Alternatives to post-processing posterior predictive p -values

JØRUND GÅSEMYR and IDA SCHEEL

Department of Mathematics, University of Oslo

Running headline: Alternatives to post-processing ppp

Abstract

The posterior predictive p -value (ppp) was invented as a Bayesian counterpart to classical p -values. The methodology can be applied to discrepancy measures involving both data and parameters, and can hence be targeted to check for various modelling assumptions. The interpretation can however be difficult since the distribution of the ppp -value under modelling assumptions varies substantially between cases. A calibration procedure has been suggested, treating the ppp -value as a test statistic in a prior predictive test. In this paper we suggest that a prior predictive test may instead be based on the expected posterior discrepancy, which is somewhat simpler, both conceptually and computationally. Since both these methods require simulation of a large posterior parameter sample for each of an equally large prior predictive data sample, we furthermore suggest to look for ways to match the given discrepancy by a computation saving conflict measure. This approach is also based on simulations, but only requires sampling from two different distributions representing two contrasting information sources about a model parameter. The conflict measure methodology is also more flexible in that it handles non-informative priors without difficulty. We compare the different approaches theoretically in some simple models and in a more complex applied example.

Key words: Bayesian model criticism, calibration, conflict measure, posterior predictive p -values, prior predictive p -values

1 Introduction

1.1 Background

The posterior predictive p -value (ppp) of Gelman *et al.* (1996) was invented as a Bayesian counterpart to classical p -values, to be used in tests for goodness of fit for a given Bayesian model. Unlike classical tests, the methodology can be applied not only to test statistics, but also to discrepancy measures $d(\mathbf{y}, \boldsymbol{\theta})$ depending on the unknown parameter vector $\boldsymbol{\theta}$, which is equipped with the prior distribution $\pi(\boldsymbol{\theta})$, as well as on observed data $\mathbf{y}$. The basic assumption is that $\mathbf{y}$ is a realisation from the model $\mathbf{Y} \sim f_{\mathbf{Y}|\boldsymbol{\theta}}$. The construction involves a posterior sample $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta} | \mathbf{y})$, and a corresponding posterior predictive random variable $\mathbf{Y}^{\text{rep}} \sim f_{\mathbf{Y}|\boldsymbol{\theta}}$. Hence, the value of the marginal distribution of $\mathbf{Y}^{\text{rep}}$ at $\mathbf{y}^{\text{rep}}$ is $\int f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}^{\text{rep}} | \boldsymbol{\theta})\pi(\boldsymbol{\theta} | \mathbf{y})d\boldsymbol{\theta}$. The formal definition of the posterior predictive p -value $ppp(\mathbf{y})$ is (cf. Equation (7) in Gelman *et al.* (1996))

$$ppp(\mathbf{y}) = P(d(\mathbf{Y}^{\text{rep}}, \boldsymbol{\theta}) \geq d(\mathbf{y}, \boldsymbol{\theta}) | \mathbf{y}). \quad (1)$$

The computation of an estimate of $ppp(\mathbf{y})$ is a straightforward by-product of the posterior simulation used for the Bayesian analysis. Indeed, based on S independent samples $(\mathbf{y}_s^{\text{rep}}, \boldsymbol{\theta}_s)$ from the distribution $f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}_s^{\text{rep}} | \boldsymbol{\theta}_s)\pi(\boldsymbol{\theta}_s | \mathbf{y})$, $s = 1, 2, \dots, S$, $ppp(\mathbf{y})$ can be estimated as the sample mean of $I(d(\mathbf{y}_s^{\text{rep}}, \boldsymbol{\theta}_s) \geq d(\mathbf{y}, \boldsymbol{\theta}_s))$, where I denotes the indicator function. Hence, the posterior predictive distribution of $d(\mathbf{Y}^{\text{rep}}, \boldsymbol{\theta})$ acts as a reference distribution for samples from the posterior distribution of $d(\mathbf{y}, \boldsymbol{\theta})$ given data $\mathbf{y}$. Any aspect of the model that might be of interest could be explored by choosing different discrepancy measures. A surprisingly low ppp -value could lead to that aspect of the model being modified.

The interpretation of this p -value is however problematic because there is no common reference distribution for the ppp -value itself, analogous to the uniform distribution for the classical p -value. The distribution of $ppp(\mathbf{Y})$ for replicates $\mathbf{Y}$ from the model is often far from uniform, usually much more concentrated around 1/2, and worse, varies substantially from case to case. Such problems have been pointed out by Bayarri & Berger (2000), Bayarri & Castellanos (2007), Hjort *et al.* (2006) and Dahl (2006) among others. Stein-

bakk & Storvik (2009) report the same kind of problems for extensions of the *ppp*-value that build on discrepancy measures defined in terms of unobserved latent variables $\mathbf{X}$ instead of observed data $\mathbf{y}$. Probably to some extent as a result of such criticism, Gelman (2013) states that caution in interpreting diagnostics that depend strongly on parameters or latent data is recommended. It is a theoretically appealing feature of Bayesian model checking that the test variable can depend on unknown quantities, but statements of fit based on such realised discrepancies can be difficult to interpret.

Box (1980) suggests using the prior predictive distribution, i.e. the distribution for hypothetical data $\mathbf{Y}$ from the given model, as a reference distribution for model evaluation. Appealing to the factorisation of the joint density for $(\mathbf{Y}, \boldsymbol{\theta})$ at $(\mathbf{y}, \boldsymbol{\theta})$ as (Equation (1.2) in Box (1980))

$$p(\mathbf{y}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta} | \mathbf{y})p(\mathbf{y}),$$

where the first factor is the basis for Bayesian inference, he argues that the second factor naturally serves as a basis for checking the validity of the model. In particular, the value of the density function can be used as a test statistic. Denoting by P^M the probability measure for $\mathbf{Y}$ according to the given model, this gives rise to the prior predictive *p*-value

$$prpp(\mathbf{y}) = P^M(p(\mathbf{Y}) \leq p(\mathbf{y})), \quad (2)$$

see Equation (2.3) in Box (1980) for an example. To address specific aspects of the model, the suggestion is to replace $\mathbf{Y}$ with targeted checking functions $g(\mathbf{y})$ in (2). If the relevant densities are unimodal, these *p*-values are tail probabilities. The resulting prior predictive *p*-values are uniformly distributed under the model by construction. The approach requires an informative and realistic prior distribution.

The approach of Box (1980) can not directly accommodate discrepancy measures that depend on the unknown parameter $\boldsymbol{\theta}$. However, there are various ways to construct test statistics based on such measures for use in a prior predictive approach. For instance, for discrepancy measures that are pivotal quantities, Johnson (2007) uses quantiles in the posterior distribution to this end, primarily to be applied in informal graphical checks and approximate numerical measures. When these checks are indecisive, he resorts to the post-

processing methodology of Hjort *et al.* (2006), see Johnson (2007), Equation (7). Comparing this methodology to other approaches is the main theme in the present paper, and the frequently cited article Hjort *et al.* (2006) is henceforth referred to as HDS.

Starting from a specific discrepancy measure $d(\mathbf{y}, \boldsymbol{\theta})$, the idea in HDS amounts to using the posterior predictive p -value $ppp(\mathbf{y})$ itself as a test statistic in a prior predictive test, to obtain what they call a calibrated posterior predictive p -value $cppp(\mathbf{y})$. The definition is (cf. Equation (4) of HDS)

$$cppp(\mathbf{y}) = P^M(ppp(\mathbf{Y}) \leq ppp(\mathbf{y})). \quad (3)$$

For discrepancy measures based on latent variables instead of data, Steinbakk & Storvik (2009) define similarly calibrated versions of the extended ppp -values.

Again, an informative and realistic prior distribution is needed for the calibrated ppp -value to make sense, although Section 9.3 in HDS contains some ideas on how to proceed in the case of a vague prior.

Also, in the general case the computation of $cppp(\mathbf{y})$ requires a double simulation scheme where a large posterior sample of parameter vectors $\boldsymbol{\theta}$ has to be simulated, for example by Markov Chain Monte Carlo (MCMC), for each of an equally large number of simulated data vectors $\mathbf{y}_i, i = 1, 2, \dots$ from the model. In very complex models, with running times of, say, an hour or more needed to obtain a posterior sample suitable for inference, this can be extremely or even prohibitively computer intensive. Accordingly, HDS states in Section 4.2 that one should look for ways to simplify the computational burden.

1.2 A computer intensive and a computation saving alternative to postprocessing of ppp -values

In many cases the potential model deviation addressed by a certain discrepancy measure $d(\mathbf{y}, \boldsymbol{\theta})$ is also reflected in a conflict between different information sources concerning a particular parameter or subvector of the parameter vector $\boldsymbol{\theta}$. A conflict measure approach, introduced previously by several authors (see below for detailed references), can then be used. This approach is based on contrasting the different information sources and leads to

a drastic reduction in the computational cost compared to the *cphp* approach. Because of its computational advantages we propose to use such a conflict measure approach whenever possible. If a conflict measure suited for the discrepancy under consideration is not available, we propose an alternative calibrated measure, which we call calibrated expected posterior discrepancy (*cepd*). This is an original suggestion of the present paper, and we start by introducing this measure.

Viewed from another angle, $ppp(\mathbf{y})$ is one particular choice among others of a test statistic obtained by integrating out $\boldsymbol{\theta}$ with respect to the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}$ from a function which typically is monotonic in $d(\mathbf{y}, \boldsymbol{\theta})$. Indeed, (1) can be written as

$$ppp(\mathbf{y}) = \int P(d(\mathbf{Y}^{\text{rep}}, \boldsymbol{\theta}) \geq d(\mathbf{y}, \boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}.$$

In the present paper we discuss the relationship to the alternative test statistic obtained by simply integrating out $\boldsymbol{\theta}$ from the discrepancy measure itself when used in a prior predictive test, i.e. replacing $ppp(\mathbf{y})$ by $E_{\boldsymbol{\theta} \mid \mathbf{y}}(d(\mathbf{y}, \boldsymbol{\theta}) \mid \mathbf{y})$ in (3). The corresponding p -value is termed the calibrated expected posterior discrepancy and denoted by $cepd(\mathbf{y})$. The formal definition is given in Theorem 1.

