Dept. of Math. University of Oslo
Statistical Research Report No 1
ISSN 0806–3842 December 2014

Uniformity of node level conflict measures in Bayesian hierarchical models based on directed acyclic graphs

Jørund Gåsemyr

December 18, 2014

Abstract

Hierarchical models defined by means of directed, acyclic graphs are a powerful and widely used tool for Bayesian analysis of problems of varying degrees of complexity. A simulation based method for model criticism in such models has been suggested by O'Hagan in the form of a conflict measure based on contrasting separate local information sources about each node in the graph. This measure is however not well calibrated. In order to rectify this, alternative mutually similar tail probability based measures have been proposed independently, and have been proved to be uniformly distributed under the assumed model in quite general normal models with known covariance matrices. In the present paper, exploiting the property of pivotality, we extend this result to a variety of models. An advantage of this is that computationally costly pre-calibration schemes needed for some other suggested methods can be avoided. Another advantage is that non-informative prior distributions can be used when performing model criticism.

Key words: Cross validation, data splitting, information contribution, MCMC, model criticism, pivotal distribution, pivotal quantity, pre-experimental distribution, p-value

1 Introduction

Bayesian hierarchical models are a very flexible and convenient tool for analysing complex data, which has become a standard methodology over the last decades due to the invention of MCMC techniques that can be handled by increasingly powerful computers. The methodology allows the modeller to represent an understanding of the underlying structure of the problem by means of a directed acyclic graph, with nodes in the graph corresponding to data or parameters, and directed edges between parameters representing conditional distributions. Analysis of such models gives intuitively appealing inference based on posterior probability distributions for the parameters. However, there are numerous possibilities for making wrong judgements in the process of specifying such a model, and methods for checking the adequacy of the model are needed. Sometimes a model can be checked indirectly by comparison with other candidate models through a model comparison technique, such as predictive methods, maximum posterior probability, Bayes factors or an information criterion. But even the winner in such a comparison may not be an adequate model, and methods for critizising models in the absence of any particular alternatives are also needed. By means of such methods, an initial candidate model can be assessed, and if necessary modified and elaborated on, leading to a new candidate model that again is checked for adequacy, and so on. This kind of pattern for the model building process is suggested in Box (1980), section 1.

There are many methods for checking the overall fit of the model or an aspect of the model of special interest. Most of these methods are based on locating a test statistic or a discrepancy measure in some kind of a reference distribution, thereby resulting in a p-value. Box (1980) uses the prior predictive distribution of some checking function or test statistic as a reference for the observed value of this checking function, see Box (1980), section 1.3. This requires an informative and realistic prior distribution, which is not always available or even desirable. Indeed, as pointed out in Bayarri and Castellanos (2007), in an early phase of the model building process it is often convenient to use noninformative or even improper priors, avoiding costly and time consuming elicitation of prior information. Moreover, even when a model has passed an initial test for adequacy, relevant prior information may not be available, and noninformative priors are used also for the inference.

Non-informative prior distributions represent no problem for the poste-

rior predictive p-value (ppp) of Gelman, Meng and Stern (1996), which uses the posterior distribution as reference. But this method can be very conservative due to double use of data, see Bayarri and Berger (2000), Bayarri and Castellanos (2007), Hjort, Dahl and Steinbakk (2005) (hereafter referred to as HDS) and Dahl (2006). HDS suggests a prior predictive calibration scheme (cppp) to remedy this, using the ppp-value as a test statistic in its own right. This method is however very computer intensive, and again realistic, informative priors are needed. The partial posterior predictive p-value of Bayarri and Berger (2000) avoids both these problems, but may be difficult to compute and interpret in hierarchical models. Some more informal graphical and numerical methods based on test statistics that are pivotal quantities are suggested in Johnson (2007). The various plots and numerical measures may help in suggesting parts of the model that may need further investigation, but the decisive characterization of a part of the model as being in discordance with the data seems to be based on a supplementary cppp-analysis as suggested in HDS. In Dey et al. (1998) a type of discrepancy measure that can be applied for each node in a graph is constructed. However, their method is also highly computer intensive. Moreover, the procedure is in principle of the prior predictive type, and requires informative priors in order to make sense.

In the present paper we focus on methods for checking the modelling assumptions at each node of the graphical network. Such methods may identify parts or building blocks of the model that are in discordance with reality, and can give useful information about where in the model adjustments or further elaboration may be needed. We adopt the basic perspective of O'Hagan (2003) (OH), which is to view any node in the graph as receiving information from two disjoint subsets of the neighbouring nodes, either in the form of a conditional probability density or a likelihood, or a combination of these two kinds of information sources. Our aim is to check for inconsistency between such subsets. OH suggests a measure of conflict based on normalizing these information sources to have equal height 1, and measuring the height of the graphs at the point where they intersect. Bayarri and Castellanos (2007) shows that this measure tends to be quite conservative. Moreover, considering a normal model, Dahl, Gåsemyr and Natvig (2007) (DGN) shows that for several reasons the measure of OH is poorly calibrated, leading to false warning probabilities that vary substantially between models. By addressing the different sources of inaccuracy, and in particular by instead normalizing

the information sources to probability densities, DGN modified the measure of OH to an approximately χ^2-distributed quantity under the assumed model. In Gåsemyr and Natvig (2009) (GN) these densities were instead used to define tail probability based conflict measures that were shown to be uniformly distributed in quite general linear normal models with fixed covariance matrices. Similar conflict measures were also defined in Marshall and Spiegelhalter (2007) (MS) in the less general setting of checking for outliers among the second level parameters in a random effects model. The conflict measures of DGN, GN and MS are excellently reviewed in Presanis et al. (2013), which also applies these conflict measures in complex cases of medical evidence synthesis. In Dias et al. (2010) this methodology is used to check for inconsistency in multiple treatment comparison of randomized clinical trials.

In the random effects model considered in MS, the nodes of interest are the group specific means. There may exist estimators that are sufficient statistics for these group specific means. In that case, outlier detection at the group level can also be based on cross validation, measuring the tail probability beyond the observed value of the statistic in the posterior predictive distribution given data from the other groups. This is considered the gold standard in MS. The aim of their alternative measure, which is well defined also in the absence of such sufficient statistics, is to match this measure as closely as possible. They show that if all conditional distributions in the model description are normal with fixed covariance matrices, the two measures match exactly. To further substantiate the sensibility of their new measure, they show in their appendix A3 that this equivalence result holds also for more general location distributions. The requirements are that the scale parameter is known, that the conditional density for the estimator given the group mean is symmetric, and that the difference between the estimator and the mean is a pivotal quantity.

In the present paper we will exploit the property of pivotality further. We show that symmetry is not needed for the above mentioned equivalence result, and that it applies beyond the case of location distributions. More importantly, we show that in various kinds of models, pivotality of conditional distributions used in the model specification implies that the conflict measures of GN are uniformly distributed under the assumed model. Furthermore, we show that this uniformity holds in models based on several frequently used distribution functions, by using data transformations and

reparametrizations. Hence, at least in these situations the measures of GN have comparable interpretations across different models, and can be used without computationally costly pre-calibrations schemes, such as the one suggested in HDS, and are therefore in particular well suited for model criticism in models using non-informative prior distributions.

The paper is organized as follows. Section 2 contains the necessary background material, including the definitions of the conflict measures given in DGN and GN. This section also briefly addresses computational issues, and presents a new result that is relevant in this context. The proof is given in the appendix. Section 3 discusses the concept of pivotality in relation to the conflict measures. The uniformity results announced above are given in Sections 4, 5 an 6. The latter section also contains the extension of the above mentioned equivalence result of MS. Section 7 discusses various aspects of the theory.

