Using prior expansions for prior-data conflict checking

David J. Nott*1,2, Max Seah3, Luai Al-Labadi4, Michael Evans5, Hui Khoon Ng3,6,7, and Berthold-Georg Englert3,7,8

1Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546
2Institute of Operations Research and Analytics, National University of Singapore, Singapore 117602
3Centre for Quantum Technologies, National University of Singapore, Singapore 117543
4Department of Mathematics, University of Sharjah, P.O. Box 27272 Sharjah, United Arab Emirates
5Department of Statistics, University of Toronto, Toronto, ON M5S 3G3, Canada
6Yale-NUS College, Singapore 138614
7MajuLab, CNRS-UNS-NUS-NTU International Joint Research Unit, UMI 3654, Singapore
8Department of Physics, National University of Singapore, Singapore 117542

Abstract

Any Bayesian analysis involves combining information represented through different model components, and when different sources of information are in conflict it is important to detect this. Here we consider checking for prior-data conflict in Bayesian models by expanding the prior used for the analysis into a larger family of priors, and considering a marginal likelihood score statistic for the expansion parameter. Consideration of different expansions can be informative about the nature of any conflict, and extensions to hierarchically specified priors and connections with other approaches to prior-data conflict checking are discussed. Implementation in complex situations is illustrated with two applications. The first concerns testing for the appropriateness of a LASSO penalty in shrinkage estimation of coefficients in linear regression. Our method is compared with a recent suggestion in the literature designed to be powerful against alternatives in the exponential power family, and we use this family as the

*Corresponding author: standj@nus.edu.sg
prior expansion for constructing our check. A second application concerns a problem in quantum state estimation, where a multinomial model is considered with physical constraints on the model parameters. In this example, the usefulness of different prior expansions is demonstrated for obtaining checks which are sensitive to different aspects of the prior.

**Keywords**: Bayesian inference, LASSO, Model checking, Penalized regression, Prior data-conflict.

## 1 Introduction

A common approach to checking the likelihood in a statistical analysis is to consider model expansions motivated by thinking about plausible departures from the assumed model. Then using either formal or informal methods for model choice, we can compare the expanded model with the original one to determine whether the original model was good enough. In Bayesian analyses, information from the prior is combined with information in the likelihood, and an additional aspect of checking Bayesian models is to see whether the prior and likelihood information conflict. If the likelihood is inadequate, there will be no value of the model parameter giving a good fit to the data, whereas prior-data conflict occurs when the prior is putting all its mass out in the tails of the likelihood. Checking for prior-data conflict is important, because it is not sensible to combine conflicting sources of information without careful thought, and the influence of the prior only increases with increasing conflict. In any case something of significance has been determined when a prior-data conflict has been shown to exist.

The purpose of this work is to consider model expansion for checking for prior-data conflict, rather than for checking the likelihood. Suppose that $\theta$ is a parameter, $y$ is data, $g(\theta)$ is a prior density for $\theta$, and $p(y|\theta)$ is the sampling model. Write $g(\theta|y) \propto g(\theta)p(y|\theta)$ for the posterior density. Suppose that we have checked the likelihood component of the model and that it is thought to be adequate, so that checking for prior-data conflict is of interest. Checking the likelihood component first is important, since sound inferences cannot result from a poor model no matter what prior is used for $\theta$. Our approach to prior-data conflict checking considers embedding the original prior into a larger family, which we write as $g(\theta|\gamma)$, where $\gamma$ is some expansion parameter and the original prior is $g(\theta|\gamma_0)$ for some value $\gamma_0$. The corresponding posterior densities will be denoted by $g(\theta|y, \gamma)$. Throughout this work, $\gamma$ will be a scalar, or if we embed the prior into a family with more than one
additional parameter, we will vary these parameters one by one. If we integrate out $\theta$ from the likelihood using $g(\theta|\gamma)$, we obtain

$$p(y|\gamma) = \int g(\theta|\gamma)p(y|\theta)d\theta,$$

and we propose using the score-type statistic

$$S(y) = \frac{d}{d\gamma} \log p(y|\gamma)\bigg|_{\gamma=\gamma_0},$$

(1)

for checking for prior-data conflict. A $p$-value for the statistic (1) is computed to provide a calibration of its value by using the prior predictive density as the reference distribution. Let us assume the embedding family is such that a large value of $S(y)$ indicates conflict. Then an appropriate $p$-value for the check is

$$p_S = P(S(Y) \geq S(y_{obs})),$$

(2)

where $Y \sim p(y) = \int g(\theta)p(y|\theta)d\theta$ and $y_{obs}$ is the observed value of $y$. In the next section we will describe a framework for Bayesian model checking that explains some logical requirements that a prior-data conflict check should satisfy, and we discuss why a check based on (1) and (2) satisfies these requirements.

We show later that under appropriate regularity conditions (1) has the alternative expression

$$S(y) = \int \frac{d}{d\gamma} \log g(\theta|\gamma)\bigg|_{\gamma=\gamma_0} g(\theta|y)d\theta.$$

(3)

Expression (3) gives an intuitive meaning to the test statistic (1), as well as being useful later for computation. We can see that (1) is the posterior expectation of the rate of change of the log prior with respect to the expansion parameter. If there is a conflict, and our posterior distribution is concentrated in the tails of the prior, then the derivative of the log prior with respect to the expansion parameter will be large if the prior is changing in a direction that resolves the conflict when we vary $\gamma$ around $\gamma_0$. One of the main advantages of the score type check we propose is that by appropriate choices of the prior expansion we can obtain checks of conflict which are sensitive to different aspects of the prior. Other proposals for prior-data conflict checking do not have this feature, which will be illustrated in some of the examples.
In the next section, we discuss prior-data conflict checking and how this differs from checking the likelihood component in a model-based statistical analysis. We also review the existing literature on checking for prior-data conflict. Section 3 discusses hierarchical extensions of our check. Section 4 illustrates implementation of the checks in two complex applications. The first concerns shrinkage estimation of coefficients in linear regression with squared error loss and a LASSO penalty (Tibshirani 1996), which can be thought of equivalently as MAP estimation for a Gaussian linear regression model with a Laplace prior on the coefficients. Griffin and Hoff (2017) have recently considered a test for the appropriateness of the LASSO penalty based on the empirical kurtosis for a point estimate of the coefficients, and where their test is designed to be powerful against alternative priors in an exponential power family. Here we consider the embedding of the Laplace prior into this same family for the construction of our prior-data conflict score test, and show that our method is an attractive one in this example. The second application considered relates to a problem in quantum tomography. Here the model is a multinomial, but with physical constraints on the parameter space. We consider several different prior expansions leading to statistics that are sensitive to different aspects of the prior. Section 5 gives some concluding discussion.

2 Prior-data conflict checking

In this section we first outline a framework for Bayesian model checking due to Evans and Moshonov (2006) that helps to explain some logical requirements that should be satisfied by any check for prior-data conflict. We then review the divergence-based checks considered in Nott et al. (2016) and their connections with the score-based checks suggested here. We follow this with a review of the wider prior-data conflict checking literature.

2.1 Model factorization and Bayesian model checking

Evans and Moshonov (2006) develop a general framework for Bayesian model checking based on a factorization of the joint density for parameters and data in a Bayesian model into different terms to be used for different purposes in a Bayesian analysis. Their factorization extends an earlier similar suggestion due to Box (1980). Similar to the introduction, let $\theta$ be a parameter, $y$ be data, and $T$ be a minimal sufficient statistic. If no non-trivial minimal sufficient statistic is available, we can consider some asymptotically sufficient statistic such as the maximum likelihood estimator. We again write $p(y|\theta)$ for the sampling model, $g(\theta)$
for the prior, and consider decomposing the joint model as
\[ g(\theta)p(y|\theta) = p(t)g(\theta|t)p(y|t), \]
where all terms are densities with respect to appropriate base measures, \( p(t) = \int p(t|\theta)g(\theta)d\theta \)
denotes the prior predictive density of \( T \), \( g(\theta|t) \) is the posterior density (since \( T \) is sufficient) and \( p(y|t) \) is the density of \( y \) given \( T, \theta \) which does not depend on \( \theta \) due to the sufficiency of \( T \). This decomposition can be extended further in the case of hierarchical models or when there exist ancillary statistics but we won’t consider this now.

Evans and Moshonov (2006) observe that \( g(\theta|t) \) should be used for Bayesian inference about the parameters, and suggest that \( p(y|t) \) should be used for checking the sampling model, since \( p(y|t) \) doesn’t depend on the prior, which has nothing to do with the correctness of the sampling model. When there are ancillary statistics, they can also be used for checking the sampling model. It is logical to use the remaining term \( p(t) \) for checking for prior-data conflict, since any variation in \( y \) that does not affect the likelihood is irrelevant to whether the prior and likelihood conflict. So any discrepancy used to test for prior-data conflict should be a function of a sufficient statistic, and to exclude as much irrelevant variation as possible, a minimal sufficient statistic.

