prior satisfies our requirement, as in the limit there is no prior-data conflict. This is also true, as discussed in Example 1 (Location normal), for a sequence of $N(\theta_0, \sigma_0^2)$ priors with $\sigma_0^2 \to \infty$. Both of these sequences give the same limiting posterior inferences.

The discussion in Section 3 indicates that the definition of a noninformative prior must also involve ancillary statistics.

Example 9. *Mixture*

Consider again the context of Example 5 (Mixture) and suppose that we take the prior on θ to be the uniform prior on $(-w, w)$. Then, after conditioning on the ancillary c, the analysis of Example 8 (Location normal) shows that the improper uniform prior satisfies our requirement for noninformativity as the sequence of P-values converges to 1. Notice, however, that if p_0 was small and we didn't condition on c, then if we observed $c = 1$ the P-value would not converge to 1 as $w \to \infty$. This reinforces the necessity of conditioning on ancillaries when considering whether or not a prior-data conflict exists.

In Examples 6 (Special multinomial) and 7 (Location Cauchy) it is not entirely clear what priors will satisfy our requirement for noninformativity. This is partly a technical problem that might yield results with more work, but it is conceivable that in certain problems no such priors or sequences of priors will exist. Further, there is no reason to suppose that, even if such a prior exists, that it will be necessarily unique.

In the following Example 11 we discuss what it means for a sequence of priors to be noninformative for the location-scale normal problem, as discussed in Example 4 (Location-scale normal), in light of our requirement. This presents a more challenging problem than those considered in this section.

5 Diagnostics for Ignoring Prior-data Conflict

Suppose we have concluded that there is evidence of a prior-data conflict. The question then remains as to what to do next. In general this is a difficult question to answer. Modifying the prior in some simple way to avoid the conflict is hard to justify scientifically. One possibility, however, is that we proceed to replace an informative prior by a noninformative prior as discussed in Section 4. Of course, this approach depends on a noninformative prior existing.

One scientifically justifiable answer is that we collect more data, for we know that with sufficient data, the effect of the prior on our inferences is immaterial. Of course, circumstances can arise where collecting more data is not possible, but we would still like to know if, in a given context, we have enough data to ignore the prior-data conflict and, if not, how much more data we need.

When there is a noninformative prior, we can simply compare the posterior inferences obtained via this prior with those obtained using the informative prior. If these inferences do not differ by an amount that is considered to be of practical importance, then it seems reasonable to ignore the prior-data conflict. The amount of difference will depend on the particular application, as a small difference in one context may be considered quite large in another.

Consider the following examples:

Example 10. *Bernoulli model*

Suppose the situation is as described in Example 2 (Bernoulli). Suppose further we are interested in estimating θ and for this we use the posterior mode. Given that the posterior distribution is Beta$(\alpha+n\bar{s}, \beta+n(1-\bar{s}))$ the mode is $(\alpha+n\bar{s}-1)/(\alpha+\beta+n-2)$. A noninformative prior is, as described in Example 2, given by $\alpha = \beta = 1$. A comparison between the posterior modes leads to the difference

$$\left| \frac{\alpha + n\bar{s} - 1}{\alpha + \beta + n - 2} - \bar{s} \right| = \left| \frac{(\alpha - 1)(1 - \bar{s}) - (\beta - 1)\bar{s}}{\alpha + \beta + n - 2} \right|. \tag{7}$$

If (7) is smaller than some prescribed value δ, then we can say that any prior-data conflict can be ignored. Note that (7) is bounded above by $(\alpha + \beta - 2)/(\alpha + \beta + n - 2)$ and so we can determine n so that this difference is always smaller than δ for a given choice of prior. The value δ is such that a difference greater than this is of practical significance and not otherwise. Necessarily, δ depends upon the application and must be specified by the analyst.