2 Directed acyclic graphs and node-specific conflict

2.1 Directed acyclic graphs and Bayesian hierarchical models

A large and important class of Bayesian hierarchical models can be represented and visualized by means of directed acyclic graphs, DAGs. An example discussed extensively in OH is the random effects model with normal random effects and normal error terms, defined by

$$Y_{i,j} \sim N(\lambda_i, \sigma^2), \lambda_i \sim N(\mu, \tau^2), j = 1, \ldots, n_i, i = 1, \ldots, m. \qquad (1)$$

In general we identify the nodes or vertices of the graph with the unknown parameters $\boldsymbol{\theta}$ and the observed data $\mathbf{y}$. The latter are the realizations of the random vector $\mathbf{Y}$, and are represented by the bottom nodes, having only directed edges pointing towards them. The parameters, the components of $\boldsymbol{\theta}$, are also considered as random variables in our Bayesian framework. In general, if there is a directed edge from node a to node b, then a is a parent of b, and b is a child of a. We denote by $\mathrm{Ch}(a)$ the set of child nodes of a, and by $\mathrm{Pa}(b)$ the set of parent nodes of b. More generally, b is

a descendant of a if there is a directed path from a to b, and in that case, a is an ancestor of b. The set of descendants of a is denoted by $\mathrm{Desc}(a)$, and for convenience is defined to contain a itself. The directed edges encode conditional independence assumptions, indicating that given its parents, a node is assumed to be independent of all other non-descendants. Hence, given the vector $\boldsymbol{\mu}$ of top level nodes the joint probability distribution of all the other parameters $\boldsymbol{\nu}$ and variables $\mathbf{Y}$ is of the form

$$p(\mathbf{y}, \boldsymbol{\nu}) = \prod_{y \in \mathbf{y}} p(y|\mathrm{Pa}(y)) \prod_{\nu \in \boldsymbol{\nu}} p(\nu|\mathrm{Pa}(\nu)). \tag{2}$$

A prior distribution $\pi(\boldsymbol{\mu})$ is specified for the top level parameters $\boldsymbol{\mu}$, and the inference is based on the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$.

This set up can be generalized in various directions. The nodes may be allowed to represent vectors, both at the parameter and the data levels, and conflict analysis in such models is considered in GN. Instead of DAGs one may consider chain graphs, as described in Lauritzen (1996), with undirected edges representing mutual dependence as in Markov random fields. Scheel, Green and Rougier (2011) introduces a graphical diagnostic for model criticism in such models.

2.2 Information contributions

Although the representation of a Bayesian hierarchical model in terms of a DAG is not necessarily unique, and in some cases just may be a convenient way to model the uncertainty underlying the observed data, the representation is often meant to reflect an understanding of the structure of the problem. By looking for a conflict associated with the different nodes in the DAG we may therefore put our understanding of this structure to test. We may also identify parts of the model that behave in an exceptional way, possibly deciding to give this part exceptional treatment.

OH argues that for each node λ in such a model one may in general think of each neighbouring node as providing information about λ, and that it is of interest to consider the possibility of conflict between these sources of information. The parents of λ provide information through the term $p(\lambda|\mathrm{Pa}(\lambda))$, which can be thought of as a local prior information source. On the other hand, each child node γ of λ provides information through $p(\gamma|\mathrm{Pa}(\gamma))$, which

we think of as a local likelihood information source, whether γ is an observed variable or an unobserved parameter. Together these types of factors determine the full conditional distribution of λ given all the observed and unobserved variables in the network, i.e.

$$p(\lambda|(\mathbf{y},\boldsymbol{\theta})_{-\lambda}) \propto p(\lambda|\mathrm{Pa}(\lambda)) \prod_{\gamma\in\mathrm{Ch}(\lambda)} p(\gamma|\mathrm{Pa}(\gamma)). \tag{3}$$

Here $(\mathbf{y},\boldsymbol{\theta})_{-\lambda}$ denotes the vector of all components of $(\mathbf{y},\boldsymbol{\theta})$ except for λ.

It is often of interest to contrast the local prior information with the likelihood information source formed by multiplying the factors $p(\gamma|\mathrm{Pa}(\gamma)$ for all child nodes $\gamma \in \mathrm{Ch}(\lambda)$. In DGN this product is normalized to a probability density function denoted by f_c, which we will call the likelihood information contribution, while the local prior density is denoted by f_p and called the prior information contribution. These information contributions contain unknown parameters, and are hence integrated with respect to posterior distributions for these parameters to form what we now will call integrated information contributions (iic) g_c and g_p. In this construction a key to avoid the conservatism of the OH measure is to prevent dependence between the two information sources by introducing a suitable data splitting $\mathbf{Y} = (\mathbf{Y}_p, \mathbf{Y}_c)$, and condition the parameters of f_p on $\mathbf{y}_p$ and the parameters of f_c on $\mathbf{y}_c$.

Definition 1 *For a given parameter node λ, denote by $\boldsymbol{\beta}_p$ the vector whose components are $Pa(\lambda)$, and by $\boldsymbol{\beta}_c$ the vector whose components are*

$$\cup_{\gamma\in Ch(\lambda)}(\{\gamma\} \cup Pa(\gamma)) - \{\lambda\}. \tag{4}$$

Let $\mathbf{Y} = (\mathbf{Y}_p, \mathbf{Y}_c)$ be a splitting of the data $\mathbf{Y}$. Define the densities f_p, f_c, the prior respectively likelihood information contributions, by

$$f_p(\lambda;\boldsymbol{\beta}_p) = p(\lambda|\boldsymbol{\beta}_p), \quad f_c(\lambda;\boldsymbol{\beta}_c) \propto \prod_{\gamma\in Ch(\lambda)} p(\gamma|Pa(\gamma)) \tag{5}$$

Define the integrated information contribution (iic) densities g_p, g_c by

$$g_p(\lambda) = \int f_p(\lambda;\boldsymbol{\beta}_p)\pi(\boldsymbol{\beta}_p|\mathbf{y}_p)d\boldsymbol{\beta}_p, \;\; g_c(\lambda) = \int f_c(\lambda;\boldsymbol{\beta}_c)\pi(\boldsymbol{\beta}_c|\mathbf{y}_c)d\boldsymbol{\beta}_c, \tag{6}$$

and denote by G_p, G_c the corresponding cumulative distribution functions.

Note that $\boldsymbol{\beta}_c$ may contain data nodes. The second integral in (6) is then taken only with respect to the random components of $\boldsymbol{\beta}_c$, i.e. the parameters in $\boldsymbol{\beta}_c$. If $\boldsymbol{\beta}_c$ contains no parameters, then g_c and f_c coincide.

The set of information sources linked to a specific node may in general be combined in different ways, potentially revealing different types of conflict about the node. This leads to a modification of Definition 1 where $\boldsymbol{\beta}_c$ does not contain all child nodes of λ, the others being instead included in $\boldsymbol{\beta}_p$ together with their parent nodes. This is natural e.g. in the context of outlier detection among independent observations with a common mean. Note that $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_c$ may then be overlapping, containing common coparents with λ. The definitions given in (5) are then modified as follows. Denote by $\boldsymbol{\gamma}$ the vector whose components are the child nodes of λ contained in $\boldsymbol{\beta}_c$, and by $\boldsymbol{\gamma}_1$ the rest of the child nodes, i.e. the set of child nodes contained in $\boldsymbol{\beta}_p$. The information contributions are then given by

$$f_p(\lambda; \boldsymbol{\beta}_p) \propto p(\boldsymbol{\gamma}_1 | \mathrm{Pa}(\boldsymbol{\gamma}_1) p(\lambda | \mathrm{Pa}(\lambda)), \tag{7}$$

$$f_c(\lambda; \boldsymbol{\beta}_c) \propto p(\boldsymbol{\gamma} | \mathrm{Pa}(\boldsymbol{\gamma})). \tag{8}$$

In (7), $p(\lambda|\mathrm{Pa}(\lambda))$ is replaced by the prior density $\pi(\lambda)$ if λ is a top level parameter. The corresponding iic densities are defined by (6) as before.

