Checking for Prior-Data Conflict with Hierarchically Specified Priors

by

Michael Evans
Department of Statistics
University of Toronto

and

Hadas Moshonov
Department of Statistics
University of Toronto

Technical Report No. 0503, June 12, 2005

TECHNICAL REPORT SERIES
UNIVERSITY OF TORONTO
DEPARTMENT OF STATISTICS

Checking for Prior-Data Conflict with Hierarchically Specified Priors

Michael Evans and Hadas Moshonov
University of Toronto

Abstract

Priors are often specified component-wise. This may entail placing independent priors on parameter components or specifying the prior in a sequential or hierarchical fashion. We consider methods for checking for the source of any prior-data conflict in the individually specified components of the prior.

1 Introduction

Virtually all statistical analyses are dependent on choices made by the analyst. In Bayesian analyses these choices take the form of the sampling model and the prior. If inferences drawn from the model, prior and data are to have validity, then it is important that we feel confident that the choices made make sense in light of the observed data.

Model checking in this context has been discussed in Guttman (1967), Box (1980), Rubin (1984), Gelman, Meng and Stern (1996), Bayarri and Berger (2000) and Johnson (2004). Most of these papers considered the effect of both the sampling model and the prior simultaneously while Bayarri and Berger (2000) focused on the sampling model.

As pointed out in Evans and Moshonov (2005), however, there are two possible ways in which the Bayesian model can fail. The sampling model may fail, in the sense that the observed data is surprising for each of the assumed distributions in the model or, when the sampling model is appropriate, the prior may place its mass primarily on distributions in the sampling model for which the observed data is surprising. We refer to the latter failure as *prior-data conflict*. It is also argued in Evans and Moshonov (2005) that it is important to check for these possible failures separately as, when there is sufficient data, prior-data conflict can be ignored as the impact of the prior on inference disappears with increasing amounts of data.

The developments in Evans and Moshonov (2005) lead to basing the assessment of whether or not a prior-data conflict exists, on the prior predictive distribution for the minimal sufficient statistic T for the model. By this we mean that, if the observed value of T is a surprising (out in the tails) value for

the prior predictive distribution of T then we have evidence that a prior-data conflict exists. Part of the justification for this is that the prior predictive distribution of the full data given T does not depend on the prior and so can tell us nothing about the existence of prior-data conflict.

Further, it was shown in Evans and Moshonov (2005) that when there is an ancillary statistic $U(T)$, i.e., an ancillary that is a function of T, then it is necessary to replace the prior predictive of T by the conditional prior predictive of T given $U(T)$. This removes variation in the prior predictive that does not depend on the particular prior used and so results in a more accurate assessment of a prior-data conflict.

We denote the sample space for the data s by S, the parameter space by Ω, the sampling model by the collection of densities $\{f_\theta : \theta \in \Omega\}$ with respect to some support measure μ, and the prior probability measure on Ω by Π, with density π with respect to a support measure υ. The developments in Evans and Moshonov (2005) then lead to the factorization of the full joint distribution of (θ, s) as

$$P(\cdot \mid T) \times P_{U(T)} \times M_T(\cdot \mid U(T)) \times \Pi(\cdot \mid T), \tag{1}$$

where $P(\cdot \mid T)$ is the conditional distribution of the data given T, $P_{U(T)}$ is the marginal distribution of $U(T)$, $M_T(\cdot \mid U(T))$ is the conditional prior predictive of T given $U(T)$, and $\Pi(\cdot \mid T)$ is the posterior distribution of the parameter θ. As $P(\cdot \mid T)$ and $P_{U(T)}$ depend only on the model $\{f_\theta : \theta \in \Omega\}$ they are available for checking the sampling model, $M_T(\cdot \mid U(T))$ is available for checking for prior-data conflict and finally $\Pi(\cdot \mid T)$ is available for inference about θ. In effect each component of this factorization plays a role in a statistical analysis quite separate from the others. First we check for the correctness of the sampling model, if no evidence is found that the sampling model is incorrect, then we proceed to check for prior-data conflict, and finally proceed to inference about θ if no evidence of prior-data conflict is obtained.

The end result of deciding that there is evidence that the model is not appropriate for the data, or that there is evidence of a prior-data conflict can, however, vary depending on the circumstances. For example, it may be that we decide that, even though there is evidence for a prior-data conflict, the amount of data is sufficient so that it can be ignored and we can proceed to inference about θ using $\Pi(\cdot \mid T)$. Diagnostics are developed in Evans and Moshonov (2005) to assess this. A similar comment applies to checking for correctness of the sampling model. Throughout the remainder of the paper we assume that the appropriate checks have been applied to the sampling model and the outcome of this is such that we are prepared to proceed to check for prior-data conflict. We note that if our conclusion from the first stage is that the sampling model is definitely not appropriate, then it would not make any sense to proceed to checking for prior-data conflict.

As discussed in Evans and Moshonov (2005), it is also apparent that, if a so-called non-informative prior can result in a prior-data conflict, then the prior is indeed adding some information into the analysis. So the lack of any possibility for a prior-data conflict to exist, is at least a partial characterization of what

it means for a prior to be noninformative. Since the characterization of prior-data conflict depends on the prior predictive of T, and this is only a proper probability model when π is proper, our discussion of this is initially limited to proper priors. This is extended, however, to improper priors by looking at the limits of sequences of proper priors and determining when these sequences are noninformative.

In this paper we are concerned with not just determining whether or not a prior-data conflict exists, but in determining the source of that conflict. For example, suppose that the prior π is specified sequentially, or hierarchically, as

$$\pi(\theta) = \pi_1(\theta_1)\,\pi(\theta_2 \mid \theta_1) \cdots \pi(\theta_p \mid \theta_1, \theta_2, \ldots, \theta_{p-1}) \tag{2}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$. Then if we decide that a prior-data conflict exists it is reasonable to ask if it is some particular component, or components, of π that are causing this and not others. The methods discussed in Evans and Moshonov (2005) only deal with the full prior rather than individual components. Further we extend the criterion for noninformativity to the component parameter case.

The methods we develop here for checking for individual components will be seen not to apply to a general decomposition of a prior as written in (2). In particular, we will restrict our attention to exponential statistical models and group statistical models and even in those contexts only certain decompositions will be seen to be amenable to our methodology. Also we will restrict our attention here to the case where $p = 2$. Note, however, that this does not require that the θ_i be 1-dimensional parameters. For example θ_1 could be k-dimensional and θ_2 l-dimensional. We will consider more general decompositions in a further paper.