Example 11. *Location-scale normal model*

Suppose the situation is as described in Example 4 (Location-scale normal) and we are interested in a .95-HPD interval for μ. With the prior specified as in (5), the posterior distribution of μ is

$$\mu_x + \left(n + 1/\tau_0^2\right)^{-1/2} (\alpha_0 + n/2)^{-1/2} \beta_x^{1/2} t_{2\alpha_0+n},$$

where μ_x, β_x are as specified in Example 4. Therefore, a γ-HDP interval for μ is given by

$$\mu_x \pm \left(n + 1/\tau_0^2\right)^{-1/2} (\alpha_0 + n/2)^{-1/2} \beta_x^{1/2} G^{-1}_{2\alpha_0+n}((1+\gamma)/2) \tag{8}$$

where $G^{-1}_{2\alpha_0+n}$ is the inverse cdf for the $t_{2\alpha_0+n}(0,1)$ distribution.

To determine a noninformative sequence of priors, we must consider

$$M\left(m(\bar{x},s^2)\leq m(\bar{x}_0,s_0^2)\right)$$

as a function of the hyperparameters μ_0,τ_0,α_0, and β_0. From (6), and some easy algebraic manipulations, we see that

$$M\left(m(\bar{x},s^2)\leq m(\bar{x}_0,s_0^2)\right)=M\left(\begin{array}{c}\frac{n}{n-1}\frac{1}{n\tau_0^2+1}\frac{\alpha_0}{\beta_0}\left(\bar{x}-\mu_0\right)^2+\frac{\alpha_0}{\beta_0}s^2\geq\\ \frac{n}{n-1}\frac{1}{n\tau_0^2+1}\frac{\alpha_0}{\beta_0}\left(\bar{x}_0-\mu_0\right)^2+\frac{\alpha_0}{\beta_0}s_0^2\end{array}\right). \tag{9}$$

Now recall that $s^2\sim(\beta_0/\alpha_0)F_{(n-1,2\alpha_0)}$ and so the left side of the inequality in (9) converges almost surely (and so in distribution also) to a random variable with the $F_{(n-1,2\alpha_0)}$ distribution, as $\tau_0^2\to\infty,\beta_0\to 0$ with the added restriction that $\beta_0\tau_0^2\to\infty$. Then, with the additional restriction that $\alpha_0/\beta_0\to 0$, we see that (9) converges to 1. Note that since $\beta_0\to 0$ this also implies that $\alpha_0\to 0$. Therefore a sequence of priors, as specified in Example 4 (Location-scale normal), via the hyperparameters $(\mu_0,\tau_0,\alpha_0,\beta_0)$, will be noninformative for this problem whenever $\tau_0\to\infty,\beta_0\tau_0^2\to\infty$ and $\alpha_0/\beta_0\to 0$. Actually, it is apparent that we need only require that $\alpha_0/\beta_0\to 0$ to have a noninformative sequence of priors, so there are many such sequences.

The condition $\tau_0\to\infty$ is saying that the variance of the prior on μ is becoming large. Since the mean of a Gamma(α_0,β_0) distribution is α_0/β_0, the condition $\alpha_0/\beta_0\to 0$ says that the mass of the prior distribution for $1/\sigma^2$ is concentrating at 0. But since the variance of a Gamma(α_0,β_0) distribution is α_0/β_0^2, the condition $\beta_0\tau_0^2\to\infty$ implies that the concentration must occur at a suitably slow rate. For example, if we take $\beta_0=1/\tau_0$ and $\alpha_0=1/\tau_0^p$ with $p>1$, these conditions will be satisfied as $\tau_0\to\infty$. Note that with these choices the prior variance of $1/\sigma^2$ is $\alpha_0/\beta_0^2=\tau_0^{2-p}$ and so this will become infinite as $\tau_0\to\infty$ whenever $1<p<2$, will remain constant at 1 when $p=2$, and goes to 0 when $p>2$.

Under a sequence of priors, as specified above, we have that the γ-HPD interval for μ given by (8) converges to

$$\bar{x}\pm(s/\sqrt{n})(1-1/n)^{1/2}G_n^{-1}((1+\gamma)/2). \tag{10}$$

This interval differs from the classical interval by using the factor $(1-1/n)^{1/2}G_n^{-1}((1+\gamma)/2)$ instead of $G_{n-1}^{-1}((1+\gamma)/2)$. Note, however, that the $(1-1/n)^{1/2}t_n$ distribution and the t_{n-1} distribution are very similar.