2.3 Node-specific conflict measures

The conflict measure c_λ^2 of DGN is defined as

$$c_\lambda^2 = (E^{G_p}(\lambda) - E^{G_c}(\lambda))^2 / (\mathrm{var}^{G_p}(\lambda) + \mathrm{var}^{G_c}(\lambda)) \tag{9}$$

The χ_1^2-distribution is the yardstick for measuring the level of conflict associated with c_λ^2. The conflict measures of GN are based on the same iic distributions, but focus on tail behaviour, and use the uniform distribution on $[0, 1]$ as reference distribution. The set up in GN is also more general by allowing likelihood information contributions to be based on a subset of the likelihood information sources and combining the rest with the local prior, cf. (7) and (8). For a given pair G_p, G_c of iic distributions let λ_p^* and λ_c^*be independent samples from G_p and G_c respectively. Let G be the cumulative distribution function for $\delta = \lambda_p^* - \lambda_c^*$. Define

$$c_\lambda^{3+} = G(0), \quad c_\lambda^{3-} = \bar{G}(0) \stackrel{\text{def}}{=} 1 - G(0) \tag{10}$$

and

$$c_\lambda^3 = 1 - 2\min(G(0), \bar{G}(0)) = 2|G(0) - 1/2|. \tag{11}$$

The c_λ^{3+}-measure is also very similar to the P_λ^{conf} measure suggested in MS, aimed at detecting outlying groups or units in a three level hierarchical model, where the second level parameters are location parameters for group specific data. However, the measure is interpreted as a p-value, with small values indicative of conflict, and is hence aimed at detecting the same divergent behaviour as the c_λ^{3-} measure of GN.

GN also defines a measure based on defining a tail area in terms of the density g of G, namely

$$c_\lambda^4 = P^G(g(\delta) > g(0)) \tag{12}$$

In a simulation study of the c_λ^2-measure in DGN using a warning level equal to the 95 percent quantile of the χ_1^2-distribution, a false warning probability of close to 5 percent is obtained for a normal random effects model with unknown variance parameters as in (1), and also in similar random effects models with heavytailed t- and uniformly distributed random effects. Also with respect to detection power this measure performs well when compared with a calibrated version of the OH measure. In a general normal model described by a DAG, with fixed covariances and with the basic improper prior distribution 1, the c^3 and c^4-measures are equivalent. They are shown in GN to be uniformly distributed pre-experimentally, i.e. their distributions as functions of a $\mathbf{Y}$ which is distributed according to the assumed model are uniform, regardless of the true value of the basic location parameter. Another way of stating this is that we obtain a proper p-value by subtracting these measures from 1.

In the following sections, exploiting the property of pivotality, we extend the theoretical analysis of GN from normal models to models involving various non-Gaussian distributions.

2.4 Integrated information contributions as posterior distributions

In most cases the conflict measures of DGN and GN are based on samples from G_p and G_c. Definition 1 suggests obtaining such samples by running an MCMC algorithm to generate posterior samples of the unknown parameters in $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_c$ from the respective posterior distributions, and then generate samples λ_p^* and λ_c^* from the respective information contributions for each such sample. This procedure is straightforward if the information contributions are standard probability densities. If not, the generation of samples can often be handled in practice by using the fact that under certain conditions on the data splitting the distributions G_p and G_c can be interpreted as posterior distributions conditional on $\mathbf{y}_p$ and $\mathbf{y}_c$ respectively, the latter based on the improper prior $\pi(\lambda) = 1$, independently of the coparents.

Theorem 1 *Suppose that the data splitting satisfies*

$$\mathbf{Y}_c = \mathbf{Y} \cap [\cup_{\gamma \in Ch(\lambda) \cap \boldsymbol{\beta}_c} Desc(\gamma)], \quad \mathbf{Y}_p = \mathbf{Y} - \mathbf{Y}_c, \tag{13}$$

the latter expression by abuse of notation meaning the components of $\mathbf{Y}$ *not present in* $\mathbf{Y}_c$*. We then have*

$$g_p(\lambda) = \pi(\lambda | \mathbf{y}_p)$$

and, specifying as prior density

$$\pi(\lambda | Pa(\, Ch(\lambda) \cap \boldsymbol{\beta}_c) - \lambda) = 1, \tag{14}$$

$$g_c(\lambda) = \pi(\lambda | \mathbf{y}_c)$$

.

The proof is given in the appendix.

In the case when $\mathrm{Ch}(\lambda)$ consists of data nodes, this condition is satisfied if

$$\mathbf{Y}_c = \mathrm{Ch}(\lambda) = \boldsymbol{\beta}_c \cap \mathbf{Y}, \quad \mathbf{Y}_p = \mathbf{Y} - \mathbf{Y}_c.$$

In DGN this splitting was compared with two other splittings and found to be optimal with respect to detection power for the conflict measure c_λ^2, which is also a well calibrated measure under this splitting. Throughout the rest of the paper we will assume that the condition (13) is satisfied.

3 Pivotality, transformations and reparametrizations

The iic distributions and the corresponding conflict measures depend on the choice of parametrization. On the basis of experience so far, it is not unreasonable to believe that in practice the conflict measures are fairly robust to changes in parametrization. However, our focus in the present paper will be to identify circumstances under which this non-invariance can be handled in a theoretically satisfactory way.

Consider a specific node in the DAG, and denote by ϕ the parameter corresponding to this node in a standard parametrization of the model. Suppose for simplicity that $\mathbf{Y}_c = \mathrm{Ch}(\phi)$. Assume that there exists a sufficient statistic Y_c and an alternative parametrization λ, being a strictly monotonic transformation $\lambda(\phi)$, such that $Y_c - \lambda$ is a pivotal quantity, i.e. the density for Y_c given λ is of the form

$$p(y_c|\lambda) = f_{Y_c}(y_c|\lambda) = f_0(y_c - \lambda) \tag{15}$$

for some known density function f_0. In the present paper we will for convenience also term $f_{Y_c}(y_c|\lambda)$ a pivotal density and the corresponding cdf $F_{Y_c}(y|\lambda)$ a pivotal distribution function when (15) is satisfied. We will regard a parametrization satisfying (15) as a canonical or reference parametrization if it exists, as opposed to the standard parametrization involving ϕ. Accordingly, the conflict measures given in (9),(10), (11)and (12) are preferably defined in terms of this canonical parametrization.

By Theorem 1, samples λ_c^* from G_c may be obtained by MCMC as posterior samples from $\pi(\lambda|y_c)$ when the splitting satisfies (13) and, in accordance with (14), λ has the improper prior $\pi(\lambda) = 1$. However, we may alternatively run an MCMC algorithm in the standard parametrization, using the prior density $|d\lambda/d\phi|$ for ϕ, to obtain samples ϕ_c^* from $\pi(\phi|\mathbf{Y}_c)$, and then calculate $\lambda_c^* = \lambda(\phi_c^*)$. To represent the iic distribution $G_p(\lambda)$, we may calculate $\lambda_p^* = \lambda(\phi_p^*)$for samples ϕ_p^* from $\pi(\phi|\mathbf{y}_p)$. Now the c_λ^4-measure can be calculated from (12), using a kernel density estimate of $g(\delta)$ based on corresponding samples $\delta^* = \lambda_p^* - \lambda_c^*$. However, if we limit attention to the c_λ^3-measure (11) and its one-sided versions (10), we may use the samples from $\pi(\phi|\mathbf{y}_c)$ and $\pi(\phi|\mathbf{y}_p)$ directly. Indeed, assuming that λ is an increasing function of ϕ, the condition $\lambda_p^* \geq \lambda_c^*$ is equivalent to the condition $\phi_p^* \geq \phi_c^*$. Hence, the

probability $G(0)$ that $\lambda_p^* - \lambda_c^* \leq 0$ can be estimated as the proportion of sample values for which $\phi_p^* \leq \phi_c^*$.