Like the *cphp* approach, the *cepd* approach is based on prior predictive testing and hence requires double simulation in the general case. In contrast, prior predictive testing can be avoided altogether if instead we can use the conflict measure approach introduced in Marshall & Spiegelhalter (2007), Dahl *et al.* (2007) and Gåsemyr & Natvig (2009), nicely reviewed and applied in Presanis *et al.* (2013). Gåsemyr & Natvig (2017) provides a more recent review, also summarising aspects of the theory appearing in Gåsemyr (2016). The basic idea behind a conflict measure is to consider potential conflict between the different sources of information regarding $\boldsymbol{\theta}$. This approach also requires simulation in general but, whenever applicable, is computationally advantageous, since it is based on the simulation of samples of $\boldsymbol{\theta}$ from only two different distributions, each of which is defined in terms of either the prior or the likelihood, or a combination of these information sources, but involves no random draws $\mathbf{Y}$ from the model. Details on the approach are given in the context of the various examples discussed later in the paper.

A natural first step in the comparison of our suggested alternatives to the *cphp* approach is to consider the normal models forming the core of the HDS-paper, consisting of analytic calculations for location parameters in standard normal models (Sections 2, 4 and 5), and an application of the resulting theory to such a model for a speedskating data set (Section 6). These models do not really represent computational challenges, but are attractive because they allow theoretical analysis. The discrepancy measures considered turn out to address conflict between prior and likelihood concerning a location parameter. In the present paper we broaden the perspective by considering in addition discrepancies between group means in a normal random effects model, and also discrepancies concerning the precision parameter. Furthermore, we compare the different approaches in a more complex applied example. In most of the theoretical examples we find the different methods to be equivalent, implying also that the conflict measures are well calibrated in the sense of being uniformly distributed under the model. These are novel results. Such exact equivalence and uniformity can only be expected to hold in a limited number of models. Nevertheless we think that the analysis indicates a fundamentally sound scaling of the conflict measures, permitting a reasonably consistent interpretation across models. This is supported by theoretical results in more general normal models in Gåsemyr & Natvig (2009), the empirical studies in Dahl *et al.* (2007) concerning also models with non-normal distributions, and the theoretical results in Gåsemyr (2016) concerning non-symmetric distributions. We think that the decision on whether to modify the model should in any case not be based solely on the exact value of a *cphp*($\mathbf{y}$), a *cepd*($\mathbf{y}$) or a conflict measure. For convenience we often phrase our discussion in the language of hypothesis testing, but this does not necessarily reflect the decision theoretic setting for the methodologies under consideration.

The paper is organised as follows. In Section 2 we illustrate the different concepts introduced above on a simple normal example, an example that was used for the same purpose in Section 2.1, 2.2 and 4.1 in HDS. Section 2 also contains a quite useful basic result on the connection between *ppp*($\mathbf{y}$) and the posterior expectation of the discrepancy measure when used in prior predictive testing. In Section 3 we consider discrepancy concerning location as well as precision in a model where both location and precision are unknown. In Section

4 we consider the random effects normal model. Section 5 contains an applied example, focusing on discrepancy and conflict involving a multidimensional parameter in a situation where analytic solutions are unavailable. Some concluding remarks are given in Section 6. A technical lemma and some proofs are given in the Appendix.

2 A motivating example and a basic result

2.1 The *cppp* approach and the alternative *cepd* approach

Consider the simple normal model of HDS, Section 2.1, i.e.

$$Y_1, \dots, Y_n \text{ IID } \sim N(\theta, \lambda^{-1}), \theta \sim N(\theta_0, (c_0\lambda)^{-1}). \quad (4)$$

where the precision parameter λ is assumed known. As in HDS, we define a discrepancy measure by

$$d(\mathbf{y}, \theta) = n\lambda(\bar{y} - \theta)^2. \quad (5)$$

This discrepancy measure is a pivotal quantity, and it is obvious that $d(\mathbf{Y}^{\text{rep}}, \theta)$ has the distribution of a χ_1^2 -variable Z_1 given θ . Denote by Ψ_1 the cumulative distribution function (cdf) of the χ_1^2 -distribution, and let $\bar{\Psi}_1 = 1 - \Psi_1$. Also let D be a random draw from the distribution of $d(\mathbf{y}, \theta)$ given $\mathbf{y}$. Denoting by I the indicator function, it follows from (1) that

$$\begin{aligned} ppp(\mathbf{y}) &= P(Z_1 \geq D) = E_{Z_1, D}(I(Z_1 \geq D)) = E_D(E_{Z_1}(I(Z_1 \geq D))) \\ &= E_D(\bar{\Psi}_1(D)) = E_{\theta|\mathbf{y}}(\bar{\Psi}_1(d(\mathbf{y}, \theta) | \mathbf{y})) = \int \bar{\Psi}_1(d(\mathbf{y}, \theta))\pi(\theta | \mathbf{y})d\theta. \end{aligned} \quad (6)$$

The *cppp*-approach prescribes using (6) as a test statistic in a prior predictive test. This raises the following question: What is magic about the function $\bar{\Psi}_1$ in this procedure?

Question 1. *Would it be equally or more reasonable to replace $\bar{\Psi}_1$ by some other monotone function, for instance the identity function, in (6)?*

Using the identity function would lead to using the posterior expectation of the discrepancy measure itself, i.e. the quantity D_{average} in the notation of Gelman *et al.* (1996), as a test statistic. We will in fact show that in this simple case, this leads to the same test as the

cppp-approach. Indeed, we will show below that the conditions for the following theorem are satisfied in this case. Condition b) of the theorem means that the possible outcomes $\mathbf{y}$ can be ordered corresponding to an ordering of the cumulative distribution functions for $d(\mathbf{y}, \theta)$ given $\mathbf{y}$.

Theorem 1. *Let θ be a continuously distributed parameter or parameter vector in a Bayesian statistical model for data $\mathbf{Y} \in \mathcal{Y}$, and let $d(\mathbf{Y}, \theta) \geq 0$ be a continuous discrepancy measure. For $\mathbf{Y} \in \mathcal{Y}$, let $F_{\mathbf{Y}}$ denote the cdf and $f_{\mathbf{Y}}$ the density for the posterior distribution of $d(\mathbf{Y}, \theta)$ given $\mathbf{Y}$. Suppose that*

a) the distribution of $d(\mathbf{Y}^{rep}, \theta)$ given θ has a fixed cdf $H(x)$, which does not depend on neither the observed $\mathbf{y}$ nor θ , and

b) if $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, then either $F_{\mathbf{y}_1} \leq F_{\mathbf{y}_2}$ (i.e. $F_{\mathbf{y}_1}(x) \leq F_{\mathbf{y}_2}(x)$ for all $x \geq 0$), or vice versa.

Then

$$cppp(\mathbf{y}) = P^M(F_{\mathbf{Y}} \leq F_{\mathbf{y}}) = P^M(E_{\theta|\mathbf{Y}}(d(\mathbf{Y}, \theta) | \mathbf{Y}) \geq E_{\theta|\mathbf{y}}(d(\mathbf{y}, \theta) | \mathbf{y})) \stackrel{\text{def}}{=} cepd(\mathbf{y}). \quad (7)$$

The acronym *cepd* stands for calibrated expected posterior discrepancy.

Proof. For every $\mathbf{Y} \in \mathcal{Y}$ we have as in (6), using condition a),

$$ppp(\mathbf{Y}) = E_{\theta|\mathbf{Y}}(\bar{H}(d(\mathbf{Y}, \theta) | \mathbf{Y})) = \int_0^\infty \bar{H}(x) f_{\mathbf{Y}}(x) dx.$$

Denoting by h the density of H and using integration by parts this gives

$$ppp(\mathbf{Y}) = \int_0^\infty h(x) F_{\mathbf{Y}}(x) dx.$$

Using condition b) it follows that

$$ppp(\mathbf{Y}) \leq ppp(\mathbf{y}) \text{ if and only if } F_{\mathbf{Y}} \leq F_{\mathbf{y}}. \quad (8)$$

The first equality of (7) follows directly from (3) and (8). The second equality follows by noting that $E_{\theta|\mathbf{Y}}(d(\mathbf{Y}, \theta) | \mathbf{Y}) = \int_0^\infty \bar{F}_{\mathbf{Y}}(x) dx$. $\square$

If the conditions of Theorem 1 are satisfied, an estimate of $cppp(\mathbf{y})$ can be obtained as the proportion of samples $\mathbf{y}^{(r)}, r = 1, 2, \dots$ from the model for which the estimated value of $E_{\theta|\mathbf{y}^{(r)}}(d(\mathbf{y}^{(r)}, \theta) | \mathbf{y}^{(r)})$ exceeds $E_{\theta|\mathbf{y}}(d(\mathbf{y}, \theta) | \mathbf{y})$. Hence, once the need for a prior predictive

calibration based on a sample $\mathbf{y}^{(r)}, r = 1, 2, \dots$, is recognised, the conversion of a posterior sample of $d(\mathbf{y}^{(r)}, \theta)$ to a posterior predictive p -value for each $\mathbf{y}^{(r)}$ seems unnecessary. In fact, if it is difficult to sample $\mathbf{Y}^{\text{rep}}$ from H some computational saving may even be obtained by replacing $c\text{ppp}(\mathbf{y})$ by $c\text{epd}(\mathbf{y})$. If the conditions for Theorem 1 are not satisfied, the two measures may differ, but it is not at all obvious that $c\text{ppp}(\mathbf{y})$ should then necessarily be preferable. Section 3.2 provides an example pointing in the opposite direction. In addition, the $c\text{epd}$ approach is based directly on posterior samples of $d(\mathbf{y}, \theta)$, and hence in our view has the advantage of being conceptually simpler than the $c\text{ppp}$ approach, which in addition is based on random posterior predictive samples of $\mathbf{Y}^{\text{rep}}$ given θ , as well as the conversion of these samples to a p -value.

Considering our simple normal model, we have already noted that condition a) is satisfied, H being the χ_1^2 cdf Ψ_1 in this case. To prove condition b), we use an expression given in HDS for a variable distributed according to the posterior distribution of $d(\mathbf{y}, \theta)$ given $\mathbf{y}$ in this situation, namely

$$(n/(c_0 + n))[N + (\sqrt{\lambda}c_0(\bar{y} - \theta_0)/\sqrt{c_0 + n})]^2, \quad (9)$$

where $N = \sqrt{(c_0 + n)\lambda}[-\theta + ((n(\bar{y} + c_0\theta_0)/(c_0 + n))]$ is a standard normal variable given $\mathbf{y}$ and λ . We will return to this formula later, but for the moment it is sufficient to note that by Lemma 2 given in the Appendix the distribution of (9) satisfies condition b) of Theorem 1, the stochastic ordering being determined by the statistic $T(\mathbf{Y}) = (\bar{Y} - \theta_0)^2$. (The variable U of the lemma is just a constant in this case; the scalar v corresponds to $\bar{y} - \theta_0$, and the variable S corresponds to N). This proves (7), and hence answers Question 1 affirmatively in our simple normal case.