Now we want to compare the two HPD intervals to obtain our diagnostic. Perhaps the most natural comparison is to compute the length measure of the symmetric difference of the two intervals. Certainly something like this would be necessary for general regions. For intervals, however, if one can show that the differences between the corresponding endpoints are less than δ, then the length measure of the symmetric difference is also less than δ. So our diagnostic is the maximum difference between the corresponding endpoints of the intervals given by (8) and (10). If this maximum is less than δ, where δ is a value such that a difference greater than this is of practical significance and not otherwise, then we can ignore any prior-data conflict.

Other diagnostics suggest themselves. For example, in general we could compute the divergence between the posterior under the informative and noninformative prior. For this, however, we would need to first state a cut-off value for the divergence, below which we would not view the difference between the distributions as being material.

6 Factoring the Joint Distribution

As further support for the approach taken here, consider the following factorization of the joint distribution of the data and parameter. This seems to point to a logical separation between the activities of checking the sampling model, checking for prior-data conflict and inference about the model parameter.

In a Bayesian analysis with a proper prior, the full information available to the statistician for subsequent analysis is effectively the observed data s_0 and the joint distribution of (θ, s) as given by the measure $P_\theta \times \Pi$, where P_θ is the conditional distribution of the data s given θ and Π is the marginal distribution of θ. Denoting a minimal sufficient statistic by T we see that this factors as $P(\cdot \mid T) \times P_{T\theta} \times \Pi$, where $P_{T\theta}$ is the marginal probability measure of T and $P(\cdot \mid T)$ is the conditional distribution of the data s given T.

The measure $P(\cdot \mid T)$ is independent of θ and the prior and so only depends on the choice of the sampling model $\{P_\theta : \theta \in \Omega\}$. Therefore, we can compare the observed data s_0 against this distribution to assess whether or not the sampling model is reasonable.

Also we can write $P_{T\theta} \times \Pi$ as $M_T \times \Pi(\cdot \mid T)$ where M_T is the prior predictive distribution of T and $\Pi(\cdot \mid T)$ is the posterior distribution of θ. As we have discussed in this paper, comparing the observed value $T(s_0)$ with the distribution M_T is an appropriate method for assessing whether or not there is a conflict between the prior and the data.

For a maximal ancillary U that is a function of the minimal sufficient statistic T, we can write $M_T \times \Pi(\cdot \mid T)$ as $P_U \times M_T(\cdot \mid U) \times \Pi(\cdot \mid T)$, where P_U is the marginal distribution of U and $M_T(\cdot \mid U)$ is the conditional prior predictive distribution of T given U. We have that P_U depends only on the choice of the sampling model $\{P_\theta : \theta \in \Omega\}$. Therefore we can also compare the observed value $U(s_0)$ with P_U to assess whether or not the sampling model is reasonable. In a case where U is not independent of T, we have argued that it is more appropriate to compare $T(s_0)$ with $M_T(\cdot \mid U)$ than against M_T to assess whether or not a prior-data conflict exists. Conditioning on U removes inappropriate variation (variation that does not depend on the prior) from the comparison.

Finally, when we feel comfortable with both the sampling model and the prior, we can proceed to inferences about θ; for this we use the posterior distribution $\Pi(\cdot \mid T)$. Note that checking for prior-data conflict is only appropriate when we know the sampling model is reasonable, and making inferences about θ is only appropriate when we know both the sampling model is appropriate and that there is no prior-data conflict.

In Box (1980) a method was suggested for model checking based upon the full marginal prior predictive M for the data. In several examples this can be seen to give anomalous results. The approach taken in this paper is a refinement of Box's proposal, for we can factor M as $P(\cdot \mid T) \times M_T$ and see that the first component is available for checking the sampling model and the second is available for checking for prior-data conflict.

7 Hierarchically Specified Priors

Suppose we can write $\theta = (\theta_1, \theta_2)$ and the prior is specified as

$$\pi(\theta_1, \theta_2) = \pi_1(\theta_1)\pi_2(\theta_2 \mid \theta_1),$$

where π_1 is the marginal prior for θ_1 and $\pi_2(\cdot \mid \theta_1)$ is the conditional prior for θ_2 given θ_1. Of course any prior can be written this way, but here we mean that the prior π is constructed by specifying π_1 and π_2. It is natural then to see if we can generalize the methods discussed here for checking the full prior π, to checking the individual components π_1 and π_2 to help identify the source of any conflict more precisely. A prior may be specified with more than two components, but we will restrict our discussion here to the two component case and leave the more general problem to be treated elsewhere. We do not, however, restrict θ_1 and θ_2 to be 1-dimensional.