4 Data - prior conflict; application to normal variance parameters and some common survival distributions

The main result of this section is Theorem 2, stating that for a pure data - prior conflict, the c^3 - and c^4-measures are uniformly distributed if the density at the data level is pivotal. We also demonstrate how this pivotality condition can be met for a normal variance parameter as well as for some common survival distributions by means of transformations of the parameter and a sufficient statistic. We motivate the general theoretical results by focusing on the case of a normal variance parameter, and start by demonstrating the existence of a parametrization for which a sufficient statistic for this parameter has a pivotal density.

Suppose that $\mathbf{Y}_c$ consists of independent normal variables $Y_1, \ldots, Y_n$ with known expectations $\mu_1, \ldots, \mu_n$ and common, unknown variance σ^2. Define the associated sum of squares $S^2 = \sum_{i=1}^n (Y_i - \mu_i)^2$, which is a sufficient statistic for σ^2. Then, conditional on σ^2, $U = S^2/\sigma^2$ is χ^2-distributed with n degrees of freedom. Hence, U has density

$$f_U(u) = (1/2^{n/2}\Gamma(n/2))u^{(n/2)-1}\exp(-u/2)$$

.

Let $V = \log(S^2)$, and define $\rho = \log(\sigma^2)$. Hence, $U = \exp(V - \rho)$. Consequently, $du/dv = \exp(v - \rho)$, and the density for V is

$$f_V(v) = (1/2^{n/2}\Gamma(n/2))\exp((n/2)(v-\rho))\exp(-\exp(v-\rho)/2). \qquad (16)$$

Hence, $V - \rho$ is a pivotal quantity, and the density for V is indeed pivotal in the parametrization $\rho = \log(\sigma^2)$ according to our terminology.

Part of the proof for Theorem 2 reappears in other arguments as well, and is hence formalized as a lemma.

Lemma 1 *Suppose the child node part of the data splitting is a scalar Y_c with density of the form*

$f_{Y_c}(y|\lambda) = f_0(y - \lambda)$,

where f_0 is a known density function. If λ_p^* and λ_c^* are independent variables from the iic distributions G_p and G_c respectively, and g is the density function for $\delta = \lambda_p^* - \lambda_c^*$, then given $Y_c = y$ we have

$$g(\delta) = \int f_{Y_c}(\delta + y|\lambda) g_p(\lambda) d\lambda$$

.

Proof: Define the density $g_0(x) = f_0(-x)$. Then by the special case of (6) where $\boldsymbol{\beta}_c = y_c$ contains no random parameters, the iic density $g_c(\lambda)$ given $Y_c = y$ is proportional to

$f_{Y_c}(y|\lambda) = f_0(y - \lambda) = g_0(\lambda - y)$.

Since $\int g_0(\lambda - y)d\lambda = 1$, we have $g_c(\lambda) = g_0(\lambda - y)$. Hence, noting that $\delta = \lambda_p^* - \lambda_c^*$ implies that $\lambda_c^* = \lambda_p^* - \delta$, we have

$g(\delta) = \int g_c(\lambda - \delta) g_p(\lambda) d\lambda = \int g_0(\lambda - \delta - y) g_p(\lambda) d\lambda =$

$\int f_0((\delta + y) - \lambda) g_p(\lambda) d\lambda = \int f_{Y_c}(\delta + y|\lambda) g_p(\lambda) d\lambda$,

as asserted.

Now consider the model

$\mathbf{Y} \sim F_{\mathbf{Y}}(\mathbf{y}|\lambda), \;\; \lambda \sim F_\lambda(\lambda)$,

where F_λ is an arbitrary informative prior distribution. Here we think of this prior distribution as representing part of the randomness in the data generating process, rather than subjective uncertainty about the location of a fixed but unknown λ. An alternative perspective on this is discussed in section 7. The corresponding densities are denoted by $f_{\mathbf{Y}}$ and f_λ. If contrasting the prior density with the likelihood $f_{\mathbf{Y}}(\mathbf{y}|\lambda)$ indicates a discrepancy between the prior and likelihood information contributions, we will use the term data-prior conflict for this discrepancy. The following theorem deals with this kind of conflict. Note that in this situation the $\mathbf{Y}_p$-part of the data splitting is empty.

Theorem 2 *Suppose the conditional density for the scalar variable Y given the parameter λ is of the form $f_Y(y|\lambda) = f_0(y - \lambda)$, and that λ is generated from an arbitrary informative prior density $f_\lambda(\lambda)$. Then the data-prior conflict measures about λ are pre-experimentally uniformly distributed both for the c_λ^3- and c_λ^4-measures.*

Proof: According to the model the density for Y is $f(y) = \int f_Y(y|\lambda f_\lambda(\lambda)d\lambda$. Let F be the corresponding cdf. Noting that in this situation, the parent node iic density g_p coincides with the prior density f_λ, it follows from Lemma 1 that

$g(\delta) = \int f_Y(\delta + y|\lambda)f_\lambda(\lambda)d\lambda = f(\delta + y)$. It follows that $G(0) = F(y)$. Since $F(Y)$ is uniformly distributed if Y is distributed according to the model, we have that $G(0)$ is pre-experimentally uniform. Consequently, by (10) and (11) the c^3_λ-measures are uniform.

Let $I(\cdot)$ be the indicator funtion. the c^4_λ-measure of (12) is

$P^G(g(\delta) > g(0)) = \int I(g(\delta) > g(0))g(\delta)d\delta = \int I(f(\delta+y) > f(y))f(\delta+y)d\delta =$

$\int I(f(x) > f(y))f(x)dx = P^F(f(X) > f(y)) = 1 - R(f(y)),$

where R is the cdf of $f(X)$ when $X \sim F$. The pre-experimental uniformity follows, since also $Y \sim F$, and hence $f(Y) \sim R$. This completes the proof.

Returning to the normal case introduced at the beginning of this section, we can apply Theorem 2 to the pivotal quantity $V - \rho$, concluding that regardless of the prior distribution of ρ, all our conflict measures for ρ are uniform. As shown at the end of the previous section, with σ^2 corresponding to ϕ and ρ corresponding to λ, for the c^3-measure, the theorem also applies in the standard parametrization if we define $g_c(\sigma^2)$ as the posterior density based on the improper prior $\pi(\sigma^2) = 1/\sigma^2)$ instead of $\pi(\sigma^2) = 1$. The iic density g_c for σ^2 is then inverse gamma with shape parameter $n/2$ and scale parameter $S^2/2$.