Evans and Moshonov (2006) suggested using the prior predictive \( p \)-value
\[ p_{EM} = P(p(T) \leq p(t_{\text{obs}})) \] 
(4)
to check for conflict, where \( T \sim p(t) \) is a sample from the prior predictive for the minimal sufficient statistic \( T \) and \( t_{\text{obs}} \) is the observed value. Since \( T \) determines the likelihood function, (4) is measuring how surprising the observed likelihood is, in terms of the minimal sufficient statistic, under the prior. So (4) will be small if the observed value \( t_{\text{obs}} \) is in the tails of the prior predictive density for \( T \). Conditioning on ancillary statistics when they are available can also be considered in (4), as well as extensions to priors which are elicited hierarchically.

2.2 Conflict checks based on relative belief

Evans and Jang (2010) note that (4) is not invariant to the minimal sufficient statistic chosen, and suggest an invariantized version of the check. Nott et al. (2016) considered an alternative check based on prior to posterior Rényi divergences (Rényi 1961) which is also
invariant. We describe this approach in some detail since it is closely related to our proposed score checks. For the method of Nott et al. (2016), a conflict p-value is computed as

\[ p_\alpha = P(R_\alpha(Y) \geq R_\alpha(y_{obs})) \]  

(5)

where \( Y \) denotes a draw from the prior predictive distribution for \( y \), \( p(y) = \int g(\theta)p(y|\theta)d\theta \), and \( R_\alpha(y) \) denotes the prior to posterior Rényi divergence for data \( y \),

\[ R_\alpha(y) = \frac{1}{\alpha - 1} \log \int \left\{ \frac{g(\theta|y)}{g(\theta)} \right\}^{\alpha - 1} g(\theta|y)d\theta, \]

where \( \alpha > 0 \) and the case \( \alpha = 1 \) is defined by taking a limit \( \alpha \to 1 \), which corresponds to the Kullback-Leibler divergence.

Nott et al. (2016) note connections between their suggested check and the relative belief framework for inference (Evans 2015; Baskurt and Evans 2013). For a parameter of interest \( \psi(\theta) \), the relative belief function for \( \psi \) is

\[ RB(\psi|y) = \frac{g(\psi|y)}{g(\psi)}, \]

where \( g(\psi|y) \) is the posterior distribution for \( \psi \), and \( g(\psi) \) is the prior. If the relative belief is larger than 1 at a given \( \psi \), this means there is evidence for that value, whereas if it is less than 1 there is evidence against. We see that \( R_\alpha(y_{obs}) \) is a measure of the average evidence in \( y_{obs} \) or equivalently the average change in beliefs from \textit{a priori} to \textit{a posteriori}. So (5) is a measure of how much beliefs about \( \theta \) have changed from prior to posterior compared with what is expected under the prior, and is hence a measure of how surprising the data are under the prior. It is also noteworthy that relative belief inferences have been shown to possess optimal robustness to the prior properties but this robustness decreases with increasing prior data conflict, see (Al Labadi and Evans 2017). The case \( \alpha = 2 \) gives the posterior mean of the relative belief function, whereas \( \alpha \to \infty \) corresponds to the maximum relative belief. Instead of computing (5) based on the prior and posterior for \( \theta \), one can also consider the prior and posterior for a parameter of interest \( \psi \). Nott et al. (2016) also discuss hierarchical versions of their check, and this extension is also considered here later.

Because the discrepancy for the check \( R_\alpha(y) \) depends on the data only through the posterior, this discrepancy is a function of any minimal sufficient statistic, and it is invariant to the choice of sufficient statistic. There is also a connection between the check (5) and the
Jeffreys’ prior. Nott et al. (2016) show that the limiting form of the p-value is
\[ P \left( g(\theta^*)|I(\theta^*)|^{-1/2} \geq g(\theta)|I(\theta)|^{-1/2} \right) \]  (6)
where \( \theta^* \) denotes the true parameter and \( \theta \sim g(\theta) \). The p-value \( \text{(6)} \) is a measure of how far out in the tails of the prior density the true parameter is (with the prior expressed as a density with respect to the Jeffreys’ prior as support measure). The limiting value is 1 if \( g(\theta) \) is taken as the Jeffreys’ prior, which gives a heuristic reason why the Jeffreys’ prior is non-informative, since asymptotically a conflict is not possible. A similar result holds for the check of Evans and Moshonov (2006) but where the prior density is expressed with respect to the Lebesgue measure as support measure. The hierarchical extension of the checks of Nott et al. (2016) also has some relationship to reference priors (Berger et al. 2009; Ghosh 2011).

2.3 Connections between the relative belief and score checks

The score based statistic \( \text{(1)} \) is closely related to the divergence based check of Nott et al. (2016) for a number of reasons. First, it shares the property with \( \text{(5)} \) of depending on the data only through the posterior. This follows from the expression \( \text{(3)} \) for \( S(y) \), an expression which is derived from Fisher’s identity (see, for example, Cappé et al. (2005) equation (10.12)). Fisher’s identity applies when we have some model for data \( y \), with latent variables \( z \) and a parameter \( \eta \). There is a joint model for \( (y, z) \) given \( \eta \), \( p(y, z|\eta) \) say, and \( p(y|\eta) \) is obtained by integrating out the latent variables, \( p(y|\eta) = \int p(y, z|\eta)dz \). Fisher’s identity states that under appropriate regularity conditions
\[ \nabla_\eta \log p(y|\eta) = \int (\nabla_\eta \log p(y, z|\eta)) p(z|y, \eta)dz. \]
Using this formula and identifying \( \theta \) with \( z \) and \( \gamma \) with \( \eta \), the expression \( \text{(3)} \) for \( S(y) \) follows, provided that the differentiation under the integral sign required for Fisher’s identity is valid. Because the score check depends on the data only through the posterior, the statistic \( S(y) \) depends only on the data through the value of a sufficient statistic, and it is invariant to the choice of that statistic. This is desirable for a prior-data conflict check as discussed in Section 2.1. Furthermore, to apply the method it is not required to identify any non-trivial minimal sufficient statistic, since \( S(y) \) is computed directly from the posterior distribution.

As well as depending on the data only through the posterior, the score based check is
similar to the divergence based one through its connection with relative belief based inference. By rearranging Bayes’ rule with the prior \( g(\theta | \gamma) \),

\[
p(y|\gamma) = \frac{g(\theta | \gamma)p(y|\theta)}{g(\theta | y, \gamma)},
\]

so that

\[
\frac{d}{d\gamma} \log p(y|\gamma) = -\frac{d}{d\gamma} \log \frac{g(\theta | y, \gamma)}{g(\theta | \gamma)}.
\]

Hence \( S(y) \) is the derivative with respect to the expansion parameter at \( \gamma_0 \) of the negative log relative belief, evaluated at any \( \theta \). Since the right-hand side does not depend on \( \theta \), we can also average over any distribution on \( \theta \) and choosing \( g(\theta | y, \gamma) \),

\[
\frac{d}{d\gamma} \log p(y|\gamma) = -\frac{d}{d\gamma} \int \log \frac{g(\theta | y, \gamma)}{g(\theta | \gamma)} g(\theta | y, \gamma) d\theta.
\]

Hence the score-based check statistic is the negative of the derivative with respect to \( \gamma \) at \( \gamma_0 \) of the Kullback-Leibler divergence statistic of Nott et al. (2016).

A further connection between the two approaches emerges by considering the expansion \( g(\theta | \gamma) = (1 - \gamma)g(\theta) + \gamma q(\theta) \) for a fixed prior \( q(\theta) \). Using (3),

\[
S(y) = E\left( \frac{q(\theta) - g(\theta)}{g(\theta)} \mid y \right)
= E\left( \frac{q(\theta)}{g(\theta)} \mid y \right) - 1.
\]

If the Jeffreys’ prior is proper, then taking \( q(\theta) \) to be the Jeffreys’ prior, (2) becomes

\[
ps = P\left( E\left( \frac{1}{g(\theta)|I(\theta)|^{-1/2}} \mid y_{\text{obs}} \right) \leq E\left( \frac{1}{g(\theta)|I(\theta)|^{-1/2}} \mid Y \right) \right),
\]

for \( Y \sim p(y) \), and in the asymptotic limit this is equivalent to the \( p \)-value (6) obtained using the divergence based check.