In Evans and Moshonov (2005) the question of how to check the individual components was considered and a partial solution developed. The solution was partial in the sense that not all such decompositions of π are amenable to the methodology. We note, however, that we can always check the full prior using the methods we have presented so far.

As in the previous sections we suppose that T is the minimal sufficient statistic for the model and $U(T)$ is maximal ancillary for θ. Initially we consider checking π_2 for conflict with the data. Suppose we can find a statistic $V(T)$ that is *ancillary for* θ_2 in the sense that the sampling distribution of $V(T)$ is independent of θ_2 and depends on θ_1. If we conclude that the observed value $T(s_0)$ is a surprising value from $M_T(\cdot \mid U(T(s_0)))$, this could arise because $V(T(s_0))$ is a surprising value from $M_{V(T)}(\cdot \mid U(T(s_0)))$, the conditional prior predictive of $V(T)$ given $U(T)$. The following result is established (the proof is similar to that of Theorem 1) in Evans and Moshonov (2005).

Theorem 3. If $V(T)$ is ancillary for θ_2, then $M_{V(T)}(\cdot \mid U(T))$ does not depend on π_2.

Therefore, a surprising value of $V(T(s_0))$ cannot be due to a conflict with π_2. Now we argue, just as in Section 3, for ancillary statistics. To assess whether or not $T(s_0)$ is conflicting with π_2, the sensible thing to do is to remove the variation in $M_T(\cdot \mid U(T))$ due to $V(T)$ when making the comparison; we do this by conditioning on $V(T)$—namely, we compare $T(s_0)$ to $M_T(\cdot \mid U(T(s_0)), V(T(s_0)))$. Of course we want to remove the maximum amount of this variation and so we take $V(T)$ to be a maximal ancillary for θ_2. Again there could be a number of distinct maximal ancillaries for θ_2 and these all provide valid checks.

We have required that the marginal distribution of $V(T)$ depend on θ_1 and so we can compare $V(T(s_0))$ to $M_{V(T)}(\cdot \mid U(T(s_0)))$ to assess whether or not there is any conflict with π_1. Notice that Theorem 3 implies that this comparison does not depend in any way upon π_2. In contrast $M_T(\cdot \mid U(T(s_0)), V(T(s_0)))$ will generally depend upon π_1 as well as π_2. This suggests that we first check for prior-data conflict with π_1 by comparing $V(T(s_0)$ to $M_{V(T)}(\cdot \mid U(T(s_0)))$ and, if no conflict is found, then proceed to check for prior-data conflict with π_2 by comparing $T(s_0)$ to $M_T(\cdot \mid U(T(s_0)), V(T(s_0)))$. This is analogous to the proviso we stated initially that we don't check for prior-data conflict unless we have first agreed that the sampling model makes sense.

We see that this approach is based on a further factorization of the prior predictive, as $M_T(\cdot \mid U(T)) = M_{V(T)}(\cdot \mid U(T)) \times M_T(\cdot \mid U(T), V(T))$. The first factor is concerned with checking π_1 and the second is concerned with checking π_2. We consider an example.

Example 12. *Location-scale normal model*

Suppose the situation is as described in Example 4 (Location-scale normal) and recall that $T = (\bar{x}, s^2)$. Here π_1 is the prior on $\theta_1 = \sigma^2$ and $\pi_2(\cdot \mid \theta_1)$ is the conditional prior on $\theta_2 = \mu$ given σ^2. We see immediately that s^2 is ancillary for μ. The marginal prior predictive of s^2 is

given by $s^2 \sim (\beta_0/\alpha_0)F_{(n-1,2\alpha_0)}$ and the conditional prior predictive of $\bar{x}$ given s^2 is distributed as $t_{n+2\alpha_0-1}(\mu_0, \tilde{\sigma})$, where

$$\tilde{\sigma}^2 = \left\{\tau_0^2\left(n\tau_0^2+1\right)\left(2\beta_0+(n-1)\,s^2\right)\right\}/\{n\tau_0^2\,(n+2\alpha_0-1)\}.$$