Theorem 2 also applies if $\mathbf{Y}$ consists of n independent observations from an exponential distribution with failure rate β. In that case, the sum Y of the observations is a sufficient statistic. It is well known that Y is gamma distributed with shape parameter n and scale parameter β. Hence, the density for Y is proportional in both y and β to

$\beta^n y^{n-1} \exp(-\beta y).$

Define $V = \log(Y), \rho = -\log(\beta)$. Then $dy/dv = \exp(v)$, and hence the density for V is proportional to

$$\exp(n(v - \rho)) \exp(-\exp(v - \rho))$$

,

showing that $V - \rho$ is a pivotal quantity. The prior density for β that should

be used when the original parametrization is applied, is

$\pi(\beta) = -d\rho/d\beta = 1/\beta,$

and the corresponding posterior density = iic density g_c is gamma with shape parameter n and scale parameter Y. Essentially the same argument can be used if $\mathbf{Y}$ consists of n independent observations from a gamma distribution with known shape parameter α and unknown scale parameter β, using that the sum Y of the observations is gamma distributed with shape parameter $n\alpha$ and scale parameter β in that case. This argument also covers the case when the components of $\mathbf{Y}$ have a common inverse gamma distribution, since their inverses then are gamma distributed with the same parameters.

Furthermore, if each component $Y_i, i = 1, \ldots, n$ of $\mathbf{Y}$ has a Weibull density

$\beta\alpha y^{\alpha-1} \exp(-\beta y^{\alpha})$

with α known, then $Z_i = Y_i^{\alpha}$ is exponentially distributed with failure rate β. Hence, referring to the exponential case, $V = \log(\sum_{i=1}^{n} y_i^{\alpha})$ has a pivotal density expressed by means of the parameter $\rho = -\log(\beta)$. Again, $\pi(\beta) = 1/\beta$ is the relevant prior distribution in the usual parametrization.

Finally, we note that if $Y_i, i = 1, \ldots, n$ are independent, lognormally distributed variables, then obviously $V = \sum_{i=1}^{n} \log(Y_i)$ is a normally distributed variable which is sufficient and has a pivotal density. Hence, Theorem 2 can be used also in this case.

5 Data - data conflict

Suppose all components of $\mathbf{Y}$ have distributions determined by the same parameter λ. Suppose we want to contrast information contributions from separate parts of $\mathbf{Y}$ about λ, and define the splitting $(\mathbf{Y}_p, \mathbf{Y}_c)$ accordingly. Focusing on this kind of possible conflict, we assume complete prior ignorance about λ, and accordingly assume that λ has the improper prior $\pi(\lambda) = 1$. Hence, recalling (7) and (8) we contrast the information in $f_c(\lambda; \mathbf{Y}_c)$ with that in $f_p(\lambda; \mathbf{Y}_p)$. Since there is no prior information incorporated in f_p , the two information contributions in principle play symmetric roles. It is therefore natural to use the term data - data conflict in this context. However, as a particular application one may think of $\mathbf{Y}_c$ as a scalar variable representing a

possible outlier in order to justify the p vs. c, i.e. prior vs. child, notation also in this situation . Theorem 3 below shows that the c^3- and c^4-measures are pre-experimentally uniformly distributed in this case if there exists sufficient statistics Y_c respectively Y_p for $\mathbf{Y}_c$ respectively $\mathbf{Y}_p$ for which $Y_c - \lambda$ and $Y_p - \lambda$ are pivotal quantities.

Theorem 3 *Suppose that the conditional densities for the scalar variables Y_p and Y_c given the parameter λ are of the form*

$f_{Y_p}(y|\lambda) = f_{p,0}(y-\lambda), \;\; f_{Y_c}(y|\lambda) = f_{c,0}(y-\lambda).$

Assume λ has the improper prior $\pi(\lambda) = 1$. Then the data - data conflict measures about λ are pre-experimentally uniformly distributed both for the c^3_λ- and c^4_λ-measures.

Proof: With no nuisance parameters and with the improper prior $\pi(\lambda) = 1$, the iic distribution G_p for λ has density $g_{p,0}(\lambda - y_p)$, where $g_{p,0}(x) = f_{p,0}(-x)$. By Lemma 1 we have

$g(\delta) = \int f_{Y_c}(\delta + y|\lambda) g_p(\lambda) d\lambda = \int f_{c,0}(\delta + y_c - \lambda) g_{p,0}(\lambda - y_p) d\lambda =$

$\int f_{c,0}(\delta + y_c - y_p - (\lambda - y_p)) g_{p,0}(\lambda - y_p) d\lambda =$

$f_{c,0} * g_{p,0}(\delta + y_c - y_p).$

Defining $F_{c,0} * G_{p,0}(x) = \int_{-\infty}^{x} f_{c,0} * g_{p,0}(u) du$, it follows that $G(0) = F_{c,0} * G_{p,0}(y_c - y_p)$. Now $F_{c,0} * G_{p,0}$ is the cdf for a variable of the form $Z_c + U$, where Z_c and U are independent, and where $Z_c \sim F_{c,0}(z_c), U \sim G_{p,0}(u)$, i.e. of $Z_c - Z_p$, where $Z_p \sim F_{p,0}(z_p)$. Hence, $F_{c,0} * G_{p,0}(Z_c - Z_p)$ is uniformly distributed. We denote by λ_0 the true, unknown value of λ. Since clearly Z_p, Z_c have the same distributions as $Y_p - \lambda_0, Y_c - \lambda_0$ respectively, it follows that also $G(0) = F_{c,0} * G_{p,0}(Y_c - Y_p)$ is uniformly distributed pre-experimentally. This takes care of the c^3_λ-measures.

From the equation

$G(\delta) = F_{c,0} * G_{p,0}(\delta + y_c - y_p)$

and the fact that $F_{c,0} * G_{p,0}$ is the distribution function for $Y_c - Y_p$, the uniformity of the c^4_λ-measure follows by the same proof as in the previous section, by replacing f by $f_{c,0} * g_{p,0}$ and y by $y_c - y_p$ in the proof. This completes the proof.

As an application, consider a model where $\mathbf{Y}_r$ consists of independent normal variables with common variance $\sigma^2, r = p, c$. Assume doubt can be raised as to whether the two sets of variables have the same variance. Then the group specific sums of squares S_p^2 and S_c^2 are sufficient statistics, and the differences $\log(S_p^2) - \log(\sigma^2)$ and $\log(S_c^2) - \log(\sigma^2)$ are the pivotal quantities needed to make Theorem 3 applicable.

Theorem 3 can also be applied if the components of $\mathbf{Y}_c$ and $\mathbf{Y}_p$ are log-normally or exponentially distributed, or gamma, inverse gamma or Weibull with known shape parameter, since pivotal quantities based on sufficient statistics exist for these distributions.

In general, sufficient statistics do not necessarily exist. However, if all components of $\mathbf{Y}_p$ and $\mathbf{Y}_c$ have pivotal densities, the following lemma can be used to show that at least there exist pivotal quantities based on summary statistics Y_p and Y_c for respectively $\mathbf{Y}_p$ and $\mathbf{Y}_c$.

Lemma 2 *Suppose $Y_1, \ldots, Y_n$ are independent given λ, and that Y_i has density $f_i(y_i|\lambda) = f_{i,0}(y_i - \lambda), i = 1, \ldots, n$. Suppose λ has the improper prior 1, and define $Y = E(\lambda|\mathbf{Y})$. Let λ_0 be the true value of λ. Then the distribution of $Y - \lambda_0$ does not depend on λ_0, and is hence a pivotal quantity.*

Proof: We have

$$Y - \lambda_0 = (\int (\lambda - \lambda_0) \prod f_i(Y_i|\lambda) d\lambda) / (\int \prod f_i(Y_i|\lambda) d\lambda) =$$

$$(\int (\lambda - \lambda_0) \prod f_{i,0}((Y_i - \lambda_0) - (\lambda - \lambda_0)) d\lambda) / (\int \prod f_{i,0}((Y_i - \lambda_0) - (\lambda - \lambda_0)) d\lambda) =$$

$$(\int \eta \prod f_{i,0}((Y_i - \lambda_0) - \eta) d\eta) / (\int \prod f_{i,0}((Y_i - \lambda_0) - \eta) d\eta).$$

Since the distribution of each of the variables $Y_i - \lambda_0$ is independent of λ_0, it follows that the same is true for $Y - \lambda_0$, as asserted.