In general, a significant computational saving can be achieved if the stochastic ordering of Theorem 1 is determined by such a statistic $T(\mathbf{Y})$, computable without involving the posterior distribution of $d(\mathbf{Y}, \theta)$, which by Theorem 1 satisfies

$$c\text{ppp}(\mathbf{y}) = c\text{epd}(\mathbf{y}) = P^{\text{M}}(T(\mathbf{Y}) \geq T(\mathbf{y})). \quad (10)$$

For model (4) with discrepancy (5), we have in particular

$$c\text{ppp}(\mathbf{y}) = c\text{epd}(\mathbf{y}) = P^{\text{M}}((\bar{Y} - \theta_0)^2 \geq (\bar{y} - \theta_0)^2). \quad (11)$$

Thus, just a prior predictive test based on this statistic would be sufficient for the calculation of $c_{ppp}(\mathbf{y})$ and $c_{epd}(\mathbf{y})$ in this case. Of course, as also demonstrated in HDS, Equation (12), not even a prior predictive sample of $\mathbf{Y}$'s would be necessary in this case. Indeed, arguing directly instead of via the full distribution of $ppp(\mathbf{Y})$ as in HDS, $\bar{Y} - \theta_0$ is normally distributed with expectation 0 and precision $((1/c_0\lambda) + (1/n\lambda))^{-1} = c_0n\lambda/(c_0 + n)$ under the model. It follows from (11) that

$$c_{ppp}(\mathbf{y}) = c_{epd}(\mathbf{y}) = \bar{\Psi}_1((c_0n\lambda/(c_0 + n))(\bar{y} - \theta_0)^2). \quad (12)$$

If an exceptionally small value of $c_{ppp}(\mathbf{y})/c_{epd}(\mathbf{y})$ is observed, the standard action would be to re-examine the reasoning behind the choice of the prior expectation θ_0 . The prior predictive methodology underlying (12) is based on informative priors, but if (12) nevertheless is calculated and found to be small under an intended non-informative prior, reducing c_0 to reduce prior precision would be more relevant. Although the prior might work satisfactorily as a non-informative prior even when the observed data are quite unlikely under the model (as discussed in Section 4 of Gelman (2013)), such an adjustment of the prior precision could hardly do any harm.

2.2 The conflict measure approach

It would be overoptimistic to hope to identify a statistic such as $(\bar{Y} - \theta_0)^2$ in every possible situation. Therefore, a second natural question arises:

Question 2. *Is there a generic diagnostic approach to identification of model deviations that serves the same purpose as the c_{ppp} approach, i.e. turning a certain data- and parameter dependent discrepancy measure into a useful well calibrated parameter-free diagnostic measure, but with a lighter computational burden?*

In the present example we note that $\mathbf{y}$ provides information about θ by pointing to a likely location $\bar{y}$ for θ , while the prior is another source of information pointing to θ_0 as the likely location. By (9), the value of $d(\mathbf{y}, \theta)$ for random samples θ from the posterior distribution will typically be larger the less in agreement these two sources of information are. Applying

this kind of reasoning more generally, a possibility in many cases might be to examine the degree of conflict between information about θ from data $\mathbf{y}$ vs. prior information, without involving the discrepancy measure $d(\mathbf{y}, \theta)$ at all. The idea of assessing the degree of conflict between such information sources is pursued in O’Hagan (2003), which has been a source of inspiration for the measures introduced in Dahl *et al.* (2007), Marshall & Spiegelhalter (2007) and Gsemyr & Natvig (2009), reviewed and applied in Presanis *et al.* (2013) and further analysed in Gsemyr (2016). See also the recent review Gsemyr & Natvig (2017). These measures can be applied for any parameter in a Bayesian model. In general, the methodology admits grouping the various information sources about a specific parameter into two disjoint subsets in different ways, and by contrasting such groups allows assessment of different kinds of possible conflict concerning this parameter. The general philosophy behind this kind of grouping is to avoid using the same piece of information twice.

For the kind of data - prior conflict that is relevant in the present context, the idea is to normalise the likelihood $p(\mathbf{y} \mid \theta)$ to a density, generically denoted by $g_c(\theta)$ and called the likelihood information contribution. Alternatively, g_c can be viewed as the posterior density for θ given $\mathbf{y}$, based on the improper prior 1 for θ . Due to non-uniqueness with respect to parametrisation, other non-informative priors are preferable in some cases. This is discussed in Gsemyr (2016). Relevant for the conflict studied in Section 3.2 is the recommendation to use the improper prior $1/\theta$ if the likelihood is proportional to a gamma density for θ . Having defined g_c , samples θ_c^* are simulated from this density. Also, independent samples θ_p^* are generated from the prior density $g_p(\theta) = p(\theta \mid \theta_0) = \pi(\theta)$, the prior information contribution. In general, g_p is not necessarily a prior distribution (see e.g. the example of Section 4 where g_p is a posterior distribution given part of the data), but note that when it is, samples may usually be generated directly, without MCMC. The indices p and c stand for ”parent” and ”child” respectively, referring to a graphical representation of the model such as the one depicted in Figure 1, which represents the model discussed in the next section. In the model (4), θ_0 is a parent node of θ , while the data $\mathbf{y}$ consists of child nodes of θ . With this representation in mind, we also refer to g_p and g_c as parent and child node information contributions respectively.

In general, conflict measures can be defined for any node in such a graphical model description. This includes bottom nodes representing parts of or all the observed data $\mathbf{y}$, leading to conflict measures corresponding to cross validation or predictive model evaluation. It also includes top nodes representing fixed basic parameters such as θ_0 in (4). The prior information contribution is then represented by this fixed parameter value instead of a random sample. The likelihood information contribution $p(\theta \mid \theta_0)$, which is the basis for defining g_c , contains the unknown parameter θ , which in this context appears as a nuisance parameter. In general, if nuisance parameters are present in the likelihood and/or prior information contributions, they are integrated out in the formation of g_c and g_p , which are then referred to as integrated information contributions. How this is done is indicated in the context of the applications discussed in the next sections. We use conflict about a top node in the model analysed in Section 3.1, extending this sketchy described set up to a similar model with unknown precision. Both data node conflict and top node conflict are introduced and discussed in Gåsemyr & Natvig (2009) and summarised in Gåsemyr & Natvig (2017).

In Gåsemyr & Natvig (2009) the distribution function G for the difference $\delta = \theta_p^* - \theta_c^*$ is used to define several conflict measures. The following conflict measure uses the density g corresponding to G in analogy with the *prpp*-value (2) proposed by Box (1980):

$$c_\theta^4 = P^G(g(\delta) \geq g(0)). \quad (13)$$

This measure may be estimated by means of a kernel density estimate for g obtained from a random sample of δ . A large value of c_θ^4 indicates that $\delta = 0$, corresponding to identical samples $\theta_p^* = \theta_c^*$, is located in the tail of G . The following proposition gives the c_θ^4 -measure for our present example, summarises the results for the approaches of Section 2.1 and shows that all approaches give equivalent results. The proof is given in the Appendix.

Proposition 1. *In the normal model (4) we have*

$$c_\theta^4 = \Psi_1((c_0 + n)^{-1} c_0 n \lambda (\theta_0 - \bar{y})^2).$$

Also, with the discrepancy measure (5),

$$c_{ppp}(\mathbf{y}) = c_{epd}(\mathbf{y}) = P^M((\bar{Y} - \theta_0)^2 \geq (\bar{y} - \theta_0)^2) = 1 - c_\theta^4.$$

In general, no analytic answer as in Proposition 1 exists, and the conflict measure approach requires generation of samples from two different distributions. But this is a vastly smaller computational burden than the double simulation needed to calculate the calibrated p -values when analytic solutions are unavailable, cf. the last paragraph of Section 1.1. Since efficient routines for density estimation exist, we think in view of Proposition 1 that it is fair to say that the conflict measure approach suggests itself as a possible, perhaps not completely generic, but at least partial answer to Question 2. Moreover, an alternative conflict measure built directly on the cumulative distribution function G of δ , and hence requiring no density estimate of g , is

$$c_\theta^3 = 1 - 2 \min(G(0), \bar{G}(0)). \quad (14)$$

Comparing (13) and (14), it is obvious that these measures coincide if the density g is symmetric and unimodal, as is the case in our present example.

The measure (14) is easily estimated by counting the proportion of sample values of δ that exceed 0. The advantage of (13) is, however, that it can be generalised to the case when θ is a vector. In fact, a linear regression model with a p -dimensional parameter vector β as discussed in Section 5 of HDS, could have been chosen as a basic example, both for the present and the next section. We chose the simpler model to focus more on ideas than on mathematical technicalities. Since the symmetry of the normal density might be suspected to be essential for the neat relationships demonstrated in Proposition 1 concerning the location parameter, some effort will instead be devoted to analyse discrepancy and conflict concerning the precision parameter (see Section 3.2). In Section 3 we also indicate how nuisance parameters are handled. Simulation based conflict analysis for a parameter vector is used and compared to the *cppp* and *cepd* approaches in the example in Section 5, where analytic solutions are unavailable. This example shows that computationally the c_θ^4 measure compares quite favourably to the two calibration based measures even when density estimation for a ten dimensional vector is involved.

Remark 1. *It is seen from Proposition 1 that for given values of θ_0, n and $\bar{y}$, the conflict increases and the *cppp*-value decreases as c_0 increases. Thus, the more informative the prior*

distribution is, the greater is the conflict. This is a desirable behaviour. It allows us to avoid conflict by not trusting the prior guess of location too much, i.e. by choosing a vague prior. But it is in strange contrast with the suggestion in Section 10.1 of HDS of choosing the most conservative prior that still does not clash with the data. This suggestion may be motivated by numerical results, given in Section 9.1 of HDS, for an example where, we think rather untypically, the cppp-value increases as the prior precision increases.

3 Discrepancies in a normal-gamma model

In this section we consider the situation when both location and scale parameters are unknown. We will consider discrepancies both for location with precision as a nuisance parameter and vice versa.