Although there is a certain natural logic to the methods developed here, we can't say that there isn't an approach capable of handling a general decomposition. It is also possible, however, that the methods developed here do place some restrictions on how we decompose a prior for hierarchical specification, at least when we want to assess individual components for conflict with the data. In any case, the methods developed in Evans and Moshonov (2005) are always available for checking the full prior.

2 Checking Prior Components

Suppose that the prior has been specified as in (2). Rather than assessing whether or not the full prior is in conflict with the data we consider instead the problem of assessing whether or not each component, as specified in (2), is in conflict. This approach results in a more refined understanding of how a prior conflicts with the data when such a conflict exists.

When assessing prior-data conflict for the full prior, we first reduced to the prior predictive distribution of the minimal sufficient statistic T. This reduction, as argued in Evans and Moshonov (2005), was based in part on the fact that the prior predictive distribution of the data given T does not depend on the prior and so the data beyond the value of $T(s)$, can tell us nothing about whether

a prior-data conflict exists. To determine whether or not a component of the prior is in conflict it seems sensible then to first reduce to the prior predictive distribution of T, denoted as

$$M_T(B) = \int_B \int_\Omega f_{\theta T}(t)\pi(\theta)\ \upsilon(d\theta)\ \lambda(dt) = \int_B m_T(t)\ \lambda(dt),$$

where λ is a support measure on the target space for T.

To assess whether or not the full prior is in conflict we then compare the observed value $T(s)$ with M_T to see if it is a surprising value and, if so, conclude that a prior-data conflict exists. In addition, if $U(T)$ is ancillary for θ then we replace M_T in this comparison by the conditional distribution of T given $U(T)$, namely $M_T(\cdot \mid U(T))$. This has the effect of removing variation in M_T that is only dependent on the sampling model and so makes for a more accurate assessment of whether or not $T(s)$ is surprising. Further we take $U(T)$ to be a maximal ancillary so that the maximum amount of unrelated variation is removed from the comparison. More details on this, and many examples, can be found in Evans and Moshonov (2005).

We generalize checking for prior-data conflict for the full prior to components by generalizing the concept of ancillarity according to the following definition.

Definition 1. A function V of the minimal sufficient statistic T is said to be *ancillary for* θ_2 if the distribution of $V(T)$ depends on θ_1 but not θ_2.

Note that if $U(T)$ is ancillary for θ and $V(T)$ is ancillary for θ_2 then the conditional distribution of $V(T)$ given $U(T)$ also does not depend upon θ_2. We have also required that the distribution of $V(T)$ depend on θ_1 so that $V(T)$ is not ancillary for the full parameter.

So suppose $\theta = (\theta_1, \theta_2)$ and $\pi(\theta) = \pi_1(\theta_1)\,\pi_2(\theta_2 \mid \theta_1)$. This kind of decomposition typically arises in hierarchical modeling when θ_2 is the parameter of interest (not necessarily 1-dimensional) and θ_1 is comprised of either nuisance parameters or hyperparameters. In many situations θ_2 is of central interest so that this decomposition is not arbitrary.

Initially we consider checking for any prior-data conflict with the prior π_2 on θ_2. If $V(T)$ is ancillary for θ_2 then, just as with an ancillary for the full parameter, we could decide that $T(s)$ is a surprising value from the prior predictive distribution $M_T(\cdot \mid U(T))$, simply because $V(T(s))$ is a surprising value from its conditional prior predictive distribution $M_{V(T)}(\cdot \mid U(T))$. While this is appropriate in assessing whether any prior-data conflict exists, such a conflict cannot be caused by the component π_2, because the prior predictive distribution (conditional on $U(T)$ or otherwise) of $V(T)$ does not depend on π_2. Accordingly, to assess whether or not there is any prior-data conflict caused by the choice of π_2 we must compare $T(s)$ with $M_T(\cdot \mid U(T), V(T))$, as this removes the variation in $M_T(\cdot \mid U(T))$ due to $V(T)$.

Again, we want to remove the maximum amount of this ancillary for θ_2 variation in making this comparison so we choose $V(T)$ to be a maximal ancillary for θ_2, namely, require that $V(T)$ not be a function of any statistic that is also ancillary for θ_2. As discussed in Lehmann and Scholz (1992) for full ancillaries,

there is currently no general method for constructing maximal ancillaries for θ_2 or even proving that a given ancillary is maximal. Further it is conceivable that there is more than one maximal ancillary for θ_2. We note, however, as in Evans and Moshonov (2005), the lack of a unique maximal ancillary for θ_2 does not cause a problem in this context, because two different maximal ancillaries $V_1(T)$ and $V_2(T)$ simply represent different methods of removing variation from the full prior predictive and so provide different assessments. In other words, if the assessment via $M_T(\cdot \,|\, U(T), V_1(T))$ provides evidence of a conflict and that via $M_T(\cdot \,|\, U(T), V_2(T))$ does not (or conversely), this is not a problem as we would conclude that a prior-data conflict due to π_2 exists. In such a situation these tests do not contradict one another, they simply represent different methods of slicing up the conditional prior predictive $M_T(\cdot \,|\, U(T))$ to see if $T(s)$ looks anomalous.

We also note that the conditional prior predictive distribution for $V(T)$, namely, $M_{V(T)}(\cdot \,|\, U(T))$, is available for checking whether or not there is any conflict due to π_1. We have the following result.

Theorem 1. If $V(T)$ is ancillary for θ_2 then $M_{V(T)}(\cdot \,|\, U(T))$ does not depend on π_2.

Proof: Let $\Omega_2(\theta_1) = \{\theta_2 : (\theta_1, \theta_2) \in \Omega\}$, then we have that

$$\begin{aligned}
& M_{V(T)}(B \,|\, U(T)) \\
= & \int_{\Omega} P_{V,\theta}(B \,|\, U(T)) \pi(\theta)\, \upsilon(d\theta) \\
= & \int_{\Omega_1} \int_{\Omega_2(\theta_1)} P_{V,\theta_1}(B \,|\, U(T)) \pi_1(\theta_1)\, \pi_2(\theta_2 \,|\, \theta_1)\, \upsilon_2(d\theta_2)\, \upsilon_1(d\theta_1) \\
= & \int_{\Omega_1} P_{V,\theta_1}(B \,|\, U(T)) \pi_1(\theta_1) \left(\int_{\Omega_2(\theta_1)} \pi_2(\theta_2 \,|\, \theta_1)\, \upsilon_2(d\theta_2) \right) \upsilon_1(d\theta_1) \\
= & \int_{\Omega_1} P_{V,\theta_1}(B \,|\, U(T)) \pi_1(\theta_1)\, \upsilon_1(d\theta_1)
\end{aligned}$$

which establishes the result. ∎

We see then that checking for the individual prior components has lead to a further factorization of $M_T(\cdot \,|\, U(T))$ as

$$M_T(\cdot \,|\, U(T)) = M_{V(T)}(\cdot \,|\, U(T)) \times M_T(\cdot \,|\, U(T), V(T)). \tag{3}$$

The availability of this decomposition is dependent on the existence of a statistic $V(T)$ that is ancillary for the parameter of interest θ_2 and in general there is nothing to guarantee this. In Sections 3 and 4, however, we examine some quite general contexts where this is the case.