With appropriate pivotality conditions for each component of the vectors $\mathbf{Y}_c, \mathbf{Y}_p$, depending on the same parameter λ, Lemma 2 applies to $Y_r = E(\lambda|\mathbf{Y}_r), r = c, p$. Assuming $\pi(\lambda) = 1$, the conflict measures between g_c and g_p could intuitively be approximated by similarly defined conflict measures between $\pi(\lambda|y_c)$ and $\pi(\lambda|y_p)$. Applying Theorem 3 it follows that these conflict measures are uniformly distributed. A similar argument can be used also in connection with Theorem 2.

6 Random effects models

Suppose λ is a group specific parameter in a random effects model, but not necessarily with normal conditional distributions as in (1). Let $\mathbf{Y}_c$ consist of the variables whose realizations are the observations for the individuals belonging to this group. Let $\mathbf{Y}_p$ be the corresponding vector for individuals belonging to all other groups in the model. In this situation MS defines a conflict p-value ${p_\lambda}^{\text{conf}}$, based on independent samples from $\pi(\lambda|\mathbf{y}_p)$ and $\pi(\lambda|\mathbf{y}_c)$ in the same way as the c^{3+}-measure of GN based on samples from g_p and g_c. In this situation, by Theorem 1 $g_p(\lambda) = \pi(\lambda|\mathbf{y}_p)$. If the posterior distribution $\pi(\lambda|\mathbf{y}_c)$ is based on the improper prior $\pi(\lambda) = 1$, then also $g_c(\lambda) = \pi(\lambda|\mathbf{y}_c)$. Hence, the two conflict measures are identical in this case, i.e.

$$ {p_\lambda}^{\text{conf}} = c_\lambda^{3+}. \tag{17} $$

Suppose now that Y_c is a statistic for $\mathbf{Y}_c$ with density $f_{Y_c}(\cdot|\lambda)$ and cdf $F_{Y_c}(\cdot|\lambda)$. MS then also defines the cross-validatory p-value

$$ {P_\lambda}^{\text{mix}} = \int F_{Y_c}(y_c|\lambda)\pi(\lambda|\mathbf{y}_p)d\lambda. \tag{18} $$

In GN this quantity is considered as a special case of the c^{3+}-measure, viewing Y_c as a node in the DAG for which the Dirac measure at the observed value y_c provides a degenerate point mass information contribution, and accordingly denoted by $c_{Y_c}^{3+}$. Hence,

$$ c_{Y_c}^{3+} = {P_\lambda}^{\text{mix}} \tag{19} $$

In Appendix A3 MS shows that if λ is a location parameter with prior density $\pi(\lambda) = 1$, and if Y_c is a sufficient statistic whose density is symmetric, and for which $Y_c - \lambda$ is a pivotal quantity, then

$$ {P_\lambda}^{\text{mix}} = {p_\lambda}^{\text{conf}}\ . \tag{20} $$

Bearing in mind equations (17) and (19), part a) of the following theorem says that the identity (20) holds even if the density for Y_c is not symmetric. Part b) represents an extension of Theorem 3 to a genuinely hierarchical model.

Theorem 4 *a) Suppose the conditional density for the scalar variable Y_c given the parameter λ is of the form $f_{Y_c}(y|\lambda) = f^2_{c,0}(y-\lambda)$. Then*

$$c^{3+}_{Y_c} \stackrel{\text{def}}{=} \int F_{Y_c}(y_c|\lambda)\pi(\lambda|\mathbf{y}_p)d\lambda = c^{3+}_{\lambda}$$

.

b) Suppose in addition that λ and the scalar variable Y_p are independent given the parameter μ, whose prior distribution is $\pi(\mu) = 1$, and have conditional densities of the form

$f_\lambda(\lambda|\mu) = f^1_{c,0}(\lambda-\mu)$, $f_{Y_p}(y_p|\mu) = f_{p,0}(y_p-\mu)$.

Then the conflict measures c^3_λ and c^4_λ are pre-experimentally uniformly distributed.

Proof: It follows from Lemma 1 that

$g(\delta) = \int f_{Y_c}(\delta + y_c|\lambda)g_p(\lambda)d\lambda$.

Hence,

$c^{3+}_\lambda = G(0) = \int F_{Y_c}(y_c|\lambda)g_p(\lambda)d\lambda$.

Part a) follows, since by Theorem 1

$g_p(\lambda) = \pi(\lambda|\mathbf{y}_p)$.

To prove part b), define $g_{p,0}(x) = f_{p,0}(-x)$. We then have that $\pi(\mu|y_p) = g_{p,0}(\mu - y_p)$. It follows that

$g_p(\lambda) = \int f_\lambda(\lambda|\mu)\pi(\mu|y_p)d\mu = \int f^1_{c,0}(\lambda-\mu)g_{p,0}(\mu-y_p)d\mu =$

$\int f^1_{c,0}(\lambda - y_p - (\mu - y_p))g_{p,0}(\mu-y_p)d\mu = f^1_{c,0} * g_{p,0}(\lambda - y_p)$.

Arguing as in the proof of Theorem 3, with $f^2_{c,0}$ replacing $f_{c,0}$ and $f^1_{c,0} * g_{p,0}$ in place of $g_{p,0}$, we therefore obtain

$g(\delta) = f^2_{c,0} * (f^1_{c,0} * g_{p,0})(\delta + y_c - y_p)$.

It follows that

$G(0) = \int_{-\infty}^0 g(\delta)d\delta = (F^2_{c,0} * F^1_{c,0}) * G_{p,0}(y_c - y_p)$.

Denoting the true value of μ by μ_0, we have that $F^2_{c,0} * F^1_{c,0}(y_c - \mu_0)$ and $F_{p,0}(y_p - \mu_0)$ are the true distribution functions of Y_c and Y_p respectively.

Again we may argue as in the proof of Theorem 3 to show that $G(0)$ is pre-experimentally uniformly distributed, proving the result for the c^3-measures. The result for the c^4-measure is also proved as in the proof of Theorem 3.

In the case of a random effects model where all conditional densities in the model description are pivotal, we may use the construction of Lemma 2 to obtain summary statistics satisfying the conditions of part b). To see this, suppose that the parameter λ of part b) represents one out of $k+1$ groups, and that $Y_c = E(\lambda|\mathbf{Y}_c)$ is the statistic representing the data for this group. Then Y_c has a pivotal density by Lemma 2. For $i = 1, \ldots, k$ let $\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})$, and let $\mathbf{Y}_p = (\mathbf{Y}_1, \ldots, \mathbf{Y}_k)$. Allowing for individual covariates, we assume that for $i = 1, \ldots, k, j = 1, \ldots, n_i$ the variables $Y_{i,j}$ have densities of the form

$$f_{Y_{i,j}}(y|\lambda_i) = f^2_{i,j,0}(y - \lambda_i)$$

and are independent given $\lambda_1, \ldots, \lambda_k$. Furthermore, allowing for covariates also at the group level, we assume that $\lambda_1, \ldots, \lambda_k$ are independent given μ, with densities

$$f_{\lambda_i}(\lambda_i|\mu) = f^1_{i,0}(\lambda_i - \mu).$$

Define $Y_i = E(\lambda_i|\mathbf{Y}_i), i = 1, \ldots, k$. Then, by Lemma 2 the density of Y_i at y given λ_i is of the form $f^2_{i,0}(y - \lambda_i)$. It then follows that the corresponding density given μ is

$$\int f^2_{i,0}(y - \lambda) f^1_{i,0}(\lambda - \mu) d\lambda = f^2_{i,0} * f^1_{i,0}(y - \mu),$$

so that $Y_i - \mu$ is a pivotal quantity. Also, the variables $Y_1, \ldots, Y_k$ are independent given μ. Then, using Lemma 2 again, it follows that the statistic $Y_p = E(\mu|Y_1, \ldots, Y_k)$ has the property that $Y_p - \mu$ is pivotal. Hence, the conditions of part b) of Theorem 4 are satisfied, as asserted.