Now consider the numerical values of the hyperparameters and data prescribed in Example 4. To assess if there is any conflict with π_1, we compare $s^2/5 = 0.1673126$ with the $F(19,2)$ distribution. In Example 4 we computed this P-value to be .47832 and this doesn't indicate any prior-data conflict. To assess if there is any conflict with π_2, we compare $(\bar{x}-\mu_0)/\tilde{\sigma} = -31.81617$ to the $t_{21}\,(0,1)$ distribution. This is clearly a very extreme value and in fact the two-sided P-value is 0 to 7 decimals. This check has appropriately detected the discrepancy between the prior and the location of the data. In Example 4 we also considered using the marginal prior predictive distribution of $\bar{x}$ (a Student(2) distribution) to assess whether or not there was any prior-data conflict. This resulted in a P-value of .0021 and so gave evidence of a prior-data conflict. Intuitively, this seemed to indicate a problem with the specification of the part of the prior for μ, but there was no clear rationale for this. We see now, however, that the hierarchical approach indicates a very clear problem with the specification of π_2 and does so much more dramatically.

We might ask, however, under what circumstances is it possible for the check for π_2 to be independent of π_1, so we could do the checks in any order. One set of such conditions can arise whenever there is a statistic $V(T)$ that is sufficient-ancillary for (θ_1,θ_2), namely, $V(T)$ is ancillary for θ_2 and the conditional distribution of T given $V(T)$ does not depend on θ_1. The following result is proved in Evans and Moshonov (2005).

Theorem 4. Suppose that $\Omega = \Omega_1 \times \Omega_2$ with θ_1 and θ_2 a priori independent, and $V(T)$ is sufficient-ancillary for (θ_1,θ_2). Then $M_T(\cdot \mid U(T), V(T))$ is independent of π_1.

In general, as noted in Fraser (1979), it is difficult to find sufficient-ancillary statistics. For example, s^2 is not sufficient-ancillary in Example 12 (Location-scale normal). In Evans and Moshonov (2005) the multinomial model was shown to possess a sufficient-ancillary statistic after a reparameterization.

The need to specify an order for the checking, except under the rather specialized conditions in Theorem 4, and further the need for a statistic $V(T)$ ancillary for θ_2, show that checking for individual components may not be available generally. There is, however, a fairly wide class of models and decompositions where such a $V(T)$ exists. In Evans and Moshonov (2005), an explicit construction is given for $V(T)$ when the basic statistical model corresponds to a group model with a specific structure. In particular, when the parameter space is a group that can be written as $\Omega = \Omega_2\Omega_1$ with $\theta_1 \in \Omega_1, \theta_2 \in \Omega_2$ and Ω_1, Ω_2 subgroups with the product being a semidirect product, then $V(T)$ is easily obtained. Example 12 (Location-scale normal) exhibits this structure and the construction procedure leads to $V(\bar{x}, s^2) = s$. In fact, $V(\bar{x}, s^2) = s$ works for any location-scale model, with the same prior decomposition, and there are other possible choices. The location-scale Cauchy model is analyzed for prior-data conflict in Evans and Moshonov (2005) when θ_1 is the scaling parameter and θ_2 is the location parameter.

As documented in Fraser (1979), there are many models in statistics that have this structure. For example, suppose we have a regression model where a basic observation $y \in R^1$ has $E(y) = x^t\theta_2, Var(y) = \theta_1$ with $\theta_2 \in R^p$ unknown, $\theta_1 > 0$ unknown and the distribution otherwise fully specified (e.g., normal). Then taking the group product to be $(\theta_1,\theta_2)\,(\theta_1',\theta_2') =$

$(\theta_1\theta_1', \theta_2+\theta_1\theta_2')$ we have that this structure obtains. This can also be generalized to multivariate regression with any error distribution.