It is notable that the existence of an appropriate factorization is not dependent on the form of the prior, only on how we specify the prior sequentially. For example, if we specify that θ_1 and θ_2 are a priori independent there is no simplification and we still need to specify that θ_2 is the parameter of primary

interest. Further, if instead we specify the prior as $\pi(\theta) = \pi_2(\theta_2)\,\pi_1(\theta_1 \mid \theta_2)$, then the analysis will be different. Overall the methods we are developing here are meant to be applicable when we have specified the prior according to a specific hierarchical structure and we have reasons for choosing this. If we do not have such a structure, then we can still use the methods of Evans and Moshonov (2005) to assess the overall prior.

Suppose now that we have a statistic $V(T)$ that is ancillary for θ_2 and sufficient for θ_1, i.e., $V(T)$ that is ancillary for θ_2 and the conditional distribution of T given $V(T)$ does not depend on θ_1. Such a statistic is called sufficient-ancillary for (θ_1, θ_2) in Fraser (1979). In this case, when θ_1 and θ_2 are a priori independent, we have the following result.

Theorem 2. Suppose that $\Omega = \Omega_1 \times \Omega_2$, the prior density is given by $\pi(\theta_1, \theta_2) = \pi_1(\theta_1)\,\pi_2(\theta_2)$ and that $V(T)$ is sufficient-ancillary for (θ_1, θ_2). Then $M_T(B \mid U(T), V(T))$ is independent of π_1.

Proof: Denoting the range space for V by $\mathcal{V}$, and the support measure on Ω_i by υ_i, we have that $M_T(\cdot \mid U(T), V(T))$ satisfies,

$$\begin{aligned}
& M_T(B \mid U(T)) \\
= & \int_{\mathcal{V}} M_T(B \mid U(T), V(T) = v)\, M_{V(T)}(dv \mid U(T)) \\
= & \int_{\Omega} P_{T,\theta}(B \mid U(T))\pi(\theta)\ \upsilon(d\theta) \\
= & \int_{\Omega_2}\int_{\Omega_1} P_{T,\theta_1,\theta_2}(B \mid U(T))\pi_1(\theta_1)\,\pi_2(\theta_2)\ \upsilon_1(d\theta_1)\ \upsilon_2(d\theta_2) \\
= & \int_{\Omega_2}\int_{\Omega_1} \left(\int_{\mathcal{V}} P_{T,\theta_2}(B \mid U(T), V(T) = v)\, P_{V(T),\theta_1}(dv \mid U(T))\right) \times \\
& \qquad \pi_1(\theta_1)\,\pi_2(\theta_2)\ \upsilon_1(d\theta_1)\ \upsilon_2(d\theta_2) \\
= & \int_{\mathcal{V}}\int_{\Omega_2} P_{T,\theta_2}(B \mid U(T), V(T) = v)\pi_2(\theta_2)\ \upsilon_2(d\theta_2) \times \\
& \qquad \int_{\Omega_1} P_{V(T),\theta_1}(dv \mid U(T))\pi_1(\theta_1)\ \upsilon_1(d\theta_1) \\
= & \int_{\mathcal{V}}\int_{\Omega_2} P_{T,\theta_2}(B \mid U(T), V(T) = v)\pi_2(\theta_2)\ \upsilon_2(d\theta_2)\ M_{V(T)}(dv \mid U(T)).
\end{aligned}$$

Therefore

$$M_T(B \mid U(T), V(T) = v) = \int_{\Omega_2} P_{T,\theta_2}(B \mid U(T), V(T) = v)\pi_2(\theta_2)\ \upsilon_2(d\theta_2)$$

almost surely and this is independent of π_1 as claimed. ∎

From Theorem 2 we see that the check for π_2 is independent of how we specify π_1 whenever the conditions of the theorem obtain. As noted in Fraser (1979), however, situations where we have a sufficient-ancillary statistic for (θ_1, θ_2) are relatively hard to find although we provide one context in Example 1. Of course

the lack of dependence of the check for π_2 on π_1 is also dependent on specifying independent priors for θ_1 and θ_2.

So in general, when we specify $\pi(\theta) = \pi_1(\theta_1)\,\pi_2(\theta_2 \mid \theta_1)$ the check for π_2 is really checking π_1 in part, even though conditioning on an ancillary for θ_2 is helping to remove this dependence somewhat. For this reason it seems that there is a natural order to the checking unless the conditions of Theorem 2 hold. First we check π_1 and we note by Theorem 1 that this does not depend on how we specify π_2. If we obtain no evidence of any prior-data conflict for π_1, then we proceed to check π_2. This is similar to the restriction that we first check for the correctness of the sampling model, and only proceed to check for prior-data conflict, when we have no evidence against the sampling model.

3 Checking Prior Components with Exponential Models

Consider the following examples where can apply the methodology discussed in Section 2.

Example 1. *Multinomial model*

Suppose we observe $(f_1, f_2) \sim \text{Multinomial}(n, \theta_1, \theta_2, 1-\theta_1-\theta_2)$ with

$$\Omega = \{(\theta_1, \theta_2) : \theta_1, \theta_2 \geq 0, 0 \leq \theta_1 + \theta_2 \leq 1\}.$$

Then $T = (f_1, f_2)$ is a minimal sufficient statistic. If we prescribe the prior as $\pi(\theta_1, \theta_2) = \pi_1(\theta_1)\,\pi_2(\theta_2 \mid \theta_1)$, then $V(T) = f_1 \sim \text{Binomial}(n, \theta_1)$ is ancillary for θ_2. So we can check for any prior-data conflict with π_1 by using the prior predictive of f_1 and then check for any prior-data conflict with π_2 by using the conditional prior predictive of (f_1, f_2) given f_1.

For example, suppose that θ_1 has a $\text{Beta}(\alpha_1, \beta_1)$ distribution and

$$\frac{\theta_2}{1-\theta_1} \mid \theta_1 \sim \text{Beta}(\alpha_2, \beta_2).$$

Note that the joint prior distribution of (θ_1, θ_2) is $\text{Dirichlet}(\alpha_1, \alpha_2, \beta_2)$ if and only if $\beta_1 = \alpha_2 + \beta_2$.