The model under consideration is the same as in Section 2.3 in HDS, with generalisation to a regression model in Section 5.3 of that paper:

$$Y_1, \dots, Y_n \text{ IID } \sim N(\mu, \lambda^{-1}), \mu \sim N(\mu_0, (c_0\lambda)^{-1}), \lambda \sim \text{gamma}(\lambda; a/2, b/2), \quad (15)$$

where $\text{gamma}(\lambda; a/2, b/2)$ denotes the density for the gamma distribution with shape parameter $a/2$ and scale parameter $b/2$. This model is shown graphically in Figure 1. Although linking the uncertainties concerning the prior and data precisions in this way is not always reasonable, we want to analyse this model in order to stay close to the treatment in HDS. According to (15) the prior density for (μ, λ) is

$$\begin{aligned} \pi(\mu, \lambda) &= \phi(\mu; \mu_0, (c_0\lambda)^{-1}) \text{gamma}(\lambda; a/2, b/2) \\ &\propto \lambda^{(a-1)/2} \exp(-\lambda(b + c_0(\mu - \mu_0)^2)/2), \end{aligned} \quad (16)$$

where $\phi(\mu; \mu_0, (c_0\lambda)^{-1})$ denotes the normal density with expectation μ_0 and precision $c_0\lambda$. We will assume that a is an integer. The likelihood is

$$\lambda^{n/2} \exp(-\lambda(S^2(\mathbf{y}) + n(\bar{y} - \mu)^2)/2), \quad (17)$$

where $S^2(\mathbf{y}) = \sum_{i=1}^n (y_i - \bar{y})^2$. Define

$$\hat{\mu} = E(\mu \mid \mathbf{y}) = (c_0\mu_0 + n\bar{y})/(c_0 + n), R_0^2(\mathbf{y}) = (c_0n/(c_0 + n))(\bar{y} - \mu_0)^2. \quad (18)$$

One may then verify that

$$c_0(\mu - \mu_0)^2 + n(\bar{y} - \mu)^2 = (c_0 + n)(\mu - \hat{\mu})^2 + R_0^2(\mathbf{y}).$$

Inserting this identity in the product of (16) and (17) it follows that the posterior density for (μ, λ) is

$$\pi(\mu, \lambda \mid \mathbf{y}) = \phi(\mu; \hat{\mu}, ((c_0 + n)\lambda)^{-1}) \text{ gamma}(\lambda; (a + n)/2, (b + S^2(\mathbf{y}) + R_0^2(\mathbf{y}))/2). \quad (19)$$

[Figure 1 about here.]

Section 3.1 deals with *cphp* and *cepd* analysis as well as conflict analysis concerning the location parameter in this model, and shows that all these approaches are equivalent. Section 3.1 also indicates how nuisance parameters are treated in conflict analysis. Discrepancy and conflict for the precision parameter are treated in Section 3.2.

3.1 Discrepancy and conflict analysis for location

The unknown parameters are now $\boldsymbol{\theta} = (\mu, \lambda)$. Following HDS we still consider the discrepancy measure $d(\mathbf{y}, \boldsymbol{\theta}) = n\lambda(\bar{y} - \mu)^2$. We first note that the randomness of λ makes no difference concerning the χ_1^2 -distribution of $d(\mathbf{Y}^{\text{rep}}, \boldsymbol{\theta})$, making condition a) of Theorem 1 satisfied. We claim that condition b) is satisfied as well, implying that $cphp(\mathbf{y}) = cepd(\mathbf{y})$. To see this, as in (9) let N be the variable obtained by standardising the normal variable $-\mu$ conditional on λ and $\mathbf{y}$, i.e. $N = \sqrt{(c_0 + n)\lambda}[-\mu + ((n\bar{y} + c_0\mu_0)/(c_0 + n))]$. Then N is independent of λ . We note that by (9) the posterior distribution of $d(\mathbf{y}, \boldsymbol{\theta})$ is that of

$$(n/(c_0 + n))[N + (\sqrt{\lambda}c_0(\bar{y} - \mu_0)/\sqrt{c_0 + n})]^2.$$

In the second summand of this expression we have by (19) that the distribution of λ given $\mathbf{y}$ is gamma with parameters $(a + n)/2$ and $(b + S^2(\mathbf{y}) + R_0^2(\mathbf{y}))/2$. Therefore $\lambda(b + S^2(\mathbf{y}) + R_0^2(\mathbf{y})) \stackrel{\text{def}}{=} Z_{a+n}$ is χ_{a+n}^2 -distributed given $\mathbf{y}$. The second summand can therefore be written as

$$(\sqrt{\lambda}c_0(\bar{y} - \mu_0)/\sqrt{c_0 + n}) = (\sqrt{Z_{a+n}}/(b + S^2(\mathbf{y}) + R_0^2(\mathbf{y}))^{1/2})(c_0(\bar{y} - \mu_0)/\sqrt{c_0 + n}).$$

Using that N has a symmetric and unimodal distribution and Z_{a+n} is positive, it follows from Lemma 2 given in the Appendix that $d(\mathbf{y}, \boldsymbol{\theta})$ given $\mathbf{y}$ is stochastically increasing in $|\bar{y} - \mu_0| / (b + S^2(\mathbf{y}) + R_0^2(\mathbf{y}))^{1/2}$, or by (18) equivalently in

$$R_0^2(\mathbf{y}) / (b + S^2(\mathbf{y}) + R_0^2(\mathbf{y})) = [1 + ((b + S^2(\mathbf{y}) / R_0^2(\mathbf{y}))^{-1})]^{-1}.$$

From this we finally conclude that $d(\mathbf{y}, \boldsymbol{\theta})$ is stochastically increasing in

$$T(\mathbf{y}) = R_0^2(\mathbf{y}) / (b + S^2(\mathbf{y})). \quad (20)$$

This proves our claim, and we conclude that Theorem 1 applies. Moreover, the required stochastic ordering is determined by the statistic $T(\mathbf{Y})$, which therefore satisfies (10).

The following proposition gives in addition an analytic expression for $c_{ppp}(\mathbf{y}) = c_{epd}(\mathbf{y})$.

Proposition 2. *For the location discrepancy in the normal gamma model (15) we have*

$$\begin{aligned} c_{ppp}(\mathbf{y}) = c_{epd}(\mathbf{y}) &= P^M(R_0^2(\mathbf{Y}) / (b + S^2(\mathbf{Y})) \geq R_0^2(\mathbf{y}) / (b + S^2(\mathbf{y}))) \\ &= \bar{F}_{1, a+n-1}((a+n-1)R_0^2(\mathbf{y}) / (b + S^2(\mathbf{y}))) \end{aligned}$$

where $F_{1, a+n-1}$ denotes the cdf of the Fisher distribution with 1 and $a+n-1$ degrees of freedom.

Proof. The distribution of $T(\mathbf{Y})$ under the model is the distribution of $\lambda R_0^2(\mathbf{Y}) / (\lambda(b + S^2(\mathbf{Y})))$. Here λb is χ_a^2 -distributed, $\lambda S^2(\mathbf{Y})$ is χ_{n-1}^2 -distributed, independently of λ , and by (18) the numerator $\lambda R_0^2(\mathbf{Y})$ is χ_1^2 -distributed, independently of both terms in the denominator. It follows that $(a+n-1)T(\mathbf{Y})$ is Fisher-distributed with 1 and $a+n-1$ degrees of freedom. Hence, by Theorem 1, using also (20),

$$c_{ppp}(\mathbf{y}) = c_{epd}(\mathbf{y}) = P^M(T(\mathbf{Y}) \geq T(\mathbf{y})) = \bar{F}_{1, a+n-1}((a+n-1)R_0^2(\mathbf{y}) / (b + S^2(\mathbf{y}))) \quad (21)$$

□

We will now establish the equivalence between the c_{ppp}/c_{epd} approach and analysis based on conflict measures $c_{\mu_0}^3$ and $c_{\mu_0}^4$, questioning the validity of the modelling assumption $E(\mu) = \mu_0$, for model (15). Conflict concerning fixed top level nodes is introduced at the end of Section 5 of Gåsemyr & Natvig (2009), and summarised in Gåsemyr & Natvig (2017), Section 4.4. We choose to use this framework here, partly in order to highlight this as a

useful variant of conflict analysis, and partly because it allows derivation of exact analytic results.

Considering the possibility that the modelling assumption that $E(\mu) = \mu_0$ is incorrect, we assume that there exists a true value $\tilde{\mu}$, possibly different from μ_0 , for this expectation. This means that according to (15) $\tilde{\mu}$ has a Dirac prior with support in μ_0 . Hence, we may write (15) as

$$Y_1, \dots, Y_n \text{ IID } \sim N(\mu, \lambda^{-1}), \mu \sim N(\tilde{\mu}, (c_0\lambda)^{-1}), \tilde{\mu} \sim \Delta_{\mu_0}(\tilde{\mu}), \lambda \sim \text{gamma}(\lambda; a/2, b/2).$$

Recall the notion of information contributions received from neighbouring nodes as described prior to (13). The prior information contribution G_p for $\tilde{\mu}$ coincides with the prior distribution and is deterministic, represented by $\tilde{\mu}_p = \mu_0$. We also introduce an integrated likelihood information contribution G_c for $\tilde{\mu}$, based on integrating out the nuisance parameters from the local likelihood information contribution. Details are given below. The distribution G_c is represented by independent random samples $\tilde{\mu}_c^*$. We contrast the prior and likelihood information contributions about $\tilde{\mu}$ through the distribution G of $\delta = \tilde{\mu}_p - \tilde{\mu}_c^* = \mu_0 - \tilde{\mu}_c^*$.

We now describe how to treat the nuisance parameters in order to form the integrated likelihood information contribution G_c . The local likelihood information contribution, being a density proportional to the local likelihood $\phi(\mu; \tilde{\mu}, (c_0\lambda)^{-1})$, is by symmetry $\phi(\tilde{\mu}; \mu, (c_0\lambda)^{-1})$. This density depends on the unknown nuisance parameters (μ, λ) . The integrated likelihood information contribution g_c is obtained by integrating out (μ, λ) from the local likelihood information contribution with respect to the relevant posterior distribution. The principle of using information only once indicates that in this calculation we use the non-informative prior 1 for μ , since μ_0 is used in the contrasting Dirac measure acting as g_p . For λ we want to use all available information, both from the prior and the data, and accordingly use the prior $\pi(\lambda) = \text{gamma}(\lambda; a/2, b/2)$. This gives the posterior

$$\pi(\mu, \lambda \mid \mathbf{y}) = \phi(\mu; \bar{y}, (n\lambda)^{-1}) \text{gamma}(\lambda; (a + n - 1)/2, (b + S^2(\mathbf{y})/2) \quad (22)$$

for (μ, λ) . With this set up we have an analytic solution for $c_{\mu_0}^3 = c_{\mu_0}^4$ given in the following lemma, which is proved in the Appendix.