There is another context where priors are specified hierarchically—namely, hierarchical modeling. We make a slight change of notation here and suppose that $\theta_2 \in \Omega_2$ denotes the sampling model parameter and $\theta_1 \in \Omega_1$ denotes a set of hyperparameters that specify the prior for θ_2 as $\pi_2(\cdot \mid \theta_1)$. Further, suppose we have completed the specification of the prior via π_1. This induces the prior measure $\Pi_2^*(A) = \int_{\Omega_1} \Pi_2(A \mid \theta_1)\, \Pi_1(d\theta_1)$ for $A \subset \Omega_2$ and all the methods we have been discussing, based on the minimal statistic T for the model $\{P_{\theta_2} : \theta_2 \in \Omega_2\}$, are available to check whether or not Π_2^* conflicts with the data. Again we might like to check for conflicts with the individual components of the prior, but the situation is different than the previous problem because θ_1 is not part of the model parameter. Therefore the methods we have discussed in this section so far, are not available for this problem and a different approach is needed.

The joint distribution of (θ_1, θ_2, s) can be factored as $P(\cdot \mid T) \times P_{T\theta_2} \times \Pi_2(\cdot \mid \theta_1) \times \Pi_1$ and we recall that $P(\cdot \mid T)$ is used to check the sampling model $\{P_{\theta_2} : \theta_2 \in \Omega_2\}$. Now observe that there is another way to at least formally generate a model for s from the joint distribution—namely, put

$$\begin{aligned} M_{\theta_1}(ds) &= \int_{\Omega_2} P_{\theta_2}(ds)\, \Pi_2(d\theta_2 \mid \theta_1) = P(ds \mid T)(t) \int_{\Omega_2} P_{T\theta_2}(dt)\, \Pi_2(d\theta_2 \mid \theta_1) \\ &= P(ds \mid T)(t) \times M_{T\theta_1}(dt). \end{aligned}$$

This model is only formal as, strictly speaking, when the model $\{P_{\theta_2} : \theta_2 \in \Omega_2\}$ is correct, it is not the case that $s \sim M_{\theta_1}$ for some $\theta_1 \in \Omega_1$, except in certain special circumstances. Note that M_{θ_1} is the conditional prior predictive distribution for s given θ_1 and $M_{T\theta_1}$ is the conditional prior predictive distribution for T given θ_1.

Now consider the model for T given by $\{M_{T\theta_1} : \theta_1 \in \Omega_1\}$ and let $V(T)$ be a minimal sufficient statistic for this model. Then we can factor $M_{T\theta_1}$ as $M(\cdot \mid V) \times M_{V\theta_1}$, where $M(\cdot \mid V)$ is the conditional prior predictive distribution of T given V, and $M_{V\theta_1}$ is the conditional prior predictive distribution of V given θ_1. Then the joint distribution of (θ_1, s) can be factored as

$$P(\cdot \mid T) \times M(\cdot \mid V) \times M_V \times \Pi_1(\cdot \mid V), \tag{11}$$

where M_V is the prior predictive distribution of V, and $\Pi_1(\cdot \mid V)$ is the posterior distribution of θ_1. Note that a simple argument establishes that the last three factors in (11) are the same, whether determined from the joint distribution of (θ_1, θ_2, s) or the joint distribution of (θ_1, s). In particular, the posterior distribution of θ_1 only depends on the data through $V(T(s_0))$.

Now, using the arguments developed in this paper, consider how each of the factors in (11) is to be used. First, $P(\cdot \mid T)$ is available for checking the basic sampling model $\{P_{\theta_2} : \theta_2 \in \Omega_2\}$. If no evidence is found against the model, we can proceed to check the model $\{M_{T\theta_1} : \theta_1 \in \Omega_1\}$ for T using $M(\cdot \mid V)$ and note that this does not depend on Π_1. If evidence is found against this model then, because we have accepted the sampling model, and so consequently the model $\{P_{T\theta_2} : \theta_2 \in \Omega_2\}$ for T, this must occur because of a conflict between the observed value $T(s_0)$ and Π_2. If we find no evidence against $\{M_{T\theta_1} : \theta_1 \in \Omega_1\}$, then we can check for a conflict with Π_1 using M_V. Finally, if there is no conflict with Π_1, then $\Pi_1(\cdot \mid V)$ is available for inference about θ_1. Of course, if there is no conflict with Π_1 and Π_2, then we can also make inference about the parameter of interest θ_2. We consider an example.