The joint prior predictive of (f_1, f_2) is given by

$$\begin{aligned} m(f_1, f_2) &= \left\{ \binom{n}{f_1} \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(f_1+\alpha_1)\Gamma(n-f_1+\beta_1)}{\Gamma(n+\alpha_1+\beta_1)} \right\} \times \\ &\quad \left\{ \binom{n-f_1}{f_2} \frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} \frac{\Gamma(f_2+\alpha_2)\Gamma(n-f_1-f_2+\beta_2)}{\Gamma(n-f_1+\alpha_2+\beta_2)} \right\} \quad (4) \end{aligned}$$

By considering a simple binomial model with a beta prior it is easy to see that each of the factors in (4) is a probability function and so the first factor is the prior predictive for f_1 and the second factor is the conditional prior predictive for f_2 given f_1.

Notice that if we reparameterize this model as $(\psi_1, \psi_2) = (\theta_1, \theta_2/(1-\theta_1))$ then f_1 is sufficient-ancillary for (ψ_1, ψ_2). Further the above prior on (θ_1, θ_2) induces independent priors on ψ_1 and ψ_2 so that Theorem 2 applies. This tell us that the conditioning on f_1 to check for conflict with π_2 has completely removed the dependence on π_1.

Suppose now that $n = 20, (\alpha_1, \beta_1) = (3, 15), (\alpha_2, \beta_2) = (3, 4)$ and we observe $(f_1, f_2) = (7, 12)$. Then the prior predictive for f_1 is plotted in Figure 1. We see that $f_1 = 7$ is not very extreme for this distribution. In fact the probability of getting a value as least as far out in the tails as this value is 0.103. Therefore we have no evidence of any prior-data conflict with π_1.

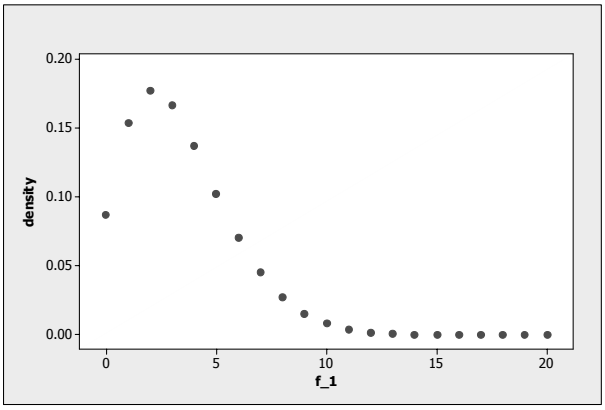

Figure 1: Prior predictive density for f_1 in Example 1.

We now proceed to assess whether or not $f_2 = 12$ is a reasonable value from its conditional prior predictive given $f_1 = 7$. The conditional prior predictive is plotted in Figure 2. The probability of getting a value as least as far out in the tails as $f_2 = 12$ is 0.017. Therefore we have evidence of prior-data conflict with π_2.

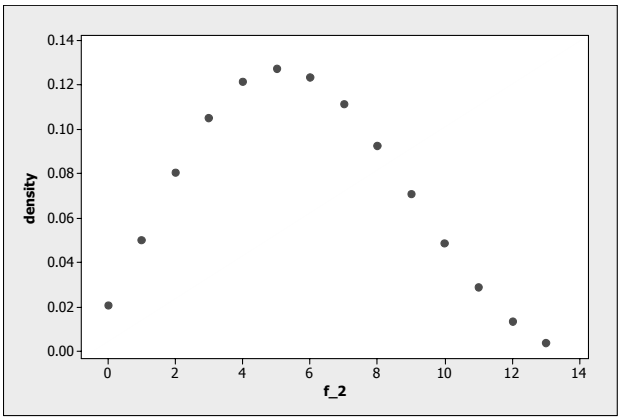

Figure 2: Conditional prior predictive of f_2 given that $f_1 = 7$.

Now suppose we choose a uniform prior on for θ_2 given the value of θ_1. This entails choosing $\alpha_2 = \beta_2 = 1$. From (4) we have that the conditional prior

predictive for f_2 given f_1 is uniform on $\{0,\ldots,n-f_1\}$. The implication of this is that we would never obtain evidence of a prior-data conflict. As discussed in Evans and Moshonov (2005) this is what we would expect from a prior that is noninformative. Similarly, when we choose $\alpha_1 = \beta_1 = 1$, we get that the prior predictive for f_1 is uniform on $\{0,\ldots,n\}$ and so the component prior is noninformative for θ_1. ∎

Example 2. *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 (σ^2, μ) with $\bar{x} \sim N(\mu, \sigma^2/n)$ independent of $s^2 \sim (\sigma^2/(n-1))\chi^2_{(n-1)}$. We see immediately that s^2 is ancillary for μ. Note that s^2 is not sufficient-ancillary for this model so Theorem 2 is not applicable.

Suppose we put the conjugate prior on (σ^2, μ) given by

$$\begin{aligned} \frac{1}{\sigma^2} &\sim \text{Gamma}(\alpha_0, \beta_0) \\ \mu \,|\, \sigma^2 &\sim N(\mu_0, \tau_0^2\sigma^2). \end{aligned}$$

So here we have that $\theta = (\sigma^2, \mu)$ and π_1 is an inverse gamma density while $\pi_2\left(\mu \,|\, \sigma^2\right)$ is the $N(\mu_0, \tau_0^2\sigma^2)$ density.

The joint prior predictive distribution of $(\bar{x}, s^2)$ is given by

$$\begin{aligned} m(\bar{x}, s^2) &= \int_{-\infty}^{\infty}\int_0^{\infty} f(\bar{x}, s^2 \,|\, \sigma^2, \mu)\pi(\mu \,|\, \sigma^2)\pi(\sigma^2)\, d\sigma^2\, d\mu \\ &= \frac{\Gamma\left(\frac{n}{2} + \alpha_0\right)}{\Gamma(\alpha_0)}\left(n + \frac{1}{\tau_0^2}\right)^{-1/2}\frac{(2\pi)^{-n/2}\,\beta_0^{\alpha_0}}{\tau_0}\left(\beta_x\right)^{-(n/2)-\alpha_0} \end{aligned} \tag{5}$$

where

$$\beta_x = \beta_0 + \frac{1}{2}\frac{\mu_0^2}{\tau_0^2(n\tau_0^2+1)} + \frac{n-1}{2}s^2 + \frac{n}{2}\frac{1}{n\tau_0^2+1}(\bar{x}-\mu_0)^2. \tag{6}$$

We can easily determine that the marginal prior predictive of s^2 is given by $s^2 \sim (\beta_0/\alpha_0)F_{(n-1,2\alpha_0)}$. From (5) and (6) we see that the conditional prior predictive of $\bar{x}$ given s^2 is distributed as $\mu_0 + \tilde{\sigma}t$ where

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

and $t \sim \text{Student}(n + 2\alpha_0 - 1)$.