Lemma 1. Let $F_T(\cdot; a + n - 1)$ denote the cdf of the standard t -distribution with $a + n - 1$ degrees of freedom. Define the linear function ξ by

$$\xi(\mu) = [(a + n - 1)c_0n/((c_0 + n)(b + S^2(\mathbf{y})))]^{1/2}(\mu - \bar{y}). \quad (23)$$

Then

$$\begin{aligned} c_{\mu_0}^3 &= c_{\mu_0}^4 = 1 - 2 \min(F_T(\xi(\mu_0); a + n - 1), \bar{F}_T(\xi(\mu_0); a + n - 1)) \\ &= F_T(|\xi(\mu_0)|; a + n - 1) - F_T(-|\xi(\mu_0)|; a + n - 1). \end{aligned} \quad (24)$$

Now we have the following proposition stating the equivalence between the *cppp/cepd* approach and analysis based on $c_{\mu_0}^3$ and $c_{\mu_0}^4$, which extends the last result from Proposition 1 to the case of an unknown precision parameter.

Proposition 3. For the location discrepancy in the normal gamma model (15) we have

$$cppp(\mathbf{y}) = cepd(\mathbf{y}) = 1 - c_{\mu_0}^3 = 1 - c_{\mu_0}^4,$$

Proof. Indeed, since the square of a t -distributed variable with $a + n - 1$ degrees of freedom is Fisher distributed with 1 and $a + n - 1$ degrees of freedom, it follows from (24) that

$$c_{\mu_0}^3 = c_{\mu_0}^4 = F_{1,a+n-1}(\xi(\mu_0)^2). \quad (25)$$

The result then follows from (18), (23) and (25). $\square$

Remark 2. Our Proposition 2 concerns a special case of the normal gamma p -dimensional regression model treated in Section 5.3 of HDS, where their Proposition 2 gives a formula for the *ppp*-distribution. In the one-dimensional case treated here and in Section 2.3 of HDS, this proposition would imply that

$$cppp(\mathbf{y}) = Pr(F_{1,1}(1 + (c_0/n), (c_0/n)Z_1) \leq ppp(\mathbf{y})),$$

where $F_{1,1}(\cdot, \cdot)$ refers to the non-central Fisher cdf with the eccentricity parameter as its second argument, while Pr relates to a χ_1^2 -distribution for the variable Z_1 . It is difficult to see this as being consistent with our Proposition 2, and in fact there seems to be an error in the proof in HDS which raises serious doubt about the cited result. Indeed, a key point in the proof is the cancellation of the precision parameter λ , which appears both in the numerator

and the denominator of the expression for the eccentricity parameter. However, λ in the denominator is part of the prior predictive generation of $\mathbf{Y}$ and is from the prior distribution, while the λ in the numerator is part of the posterior distribution of the discrepancy measure and is itself from the posterior. Hence, this cancellation is illegal. We note that the main applied example in HDS, an analysis of speedskating data in Section 6, is based on the cited proposition.

3.2 Discrepancy and conflict for the precision parameter in the normal gamma model

We consider the discrepancy measure

$$d(\mathbf{y}, \boldsymbol{\theta}) = \lambda S^2(\mathbf{y}, \mu) \stackrel{\text{def}}{=} \lambda \sum_{i=1}^n (y_i - \mu)^2, \quad (26)$$

aiming to reveal a possibly too narrow prior for the precision parameter. Then $d(\mathbf{Y}^{\text{rep}}, \boldsymbol{\theta})$ is χ_n^2 -distributed given $\boldsymbol{\theta}$, independently of $\boldsymbol{\theta}$. Similar to (6) we therefore have

$$ppp(\mathbf{y}) = E_{\boldsymbol{\theta}|\mathbf{y}}(\bar{\Psi}_n(d(\mathbf{y}, \boldsymbol{\theta}) | \mathbf{y})), \quad (27)$$

to be used as a test statistic in a prior predictive test according to the *cppp* procedure. The *cepd* approach is based on replacing the non-linear $\bar{\Psi}_n$ in (27) by the identity function, and in the following we will show that this puts this approach to an advantage. We can write $d(\mathbf{y}, \boldsymbol{\theta})$ as $\lambda(\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2)$. The posterior expectation of this quantity, which is the alternative test statistic used in (7) to define *cepd*($\mathbf{y}$), may be written

$$E_{\lambda|\mathbf{y}}[\lambda(S^2(\mathbf{y}) + n(\bar{y} - \hat{\mu})^2) | \mathbf{y}] + (n/(c_0 + n))E_{\boldsymbol{\theta}|\mathbf{y}}((c_0 + n)\lambda(\hat{\mu} - \mu)^2 | \mathbf{y}).$$

Here $\hat{\mu} = E(\mu | \mathbf{y})$, and μ has precision $\lambda(c_0 + n)$ given $\mathbf{y}$ and λ . It follows that the last of these two expectations is constant and equal to $n/(c_0 + n)$, independently of $\mathbf{y}$, so for the calculation of *cepd*($\mathbf{y}$) we could take as our test statistic the first one, which by (18) equals

$$E_{\lambda|\mathbf{y}}(\lambda(S^2(\mathbf{y}) + n[c_0(\bar{y} - \mu_0)/(c_0 + n)]^2) | \mathbf{y}) = E_{\lambda|\mathbf{y}}(\lambda(S^2(\mathbf{y}) + (c_0/(c_0 + n))R_0^2(\mathbf{y})) | \mathbf{y}). \quad (28)$$

Now by (19) λ is gamma distributed with shape parameter $(1/2)(a + n)$ and scale parameter $(1/2)\beta(\mathbf{y}) \stackrel{\text{def}}{=} (1/2)(b + S^2(\mathbf{y}) + R_0^2(\mathbf{y}))$. Assuming as before that a is an integer, it follows

that, independently of $\mathbf{y}$, $\lambda\beta(\mathbf{y})$ equals a χ_{a+n}^2 -distributed variable Z_{a+n} , so that (28) can be written as

$$\begin{aligned} E_{Z_{a+n}|\mathbf{y}}(Z_{a+n})(S^2(\mathbf{y}) + (c_0/(c_0 + n))R_0^2(\mathbf{y}))/\beta(\mathbf{y}) \\ = (a + n)(S^2(\mathbf{y}) + (c_0/(c_0 + n))R_0^2(\mathbf{y}))/[b + S^2(\mathbf{y}) + R_0^2(\mathbf{y})]. \end{aligned}$$

Writing the denominator as

$$[(S^2(\mathbf{y}) + (c_0/(c_0 + n))R_0^2(\mathbf{y}))] + [b + (n/(c_0 + n))R_0^2(\mathbf{y})]$$

we see that this fraction is monotonic in

$$T(\mathbf{y}) = (S^2(\mathbf{y}) + (c_0/(c_0 + n))R_0^2(\mathbf{y}))/[b + (n/(c_0 + n))R_0^2(\mathbf{y})]. \quad (29)$$

In conclusion, we get the same prior predictive test by using $T(\mathbf{y})$ as a test statistic as by using $E_{\theta|\mathbf{y}}(D(\mathbf{y}, \theta) | \mathbf{y})$. This is summarised as part a) of Proposition 4 below.

Due to the existence of the analytically calculable statistic $T(\mathbf{y})$ of (29) no posterior sampling of θ is required for the *cepd* approach. For the *cphp* approach such a computation saving simplification can not be achieved. Due to the non-linearity of $\bar{\Psi}_n$ the cross term in

$$((\bar{y} - \hat{\mu}) + (\hat{\mu} - \mu))^2 = (\bar{y} - \hat{\mu})^2 + 2(\bar{y} - \hat{\mu})(\hat{\mu} - \mu) + (\hat{\mu} - \mu)^2$$

does not vanish when expanding the argument for $\bar{\Psi}_n$ in (27). The contribution of this term to the posterior distribution of $d(\mathbf{y}, \theta)$ seems to make it impossible to verify condition b) of Theorem 1, or to identify any analytically computable statistic that could be used to avoid posterior sampling in the calculation of *cphp*($\mathbf{y}$). Hence, the *cphp* approach requires double simulation and is clearly at a disadvantage from a computational point of view. Moreover, the above cross term is proportional to $(\bar{y} - \mu_0)$ times a standard normal variable, and it is difficult to see this as doing anything but adding unnecessary noise to the diagnostic.

A peculiar aspect of (29) is that the term $R_0^2(\mathbf{y})$ appears both in the numerator and denominator, but with different coefficients adding up to 1. One interpretation could be that it contributes less to the discrepancy, the less certain we are about the location μ_0 . However, the degree of uncertainty concerning this location is already taken into account in the definition of $R_0^2(\mathbf{y})$ as $(\bar{y} - \mu_0)^2/((1/c_0) + (1/n))$. It might therefore be desirable to have coefficient 1 in the numerator and 0 in the denominator for $R_0^2(\mathbf{y})$. We will now show that this in fact is achieved by instead using the conflict measure approach.

In this situation a conflict measure corresponding to a one-sided test is required, namely the following one-sided version of (14):

$$c_\lambda^{3-} = \bar{G}(0) = P^G(\delta \geq 0),$$

where $\delta = \lambda_p^* - \lambda_c^*$, the difference between independent samples from prior and likelihood integrated information contributions g_p and g_c specified in the proof below, and G is the cdf of these differences.

Proposition 4. *Consider the normal gamma model (15) with a an integer. Then*

a) with the discrepancy measure (26) we have that $E_{\theta|\mathbf{y}}(d(\mathbf{y}, \theta) \mid \mathbf{y})$ is monotonic in

$$(S^2(\mathbf{y}) + (c_0/(c_0 + n))R_0^2(\mathbf{y}))/ (b + (n/(c_0 + n))R_0^2(\mathbf{y})), \text{ and}$$

$$\begin{aligned} \text{cepd}(\mathbf{y}) &= P^M((S^2(\mathbf{Y}) + (c_0/(c_0 + n))R_0^2(\mathbf{Y}))/ (b + (n/(c_0 + n))R_0^2(\mathbf{Y}))) \\ &\geq (S^2(\mathbf{y}) + (c_0/(c_0 + n))R_0^2(\mathbf{y}))/ (b + (n/(c_0 + n))R_0^2(\mathbf{y})). \end{aligned}$$

b) c_λ^{3-} is monotonic in $(S^2(\mathbf{y}) + R_0^2(\mathbf{y}))/b$, takes the value

$$c_\lambda^{3-} = F_{n,a}(a(S^2(\mathbf{y}) + R_0^2(\mathbf{y}))/nb)$$

and is uniformly distributed under the model.