Example 13. *Random effects*

Suppose that $s = (s_1, \ldots, s_n)$ is a sample from the $N_p(\mu, I)$ distribution where $\mu \sim N_p(\mu_0, \sigma^2 I)$ with $\mu_0 \in R^p$ fixed and $\sigma^{-2} \sim \chi^2_{\alpha_0}$ with α_0 fixed. So here, $\theta_2 = \mu$ and $\theta_1 = \sigma^2$. Then $T = \bar{s}$, $M_{T\theta_1}$ is the $N_p(\mu_0, (\theta_1 + 1/n)I)$ distribution and $V = (\bar{s} - \mu_0)^t(\bar{s} - \mu_0)$.

Given $V = v$, the conditional distribution of $\bar{s}$ is uniform on the sphere of radius $v^{1/2}$ centered at μ_0. So in this case $M(\cdot \mid V)$ would seem to imply that we will never find evidence against the $N_p(\mu_0, \sigma^2 I)$ factor in the prior, at least when we compute P-values as we have prescribed. At first this seems anomalous, but consider that this factor allows for any value for σ^2; by an appropriate choice of σ^2 we can avoid any prior-data conflict, in the sense that the likelihood and prior support for μ will overlap. Whether or not we have specified appropriate possible values for σ^2 will depend only on the prior Π_1 for σ^2. Contrast this with Example 12, where the sampling model depends on both θ_1 and θ_2 while the sampling model only depends on θ_2 here. In Example 12 there will also be values of θ_1 such that the prior for θ_2 will not conflict with the data, but these values of θ_1 may not be realistic in light of the likelihood. So in Example 12 a conflict may exist with Π_2 irrespective of the prior placed on θ_1.

We have that the sampling distribution of $(\sigma^2 + 1/n)^{-1}(\bar{s} - \mu_0)^t(\bar{s} - \mu_0)$ is χ^2_p. So M_V has density m_V given by

$$\frac{\Gamma(p/2 + \alpha_0/2)}{\Gamma(p/2)\,\Gamma(\alpha_0/2)} v^{p/2-1} \int_0^\infty \left(\frac{1}{1+u/n}\right)^{p/2} \exp\left\{-\frac{1}{2}\frac{v}{1+u/n}\right\} g(u)\,du,$$

where g is the $\chi^2_{p+\alpha_0}$ density. Although nonstandard, this is easily numerically evaluated on a grid of values and the relevant P-value computed.

8 Conclusions

This paper has been concerned with checking whether or not there is any conflict between the prior distribution Π and the observed data s_0. Conflict here means that the prior distribution assigns most of its mass to θ values prescribed by the sampling model $\{P_\theta : \theta \in \Omega\}$ for which the observed data is surprising. Various considerations lead us to conclude that the appropriate approach to making this comparison is to compare the observed value $T(s_0)$ of a minimal sufficient statistic T with the conditional prior predictive distribution of T given any maximal ancillary U. If many such nonequivalent maximal ancillaries are available, then all such comparisons are deemed appropriate.

We note that the above analysis implicitly assumes that the sampling model for the data is correct. Of course, this is an assumption and must also be checked. We have not discussed this issue here but, as noted in the Introduction, there are many methods available for this and we have assumed that this process has been carried out first. Given the different consequences of sampling model failure and prior-data conflict, and the ability of a Bayesian analysis to avoid the consequences of a prior-data conflict when there is sufficient data, it is felt that it is important to assess these potential failures separately.

As mentioned in the Introduction some may feel that when a prior reflects the subjective beliefs of an analyst, then they do not feel compelled to check for prior-data conflict. Furthermore, it might be argued that posterior predictive model checks could be used to check for the plausibility of inferences rather than the diagnostic approach that we have proposed. We note, however, that at least part of the motivation for the approach we are advocating for model checking and checking for prior-data conflict, as reflected in the factorization discussed in Section 6, is the avoidance of the kind of phenomena discussed in Bayarri and Berger (2000)

that led them to argue against the general use of posterior predictive checks when checking the sampling model. We acknowledge that there are currently differing viewpoints on these issues, but feel that our approach of separating model checking and checking the prior has some attractive features, as this paper has hopefully demonstrated.