For example, suppose that $\mu = 0, \sigma^2 = 1$ and that for a sample of size $n = 20$ from this distribution we obtained $\bar{x}_0 = 0.0358324$ and $s_0^2 = 0.836563$. For the prior we specify $\tau_0^2 = 1$, $\mu_0 = 50$, $\alpha_0 = 1$ and $\beta_0 = 5$.

To assess if there is any conflict with π_1 we compare $s^2/5 = 0.1673126$ with the $F(19, 2)$ distribution. Computing the P-value by 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

$$P\left(F(19,2) \leq 0.1673126\right) + P\left(F(19,2) > 1.5295\right) = .47832,$$

which does not indicate any prior-data conflict.

In Evans and Moshonov (2005) it was determined that for a sequence of priors in this example to be noninformative, when checking the full prior for any prior-data conflict, it was necessary that $\alpha_0/\beta_0 \to 0$. This implied that the relevant P-value converges to 1 no matter what data is observed and so no evidence of any conflict will ever be obtained in the limit. When checking for any prior-data conflict with π_1 we see that $P((\beta_0/\alpha_0)F_{(n-1,2\alpha_0)} \leq s_0^2) \to 0$ and $P((\beta_0/\alpha_0)F_{(n-1,2\alpha_0)} \geq s_0^2) \to 1$ as $\alpha_0/\beta_0 \to 0$. Therefore such a sequence of priors will satisfy the natural requirement of noninformativity for θ_1 that the sequence not conflict with any data in the limit.

To assess if there is any conflict with π_2 we compare

$$\frac{\bar{x} - \mu_0}{\tilde{\sigma}} = \frac{0.0358324 - 50}{1.1389099} = -43.870167$$

to the Student(21) 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.

Evans and Mosohonov (2005) also considered assessing whether any prior-data conflict existed, with respect to the full prior, by comparing the observed value of $\bar{x}$ with its marginal prior predictive. With this data this comparison resulted in a P-value of .0021 using the marginal prior predictive (a Student(2) distribution). Intuitively this check seemed to be assessing whether the prior for μ was in conflict, but there was no clear rationale for saying this. This paper is concerned with developing methods appropriate to this kind of assessment and we note that the conditional check has given a much more extreme P-value reflective of just how extreme the choice of the prior is when we put $\mu_0 = 50$.

As $\tau_0^2 \to \infty$ we have that $\tilde{\sigma}^2 \to \infty$ for every value of s^2. This implies that

$$P\left(|t| \geq \left|\frac{\bar{x}_0 - \mu_0}{\tilde{\sigma}}\right|\right) \to 1$$

as $\tau_0^2 \to \infty$. Therefore, provided that $\tau_0^2 \to \infty$ we have that such a sequence of priors will be noninformative for θ_2.

In Evans and Moshonov (2005) it was shown that a necessary requirement for a sequence of priors to be noninformative for the full parameter (σ^2, μ) is that $\alpha_0/\beta_0 \to 0$. The analysis here shows that if we require that the sequence of priors be noninformative for μ by itself, then we need to require that $\tau_0^2 \to \infty$. If we only require noninformativity for μ, there is no need to require that $\alpha_0/\beta_0 \to 0$. If we require noninformativity for σ^2, however, then we do need that $\alpha_0/\beta_0 \to 0$. So the situation is somewhat different depending on what we want the sequence of priors to be noninformative for. As noted in Evans and Moshonov (2005) the

requirement of no prior-data conflict is only a partial characterization for non-informativity. Our analysis here indicates that in addition to the requirement that $\alpha_0/\beta_0 \to 0$ we really do want $\tau_0^2 \to \infty$ as well, at least when the prior is specified hierarchically, and we want the sequence to be noninformative with respect to both parameters. ∎

Both examples can be generalized in the sense that it is possible to find a function V of the minimal sufficient statistic T so that V is ancillary for θ_2. For example, Example 1 can be generalized to the Multinomial$(n, p_1, \ldots, p_k)$ case where we put $\theta_1 = (p_1, \ldots, p_l)$ for $l < k-1$ and $\theta_2 = (p_{l+1}, \ldots, p_{k-1})$. Example 2 can be generalized in several ways. For example, if we have a linear model $y = \beta_1 x_1 + \cdots + \beta_k x_k + z$ with $z \sim N(0, \sigma^2)$, then we can take $\theta_1 = \sigma^2$ and $\theta_2 = (\beta_1, \ldots, \beta_k)$. If we are sampling from a p-dimensional normal $N_p(\mu, \Sigma)$, we can $\theta_1 = \Sigma$ take and $\theta_2 = \mu$.

All of the above examples are models that have exponential form. Consider then a model whose density takes the form

$$f_{(\psi_1,\ldots,\psi_p)}(s) = c(\psi_1, \ldots, \psi_p) h(s) \exp\{\psi_1 T_1(s) + \cdots + \psi_k T_k(s)\}$$

where $\psi = (\psi_1, \ldots, \psi_p) \in \Psi$ with Ψ an open subset of R^p and suppose that the functions $T_1, \ldots, T_p$ are linearly independent. Then $T = (T_1, \ldots, T_p)$ is a complete minimal sufficient for $(\psi_1, \ldots, \psi_p)$. Note that by Basu's theorem any ancillary U for ψ is independent of T and so we can ignore conditioning on U when checking for prior-data conflict.

To apply the approach of Section 2 to some function θ_2 of $(\psi_1, \ldots, \psi_p)$ we need to find a function V of T ancillary for θ_2. In general it isn't obvious how to do this. Indeed it is probably not the case that such a V will exist for an arbitrary θ_2 and currently we lack a characterization of such parameters. The following section gives another general context where it is easier to find such a decomposition.

4 Checking Prior Components with Group Models

Consider the following example where can apply the methodology discussed in Section 2.

Example 3. *Location-scale Cauchy*

Suppose we have a sample $s = (s_1, \ldots, s_n)$ from a distribution with density at x equal to

$$\frac{1}{\pi\sigma\left(1 + (x-\mu)^2/\sigma^2\right)}$$

where $\theta = (\sigma, \mu)$ and $\mu \in R^1$, $\sigma > 0$ are unknown. It is known that the order statistic $T(s) = (s_{(1)}, \ldots, s_{(n)})$ is a minimal sufficient statistic for this model.