## 2.4 Other approaches to prior-data conflict checking

There is an extensive existing literature on prior-data conflict checking. One class of approaches involves somehow converting the likelihood and prior information into something comparable, either through renormalization or converting the likelihood to a posterior through
a non-informative prior. A recent example of this approach is Presanis et al. (2013), where conflicts are examined locally at any node or group of nodes in a directed acyclic graph. Their work unifies and generalizes a number of previous suggestions (O’Hagan 2003; Dahl et al. 2007; Marshall and Spiegelhalter 2007; Gåsemyr and Natvig 2009). Scheel et al. (2011) consider a related approach where the model is formulated as a chain graph and at a certain node a marginal posterior distribution based on a local prior and lifted likelihood are compared. Bousquet (2008) considers an approach where ratios of prior-to-posterior Kullback-Leibler divergences are calculated, for the prior to be checked and a non-informative prior. Hierarchical extensions are also discussed. Reimherr et al. (2014) consider the difference in information required to be put into a likelihood function to obtain the same posterior uncertainty for a proper prior used in an analysis relative to a non-informative baseline prior that would be used if little prior information were available. Another method similar to those of Evans and Moshonov (2006) and Nott et al. (2016) in not requiring the formulation of any non-informative prior, is described in Dey et al. (1998), where vectors of quantiles of the posterior distribution itself are used in a Monte Carlo test using a prior predictive reference distribution. Bayarri and Castellanos (2007) review and evaluate various methods for checking the second level of hierarchical models, and advocate the partial posterior predictive p-value approach (Bayarri and Berger 2000). General discussions of Bayesian model checking which are not specifically concerned with checking for prior-data conflict are given in Gelman et al. (1996), Bayarri and Berger (2000) and Evans (2015). The method we propose here is a useful addition to the above proposals because the use of an appropriate encompassing family of priors for constructing the check gives some guidance for how to construct checks that are sensitive to conflicts in different aspects of the prior; furthermore, it does not rely on the construction of any non-informative prior for its application, which can sometimes be difficult.

There is also a growing literature on the question of what to do in a Bayesian analysis if a prior-data conflict is found. We do not discuss this further, but see Evans and Moshonov (2006) Evans and Jang (2011b) Held and Sauter (2017) and Bickel (2018) among others for some recent discussion of this issue.

3 Hierarchical extension of the score based check

When a prior distribution is elicited hierarchically, it is desirable to check the different parts of the prior as they are elicited to determine the source of the problem if a conflict occurs.
We now describe how to do this with the proposed score based checks. Let \( \theta \) be partitioned as \( \theta = (\theta_1^T, \theta_2^T)^T \) and suppose the prior has been specified as \( g(\theta) = g(\theta_1)g(\theta_2|\theta_1) \). The discussion can be generalized to the case where \( \theta \) is partitioned into more than two parts.

First, consider an expansion of the form \( g(\theta|\gamma^{(1)}) = g(\theta_1)g(\theta_2|\theta_1, \gamma^{(1)}) \), where the marginal prior for \( \theta_1 \) is held fixed but the conditional prior \( g(\theta_2|\theta_1) \) is embedded into \( g(\theta_2|\theta_1, \gamma^{(1)}) \) with \( g(\theta_2|\theta_1, \gamma_0^{(1)}) = g(\theta_2|\theta_1) \). Consider

\[
p(y|\theta_1, \gamma^{(1)}) = \int p(y|\theta)g(\theta_2|\theta_1, \gamma^{(1)})d\theta_2,
\]

and define

\[
S^{(1)}(y, \theta_1) = \frac{d}{d\gamma^{(1)}} \log p(y|\theta_1, \gamma^{(1)}) \bigg|_{\gamma^{(1)} = \gamma_0^{(1)}},
\]

and

\[
S^{(1)}(y) = E \{ S^{(1)}(y, \theta_1)|y_{\text{obs}} \}.
\]

We propose to check for conflict for the conditional prior \( g(\theta_2|\theta_1) \) using the \( p \)-value

\[
p_{S1} = P(S^{(1)}(Y) \geq S^{(1)}(y_{\text{obs}})) ,
\]

where \( Y \sim m(y) = \int g(\theta_2|\theta_1)g(\theta_1|y_{\text{obs}})p(y|\theta)d\theta \). Here it has been assumed again in the calculation of the \( p \)-value that the embedding prior family is such that a large value of \( S^{(1)}(y) \) indicates conflict.

The justification for this check is that in checking \( g(\theta_2|\theta_1) \) we should consider an appropriate check for this prior as if \( \theta_1 \) is fixed (which leads to the statistic \( S^{(1)}(y, \theta_1) \)) but then to eliminate the unknown \( \theta_1 \) we take the expectation with respect to \( \theta_1 \) under the posterior given \( y_{\text{obs}} \). So we see if there is a conflict involving \( g(\theta_2|\theta_1) \) for \( \theta_1 \) values that reflect knowledge of \( \theta_1 \) under \( y_{\text{obs}} \). The reference distribution for the check also reflects knowledge of \( \theta_1 \) under \( y_{\text{obs}} \) but using the conditional prior of \( \theta_2 \) given \( \theta_1 \) in generating predictive replicates.

To check \( g(\theta_1) \), we now consider a different expansion of the joint prior, \( g(\theta_1|\gamma^{(2)})g(\theta_2|\theta_1) \) where \( g(\theta_1|\gamma_0^{(2)}) = g(\theta_1) \) and then with \( p(y|\gamma^{(2)}) = \int p(y|\theta)g(\theta_1|\gamma^{(2)})g(\theta_2|\theta_1)d\theta \) consider the statistic

\[
S^{(2)}(y) = \frac{d}{d\gamma^{(2)}} \log p(y|\gamma^{(2)}) \bigg|_{\gamma^{(2)} = \gamma_0^{(2)}},
\]

10
and a $p$-value for the check of $g(\theta_2)$

$$p_{S_2} = P(S^{(2)}(Y) \geq S^{(2)}(y_{\text{obs}})),$$

with $Y \sim m(y) = \int p(y|\theta)g(\theta)d\theta$. It is again assumed in the $p$-value computation that the embedding prior family is such that large $S^{(2)}(y)$ indicates conflict.

These checks are similar to those considered in [Nott et al. (2016)] for their divergence based check. As explained there, in models with particular additional structure, the checks above can be modified in various ways. For example, in hierarchical models with observation or cluster specific parameters, cross-validatory versions of the check can be considered, as well as versions of partial posterior predictive checks (Bayarri and Berger 2000) in constructing $S^{(1)}(y)$ and its reference distribution. If there are sufficient or ancillary statistics at different levels this can be exploited also ([Evans and Moshonov 2006] Nott et al. 2016).

### 4 Simple examples

It is insightful to consider properties of the check (2) first in some simple examples, where calculations can be performed analytically. The examples we discuss were also given in [Evans and Moshonov (2006)] and [Nott et al. (2016)], and we make some connections with the checks (4) and (5) that they propose. In the examples below, we use the following notation, which was also used in [Nott et al. (2016)]. If $S_1(y)$ and $S_2(y)$ are two discrepancies for a Bayesian model check, and one is a monotone function of the other (as a function of $y$), we will write $S_1(y) \doteq S_2(y)$. Note that prior predictive checks based on these discrepancies will give the same result, if the appropriate tail probability is calculated.

**Example 4.1. Normal location model.**

Let $y_1, \ldots, y_n$ be a random sample, $y_i \sim N(\theta, \sigma^2)$, where $\sigma^2 > 0$ is a known variance and $\theta$ is an unknown mean. The sample mean is sufficient for $\theta$ and normally distributed, so it suffices to consider the case $n = 1$ and this will be assumed in what follows. We write $y_{\text{obs}}$ for the observed value of $y$. Suppose the prior for $\theta$ is normal, $N(\mu_0, \tau_0^2)$, where $\mu_0$ and $\tau_0^2$ are fixed hyperparameters. Next, expand the prior to $N(\mu_0, \tau^2)$, where $\tau^2$ is allowed to vary. Clearly $p(y|\tau^2)$ is a normal density, with mean $\mu_0$ and variance $\sigma^2 + \tau^2$, and hence

$$\log p(y|\tau^2) = -\frac{1}{2} \log 2\pi(\sigma^2 + \tau^2) - \frac{(y - \mu_0)^2}{2(\sigma^2 + \tau^2)},$$


which gives
\[
\frac{d^2}{d\tau^2} \log p(y|\tau^2) \bigg|_{\tau^2 = \tau_0^2} = \frac{(y - \mu_0)^2}{2(\sigma^2 + \tau_0^2)} \approx (y - \mu_0)^2.
\]

So to calculate the prior predictive \(p\)-value for the check we compare \((y_{\text{obs}} - \mu_0)^2\) to its prior predictive density. This is the same check obtained by Evans and Moshonov (2006) and Nott et al. (2016) using (4) and (5) and the corresponding \(p\)-value is (Evans and Moshonov 2006, p. 897)
\[
2 \left(1 - \Phi \left(\frac{|y_{\text{obs}} - \mu_0|}{\sqrt{\sigma^2 + \tau_0^2}}\right)\right).
\]