Proof. Part a) is already proved. Consulting Figure 1 it is seen that under model (15) the prior information contribution $g_p(\lambda)$ is based on the information from the parent nodes a and b of λ , and equals $g_p(\lambda) = \pi(\lambda) = \text{gamma}(\lambda; a/2, b/2)$. All other information sources are used in the formation of the integrated likelihood information contribution $g_c(\lambda)$. Following Gåsemyr (2016), Section 4, g_c is obtained as a posterior density based on the improper prior $\pi(\lambda) = 1/\lambda$. The relevant joint posterior density for (μ, λ) given $\mathbf{y}$ is then given by (19) with $a = b = 0$. Integrating out μ , we obtain $g_c(\lambda) = \text{gamma}(\lambda; n/2, (S^2(\mathbf{y}) + R_0^2(\mathbf{y}))/2)$. Noting that $\delta \geq 0$ if and only if $\lambda_c^*/\lambda_p^* \leq 1$, the first two assertions of part b) follow by noting that the independent samples λ_p^* and λ_c^* give rise to independent variables $b\lambda_p^*$ and $(S^2(\mathbf{y}) + R_0^2(\mathbf{y}))\lambda_c^*$ from the χ_a^2 - and χ_n^2 - distributions respectively. To prove the last assertion, we multiply both numerator and denominator of the argument for $F_{n,a}$ with a λ from the model, and

see that this argument equals (a/n) times a fraction whose numerator and denominator also under P^M are independent χ^2 -distributed variables with n and a degrees of freedom respectively. $\square$

Remark 3. *The diagnostic $c_{ppp}(\mathbf{y})$ is always uniformly distributed under the model by construction. In all cases treated in the present paper except the one under consideration in this subsection, equality of the relevant conflict measure with $1 - c_{ppp}(\mathbf{y})$ therefore implies that also the conflict measure is uniformly distributed. Proposition 4 shows that in the normal gamma model the relevant conflict measure $c_{\chi^3}^-$ is uniformly distributed although it deviates from $1 - c_{ppp}(\mathbf{y})$ in this case. Uniformity of the c_{χ^3} - and c_{χ^4} - conflict measures pre-experimentally, i.e. under the assumed model, is proved in various situations in Gåsemyr & Natvig (2009) and Gåsemyr (2016), but only in models without nuisance parameters.*

4 Data - data conflict as opposed to data - prior conflict

The *c_{ppp}* methodology is based on prior predictive testing and accordingly seems particularly well suited for analysing discrepancy between data and the prior distribution, the latter representing either an assumed mechanism for generating lower level parameters, or prior belief about the true value of certain fixed but unknown parameters. This corresponds to what is called data - prior conflict within the context of conflict analysis. The prior distribution plays a subordinate role when checking an exchangeability assumption for group level parameters in a hierarchical model, or more generally, checking for divergent behaviour of different subsets of the data. Within the context of conflict analysis, this kind of discrepancy can be addressed using non-informative or even improper priors, and the term data - data conflict is therefore used. In contrast, for the *c_{ppp}* analysis to make sense, proper and at least somewhat informative priors are needed also in this setting. As an example we will consider checking for outlying subgroups in a normal random effects model, which has served as a test case for tools for model checking in O'Hagan (2003), Marshall & Spiegelhalter (2007), Bayarri & Castellanos (2007), Dahl *et al.* (2007), Johnson (2007), Gåsemyr & Natvig (2009) and Scheel *et al.* (2011). For simplicity, here we will assume fixed and known variance

parameters. For $i = 1, \dots, k, j = 1, \dots, n_i$, assume

$$Y_{i,j} \sim N(\mu_i, \lambda^{-1}), \mu_i \sim N(\mu, (c_i \lambda)^{-1}), \mu \sim N(\mu_{\text{pri}}, (c \lambda)^{-1}). \quad (30)$$

In order to check whether group k may be an outlier we use the discrepancy measure

$$d(\mathbf{y}, \boldsymbol{\theta}) = d(\mathbf{y}_k, \mu_k) = n_k \lambda (\bar{y}_k - \mu_k)^2. \quad (31)$$

Here, $\bar{y}_k$ is the average of the components of $\mathbf{y}_k = (y_{k,1}, \dots, y_{k,n_k})$.

For the conflict analysis, we contrast the likelihood or child node information contribution from data $\mathbf{y}_k$ about μ_k , normalised to the density

$$g_c(\mu_k) = \sqrt{n_k \lambda} \phi(\sqrt{n_k \lambda}(\mu_k - \bar{y}_k)),$$

with the prior or parent node information contribution from μ . Here $\mathbf{y}_k$, consisting of the child node data of μ_k , represents the first part of a splitting of the data into two disjoint subvectors $(\mathbf{y}_c, \mathbf{y}_p)$. The vector $\mathbf{y}_p$, consisting of all data non-descendant to μ_k , is used in the formation of the integrated prior or parent node information contribution G_p about μ_k . The purpose of the data splitting is to avoid using data twice, in accordance with the general philosophy of the conflict analysis. In this context μ , being unknown, is a nuisance parameter and is integrated out from the information contribution about μ_k from μ with respect to the posterior distribution of μ given $\mathbf{y}_p$. The resulting density for μ_k is $g_p(\mu_k) = \pi(\mu_k | \mathbf{y}_p)$. Based on the distribution G of the difference between independent samples from these integrated information contributions, conflict measures for μ_k are defined as in (13) and (14). This is described in more detail and greater generality in Gåsemyr & Natvig (2009) and Gåsemyr (2016). With this set up we have the following proposition, proved in the appendix.

Proposition 5. *In the normal random effects model (30), with the discrepancy measure (31), we have*

$$c_{ppp}(\mathbf{y}) = c_{epd}(\mathbf{y}) = 1 - c_{\mu_k}^4 = 1 - c_{\mu_k}^3.$$

In a hierarchical model like (30) conflict measures can be defined for any node in the graphical representation by dividing the local parent and child information sources into two subgroups, see Section 5 of Gåsemyr & Natvig (2009) for a general formulation. For instance,

the possibility of group k being an outlier could alternatively be examined by contrasting the information from μ_k about μ with the information about μ from μ_{pri} and $\mu_1, \dots, \mu_{k-1}$. The natural data splitting for this conflict would be the same as above. The resulting conflict measure would be the natural choice in order to match the discrepancy measure $c_k \lambda(\mu_k - \mu)^2$. The *cphp* methodology has been extended to this kind of discrepancy measure, depending on parameters and latent variables but not directly on data, in Steinbakk & Storvik (2009). A corresponding extension can be done for the *cepd* approach as well. It can be shown that the different approaches give equivalent results also in this case.

It is interesting to consider the discussion in Section 9.3 in HDS in the light of Proposition 5. HDS makes a sharp distinction between priors that can reasonably be used for sampling, and priors under which sampling and hence the *cphp* approach makes no sense. Improper priors obviously fall under the latter category, but so do also vague proper priors. Paradoxically, Proposition 5 provides an example where arbitrarily vague proper priors give perfectly reasonable results. Indeed, it follows from the proof of Proposition 5 (see Equations (34) and (37)) that reducing the prior precision by letting $c \rightarrow 0$ just removes the influence of the prior expectation μ_{pri} on *cphp*($\mathbf{y}$). This is perfectly sensible if the prior knowledge is very imprecise and we are checking for a conflict of the data - data type. This limiting behaviour is completely in line with the idea behind the conflict measure approach. The parent node information contribution about μ_k should contain no contribution from a prior expectation if nothing is known about the location of μ a priori.

In the above example we were able to show theoretically that the influence of the prior expectation vanishes with decreased precision in such a way that the *cphp*-value remains sensible. In more complex models this would be very difficult or impossible to prove, and might also not be true in some cases. The advice given in HDS is to base the *cphp* analysis on a proper prior that seems reasonable, although it is much more precise than one's actual prior beliefs. This informative prior is used for model checking, while afterwards the inference is based on a different, vague prior. If this procedure was used in our present example, the result of the model check would depend on a somewhat arbitrarily chosen value of μ_{pri} . In our view this is an undesirable consequence of the approach, and methods avoiding this are

preferable.

Another matter of concern in situations with vague prior information and informative checking priors is the relevance of the uniformity of $c_{ppp}(\mathbf{y})$ under the model. The model that ensures uniformity has the informative checking prior, not the vague prior reflecting one's true beliefs. In our view, with very imprecise prior information one should rather aim for a uniform distribution of a diagnostic measure under the true value $\mu = \mu_{\text{true}}$ if possible, regardless of the value of $\mu = \mu_{\text{true}}$. In Gåsemyr & Natvig (2009) it is shown that this is in fact achieved in the above model, and in more general hierarchical normal models with known variances, for the c^3 - and c^4 -measures, if the improper prior 1 is used for the basic location parameter. If the prior is informative, we may think of it as resulting from an experiment in the past, prior to which there was a state of complete ignorance about μ , consisting of $c = n_0$ independent observations $y_{0,1}, \dots, y_{0,n_0}$ with precision λ , for which the average is $\bar{y}_0 = \mu_{\text{pri}}$. Viewed from a time point before this hypothetical experiment, the uniformity property holds. Another way of looking at this is that since the conflict measures obtain a sensible calibration or scaling when all sources of information are data based, a sensible scaling is obtained in general because we take prior beliefs and experimental data into account in a unified way.

5 Application to failure data for power plant pumps

In this section we consider a model where no analytical results are available for $c_{ppp}(\mathbf{y})$, $c_{epd}(\mathbf{y})$ and c_{θ}^4 , and where the parameter of interest is a vector. We find numerical values of $c_{ppp}(\mathbf{y})$, $c_{epd}(\mathbf{y})$ and c_{θ}^4 for a particular dataset, based on simulations.

We consider a model and data which are presented as an example called “Pump” in Volume I of the examples accompanying WinBUGS (Lunn *et al.*, 2000), and originates from Gaver & O’Muircheartaigh (1987). The data consist of the numbers of failures of $n = 10$ power plant pumps, as well as exposure in terms of the operation times. Pump i has Y_i failures, operation time t_i (in 1000s of hours) and failure rate θ_i . The data can be seen in

Table 1. The model implemented in the WinBUGS examples is

$$\begin{aligned} Y_i \mid \theta_i &\sim p(Y_i \mid \theta_i) = \text{poisson}(\theta_i t_i), \quad \theta_i \mid \alpha, \beta \sim \pi(\theta_i \mid \alpha, \beta) = \text{gamma}(\alpha, \beta), \quad i = 1, \dots, n, \\ \alpha &\sim \pi(\alpha) = \exp(1), \quad \beta \sim \pi(\beta) = \text{gamma}(a, b), \end{aligned}$$

where the gamma distribution for β has shape parameter $a = 0.1$ and rate parameter $b = 1$.

[Table 1 about here.]

In this paper we consider the same model, except for the parameters of the prior distribution for β . Our focus lies on the whole vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$. Hence all the data are children of $\boldsymbol{\theta}$, and c_θ^4 will be constructed to investigate potential data-prior conflict for this vector. Hence for all three measures $c_{ppp}(\mathbf{y})$, $c_{epd}(\mathbf{y})$ and c_θ^4 , prior distributions that it makes sense to sample from are needed. We use two different sets of prior parameter specifications for β , one that is not too restrictive with $a = 1$ and $b = 0.05$, and one on purpose intended to cause conflict with the data, with $a = 20$ and $b = 1$.