Putting $r = s_{(n)} - s_{(1)}$, we have that

$$U\left(T\left(s\right)\right) = \left(\frac{s_{(2)} - s_{(1)}}{r}, \cdots, \frac{s_{(n-1)} - s_{(1)}}{r}\right) \tag{7}$$

is ancillary for $\theta = (\mu, \sigma)$, since U is invariant under location-scale transformations, and we note that $U \in R^{n-2}$. The conditional distribution of T given U can be expressed as the conditional distribution of $(s_{(1)}, r)$ given $U\left(T\right)$. To obtain this we make the following transformation $(s_{(1)}, ..., s_{(n)}) \rightarrow (v_1, ..., v_n)$ defined as follows

$$v_1 = s_{(1)}, v_2 = r = s_{(n)} - s_{(1)}, v_{i+1} = \left(s_{(i)} - s_{(1)}\right)/r,$$

for $i = 2, \ldots, n-1$. Expressing the order statistic in terms of those new variables we have $s_{(1)} = v_1, s_{(2)} = v_1 + v_2 v_3, ..., s_{(i)} = v_1 + v_2 v_{i+1}$, for $i = 2, ..., n-1$ and $s_{(n)} = v_1 + v_2$. The Jacobian of this transformation is then given by $v_2^{n-2} = r^{n-2}$. The joint density of $(s_{(1)}, r, U) = (v_1, \ldots, v_n)$ is then the joint density of the order statistic at the above values times r^{n-2}. This implies that the conditional distribution of $\left(s_{(1)}, r\right)$ given U satisfies

$$\begin{aligned} & f_{\sigma,\mu}(s_{(1)}, r \,|\, U) \\ \propto \;& \sigma^{-2}\left(\frac{r}{\sigma}\right)^{n-2}\left(1 + \left(\frac{s_{(1)} - \mu}{\sigma}\right)^2\right)^{-1}\left(1 + \left(\frac{r}{\sigma} + \frac{s_{(1)} - \mu}{\sigma}\right)^2\right)^{-1} \\ & \times \prod_{i=3}^{n}\left(1 + \left(\frac{r}{\sigma}v_i + \frac{s_{(1)} - \mu}{\sigma}\right)^2\right)^{-1}. \end{aligned} \tag{8}$$

If we integrate out $s_{(1)}$ from (8) we see that what remains does not involve μ and so we have immediately that r is ancillary for μ.

Denoting the prior on (σ, μ) by $\pi\left(\sigma, \mu\right) = \pi_1\left(\sigma\right)\pi_2\left(\mu \,|\, \sigma\right)$ we then have that the conditional prior predictive density of $\left(s_{(1)}, r\right)$ given U satisfies

$$m(s_{(1)}, r \,|\, U) = \int_0^\infty \int_{-\infty}^\infty f_{\mu,\sigma}(s_{(1)}, r \,|\, U)\pi_1\left(\sigma\right)\pi_2\left(\mu \,|\, \sigma\right)\, d\mu\, d\sigma. \tag{9}$$

In (9) we transform from (σ, μ) to $x = \left(s_{(1)} - \mu\right)/\sigma$ and $y = r/\sigma$ and use (8) to obtain,

$$\begin{aligned} & m(s_{(1)}, r \,|\, U) \\ = \;& \int_0^\infty \int_{-\infty}^\infty \left(r^2/y^3\right) f_{s_{(1)} - xr/y, r/y}(s_{(1)}, r \,|\, U)\pi\left(s_{(1)} - xr/y, r/y\right)\, dx\, dy \\ \propto \;& \int_0^\infty \int_{-\infty}^\infty y^{n-2}\left(1 + x^2\right)^{-1}\left(1 + (y + x)^2\right)^{-1}\prod_{i=3}^{n}\left(1 + (yv_i + x)^2\right)^{-1} \\ & \times y^{-1}\pi_1\left(r/y\right)\, \pi_2\left(s_{(1)} - rx/y \,|\, r/y\right)\, dx\, dy. \end{aligned} \tag{10}$$

Integrating $s_{(1)}$ and r out of (10), using the fact that $\pi_2\left(s_{(1)} - rx/y \,|\, r/y\right)$ is a density in $s_{(1)}$ and $y^{-1}\pi_1\left(r/y\right)$ is a density in r, we obtain the inverse normalizing constant as

$$\begin{aligned} c &= \int_0^\infty \int_{-\infty}^\infty y^{n-2}\left(1+x^2\right)^{-1}\left(1+(y+x)^2\right)^{-1}\prod_{i=3}^n\left(1+(yv_i+x)^2\right)^{-1} dx\, dy. \\ &= \int_0^\infty \int_{-\infty}^\infty k(x,y)\, dx\, dy. \end{aligned}$$

Observe that the marginal prior predictive density for r is obtained by integrating $s_{(1)}$ out of (10), to obtain

$$m_1(r \,|\, U) = \int_0^\infty g(y \,|\, U) y^{-1}\pi_1\left(r/y\right)\, dy \tag{11}$$

where

$$g(y \,|\, U) = c^{-1}\int_{-\infty}^\infty k(x,y)\, dx \tag{12}$$

is the conditional density of y given U. From (11) we see that $m_1(r \,|\, U)$ is a scale probability mixture of the prior on σ. Now note that

$$\begin{aligned} & m_2(s_{(1)} \,|\, r, U) \\ = \;& \left(m_1(r \,|\, U)\right)^{-1}\int_0^\infty \left(c^{-1}\int_{-\infty}^\infty k(x,y)\pi_2\left(s_{(1)} - rx/y \,|\, r/y\right)\, dx\right) \times \\ & \qquad y^{-1}\pi_1\left(r/y\right)\, dy \\ = \;& \left(m_1(r \,|\, U)\right)^{-1}\int_0^\infty g(s_{(1)}, y \,|\, U) y^{-1}\pi_1\left(r/y\right)\, dy \end{aligned}$$

In terms of organizing the computations we first tabulate $g(\cdot \,|\, U)$, which also requires the computation of c, next we tabulate $m_1(\cdot \,|\, U)$ and use this, together with the observed value of r, to assess whether or not there is any prior-data conflict with π_1. Finally we need to tabulate $\int_0^\infty g(\cdot, y \,|\, U) y^{-1}\pi_1\left(r/y\right)\, dy$ to obtain the conditional prior predictive of $s_{(1)}$ and use this, together with the observed value of $s_{(1)}$, to assess whether or not there is any prior-data conflict with π_2.

To evaluate a P-value associated with this conditional prior predictive we must integrate numerically in this example. For example, consider the following ordered sample of $n = 10$ from the Cauchy distribution with $\mu = 0$ and $\sigma = 1$.

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

Then the observed values of the relevant statistics are $s_{(1)} = -4.48290, r = 23.2101$ and the values of $v_3, \ldots, v_{10}$ are given in the following table.