**Example 4.2. Binomial model**

Let \(y \sim \text{Binomial}(n, \theta)\) where \(\theta\) is unknown with a prior \(g(\theta)\) that is a beta distribution, \(\text{Beta}(a, b)\). The expansion we consider here is a geometric mixture of the \(\text{Beta}(a, b)\) prior and the Jeffreys’ prior, which is \(\text{Beta}(1/2, 1/2)\). We denote the mixing parameter by \(\gamma\) and
\[
g(\theta|\gamma) \propto \left\{\theta^{a-1}(1-\theta)^{(b-1)}\right\}^\gamma \times \left\{\theta^{-1/2}(1-\theta)^{-1/2}\right\}^{1-\gamma},
\]
which is a beta prior, \(g(\theta|\gamma) = \text{Beta}(\gamma a + (1 - \gamma)/2, \gamma b + (1 - \gamma)/2)\). Hence \(p(y|\gamma)\) is a beta-binomial probability function,
\[
p(y|\gamma) = \binom{n}{y} \frac{B(y + \gamma a + (1 - \gamma)/2, n - y + \gamma b + (1 - \gamma)/2)}{B(\gamma a + (1 - \gamma)/2, \gamma b + (1 - \gamma)/2)}.
\]

Taking logs and differentiating with respect to \(\gamma\), we obtain
\[
\frac{d}{d\gamma} \log p(y|\gamma) \bigg|_{\gamma = 1} \approx (a - 1/2) \log \tilde{\theta}_n + (b - 1/2) \log (1 - \tilde{\theta}_n) + O\left(\frac{1}{n}\right), \quad (9)
\]
where \(\tilde{\theta}_n = (y + a)/(n + a + b)\) is the posterior mean of \(\theta\) under the prior \(g(\theta)\). Equation (9)
is equivalent to
\[
\frac{d}{d\gamma} \log p(y|\gamma) \bigg|_{\gamma=1} = \log g(\bar{\theta}_n) + \frac{1}{2} \log|I(\bar{\theta}_n)| + O\left(\frac{1}{n}\right),
\]
(10)
where \( I(\theta) = n/(\theta(1 - \theta)) \) is the Fisher information. This matches the asymptotic form of the check \([5]\) considered in Section 4 of \cite{nott2016} We have already established in Section 3 that an arithmetic mixture involving the Jeffreys’ prior would lead to a similar result. On the other hand, if instead of considering a geometric mixture of the Jeffreys’ prior with the Beta\((a, b)\) prior we instead consider a geometric mixture of the uniform distribution with Beta\((a, b)\) instead, then we obtain, using a similar argument,
\[
\frac{d}{d\gamma} \log p(y|\gamma) \bigg|_{\gamma=1} = (a - 1) \log \tilde{\theta}_n + (b - 1) \log(1 - \tilde{\theta}_n) + O\left(\frac{1}{n}\right),
\]
(11)
so that
\[
\frac{d}{d\gamma} \log p(y|\gamma) \bigg|_{\gamma=1} = \log g(\bar{\theta}_n) + O\left(\frac{1}{n}\right),
\]
(12)
and this is a discrepancy that is asymptotically equivalent to the check suggested by \cite{evans2006} (see also \cite{evans2011}). Again, it is easy to see following the argument of Section 3 that an arithmetic mixture involving the uniform will lead to the same result. So for appropriate expansions of the Beta\((a, b)\) prior we can obtain checks asymptotically equivalent to both \((4)\) and \((5)\).

**Example 4.3. Normal location-scale model, hierarchically structured check**

Consider \( y = (y_1, \ldots, y_n) \sim N(\mu_{1n}, \sigma^2 I_n) \), where \( \mu \) and \( \sigma^2 \) are both unknown, \( 1_n \) denotes an \( n \)-vector of ones and \( I_n \) denotes the \( n \times n \) identity matrix. The prior is \( g(\mu, \sigma^2) = g(\mu|\sigma^2)g(\sigma^2) \), where \( g(\sigma^2) = IG(a, b) \) \((IG(a, b)\) denotes the inverse gamma density with parameters \( a \) and \( b \)) and \( g(\mu|\sigma^2) = N\left(\mu_0, \frac{\sigma^2}{\lambda_0}\right) \). Here \( \mu_0, \lambda_0, a \) and \( b \) are fixed hyperparameters. This is an example of a hierarchically specified prior, and \( g(\mu, \sigma^2) \) is the conjugate normal inverse gamma prior for this problem.

Consider checking the mean component of the prior, \( g(\mu|\sigma^2) \). We expand the prior \( g(\mu|\sigma^2) \) to \( g(\mu|\sigma^2, \lambda) = N\left(\mu_0, \frac{\sigma^2}{\lambda}\right) \), where now \( \lambda \) is allowed to vary \((i.e. it is no longer fixed at \( \lambda_0 \)). We have
\[
y|\sigma^2, \lambda \sim N\left(\mu_{01n}, \sigma^2 \left( I_n + \frac{1}{\lambda} E_n \right)\right),
\]

13
where $E_n$ denotes an $n \times n$ matrix of ones. Note that
\[
\left( I_n + \frac{1}{\lambda} E_n \right)^{-1} = I_n - \frac{1}{\lambda + n} E_n,
\]
and
\[
\log p(y|\sigma^2, \lambda) = -\frac{n}{2} \log 2\pi \sigma^2 - \frac{1}{2} \log |I_n + \frac{1}{\lambda} E_n| - \frac{1}{2\sigma^2} (y - \mu_0 1_n)^T (I_n + \lambda^{-1} E_n)^{-1} (y - \mu_0 1_n)
\]
\[
= -\frac{n}{2} \log 2\pi \sigma^2 - \frac{1}{2} \log |I_n + \frac{1}{\lambda} E_n| - \frac{1}{2\sigma^2} \text{tr} \left( (y - \mu_0 1_n)(y - \mu_0 1_n)^T (I_n - (\lambda + n)^{-1} E_n) \right),
\]
where $\text{tr}(\cdot)$ denotes the matrix trace. This gives
\[
\frac{d}{d\lambda} \log p(y|\sigma^2, \lambda) = \frac{1}{2\lambda^2} \text{tr} \left( (I_n + \lambda^{-1} E_n)^{-1} E_n \right) - \frac{1}{2\sigma^2} (y - \mu_0 1_n)^T ((\lambda + n)^{-2} E_n)(y - \mu_0 1_n).
\]
Noting that
\[
(y - \mu_0 1_n)^T ((\lambda + n)^{-2} E_n)(y - \mu_0 1_n) = \frac{n^2 (\bar{y} - \mu_0)^2}{(\lambda + n)^2},
\]
we obtain
\[
S^{(1)}(y) \doteq (\bar{y} - \mu_0)^2,
\]
which is the Kullback-Leibler based check considered in Nott et al. (2016). Nott et al. (2016) also note that the check is very similar to the one suggested in Evans and Moshonov (2006), p. 909.

In this example we could have expanded the prior $g(\mu|\sigma^2)$ by allowing the mean rather than the scale hyperparameter to vary, embedding $g(\mu|\sigma^2)$ into the family $g(\mu|\sigma^2, \mu') = N(\mu', \sigma^2/\lambda_0)$, where $\mu'$ is not necessarily equal to $\mu_0$. If we do this, we obtain $S^{(1)}(y) \doteq (\bar{y} - \mu_0)$, and computation of a two-sided $p$-value gives that this is equivalent to a check using the statistic $(\bar{y} - \mu_0)^2$, so that the two different ways of expanding the prior lead to the same result in this case. An example where two different embedding families lead to useful and quite different answers is considered later.
5  More complex examples

5.1  Checking the appropriateness of a LASSO penalty

As a more complex application, a problem discussed in Griffin and Hoff (2017) is now con-
dered. The goal is to assess whether or not a penalty term, used in a penalized regression to
induce sparsity, is contradicted by the data. Since the use of the penalty term they consider
is equivalent to employing a prior together with MAP estimation, checking the penalty term
can be addressed by prior-data conflict checking, and that is how we approach it here.

Example 5.1. Many means problem with LASSO penalty

To start we restrict to the many means context with no predictors, since the analysis is
easier and the behavior reflects what happens in the more general situation. Suppose \( \bar{x} \sim N(\mu, (\sigma^2/m)I_n) \) is observed with \( \mu \in \mathbb{R}^n \) and there is a belief that \( \mu \) is sparse, namely, many
of the means satisfy \( \mu_i = 0 \). It is also assumed that \( \sigma^2 = 1 \) is known, as nothing material
beyond computational complexity is added to the analysis by placing a prior on this quantity.

For the prior on \( \mu \), consider a product prior where each \( \mu_i \) has density

\[
g(\nu | \tau, q) = \frac{q}{2\tau} \left( \frac{\Gamma(3/q)}{\Gamma(1/q)^3} \right)^{1/2} \exp \left\{ - \left( \frac{\Gamma(3/q)}{\Gamma(1/q)} \right)^{q/2} \frac{|\nu|}{\tau} \right\}, \tag{13}
\]

for \( \nu \in \mathbb{R}^1 \). This is the exponential power family of priors that was considered in Griffin and
Hoff (2017), and it can be shown that if \( \mu_i \) has prior \( (13) \), then \( E(\mu_i) = 0 \) and \( \text{Var}(\mu_i) = \tau^2 \).