For $c_{ppp}(\mathbf{y})$ and $c_{epd}(\mathbf{y})$, a natural discrepancy measure is

$$d(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^n \frac{(y_i - \theta_i t_i)^2}{\theta_i t_i}.$$

For the calculation of $c_{ppp}(\mathbf{y})$ and $c_{epd}(\mathbf{y})$, we used Gibbs sampling, combined with a Metropolis step for sampling α . A double posterior MCMC simulation scheme is needed for the calibration. We used $M = 10.000$ MCMC samples (after burn-in) for posterior inference, first once for computing $ppp(\mathbf{y})$ and $E(d(\mathbf{y}, \boldsymbol{\theta}) \mid \mathbf{y})$, and then repeated $R = 10.000$ times. For $r = 1, \dots, R$, data $\mathbf{y}^{(r)}$ are simulated from the model in the following way: $\alpha^{(r)}$ and $\beta^{(r)}$ are drawn from $\pi(\alpha)$ and $\pi(\beta)$, then for $i = 1, \dots, n$, $\theta_i^{(r)}$ is drawn from $\pi(\theta_i \mid \alpha^{(r)}, \beta^{(r)})$ and hence $y_i^{(r)}$ is drawn from $p(Y_i \mid \theta_i^{(r)})$. Then, for each r , an MCMC is run, $ppp(\mathbf{y}^{(r)})$ and $E(d(\mathbf{y}^{(r)}, \boldsymbol{\theta}) \mid \mathbf{y}^{(r)})$ are computed, and hence the calibrated $c_{ppp}(\mathbf{y})$ and $c_{epd}(\mathbf{y})$ can be computed. This took 7 hours on a desktop computer.

For the calculation of c_θ^4 , no MCMC is needed. Instead, it is based on samples $\theta_i^{c(s)} \sim g_c(\theta_i)$ and $\theta_i^{p(s)} \sim g_p(\theta_i)$, for $i = 1, \dots, n$ and $s = 1, \dots, S = 250.000$. The likelihood information contribution $g_c(\theta_i)$ is obtained as the posterior distribution of θ_i based on the

improper prior $\pi(\theta_i) = 1/\theta_i$, which is $\text{gamma}(y_i, t_i)$. The prior information contribution is $g_p(\boldsymbol{\theta}) = \int \int \pi(\boldsymbol{\theta} \mid \alpha, \beta) \pi(\alpha) \pi(\beta) d\alpha d\beta$. We obtain samples from $g_p(\theta)$ by sampling $\alpha^{(s)}$ from $\pi(\alpha)$ and $\beta^{(s)}$ from $\pi(\beta)$, then $\theta_i^{p(s)}$ from $\text{gamma}(\alpha^{(s)}, \beta^{(s)})$. From this, $\delta_i^{(s)} = \theta_i^{p(s)} - \theta_i^{c(s)}$ is calculated for each $s = 1, \dots, S$, resulting in S vectors $\boldsymbol{\delta}^{(1)}, \dots, \boldsymbol{\delta}^{(S)}$. The density g of $\boldsymbol{\delta}$ is then estimated by kernel density estimation. Denoting the estimated density by $\hat{g}$, c_θ^4 can be calculated in the following way

$$c_\theta^4 = \frac{1}{S} \sum_{s=1}^S I(\hat{g}(\boldsymbol{\delta}^{(s)}) > \hat{g}(\mathbf{0})).$$

The density estimation was done using the function `kepdf` in the R package `pdfCluster` (Azzalini & Menardi, 2014). Computing c_θ^4 took 40 minutes on a desktop computer.

For $a = 1$ and $b = 0.05$ we get $c_{ppp}(\mathbf{y}) = 0.167$, $c_{epd}(\mathbf{y}) = 0.153$ and $1 - c_\theta^4 = 0.155$. For $a = 20$ and $b = 1$ we get $c_{ppp}(\mathbf{y}) = 0.0014$, $c_{epd}(\mathbf{y}) = 0.0376$ and $1 - c_\theta^4 = 0$. The numerical results are not identical for the three measures, however the conclusions drawn regarding how well the model fits the data are the same for the two sets of prior specifications. We have previously discussed the computational burden of $c_{ppp}(\mathbf{y})$ and $c_{epd}(\mathbf{y})$, caused by the need for a double simulation scheme for the calibration. In this example, $c_{ppp}(\mathbf{y})$ and $c_{epd}(\mathbf{y})$ are indeed much more computationally expensive than c_θ^4 . Of course, the computing times of $c_{ppp}(\mathbf{y})$, $c_{epd}(\mathbf{y})$ and c_θ^4 will all depend on various factors such as the model, the size of the data, the MCMC algorithm and, especially for c_θ^4 , the dimension of the parameter of interest.

6 Concluding remarks

When it is of interest to assess model adequacy in a Bayesian model by means of a certain discrepancy measure, it is tempting from a computational point of view to do this through a construction based on the posterior sample that is needed for the inference anyway. This is especially so if there are several discrepancy measures of interest for the given model. The posterior predictive p -value is such a construction. However, since there are serious problems with interpreting the ppp -value, Hjort *et al.* (2006) has suggested the more computer intensive

calibrated *cppp*-value, using a prior predictive approach. An important message of the present paper is however that the *ppp*-value itself may not be the best test statistic to use for a prior predictive approach based on a given discrepancy measure. One alternative could for instance be the posterior expectation of the discrepancy measure. The comparisons performed in this paper indicate that the resulting *cepd*-value may be a good alternative.

Since calibration of a measure based on posterior parameter samples requires double simulation in the general case, it is of interest to search for less computer intensive alternatives. In the present paper we suggest that conflict measures, which in the general case only require at most two MCMC runs to obtain samples from the involved prior and likelihood integrated information contributions, may represent such an alternative in many situations. This suggestion is supported by our finding that in some of the main examples considered in Hjort *et al.* (2006) specific conflict measures turn out to be equivalent to the *cppp*-value, while the comparison seems to be in favour of the conflict measure approach in the case analysed in Section 3.2 of the present paper. In the applied example considered in Section 5 there are small numerical differences between the measures, but they lead to the same conclusion.

The conflict measure approach can be applied to any parameter in a Bayesian model, including fixed (non-random) top level parameters, as well as to subsets or functions of the data. Several potential model deviations can be assessed for each node by different partitions of the local information sources and correspondingly adjusted data splittings. However, despite this flexibility a relevant conflict measure matching a given discrepancy function does not necessarily suggest itself as naturally as has been the case in the examples studied in this paper. With some creativity a solution to this problem may nevertheless be found in many cases. For example, if a certain discrepancy is not picked up by any model parameter, one idea could be to expand the model by including such a parameter. For instance, if skewness is a concern, as in the speed of light measurements analysed in Section 9.1 of Hjort *et al.* (2006), one idea could be to replace the normal distribution with the skew normal one, see Azzalini (1985). Then we could consider conflict between data and a fairly sharp or exact Dirac prior at 0 for the skewness parameter.

Acknowledgement

We are grateful to professor Bent Natvig, two reviewers and especially to the Associate Editor whose numerous suggestions and comments helped us to improve this paper.

References

Azzalini, A. (1985). A class of distributions which include the normal ones. *Scand. J. Stat.* **12**, 171–178.

Azzalini, A. & Menardi, G. (2014). Clustering via nonparametric density estimation: The R package pdfCluster. *Journal of Statistical Software* **57**, 1–26. <http://www.jstatsoft.org/v57/i11/>.

Bayarri, M. J. & Berger, J. O. (2000). P values in composite null models (with discussion). *J. Amer. Statist. Assoc.* **95**, 1127–1142.

Bayarri, M. J. & Castellanos, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statist. Sci.* **22**, 322–343.

Box, G. E. P. (1980). Sampling and Bayes’ inference in scientific modelling and robustness (with discussion and rejoinder). *J. Roy. Statist. Soc. ser.A* **143**, 383–430.

Dahl, F. A. (2006). On the conservativeness of posterior predictive p-values. *Statist. Probab. Lett.* **76**, 1170–1174.

Dahl, F. A., Gåsemyr, J. & Natvig, B. (2007). A robust conflict measure of inconsistencies in Bayesian hierarchical models. *Scand. J. Stat.* **34**, 816–828.

Gaver, D. P. & O’Muircheartaigh, I. G. (1987). Robust empirical Bayes analyses of event rates. *Technometrics* **29**, 1–15.

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electron. J. Stat.* **7**, 2595–2602.

- Gelman, A., Meng, X.-L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion and rejoinder). *Statist. Sinica* **6**, 733–807.
- Gåsemyr, J. (2016) Uniformity of node level conflict measures in Bayesian hierarchical models based on directed acyclic graphs. *Scand. J. Stat.* **43**, 20–34.
- Gåsemyr, J. & Natvig, B. (2009). Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models. *Scand. J. Stat.* **36**, 822–838.
- Gåsemyr, J. & Natvig, B. (2017). Node-Level Conflict Measures in Bayesian Hierarchical Models Based on Directed Acyclic Graphs. In *Bayesian Inference* (ed J. P. Tejedor), doi: 0.5772/intechopen.70058, url: <https://doi.org/10.5772/intechopen.70058> InTech, Rijeka, 2017.
- Hjort, N. L., Dahl, F. A. & Steinbakk, G. H. (2006). Post-processing posterior predictive p -values. *J. Amer. Statist. Assoc.*, **101**, 1157–1174.
- Johnson, V. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Anal.* **2**, 719–734.
- Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS- A Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.*, **10**, 325–337.
- Marshall, E. C. & Spiegelhalter, D. J. (2007). Identifying outliers in Bayesian hierarchical models. A simulation based approach. *Bayesian Anal.* **2**, 409–444.
- O’Hagan, A. (2003). HSSS model criticism (with discussion). In *Highly Structured Stochastic Systems* (eds P. J. Green, N. L. Hjort, and S. Richardson), 423–444. Oxford University Press, Oxford, 2003.
- Presanis, A. M., Ohlssen, D., Spiegelhalter, D. & De Angelis, D. (2013). Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statist. Sci.* **28**, 376–397.

Scheel, I., Green, P. & Rougier, J. C. (2011). A graphical diagnostic to identifying influential model choices in Bayesian hierarchical models. *Scand. J. Stat.*, **38**, 529–550.

Steinbakk, G. & Storvik, G. (2009). Posterior predictive p-values in Bayesian hierarchical models. *Scand. J. Stat.*, **36**, 320–336.