In the special case when $\lambda, \lambda_1, \ldots, \lambda_k$ are the group mean parameters in a normal model with fixed group-specific variances, the above construction leads to the sufficient statistic $Y_c = E(\lambda|\mathbf{Y}_c) = \bar{\mathbf{Y}}_c$, as well as $Y_p = E(\mu|\bar{Y}_1, \ldots, \bar{Y}_k)$, which is a weighted average of the observed group means, weighted by the respective precisions, which is also sufficient. Hence, the c^3_λ- and c^4_λ-measures based on the full data are uniformly distributed in this case.

In general we do not suggest actually computing statistics by means of the construction of Lemma 2, nor to base conflict analysis on them. The point is to show that our conflict measures, when based on statistics retaining much

of the information about location in the full data, are uniformly distributed. We might therefore expect that the conflict measures based on the full data are not too far from being uniform. However, it is interesting to note that if in a random effects model, such as the one described above, $Y_i = E(\lambda_i|\mathbf{Y}_i)$ is a sufficient statistic with respect to λ_i, then Y_i is sufficient also with respect to μ, whether the densities involved are pivotal or not. Indeed, the assumed sufficiency implies that $\pi(\lambda_i|\mathbf{y}_i) = \pi(\lambda_i|y_i)$. It follows that

$\pi(\mu|\mathbf{y}_i) = \int \pi(\mu|\lambda_i)\pi(\lambda_i|\mathbf{y}_i)d\lambda_i =$

$\int \pi(\mu|\lambda_i)\pi(\lambda_i|y_i)d\lambda_i = \pi(\mu|y_i),$

which depends on $\mathbf{y}_i$ only through y_i.

We conclude this section with an example of a non-Gaussian random effects model where our conflict measures are approximately uniformly distributed.

Example. Suppose that given $\lambda_1, \ldots, \lambda_k$ the variables $Y_{i,j}$ are independent and exponentially distributed with failure rates $\lambda_i, j = 1, \ldots, n_i, i = 1, \ldots, k$. Furthermore, assume that $\lambda_1, \ldots, \lambda_k$ are gamma distributed with known shape parameter α_0 and unknown scale parameter β. Let $Y_i = \sum_{j=1}^{n_i} Y_{i,j}$. Then Y_i is a sufficient statistic for $\mathbf{Y}_i$, and is gamma distributed with shape parameter n_i and scale parameter λ_i. It follows that the density of Y_i at y given α_0, β is

$p(y|\alpha_0, \beta) = \int (\lambda_i^{n_i}/\Gamma(n_i))y^{n_i-1}\exp(-\lambda_i y)(\beta^{\alpha_0}/\Gamma(\alpha_0))\lambda_i^{\alpha_0-1}\exp(-\lambda_i\beta)d\lambda_i =$

$(\beta^{\alpha_0}y^{n_i-1}/(\Gamma(n_i)\Gamma(\alpha_0)) \int \lambda_i^{n_i+\alpha_0-1}\exp(-(y+\beta)\lambda_i)d\lambda_i =$

$(\Gamma(n_i+\alpha_0)/(\Gamma(n_i)\Gamma(\alpha_0))(\beta^{\alpha_0}y^{n_i-1})/(\beta+y)^{n_i+\alpha_0}.$

Dividing both numerator and denominator in the last fraction by $y^{n_i+\alpha_0}$ we obtain

$$p(y|\alpha_0, \beta) = (\Gamma(n_i+\alpha_0)/(\Gamma(n_i)\Gamma(\alpha_0))((1/y)(\beta/y)^{\alpha_0})/(1+\beta/y)^{n_i+\alpha_0}). \quad (21)$$

This may be transformed into a pivotal density by the transformation $V_i = \log(Y_i)$ and the reparametrization $\rho = \log(\beta)$. A uniform prior distribution for ρ in this parametrization leads to the prior $\pi(\beta) = 1/\beta$ in the standard parametrization.

Define $U_i = Y_i/(n_i + \alpha_0)$. By approximating the denominator $(1 + \beta/(n_i + \alpha_0)u_i)^{n_i+\alpha_0}$ of (21) by $\exp(\beta/u_i)$ we see that U_i is approximately

inversely gamma distributed with shape parameter α_0 and scale parameter β. It follows that $\sum_{i=1}^{k} 1/U_i$ is approximately gamma distributed with shape parameter $k\alpha_0$ and scale parameter β. Hence, we have approximately that $V = -\log(\sum_{i=1}^{k} 1/U_i)$ is a sufficient statistic whose density given ρ is pivotal.

Suppose now that the parameter λ of interest has the same density as $\lambda_1, \ldots, \lambda_k$ and is the failure rate for another group of exponentially distributed variables, collected in the vector $\mathbf{Y}_c$. Defining $W =$ the logarithm of the sum of these observations, and $\psi = -\log(\lambda)$, we have that W is sufficient, and that the density for W given ψ and the density for ψ given ρ are pivotal. In view of Theorem 4b), using the canonical parametrization ρ, ψ, we conclude that c^3_ψ and c^4_ψ are approximately uniformly distributed pre-experimentally. Adhering to the standard parametrization β, λ, this applies also to the c^3_λ-measure if we define $g_c(\lambda) \propto p(\mathbf{Y}_c|\lambda)(1/\lambda)$ and $\pi(\beta) = 1/\beta$.

It is worth noting that in this example the exponential distributions could be replaced by gamma distributions with known shape parameters $\alpha, \alpha_1, \ldots, \alpha_k$ and scale parameters $\lambda, \lambda_1, \ldots, \lambda_k$. The only change in the above calculations is that n_i must be replaced by $\alpha_i n_i$.

7 Discussion

In the present paper we have exploited the property of pivotality to show that the c^3- and c^4-measures of conflict at the node level of Bayesian hierarchical models are uniformly distributed under the assumed model in a number of situations. The normal case with fixed covariance matrices was already covered in GN. Obviously, more cases can be covered by using pivotality of the skew normal distribution (Azzalini (1985)), as well as the t-distribution. Furthermore, the numerical results in DGN for the alternative c^2-measure indicate that the presence of nuisance parameters may not represent a serious obstacle. In addition, based on the construction of Lemma 2 we may expect the conflict measures to be approximately uniformly distributed in many other situations where the conditional distributions used in the model description are pivotal. Hence, it seems likely that these measures can be used in a wide variety of models without the need for computationally costly calibration.

Our results suggest defining the conflict measures in terms of a parametrization λ for which $Y_c - \lambda)$ is a pivotal quantity, if possible, and due to Theorem

1 serve as a pragmatic argument for choosing the prior density $\pi(\lambda) = 1$ in that case. This choice is also in accordance with the recommendation given in section 1.3 of Box and Tiao (1992) of choosing a locally uniform prior for a parameter with respect to which the likelihood is data translated, i.e. of the form

$$h(\lambda - f(\mathbf{y}_c))$$

,

for some function f of $\mathbf{y}_c$. This is not quite the same as pivotality of $f(\mathbf{Y}_c) - \lambda$, since the density of $f(\mathbf{Y}_c)$ may contain a multiplicative factor depending on $\mathbf{y}_c$) but not on λ, but pivotality is a special case. As discussed in section 3 of the present paper, the calculations can nevertheless be performed through posterior parameter samples arising from a more familiar, standard parametrization, based on a relevant, non-uniform prior.