When \( q = 2 \), the prior is normal, \( N(0, \tau^2) \), and when \( q = 1 \), the prior is a Laplace rescaled
by \( \tau/\sqrt{2} \). As \( q \to 0 \) this family of priors induces greater sparsity.

A question of interest is whether or not the Laplace prior obtained when \( q = 1 \) conflicts
with the data, as this corresponds to the popular LASSO penalty (Tibshirani 1996). Griffin
and Hoff (2017) effectively compare the observed value of the kurtosis statistic

\[
k(\bar{x}) = \sum_{i=1}^{n} \frac{x_i^4}{\left( \sum_{i=1}^{n} x_i^2 \right)^2},
\]

with its prior distribution when \( q = 1 \). Actually they use the prior distribution of the
kurtosis of a sample of \( n \) from the prior itself as the reference distribution for computational
reasons, but we use the more appropriate prior distribution of \( k(\bar{x}) \) for this comparison. If
the observed \( k(\bar{x}) \) lies in the tails of its prior predictive density, then this is an indication
that the double exponential prior is in conflict, and a modification of the prior is needed.
The p-value value for the check is $2 \min \{ P(k(\bar{X}) < k(\bar{x})), P(k(\bar{X}) > k(\bar{x})) \}$ where $P$ is the prior predictive measure, and if we decide a conflict occurs when this p-value is less than 0.05, then the left and right critical values for testing $q = 1$ are (1.65, 6.72) when $n = 10$, and (3.01, 10.07) when $n = 100$. So if $n = 10$ and $k(\bar{x}) < 1.65$ or $k(\bar{x}) > 6.72$, then a prior-data conflict exists.

With $g(\mu | \bar{x}, \tau, q)$ denoting the conditional posterior for $\mu$ given $\tau$ and $q$, the score function for assessing sensitivity to $q$ is

$$S(\bar{x} | \tau) = \int g(\mu | \bar{x}, \tau, 1) \frac{d}{dq} \log g(\mu | \tau, q) \bigg|_{q=1} d\mu,$$  \hspace{1cm} (14)

namely, the posterior expectation of the derivative of $g(\cdot | \tau, q)$ with respect to $q$ evaluated at $q = 1$. A simple calculation leads to

$$\frac{d}{dq} \log g(\mu | \tau, q) \bigg|_{q=1} = A(1) + B(1) \sum_{i=1}^{n} \left| \frac{\mu_i}{\tau} \right| + C(1) \sum_{i=1}^{n} \left| \frac{\mu_i}{\tau} \right| \log \left| \frac{\mu_i}{\tau} \right|,$$  \hspace{1cm} (15)

where $A(1) = 1 + 3(\psi(1) - \psi(3))/2$, $B(1) = -\Gamma(3)^{1/2}(\log \Gamma(3) + \psi(1) - 3\psi(3))/2$, and $C(1) = -\Gamma(3)^{1/2}$. As previously, $\psi(x)$ denotes the digamma function. Rather than compute the expectation in (14), this is approximated by $\hat{S}(\bar{x} | \tau)$ where the estimates $\mu_i = \bar{x}_i$ are substituted into (15). Furthermore, it is assumed hereafter that the elicited value of $\tau$ is $\tau = 1$. In general $\tau$ is chosen such that the effective support of the prior, which can be defined as a central interval containing say 0.99 of the prior probability when $q = 1$, covers all the $\mu_i$ values. As such, any prior data conflict that arises is due to the choice of $q = 1$ alone. Figure [1] is a plot of the null distribution of $\hat{S}(\bar{x} | 1)$ when $n = 10$ and $m = 20$, which leads to critical values (0.408, 1.117). When $n = 100$ and $m = 20$, the critical values are given by (0.670, 0.898).

To compare our approach to the method of Griffin and Hoff (2017), the power of the tests was compared. Figure [2] shows plots of the power functions of the kurtosis and the approximate score statistics in different situations. It is seen that the approximate score approach compares quite favorably with the method of Griffin and Hoff (2017).

**Example 5.2. Regression with LASSO penalty**

We now extend from the many means setting to a regression problem with data $y = X\beta + \sigma z$ where $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ is unknown, $z \sim N(0, I_n)$ and again $\sigma^2 = 1$ is assumed. Also for simplicity it is assumed that the $\beta_i$ can all be treated equivalently so there is no intercept term which must be treated differently. The prior on $\beta$ is taken to be a product prior with
Figure 1: Density histogram based on $10^5$ values generated from the prior distribution of the approximate statistic based on $n = 10, m = 20, \tau = 1$.

Figure 2: Power functions when using the kurtosis statistic and approximate score statistic for testing $q = 1$ when $n = 10$ (left) and $n = 100$ (right) with $m = 20$ and $\tau = 1$. 
the same prior \([13]\) placed on each \(\beta_i\). In practice, the columns of \(X\) can be standardized to each have sum 0 and unit length. With this assumption, the mean of the \(i\)th coordinate of \(y\) is \(x_i^T \beta\), where \(x_i^T\) is the \(i\)-th row of \(X\), and note that \(x_i \in [-1, 1]^p\) because of the standardization. As such, a bound on the means \(x_i^T \beta\) that holds for all \(i\) implies that \(\tau\) can be chosen to guarantee that the bounds on the means hold with high prior probability. Accordingly, our concern for prior-data conflict can focus on \(q\), and again the case \(q = 1\) is considered.

For the kurtosis and approximate score, similar formulae are obtained as with the many means case, and here the \(\beta_i\) are estimated via least-squares. When \(k > n\), so that \(X\) is no longer of full rank, the Moore-Penrose estimates are used as these minimize the length of the estimate vector and that seems appropriate when considering sparsity. Griffin and Hoff (2017) used a ridge estimator in the non-full rank case but this made little difference in the results reported here.

A simulation study was conducted as in Section 2 of Griffin and Hoff (2017). Data were generated from the regression model with \(\sigma^2 = \tau^2 = 1\), for \(n = 25, 50, 100\) and \(200\) and \(p = 25, 50, 75\) and \(100\). The entries of \(X\) were drawn from the standard normal distribution. For \(10^3\) independent replicates of \(X\) and \(\beta\) (drawn from the prior with \(q = 1\)) the power was estimated for a grid of values for \(q\) from 0.1 to 2 in steps of 0.1. The cutoff level in the test to determine the existence of a prior-data conflict was 0.05. The simulation results are given in Figure 3. It is seen that the approximate score does quite well, and in certain cases, namely, when \(p < n\), can do better than the test based on the kurtosis statistic.

The approximate score doesn’t do as well as the kurtosis statistic when \(p > n\). This is felt in part to be due to the simulation performed. For when generating \(X\) via independent standard normals the matrix is of rank \(n\) with probability 1 when \(p \geq n\). So this situation is somewhat like having \(n\) observations with \(n\) independent variables and in such a case it is not possible to criticize the model, as it will fit the data perfectly, let alone the prior. In practice, if we wish to check both the prior and likelihood components of the model, and the coefficient vector is known to be sparse, then a preliminary screening of variables (Fan and LV 2018) can reduce \(p\) to \(p < n\) before penalized regression is performed, and the score based check would seem to be preferable for checking the appropriateness of the penalty in that case. Such a screening procedure could also be implemented together with data splitting, where the screening and analysis are done using disjoint subsets of the data.
Figure 3: Simulation results to compare the powers of the kurtosis and approximate score statistic.
5.2 Checking a truncated Dirichlet prior in a constrained multinomial model for quantum state estimation

The data acquired in measurements on quantum systems are fundamentally probabilistic because — as a matter of principle, not for lack of knowledge — one can only predict the probabilities for the various outcomes but not which outcome will be observed for the next quantum system to be measured. Therefore, the interpretation of quantum-experimental data requires the use of statistical tools, where the constraints that identify the set of physically allowed probabilities must be enforced. We shall illustrate prior checking in this context for a simple example, after setting the stage by recalling some basic tenets of quantum theory.

In the formalism of quantum theory, the Hilbert space operators of a \( D \)-dimensional quantum system can be represented by \( D \times D \) matrices; for simplicity, we shall not distinguish between the operators and the matrices that represent them. There is, in particular, the statistical operator \( \rho \) that describes the state of the quantum system, which is positive-semidefinite and has unit trace. A measurement with \( K \) outcomes is specified by \( K \) positive-semidefinite probability operators (also commonly referred to as POVMs) \( \Pi_1, \Pi_2, \ldots, \Pi_K \), one for each outcome. When such a measurement is performed on independent and identically prepared systems, we get a click of one of the detectors for each of the measured systems, and the probability that the \( k \)th detector will click for the next quantum system is \( \theta_k = \text{tr}(\rho \Pi_k) \) ("Born’s rule"). The unit sum of the probabilities, \( \sum_{k=1}^{K} \theta_k = 1 \), is ensured for any \( \rho \) by the unit sum of the \( \Pi_k \)s, \( \sum_{k=1}^{K} \Pi_k = I_D \), where as previously \( I_D \) denotes the \( D \times D \) identity matrix.