The developments here have led to a number of new developments, including a role for the ancillarity concept in Bayesian analysis, and for a necessary requirement for a prior to be noninformative. While this paper has laid the foundations for this approach, further work is required to fully investigate all the implications. Also important computational issues arise in contexts where computing the conditional prior predictive densities of a minimal sufficient statistic and associated P-values are not straightforward. Typically these problems are similar to those encountered when we want to compute normalizing constants in Bayesian problems and presumably similar approaches can be brought to bear.

9 Appendix

Proof of Theorem 1:

Suppose $T : S \to \mathcal{T}$ possesses the appropriate measurability properties. Since T is sufficient we have, for measurable $B \subset S$, and any T-measurable $C \subset \mathcal{T}$, that $P_\theta(B \cap C) = \int_C P(B \mid T)(t)\ P_{\theta T}(dt)$ for T-measurable $P(B \mid T) : \mathcal{T} \to [0,1]$. Then

$$\begin{aligned} & M(B \cap C) = \int_C M(B \mid T)(t)\ M_T(dt) = \int_\Omega P_\theta(B \cap C)\,\Pi(d\theta) \\ = & \int_\Omega \int_C P(B \mid T)(t)\ P_{\theta T}(dt)\ \Pi(d\theta) = \int_C P(B \mid T)(t) \int_\Omega P_{\theta T}(dt)\ \Pi(d\theta) \\ = & \int_C P(B \mid T)(t)\ M_T(dt) \end{aligned}$$

for every T-measurable $C \subset \mathcal{T}$. This implies that $M(B \mid T)(t) = P(B \mid T)(t)$ almost everywhere with respect to M_T and we have proved the result.

Proof of Theorem 2:

Let $A = \{\theta : \pi(\theta) > 0\}$ and note that $L(\cdot \mid s_0) \propto L(\cdot \mid T(s_0))$. Then,

$$m_T(T(s_0)) = \int_\Omega f_{\theta T}(T(s_0))\,\pi(\theta)\ \upsilon(d\theta) = \int_A f_{\theta T}(T(s_0))\,\pi(\theta)\ \upsilon(d\theta) = 0,$$

since $L(\theta \mid T(s_0)) = f_{\theta T}(T(s_0)) = 0$ when $\theta \in A$.

References

Bayarri, M. J. and Berger, J. (2000). "P values for composite null models." *Journal of the American Statistical Association*, 95(452): 1127–1142.

Berger, J. and Bernardo, J. (1992). "On the development of the reference prior method." In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, 218–220. Clarendon Press, Oxford, U.K.

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

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

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Evans, M. and Moshonov, H. (2005). "Checking for prior-data conflict with hierarchically specified priors." Technical Report 0503, Dept. of Statistics, U. of Toronto, Toronto. To appear in the proceedings of the International Workshop/Conference on Bayesian Statistics and its Applications, Dept. of Statistics, Banaras Hindu U., Varanasi, India.
URL http://fisher.utstat.toronto.edu/mikevans/papers/techrep3.pdf

Fraser, D. (1979). *Inference and Linear Models*. New York: McGraw-Hill.

Gelman, A., Meng, X., and Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies (with discussion)." *Statistica Sinica*, 6: 733–808.

Guttman, I. (1967). "The use of the concept of a future observation in goodness-of-fit problems." *Journal of the Royal Statistical Society, B*, 143: 383–430.

Kass, R. E. and Wasserman, L. (1996). "The selection of prior distributions by formal rules." *Journal of the American Statistical Association*, 91(435): 1343–1370.

Lehmann, E. and Scholz, F. (1992). "Ancillarity." In Ghosh, M. and Pathak, P. K. (eds.), *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, 32–51. IMS Lecture Notes-Monograph Series, Hayward, CA.

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

Acknowledgments

Professors Jim Berger and Arnold Zellner provided some useful comments on an orally presented version of this paper. We also thank the Associate Editor for a number of constructive comments.