0.0652163	0.1547339	0.1622763	0.1694449
0.1810571	0.1951884	0.2864865	0.3718015

First we need to assess whether or not $r = 23.2101$ is a surprising value from $m_1(\cdot \,|\, U)$. For this we first compute the conditional density of y given $v_3, \ldots, v_{10}$. We use (12) for this and note that $v_3, \ldots, v_{10}$ are all positive. Thus, based on the expression for $k(x, y)$, for each $y > 0$ we can integrate over x by summing over values that lie within the effective range of the Cauchy. In Figure 3 we have plotted the conditional density of y given U. From this we can see that tabulating y in the range $(0, 100)$ will be adequate.

Now suppose that we use the prior as given in Example 8 but with $\tau_0^2 = 1$, $\mu_0 = 0$, $\alpha_0 = 2$ and $\beta_0 = 3$. The conditional prior predictive of r given U, namely $m_1(\cdot \,|\, U)$, is displayed in Figure 4. We see from this that the observed value of $r = 23.2101$ is in the central hump of the distribution and so we have no evidence of any prior-data conflict with π_2.

In Figure 5 we have plotted $m_2(\cdot \,|\, r, U)$. From this we can see that the observed value of $s_{(1)} = -4.48290$ does not lead to any prior-data conflict with π_1.

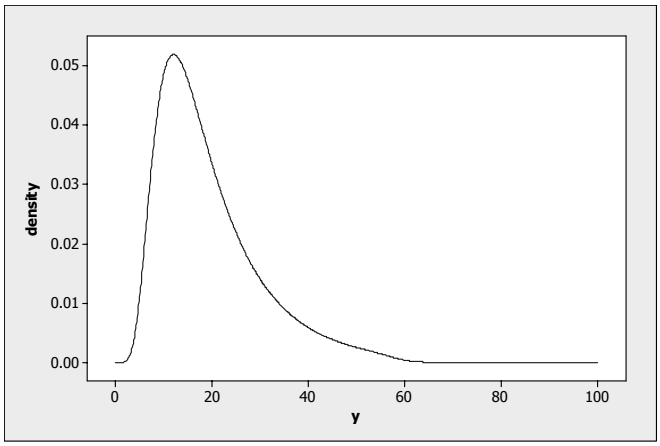

Figure 3: Conditional density $g(\cdot \,|\, U)$.

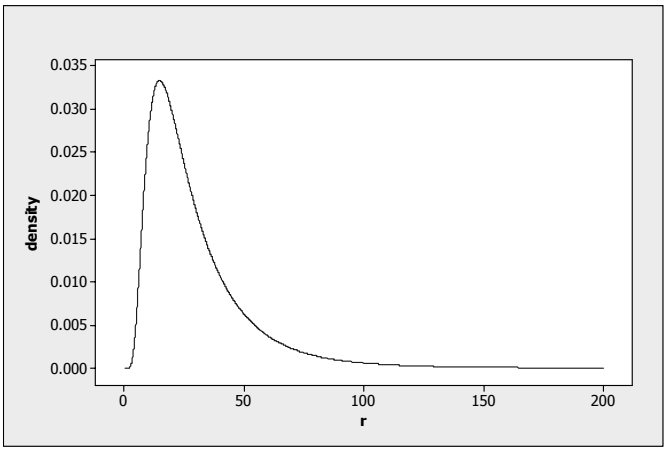

Figure 4: The conditional prior predictive density $m_1(\cdot \,|\, U)$.

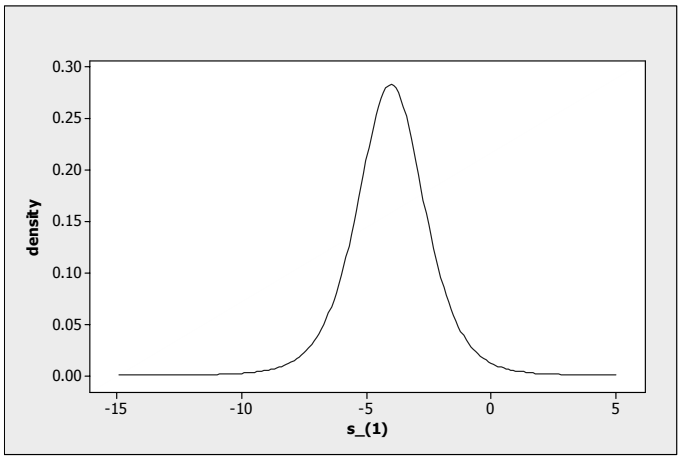

Figure 5: The conditional prior predictive $m_1(\cdot \mid r, U)$.

∎

Example 3 is a particular example of a group model. For a group model we have a group G and a family of transformations $\{W_g : g \in G\}$ acting on the sample space $\mathcal{T}$ for the minimal sufficient statistic T. We assume that this action is free, namely it satisfies $W_{g_1}(t) = W_{g_2}(t)$ if and only if $g_1 = g_2$ for any t (perhaps after removing a set of measure 0 from $\mathcal{T}$). Further we have a distribution with density f with respect to some support measure μ on $\mathcal{T}$. The sampling model is then given by the family of distributions, indexed by $g \in G$, generated via $t = W_g(z)$ where $z \sim f$. So we have that $\theta = g$ and $\Omega = G$.

In Example 3 we have the location-scale group $G = \left\{(a, c) : a \in R^1, c > 0\right\}$, with product $(a_1, c_1)(a_2, c_2) = (a_1 + c_1 a_2, c_1 c_2)$, acting on R^n via $t = W_{(a,c)}(z) = a1_n + cz$ where 1_n denotes the n-dimensional vector of 1's. The distribution of z is that of the order statistic based on a sample of n from the standard Cauchy.

Now note that we can write the order statistic $t = \left(s_{(1)}, \ldots, s_{(n)}\right)$ as $t = s_{(1)} + rU = W_{(s_{(1)}, r)}(U(t))$. Since $U(t) = (t - s_{(1)})/r$ is a maximal invariant under the action of this group, we have that U is ancillary and this result holds no matter distribution we take for z. Note that once we have specified a maximal invariant U this immediately specifies the data-dependent element in G that restores the full data from U. In Example 3 choosing U as in (7) specifies $\left[s_{(1)}, r\right] \in G$. So once we have specified a maximal invariant U, and there are many possible choices, there is a unique element $[t] \in G$ such that $t = W_{[t]}(U(t))$.