The \( K \)-tuplets of probabilities, \( \theta = (\theta_1, \theta_2, \ldots, \theta_K) \), constitute a convex set in the \((K-1)\)-simplex but usually they do not exhaust this simplex (for an example, see Figure 6 below). The permissible probabilities are those consistent with \( \rho \geq 0 \), and the actual constraints obeyed by the \( \theta_k \)s result from the properties of the \( \Pi_k \)s. We denote the convex set of permissible \( \theta \)s by \( \Theta \).

After measuring \( N \) identically prepared copies of the quantum system, thereby counting \( n_k \) clicks of the \( k \)th detector, we have the data \( y = (n_1, n_2, \ldots, n_K) \) with \( \sum_{k=1}^{K} n_k = N \). The problem of inferring the statistical operator \( \rho \) from the data \( y \) is the central theme of quantum state estimation (Paris and Řeháček 2004; Teo 2015), and Bayesian methods are well-suited for this task (Shang et al. 2013; Li et al. 2016). Enforcing \( \rho \geq 0 \), or the corresponding implied constraints on the probabilities \( \theta \), is crucial and often challenging. We shall consider a rather simple example below, with \( D = 2 \) and \( K = 3 \).

The data \( y \) are modeled as multinomial with parameter \( \theta = (\theta_1, \theta_2, \ldots, \theta_K) \in \Theta \). We
recall that a random vector $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_K)$ follows a Dirichlet distribution with parameter $(\alpha_1, \alpha_2, \ldots, \alpha_K)$, denoted by $\text{Dir}(\alpha_1, \ldots, \alpha_K)$, if it has density

$$g(\Delta) \propto \prod_{k=1}^{K} \Delta_k^{\alpha_k-1}$$

on the $(K - 1)$-dimensional region $\{\Delta : \Delta_k \geq 0 \text{ for } k = 1, \ldots, K - 1 \text{ with } \sum_{k=1}^{K-1} \Delta_k \leq 1\}$ where $\Delta_K = 1 - \sum_{k=1}^{K-1} \Delta_k$. Consider a prior for $\theta$ that is proportional to a Dirichlet prior $\text{Dir}(\alpha_0 q_1, \ldots, \alpha_0 q_K)$ on $\Theta$ where $q = (q_1, \ldots, q_K)$ is a location parameter, $\sum_{k=1}^{K} q_k = 1$, and $\alpha_0 > 0$ is an overall precision parameter. That is, the prior is

$$g(\theta|\alpha_0, q) \propto \prod_{k=1}^{K} \theta_k^{\alpha_0 q_k - 1} \text{ for } \theta \in \Theta.$$ 

The hyperparameters can be set by eliciting a point estimate of $\theta$ as the value for $q$, and calibrating the precision parameter according to the desired prior uncertainty about $\theta$.

5.2.1 Two prior expansions

We first consider two different families of expansions of the above prior, which do not attempt to check violations of the physical constraint on $\theta$. Later we consider an expansion suitable for checking the physical constraint in a simple situation ($D = 2$ and $K = 3$), and where the formula (3) does not hold without some modification. To minimize notation, in all the prior families considered below the expansion parameter in the prior is denoted by $\gamma$, although it should be noted that in different families the interpretation of this parameter differs.

In our first prior expansion, similar to our earlier binomial example, we consider a geometric mixture of the original prior with the Jeffreys prior, which is the Dirichlet prior $\text{Dir}(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})$ constrained to $\theta \in \Theta$. Mixing with the Jeffreys prior thickens the tails of the original prior, and as such this family may be helpful for constructing an overall test of conflict. This idea leads to a family that is still constrained Dirichlet, proportional to $\text{Dir}(\delta_1, \ldots, \delta_K)$ with $\delta_k = (1 - \gamma)\alpha_0 q_k + \frac{1}{2}\gamma$. The corresponding density function is denoted by

$$g^{(1)}(\theta|\alpha_0, q, \gamma) \propto \prod_{k=1}^{K} \theta_k^{\delta_k - 1} \text{ for } \theta \in \Theta.$$ 

The original prior is obtained when $\gamma = 0$.

Our second prior family also derives from considering a constrained Dirichlet density,
Dir(δ′₁, ..., δ′ₖ), with δ′ₖ = α₀ q′ₖ, q′₁ = q₁ + γ and q′ₖ = qₖ − γ/(K − 1) for k = 2, ..., K. We see that ∑ₖ=₁^K q′ₖ = 1 and ∑ₖ=₁^K δ′ₖ = α₀, so that the overall precision parameter is kept fixed while changing the location parameters by increasing q₁ by γ, with the other q′ₖ adjusted to maintain the unit-sum constraint. This family of priors is constructed to focus particularly on conflicts involving the first component θ₁ of θ, and we obtain the original prior at γ = 0. For this family we write

\[ g^{(2)}(θ|α₀, q, γ) ∝ \prod_{k=1}^{K} θ_k^{δ'_k - 1} \quad \text{for } θ ∈ Θ. \]

5.2.2 Score statistics for the checks

Consider first the family \( g^{(1)} \). Apart from terms not depending on θ, we have

\[
\frac{d}{dγ} \log g^{(1)}(θ|α₀, q, γ) = \frac{d}{dγ} \sum_{k=1}^{K} (δ_k - 1) \log θ_k, \]

\[
= \sum_{k=1}^{K} (α₀ q_k - \frac{1}{2}) \log θ_k \]

and, upon using (3),

\[ S(y) = \sum_{k=1}^{K} (α₀ q_k - \frac{1}{2}) E(\log θ_k | y). \]

On the other hand, for the family \( g^{(2)} \), we obtain that, apart from terms not depending on θ,

\[
\frac{d}{dγ} \log g^{(2)}(θ|α₀, q, γ) = \frac{d}{dγ} \left( \sum_{k=1}^{K} (δ'_k - 1) \log θ_k \right), \]

\[
= α₀ \log θ_1 - α₀ \sum_{k=2}^{K} \frac{\log θ_k}{K - 1}, \]

and using (3) again,

\[ S(y) = E(\log θ_1 | y) - \sum_{k=2}^{K} \frac{E(\log θ_k | y)}{K - 1}. \]

We see that in the case of this prior family the components are no longer treated symmetrically in the check, with conflicts involving θ₁ being the focus.
Data simulated under $g^{(1)}$  

| Power | $\gamma$ |
|-------|----------|
| 0.0   | 0.0      |
| 0.2   | 0.2      |
| 0.4   | 0.4      |
| 0.6   | 0.6      |
| 0.8   | 0.8      |
| 1.0   | 1.0      |

Data simulated under $g^{(2)}$  

| Power | $\gamma$ |
|-------|----------|
| 0.0   | 0.0      |
| 0.2   | 0.2      |
| 0.4   | 0.4      |
| 0.6   | 0.6      |
| 0.8   | 0.8      |
| 0.00  | 0.00     |
| 0.10  | 0.10     |
| 0.20  | 0.20     |
| 0.30  | 0.30     |

Figure 4: Power of checks based on families $g^{(1)}$ and $g^{(2)}$ when data are simulated under the prior predictive for $g^{(1)}$ (left) and under the prior predictive for $g^{(2)}$ (right).

5.2.3 Power comparison for the checks

We consider the case of $K = 3$ with $q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and $\alpha_0 = 30$. For the family $g^{(1)}$, $\gamma = 0$ corresponds to a Dir$(10, 10, 10)$ prior, and $\gamma = 1$ to a Dir$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ prior. For the $g^{(2)}$ expansion, $\gamma = 0$ is a Dir$(10, 10, 10)$ prior, and $\gamma = \frac{1}{3}$ is a Dir$(20, 5, 5)$ prior. We examine the power of the checks in two cases, where a $p$-value smaller than 0.05 is considered to be a conflict. In the first case, we consider simulating data under the prior predictive for the $g^{(1)}$ expansion, for values of $\gamma = i/20$, $i = 0, \ldots, 20$. Figure 4 (left) shows how the power of the two checks varies with $\gamma$, and we note that the check based on $g^{(1)}$ is more powerful, when the prior predictive for $g^{(1)}$ is used for simulating the data. The power is approximated at each value of $\gamma$ based on 500 simulations. Next, we consider simulating data under the prior predictive for $g^{(2)}$, for values of $\gamma = i/60$, $i = 0, \ldots, 20$. Again 500 simulations are performed at each $\gamma$ to compare power for the two checks, and Figure 4 (right) shows again that it is the check corresponding to the family used to generate the data that is more powerful. The different families are powerful against different kinds of conflict with the original prior, and the expansion of the prior used can be constructed with this in mind.
5.2.4 A simple quantum measurement scenario

The simplest genuine quantum system is that of a binary alternative \((D = 2)\), a qubit. Here, we have \(2 \times 2\) matrices for all operators, and the usual parameterization of the statistical operator is

\[
\rho = \frac{1}{2} \begin{pmatrix} 1 + s_3 & s_1 - is_2 \\ s_1 + is_2 & 1 - s_3 \end{pmatrix}
\]

with \(s_1^2 + s_2^2 + s_3^2 \leq 1\).