Corresponding author address:

Jørund Gåsemyr

Department of Mathematics

University of Oslo

POBox 1053 Blindern

0316 Oslo

Norway

gaasemyr@math.uio.no

Appendix

Lemma 2. *Let S, U be independent, absolutely continuously distributed random variables such that S has a symmetric and unimodal density and U is non-negative. Then the variable $|S + vU|$, depending on a real number v , is stochastically increasing in $|v|$.*

Proof. Let f be the density of S and F the corresponding cdf. Due to the symmetry of f for any real number v and any $w > 0$ we have that $P(|S + vU| \leq w) = P(|S - vU| \leq w)$. Hence, it suffices to consider positive real numbers v . For any $v, w > 0$ we have

$$P(|S + vU| \leq w) = P(-w - vU \leq S \leq w - vU) = F(w - vU) - F(-w - vU).$$

Differentiating with respect to v , we obtain by the symmetry of f

$$(d/dv)P(|S + vU| \leq w) = -f(w - vU) + f(-w - vU) = -f(w - vU) + f(w + vU).$$

Due to the unimodality and symmetry of f this is negative for any given U , hence also when integrating with respect to the distribution of U . $\square$

Proof of Proposition 1. In the normal model (4) we have that $\theta_p^* \sim N(\theta_0, (c_0\lambda)^{-1})$, while $\theta_c^* \sim N(\bar{y}, (n\lambda)^{-1})$. Hence, $\delta \sim N(\theta_0 - \bar{y}, (c_0 + n)(c_0n\lambda)^{-1})$. Hence, the density of the difference $\delta = \theta_p^* - \theta_c^*$ is

$$g(\delta) = ((c_0 + n)^{-1}c_0n\lambda)^{1/2} \phi(((c_0 + n)^{-1}c_0n\lambda)^{1/2}(\delta - (\theta_0 - \bar{y}))),$$

where ϕ denotes the standard normal density. By the symmetry and unimodality of this density it follows from (13) that

$$c_\theta^4 = P^G(g(\delta) \geq g(0)) = P^G((c_0 + n)^{-1}c_0n\lambda(\delta - (\theta_0 - \bar{y}))^2 \leq (c_0 + n)^{-1}c_0n\lambda(\theta_0 - \bar{y})^2).$$

The quantity on the left hand side of the inequality is χ_1^2 -distributed under P^G , so we get

$$c_\theta^4 = \Psi_1((c_0 + n)^{-1}c_0n\lambda(\theta_0 - \bar{y})^2).$$

Using also (12) and (11), the proposition follows. $\square$

Proof of Lemma 1. Define

$$R^2(\tilde{\mu}) = (c_0n/(c_0 + n))(\tilde{\mu} - \bar{y})^2. \quad (32)$$

Integrating out the nuisance parameters μ and λ according to (22) we then obtain

$$\begin{aligned} g_c(\tilde{\mu}) &= \int \int \phi(\tilde{\mu}; \mu, 1/(c_0\lambda)) \phi(\mu; \bar{y}, 1/(n\lambda)) \text{gamma}(\lambda; (a + n - 1)/2, (b + S^2(\mathbf{y})/2) d\mu d\lambda \\ &= \int \phi(\tilde{\mu}; \bar{y}, (c_0 + n)/(c_0n\lambda)) \text{gamma}(\lambda; (a + n - 1)/2, (b + S^2(\mathbf{y})/2) d\lambda \\ &= (c_0n/(2\pi(c_0 + n)))^{1/2} (\Gamma((a + n)/2)/\Gamma((a + n - 1)/2)) \\ &\quad \cdot ((b + S^2(\mathbf{y}))/2)^{(a+n-1)/2} / ((b + S^2(\mathbf{y}) + R^2(\tilde{\mu}))/2)^{(a+n)/2} \\ &\quad \cdot \int \text{gamma}(\lambda; (a + n)/2, (b + S^2(\mathbf{y}) + R^2(\tilde{\mu}))/2) d\lambda \\ &= (c_0n/(\pi(c_0 + n)(b + S^2(\mathbf{y}))))^{1/2} (\Gamma((a + n)/2)/\Gamma((a + n - 1)/2)) \\ &\quad \cdot (1 + R^2(\tilde{\mu})/(b + S^2(\mathbf{y})))^{-(a+n)/2} \end{aligned} \quad (33)$$

By (23), (32) and the formula for the standard t-density f_T , (33) equals

$$\xi'(\tilde{\mu}) f_T(\xi(\tilde{\mu}); a + n - 1).$$

It then follows that the variable $T = \xi(\tilde{\mu})$ is standard t-distributed with $a + n - 1$ degrees of freedom under G_c .

To generalise (13) and (14), introduce $\delta = \mu_0 - \tilde{\mu}^*$, where μ_0 represents the fixed sample value from g_p , while $\tilde{\mu}^*$ is sampled from g_c . Denote by G the cdf of δ . Then $\bar{G}(0)$ is the probability that a sampled value of $\tilde{\mu}^*$ from g_c is smaller than μ_0 , i.e. the probability $P^{G_c}(\tilde{\mu} < \mu_0)$. Since $\xi(\tilde{\mu})$ is an increasing function of $\tilde{\mu}$ this probability equals $P^{G_c}(\xi(\tilde{\mu}) < \xi(\mu_0)) = F_T(\xi(\mu_0); a + n - 1)$. The opposite inequality has probability $P^{G_c}(\xi(\tilde{\mu}) > \xi(\mu_0)) = \bar{F}_T(\xi(\mu_0); a + n - 1)$ and equals $G(0)$. Comparing with (14), this proves the lemma for $c_{\mu_0}^3$. Comparing with (13), it follows that the lemma holds also for $c_{\mu_0}^4$, because symmetry and unimodality of g_c implies that the density g of G has the same property. $\square$

Proof of Proposition 5. By a standard calculation the density $g_p(\mu_k) = \pi(\mu_k \mid \mathbf{y}_p)$ is seen to be normal. The expectation is

$$E(\mu_k \mid \mathbf{y}_p) = E(\mu \mid \mathbf{y}_p) = (c_p y_p + c \mu_{\text{pri}}) / (c + c_p) \stackrel{\text{def}}{=} \mu_0(\mathbf{y}_p), \quad (34)$$

where y_p is a linear function of $\mathbf{y}_p$, and λc_p is the precision of the corresponding random variable Y_p given μ . The variance is $(1/\lambda)[(1/c_k) + (1/(c + c_p))]$. Recalling that g_c is normal with expectation $\bar{y}_k$ and variance $1/(\lambda n_k)$, the density for the difference δ between independent samples from G_p and G_c is seen to be normal with expectation and variance respectively equal to

$$E^G(\delta) = \mu_0(\mathbf{y}_p) - \bar{y}_k, \text{var}^G(\delta) = (1/\lambda)[(1/n_k) + (1/c_k) + (1/(c + c_p))]. \quad (35)$$

Arguing as in the proof of Proposition 1 we conclude that

$$c_{\mu_k}^4 = \Psi_1(\text{var}^G(\delta)^{-1}(\mu_0(\mathbf{y}_p) - \bar{y}_k)^2), \quad (36)$$

where $\text{var}^G(\delta)$ is given by (35).

By symmetry and unimodality this also equals $c_{\mu_k}^3$.

In order to calculate the *c_{ppp}*-value in this case, we have to establish the posterior distribution of $d(\mathbf{y}, \boldsymbol{\theta})$ through calculation of the posterior distribution of μ_k given $\mathbf{y}$. But this latter distribution can be obtained by updating the posterior distribution G_p of μ_k given $\mathbf{y}_p$ with the data $\mathbf{y}_c = \mathbf{y}_k$. Denoting the precision of μ_k according to G_p by λc_0 , this is completely parallel to the updating in Section 2 of the prior with data $\mathbf{y}$. Hence, we obtain parallel to (9) that $d(\mathbf{y}, \boldsymbol{\theta}) = n_k \lambda (\bar{y}_k - \mu_k)^2$ is distributed as

$$(n_k/(c_0 + n_k))[N + (\sqrt{\lambda}c_0(\bar{y}_k - \mu_0(\mathbf{y}_p))/\sqrt{c_0 + n_k})]^2,$$

where again N is standard normal. By Lemma 2 the corresponding density satisfies part b) of Theorem 1, proving the first equality of Proposition 5. The stochastic ordering is determined by the statistic $(\bar{Y}_k - \mu_0(\mathbf{Y}_p))^2$. Completely parallel to (11) we therefore obtain

$$cPPP(\mathbf{y}) = P^M((\bar{Y}_k - \mu_0(\mathbf{Y}_p))^2 \geq (\bar{y}_k - \mu_0(\mathbf{y}_p))^2). \quad (37)$$

Under P^M the variable $T(\mathbf{Y}) = \bar{Y}_k - \mu_0(\mathbf{Y}_p)$ is normal with expectation 0. Hence, it follows from (37) that

$$cPPP(\mathbf{y}) = \bar{\Psi}_1(\text{var}(T(\mathbf{Y}))^{-1}(\bar{y}_k - \mu_0(\mathbf{y}_p))^2). \quad (38)$$

The variance of $T(\mathbf{Y})$ is

$$E(\text{var}_{\mathbf{Y}|\mu}((\bar{Y}_k - \mu_0(\mathbf{Y}_p)) \mid \mu)) + \text{var}(E_{\mathbf{Y}|\mu}((\bar{Y}_k - \mu_0(\mathbf{Y}_p)) \mid \mu)).$$

Using (34) this leads to

$$\begin{aligned} \text{var}(T(\mathbf{Y})) &= E[\text{var}_{\bar{Y}_k|\mu}(\bar{Y}_k \mid \mu) + \text{var}_{Y_p|\mu}((c_p Y_p)/(c + c_p) \mid \mu)] + \text{var}[\mu - (c_p \mu)/(c + c_p)] \\ &= (1/\lambda)\{[(1/n_k) + (1/c_k) + (c_p/(c + c_p)^2)] + [c/(c + c_p)^2]\} \\ &= (1/\lambda)[(1/n_k) + (1/c_k) + (1/(c + c_p))] = \text{var}^G(\delta), \end{aligned} \quad (39)$$

the last equality following from (35). Hence, the proposition follows from (38), (39) and (36). $\square$

Figures

Figure 1: Graphical representation of the model discussed in Section 3. Circles represent random variables, squares represent constants.

Tables

Table 1: The Pump data

Pump	1	2	3	4	5	6	7	8	9	10
t_i	94.3	15.7	62.9	126	5.24	31.4	1.05	1.05	2.1	10.5
y_i	5	1	5	14	3	19	1	1	4	22