In the location-scale group G we can write $(a, c) = (a, 1)(0, c)$ and the elements $(a, 1)$ belong to the location group $G_2 = \left\{(a, 1) : a \in R^1\right\}$ while the elements $(0, c)$ belong to the scale group $G_1 = \{(0, c) : c > 0\}$. In effect the location-scale group is a semidirect product of the location and scale groups since we can write $G = G_2 G_1$ and G_2 is a normal subgroup of G. If a group G can be written as the semidirect product $G = G_2 G_1$, then any element $g \in G$ can be written uniquely as $g = g_2 g_1$ where $g_1 \in G_1$ and $g_2 \in G_2$. For details on this see, for example, Robinson (1980). Further we have that

$W_g(t) = W_{g_2 g_1}(t) = W_{g_2} \circ W_{g_1}(t)$ and $[t] = [t]_2 [t]_1$ with $[t]_1 \in G_1$ and $[t]_2 \in G_2$.

Using the fact that, for any maximal invariant U we have $U(t) = U(W_g(t))$, observe that

$$\begin{aligned} W_g(t) &= W_{g_2 g_1}(t) \circ W_{[t]_2 [t]_1}(U(t)) = W_{g_2 g_1 [t]_2 [t]_1}(U(t)) \\ &= W_{g_2 g_1 [t]_2 g_1^{-1} g_1 [t]_1}(U(t)) = W_{g_2 g_1 [t]_2 g_1^{-1}} \circ W_{g_1 [t]_1}(U(t)) \end{aligned} \quad (13)$$

and the normality of G_2 implies that $g_1 [t]_2 g_1^{-1} \in G_2$ and so $g_2 g_1 [t]_2 g_1^{-1} \in G_2$. For the location-scale group, using the maximal ancillary equal to (7), we have that $[t]_1 = (0, r)$ and $[t]_2 = \left(s_{(1)}, 1\right)$ and so when $g = (a, c) = (a, 1)(0, c)$ we have that

$$\begin{aligned} g_2 g_1 [t]_2 g_1^{-1} &= (a, 1)(0, c)\left(s_{(1)}, 1\right)\left(0, c^{-1}\right) = \left(a + c s_{(1)}, 1\right), \\ g_1 [t]_1 &= (0, c)(0, r) = (0, cr). \end{aligned}$$

Now $t = W_{[t]}(U(t))$ and so

$$W_g(t) = W_{[W_g(t)]_2 [W_g(t)]_1}(U(W_g(t))) = W_{[W_g(t)]_2 [W_g(t)]_1}(U(t))$$

together with (13) implies that $[W_g(t)]_2 = g_2 g_1 [t]_2 g_1^{-1}, [W_g(t)]_1 = g_1 [t]_1$. In particular this implies that $V(t) = [W_g(t)]_1$ is invariant under G_2 and equivariant under G_1. From this we have the following result.

Theorem 3. Suppose that the statistical model for the minimum sufficient statistic T is given by $\{f_{\theta T} : \theta \in G\}$ where G is a group acting freely on $\mathcal{T}$ via the class of transformations $\{W_g : g \in G\}$ and $f_{\theta T}$ corresponds to the distribution of t when we write $t = W_\theta(z)$ with $z \sim f$. Further suppose that G is a semidirect product $G = G_2 G_1$ and we write $\theta = \theta_2 \theta_1$ with $\theta_i \in G_i$. For a particular choice of maximal invariant U, and thus corresponding transformation $[\cdot] : \mathcal{T} \to G$ such that $t = W_{[t]}(U(t))$, then $V(T)$ given by $V(t) = [t]_1$ is ancillary for θ_2.

Theorem 3 gives a very large class of models for which the methods of this paper are applicable for checking for prior-data conflict for the components π_1 and π_2. For example, any regression model with nonnormal error falls in this class. For a discussion of a wide variety of group models in the frequency context see Fraser (1979). For many of these models the group involved can be decomposed as a semidirect product and so the analysis of this section will apply. Of course, this requires that we specify the components of the full prior distribution in a way that conforms to this decomposition.

5 Conclusions

We have been concerned here with checking for prior-data conflict by checking individual components of the prior when the prior is specified hierarchically in two stages. So the parameter θ is decomposed into two parts θ_1 and θ_2 where θ_2 is the parameter of interest (not necessarily 1-dimensional) and θ_1 is comprised of nuisance parameters and perhaps hyperparameters. Then a prior

π_1 is specified for θ_1 and a prior $\pi_2(\cdot \mid \theta_1)$ is specified for θ_2, conditional on θ_1. Of course the prior $\pi_2(\cdot \mid \theta_1)$ can also be independent of θ. The methodology developed is a generalization of the methods discussed in Evans and Moshonov (2005). The key requirement is the existence of statistic $V(T)$ that is ancillary for θ_2. We have shown that, in a wide variety of models, such a statistic exists provided the decomposition of the prior conforms to a certain structure.

The necessity of the existence of the statistic $V(T)$ is of course a restriction on the methods we have developed. While we can't say that methods cannot be developed for general decompositions, there does appear to be a natural logic to the developments here and this might be interpreted as a restriction on the specification of priors hierarchically, at least when we want to check the individual components separately for prior-data conflict. When such a decomposition does not exist, or is not deemed suitable for the hierarchical specification, we can always resort to the methods of Evans and Moshonov (2005) for checking the full prior.

There are several aspects of our discussion here where more work is indicated. In particular we need to develop methods appropriate for the general decomposition given by (2) where $p > 2$. These will naturally be a generalization of the methods we have discussed in this paper. Further the computations involved for general models with general priors will obviously be more complicated than those we have presented here. Strictly speaking straight-forward Monte Carlo methods are not available. The computational problems are more similar to those involved in computing inverse normalizing constants rather than posterior calculations. These aspects of the problem are currently under investigation.

References

Bayarri, M.J. and Berger, J.O. (2000) P values for composite null models. Journal of the American Statistical Association, Vol. 95, 452, 1127-1142.

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

Evans, M. and Moshonov, H. (2005) Checking for prior-data conflict. Dept. of Statistics, University of Toronto, Technical Report 0413.

Fraser, D.A.S. (1979) Inference and Linear Models. McGraw-Hill.

Gelman, A., Meng, X., and Stern, H. (with discussion) (1996) Posterior predictive assessment of model fitness via realized discrepancies. 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.

Johnson, V. (2004) A Bayesian χ^2 test for goodness-of-fit. The Annals of Statistics, Vol. 32, No. 6, 2361-2384.

Lehmann, E. and Scholz, F.W. (1992) Ancillarity. Current Issues in Statistical Inference: Essays in Honor of D.Basu. Malay Ghosh and Pramod K. Pathak, Editors. IMS Lecture notes-monograph series, Hayward, CA, 32-51.

Robinson, D.J.S. (1980) A Course in the Theory of Groups, Springer-Verlag, New York.

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