If we regard, as we shall, the three real parameters \(s_1, s_2, s_3\) as Cartesian coordinates of a point, then there is a one-to-one correspondence between the quantum states of a qubit and the three-dimensional unit ball.

Two-outcome measurements on qubits realize the situation of coin tossing and do not exhibit features particular to quantum physics. We shall, therefore, consider 3-outcome measurements \((K = 3)\), for which we choose the \(\Pi_k\)s in accordance with

\[
\Pi_k = w_k \begin{pmatrix} 1 & e^{-i\phi_k} \\ e^{i\phi_k} & 1 \end{pmatrix}
\]

with \(w_k > 0\) for \(k = 1, 2, 3\).

The corresponding probabilities

\[
\theta_k = w_k(1 + s_1 \cos \phi_k + s_2 \sin \phi_k)
\]

(16)
do not involve \(s_3\), so that no information about this state parameter is gained from such a measurement, and the three \(\theta_k\)s are restricted by \(s_1^2 + s_2^2 \leq 1\). The relevant parameter space is now the unit disk in the \(s_1, s_2\) plane, the intersection of this plane with the unit ball.

The angles \(\phi_1, \phi_2, \phi_3\) divide the unit disk into three pie slices; see Figure 5. The condition

\[
\Pi_1 + \Pi_2 + \Pi_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\]
or, equivalently,

\[
\theta_1 + \theta_2 + \theta_3 = 1 \quad \text{for all } s_1, s_2
\]
determines the weights \(w_k\) and they will be positive if each slice is less than half of the pie; we take this for granted.

The symmetric case of three pie slices of equal size is that of the so-called trine measurement. Upon choosing \(\phi_1 = 0\) by convention, we then have \(\phi_2 = \frac{2}{3} \pi\) and \(\phi_3 = \frac{4}{3} \pi\), and the
Figure 5: The angles $\phi_1$, $\phi_2$, $\phi_3$ in the probabilities slice the circular unit-disk pie into three pieces, each slice smaller than half of the pie.

The probabilities of the trine measurement are

$$\theta_1 = \frac{1}{3}(1 + s_1), \quad \theta_2 = \frac{1}{6}(2 - s_1 \pm \sqrt{3}s_2),$$

which are constrained by

$$s_1^2 + s_2^2 = (3\theta_1 - 1)^2 + 3(\theta_2 - \theta_3)^2 \leq 1.$$  

In view of $\theta_1 + \theta_2 + \theta_3 = 1$, this can be equivalently, and more symmetrically, written as $\theta_1^2 + \theta_2^2 + \theta_3^2 \leq \frac{1}{2}$.

For $\phi_1 = 0$, $\phi_2 = \pi - \varphi$, $\phi_3 = \pi + \varphi$ we get a symmetrically distorted trine, for which the probabilities are

$$\theta_1 = \frac{1}{2}((\sin \gamma)^2(1 + s_1), \quad \theta_2 = \frac{1}{4}[1 + (\cos \gamma)^2 - s_1(\sin \gamma)^2 \pm 2s_2 \cos \gamma]$$

with $\cos \gamma = \tan(\frac{1}{2}\varphi)$, and the analog of (18) reads

$$\left(\frac{2\theta_1}{(\sin \gamma)^2 - 1}\right)^2 + \left(\frac{\theta_2 - \theta_3}{\cos \gamma}\right)^2 \leq 1 \quad \text{for} \quad (\theta_1, \theta_2, \theta_3) \in \Theta_\gamma,$$
where we note that the set of permissible $\theta$s depends on the value of $\gamma$. [Note: Later this distortion parameter $\gamma$ will play the role of the generic expansion parameter $\gamma$.] We recover the ideal trine for $\varphi = \frac{1}{3}\pi$ and $(\cos \gamma)^2 = \frac{1}{3}$, and the limiting cases of $\varphi = \frac{1}{3}\pi$ and $\varphi = 0$ yield degenerate 2-outcome measurements of no further interest. In the recent experiment by Len et al. (2018) different symmetrically distorted trines were realized (for measuring the polarization qubit of a photon), among them $(\cos \gamma)^2 = 0.1327$ for which $y = (n_1, n_2, n_3) = (180, 31, 30)$ were the counts of detection events. While the actual counts in the experiments were about ten times as many, namely $(1802, 315, 303)$ as communicated by author Y. L. Len, we are using these smaller numbers here because prior-data conflicts are less of an issue in data-dominated situations. However, even if posterior inferences are insensitive to a prior-data conflict in large data settings, it is still of interest to detect the conflict, since this indicates a lack of scientific understanding in setting up the model.

The probability space for the symmetrically distorted trine has a simple geometry, illustrated in Figure 6. The unit disk in the $s_1, s_2$ plane accounts for all $s_1, s_2$ pairs for which the probabilities in (19) obey the constraint in (20); that is: the unit disk represents the set $\Theta$ of permissible probabilities. The $s_1, s_2$ pairs for which one of the probabilities in (19) has a chosen value, mark a line in the $s_1, s_2$ plane, and different fixed values for the same $\theta_k$ yield a set of parallel lines. The particular three lines with $\theta_1 = 0$ or $\theta_2 = 0$ or $\theta_3 = 0$ are tangential to the unit circle and intersect where either $(\theta_1, \theta_2, \theta_3) = (1, 0, 0)$ or $(0, 1, 0)$ or $(0, 0, 1)$; the triangle thus defined is the 2-simplex for the probabilities in (19). For the ideal trine, it is an equilateral triangle; for the symmetrically distorted trine, we have an isosceles triangle with vertices at $(s_1, s_2) = (2(\sin \gamma)^{-2} - 1, 0)$ and $(-1, \mp (\cos \gamma)^{-1})$. For the general case of (16), there is an analogous construction with a triangle with no particular symmetry for the 2-simplex. Among all these triangles, the equilateral triangle for the ideal trine has the smallest area.

5.2.5 A prior family for checking the physical constraints

We now consider the symmetrically distorted trine with its $\gamma$-dependent set of permissible $\theta$s, in accordance with (20). If it is suspected that the trine measurement set-up was not properly balanced, this gives a natural family of priors for performing our check.

In such a situation where the support for the prior changes with $\gamma$, [3] can no longer be
Figure 6: For the probabilities \((\theta_1, \theta_2, \theta_3)\) of the symmetrically distorted trine measurement in (19), the physically allowed values correspond to the points on the unit disk (on and inside the black unit circle, the blue circle in Fig. 5) while the probability 2-simplex is a triangle whose sides touch the unit circle. The probabilities associated with top left and bottom left vertices of the triangle are \((\theta_1, \theta_2, \theta_3) = (0, 0, 1)\) and \((0, 1, 0)\), respectively, and the vertex on the right has \((1, 0, 0)\). The blue equilateral triangle is for the ideal-trine probabilities in (17) when \((\cos \gamma)^2 = \frac{1}{3}\), the green triangle is for \((\cos \gamma)^2 = \frac{1}{3} - \frac{1}{12}\), and the red triangle is for \((\cos \gamma)^2 = \frac{1}{3} + \frac{1}{12}\). The dashed blue lines show where \(\theta_3 = 0, 0.2, 0.4, 0.6, 0.8, 1\) for the blue triangle.
used to compute $S(y)$ without some modification. Instead of using (3), we expand (1) as

$$
S(y) = \frac{d}{d\gamma} \log p(y|\gamma) \bigg|_{\gamma=\gamma_0} = \frac{\frac{d}{d\gamma} p(y|\gamma)}{p(y|\gamma_0)} ,
$$

with

$$
p(y|\gamma) = \int_{\Theta_\gamma} p(y|\theta) g(\theta|\gamma) d\theta = \int_{\Theta_\gamma} \left( N_{n_1, n_2, n_3} \right) \theta_1^{n_1} \theta_2^{n_2} \theta_3^{n_3} g(\theta|\gamma) d\theta .
$$

The changing support makes $\frac{d}{d\gamma} p(y|\gamma)$ inconvenient to evaluate numerically. To deal with this, we switch from integrating over $\theta$ to integrating over $s_1$ and $s_2$ and use polar coordinates in the $s_1, s_2$ plane,

$$
s_1 = r \cos \phi, \quad s_2 = r \sin \phi
$$

with $0 \leq r \leq 1$ and $0 \leq \phi \leq 2\pi$, to enforce the constraints (19) for any value of $\gamma$. With this, we now have a situation where the sampling model itself changes with $\gamma$:

$$
p(y|r, \phi, \gamma) = \left( N_{n_1, n_2, n_3} \right) \theta_1^{n_1} \theta_2^{n_2} \theta_3^{n_3} \bigg|_{\theta_k \text{ from (19)}} \text{ with } s_1 + is_2 = re^{i\phi}
$$

**Dirichlet prior over the physical space** Consider the case of a Dirichlet prior over the physical space, i.e., $g(\theta|\gamma) \propto \theta_1^{n_1-1} \theta_2^{n_2-1} \theta_3^{n_3-1} I_\gamma(\theta)$, where $I_\gamma(\theta)$ is the indicator function,

$$
I_\gamma(\theta) = \begin{cases} 
1 & \text{if } \theta \text{ is permissible: } \theta \in \Theta_\gamma, \\
0 & \text{otherwise: } \theta \notin \Theta_\gamma.
\end{cases}
$$

Under the $r, \phi$ parameterization, we then have

$$
S(y) = \frac{d}{d\gamma} \log \frac{H_1(\gamma)}{H_2(\gamma)} \bigg|_{\gamma=\gamma_0}
$$
Figure 7: Probability of detecting a conflict at a p-value threshold 0.05 for data simulated under the prior predictive for different prior hyperparameter $\gamma$, for the case where $\cos^2(\gamma_0) = \frac{1}{3}$ and $N = 50$. The blue and green lines show power for increasing and decreasing $\gamma$, respectively.

where

$$H_1(\gamma) = \int \theta_1^{n_1+\alpha-1}\theta_2^{n_2+\alpha-1}\theta_3^{n_3+\alpha-1} \left| \theta_k \text{ from (19)} \right| r dr d\phi \text{ with } s_1 + is_2 = re^{i\phi}$$

and

$$H_2(\gamma) = \int \theta_1^{\alpha-1}\theta_2^{\alpha-1}\theta_3^{\alpha-1} \left| \theta_k \text{ from (19)} \right| r dr d\phi \text{ with } s_1 + is_2 = re^{i\phi}$$

The two-dimensional integrals above can be performed numerically. Figure 7 shows the power of the conflict score test when the underlying prior has a $\gamma$ value deviating from $\gamma_0$ for the case of a flat prior, where $\alpha = \{1, 1, 1\}$. Here, the two curves are obtained by flipping the sign of the score function. We see that they behave as we expect them to, one being sensitive to conflicts caused by $\gamma$ values too large, the other by values too small.

An example from a quantum experiment  As mentioned above, symmetrically distorted trines were realized in the experiment recently conducted by [Len et al. (2018)]. We now consider the data $y = (180, 31, 30)$ for the distorted trine with $(\cos \gamma)^2 = 0.1327$. Sup-
pose a prior which is flat over the symmetric trine is chosen \( (\cos \gamma)^2 = \frac{1}{3} \). Then the score-based conflict check gives a \( p \)-value of 0.00004, indicating a conflict. If instead the correct \( \gamma \) is chosen, the same test yields a \( p \)-value of 0.56.

6 Discussion

We have considered a new approach to constructing prior-data conflict checks based on embedding the prior used for the analysis into a larger family and then considering a marginal likelihood score statistic for the expansion parameter. The main advantage of this technique is that through the choice of the prior expansion we can construct checks which are sensitive to different aspects of the prior.

There are a number of ways in which our work could be extended. In Section 4, we considered checking for the appropriateness of the LASSO penalty in penalized regression, but it would be interesting also to check other commonly used sparse signal shrinkage priors. For example, the generalized Beta mixture of Gaussians family of Armagan et al. (2011) would provide a suitable prior expansion for checking the horseshoe prior (Carvalho et al. 2009; Carvalho et al. 2010) in our framework. It would also be interesting to use the score-based approach for checking priors on hyperparameters in nonparametric models like Gaussian processes. In the nonparametric setting, priors can be crucial for limiting flexibility and avoiding overfitting, but it can also be difficult to understand the predictive implications of an informative prior which makes checking the prior important.

Acknowledgements

This work is funded by the Singapore Ministry of Education and the National Research Foundation of Singapore. David Nott was supported by a Singapore Ministry of Education Academic Research Fund Tier 1 grant (R-155-000-189-114). Hui Khoon Ng is funded by a Yale-NUS College start-up grant. Michael Evans was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Al Labadi, L. and M. Evans (2017). Optimal robustness results for relative belief inferences and the relationship to prior-data conflict. Bayesian Analysis 12(3), 705–728.
Armagan, A., M. Clyde, and D. B. Dunson (2011). Generalized beta mixtures of Gaussians. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 24, pp. 523–531.

Baskurt, Z. and M. Evans (2013). Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. Bayesian Analysis 8(3), 569–590.

Bayarri, M. J. and J. O. Berger (2000). P values for composite null models (with discussion). Journal of the American Statistical Association 95, pp. 1127–1142.

Bayarri, M. J. and M. E. Castellanos (2007). Bayesian checking of the second levels of hierarchical models. Statistical Science 22, 322–343.

Berger, J. O., J. M. Bernardo, and D. Sun (2009). The formal definition of reference priors. The Annals of Statistics 37(2), 905–938.

Bickel, D. R. (2018). Bayesian revision of a prior given prior-data conflict, expert opinion, or a similar insight: a large-deviation approach. Statistics 52, 552–570.

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

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

Cappé, O., E. Moulines, and T. Ryden (2005). Inference in Hidden Markov Models (Springer Series in Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2009). Handling sparsity via the horseshoe. In D. van Dyk and M. Welling (Eds.), Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, Volume 5 of Proceedings of Machine Learning Research, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, pp. 73–80. PMLR.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. Biometrika 97(2), 465–480.

Dahl, F. A., J. Gásemyr, and B. Natvig (2007). A robust conflict measure of inconsistencies in Bayesian hierarchical models. Scandinavian Journal of Statistics 34, 816–828.

Dey, D. K., A. E. Gelfand, T. B. Swartz, and P. K. Vlachos (1998). A simulation-intensive approach for checking hierarchical models. Test 7, 325–346.

Evans, M. (2015). Measuring Statistical Evidence Using Relative Belief. Taylor & Francis.
Evans, M. and G. H. Jang (2010). Invariant p-values for model checking. The Annals of Statistics 38, 512–525.

Evans, M. and G. H. Jang (2011a). A limit result for the prior predictive applied to checking for prior-data conflict. Statistics and Probability Letters 81(8), 1034 – 1038.

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

Evans, M. and H. Moshonov (2006). Checking for prior-data conflict. Bayesian Analysis 1, 893–914.

Fan, J. and J. Lv (2018). Sure independence screening. In Wiley StatsRef: Statistics Reference Online. doi:10.1002/9781118445112.stat08043, Wiley.

Gelman, A., X.-L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, 733–807.

Ghosh, M. (2011). Objective priors: An introduction for frequentists. Statistical Science 26, 187–202.

Gåsemyr, J. and B. Natvig (2009). Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models. Scandinavian Journal of Statistics 36, 822–838.

Griffin, M. and P. D. Hoff (2017). Testing sparsity-inducing penalties. arXiv: 1712.06230.

Held, L. and R. Sauter (2017). Adaptive prior weighting in generalized regression. Biometrics 73(1), 242–251.

Len, Y. L., J. Dai, B.-G. Englert, and L. A. Krivitsky (2018). Unambiguous path discrimination in a two-path interferometer. Phys. Rev. A 98, 022110.

Li, X., J. Shang, H. K. Ng, and B.-G. Englert (2016). Optimal error intervals for properties of the quantum state. Phys. Rev. A 94, 062112.

Marshall, E. C. and D. J. Spiegelhalter (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. Bayesian Analysis 2, 409–444.

Nott, D. J., X. Wang, M. Evans, and B.-G. Englert (2016). Checking for prior-data conflict using prior-to-posterior divergences. arXiv:1611.00113.

O’Hagan, A. (2003). HSS model criticism (with discussion). In P. J. Green, N. L. Hjort, and S. T. Richardson (Eds.), Highly Structured Stochastic Systems, pp. 423–453. Oxford University Press.
Paris, M. and J. Řeháček (Eds.) (2004). *Quantum State Estimation* (1st ed.), Volume 649 of *Lecture Notes in Physics*. Springer-Verlag Berlin Heidelberg.

Presanis, A. M., D. Ohlssen, D. J. Spiegelhalter, and D. D. Angelis (2013). Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science* 28, 376–397.

Reimherr, M., X.-L. Meng, and D. L. Nicolae (2014). Being an informed Bayesian: Assessing prior informativeness and prior likelihood conflict. arXiv1406.5958.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, Calif., pp. 547–561. University of California Press.

Scheel, I., P. J. Green, and J. C. Rougier (2011). A graphical diagnostic for identifying influential model choices in Bayesian hierarchical models. *Scandinavian Journal of Statistics* 38(3), 529–550.

Shang, J., H. K. Ng, A. Sehrawat, X. Li, and B.-G. Englert (2013). Optimal error regions for quantum state estimation. *New Journal of Physics* 15(12), 123026.

Teo, Y. S. (2015). *Introduction to Quantum-State Estimation*. Singapore: World Scientific.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58, 267–288.