A strategy that may be useful in some cases is suggested by the discussion of Box and Tiao (1992), section 1.3. They argue that data translatedness can be achieved approximately in some cases by calculating the absolute value of the second derivative of the logarithm of the likelihood, evaluated at the ml-estimator $\hat{\lambda}$. Defining $\pi(\hat{\lambda})$ proportional to the square root of this quantity, an approximately data translated likelihood is obtained by extending this formula to a possibly improper prior $\pi(\lambda)$. The desired parametrization is then given by the equation

$d\psi/d\lambda = \pi(\lambda)$.

As discussed at the end of section 3, we need not solve this equation if we only want to calculate the c^3-measures (10) and (11). The procedure can be used also for discrete distributions. For a Poisson distribution with parameter λ, Box and Tiao (1992) shows that the relevant prior is $\pi(\lambda) = 1/\sqrt{\lambda}$. For a binomial distribution with parameter p, the relevant prior is $\pi(p) = (p(1-p))^{-1/2}$. MS contains a binomial case study, where conflict measures are calculated for infant mortality rates after heart surgery for each of 12 hospitals in England, and the exceptionally high mortality rate connected to one specific hospital is found to represent a significant deviation from the model.

In models where the property of pivotality can not be applied, as for instance when discrete distributions are involved, and theoretical or numerical

support of any other kind has not yet been given, the measures should be used with caution.

With the exception of Section 4, the theory of the present paper assumes non-informative prior distributions. This is in line with the argument of Bayarri and Castellanos (2007) that time consuming and costly elicitation of expert based prior distributions should be avoided in the initial phases of the model building process. In contrast, even in cases when inference is to be based on non-informative priors, either because an objective analysis is desirable, or because prior information is too weak to be of any use for the inference, HDS suggests in section 9.3 to construct informative priors solely for the purpose of allowing use of a prior predictive approach for model evaluation. We find this suggestion rather unappealing. In our framework, the conflict measures are uniformly distributed regardless of the value of the basic parameters.

As for Section 4, we study conflict linked to an informative prior assumed to be part of the data generating process. In this context, the exact value of the basic parameters is part of the modelling assumptions to be checked. Hence, throughout the paper our framework would allow the traditional frequentist interpretation that on average a certain proportion of the correct modelling assumptions that are checked for, will be found suspicious.

In the model described in Theorem 2 of section 4, the data node measures such as $c_{Y_c}^{3+}$, corresponding to ${P_\lambda}^{\text{mix}}$ of MS, cf. equations (18) and (19), can also be viewed in this way. Alternatively the same mathematical set up could be used in the context of measuring conflict between an observed value and a prior predictive distribution expressing subjective uncertainty about Y_c. A value close to 1 of this or any of the other conflict measures may indicate that this prior distribution is not well founded, for instance because the information used to construct the prior is either less relevant or weaker than believed, or because the method for translating this information into a prior distribution is inappropriate. A similar perspective can by adopted in the parameter node context of Theorem 2. Assuming hypothetically that the parameter node λ had been observable, and observed to take the value λ_{obs}, we would measure the conflict between this observed value and the predictive prior by the number $F_\lambda(\lambda_{\text{obs}}) = G_p(\lambda_{\text{obs}})$. Now λ is in reality not observable. However, in section 5 of GN it is shown that c_λ^{3+} is the expected value of this quantity with respect to the distribution $G_c(\lambda)$. The uniform distribution of

c_λ^{3+} demonstrated in Theorem 2 then arises partly from subjective, epistemic uncertainty expressed in F_λ and partly from alleatory uncertainty or natural variability expressed in F_{Y_c}. This comment applies also in the context of Section 5 and 6, if we replace the improper prior used in these sections by an informative one. In our view, uniformity of the conflict measure under all these sources of uncertainty is still the natural ideal criterion for being a well calibrated conflict measure, the fulfillment of which ensures comparable assessment of the level of conflict independently of the model, distributional assumptions, location in the network and size of the data set.

One way to harmonize use of informative, epistemic priors with the set up of Sections 5 and 6, is to assume that informative priors arise out of a state of complete ignorance at some time point in the past, after which data of the same kind as $\mathbf{y}_p$ observed prior to the present time are used to obtain an updated prior. The term "pre-experimental" should then refer to the situation at this hypothetical time point in the past.

Acknowledgements

I am greatful to professor Bent Natvig for a number of valuable comments and suggestions.

References

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

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

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

[4] 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.

[5] Box, G. E. P. & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. Wiley, New York, 1992.

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

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

[8] Dias, S., Welton, N. J., Caldwell, D. M. & Ades, A. E. (2010). Checking consistency in mixed treatment comparison meta-analysis. *Statist. Med* **29**, 932–944.

[9] Dey, D., Gelfand, A. Swartz, T. & Vlachos, P. (1998). A simulation-intensive approach for checking hierarchical models. *Test* **7**, 325–346.

[10] 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.

[11] Gåsemyr, J. & Natvig, B. (2009). Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models. *Scand. J. Statist.* **36**, 822–838.

[12] Hjort, N. L., Dahl, F. A. & Steinbakk, G. H. (2005). Post-processing posterior predictive p-values. *J. Amer. Statist. Assoc.*, **101**, 1157–1174.

[13] Johnson, V. (2007). Bayesian model assessment using pivotal quantities. *Bayesian anal* **2**, 719–734.

[14] Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford, 1996

[15] Marshall, E. C. & Spiegelhalter, D. J. (2007). Identifying outliers in Bayesian hierarchical models. A simulation based approach. *Bayesian anal* **2**, 409–444.

[16] O'Hagan, A. (2003). HSSS model criticism (with discussion). In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured*

Stochastic Systems, pages 423–444. Oxford University Press, Oxford, 2003.

[17] 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.

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

Author address:
Jørund Gåsemyr
Department of Mathematics
University of Oslo
POBox 1053 Blindern
Oslo
Norway

gaasemyr@math.uio.no

Appendix A. Interpretation of iic distributions as posterior distributions

If $\mathbf{Y}_p$ contains no descendant nodes of λ, then λ is independent of $\mathbf{Y}_p$ given $\mathrm{Pa}(\lambda)$. If we also let $\boldsymbol{\beta}_p = \mathrm{Pa}(\lambda)$, it is easy to see from (6) that

$$g_p(\lambda) = \pi(\lambda|\mathbf{y}_p). \tag{22}$$

In this Appendix we go beyond Definition 1, allowing for information contributions of the form (7) and (8), cf. also section 5 of GN. Under certain conditions on the data splitting we generalize (22) to situations where $\boldsymbol{\beta}_p$ contains some child nodes of λ, while $\boldsymbol{\beta}_c$ contains the rest of these nodes. We also show that under other conditions on the splitting,

$$g_c(\lambda) = \pi(\lambda|\mathbf{y}_c) \tag{23}$$

if λ has a uniform prior distribution. It turns out that for any way of distributing disjoint parts of $\mathrm{Ch}(\lambda)$ to $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_c$, the conditions for (22) and

(23) to hold are met simultaneously for exactly one data splitting, namely the splitting given by (13).

A1. The child node information contribution

Case 1, the data node case.

Suppose first that $\mathbf{Y}_c$ consists of the child nodes of λ in $\boldsymbol{\beta}_c$. This means that $\boldsymbol{\beta}_c = (\mathbf{Y}_c, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ consists of the coparents with λ for $\mathbf{Y}_c$. Assume, in accordance with (14), that $\pi(\lambda) = \pi(\lambda|\boldsymbol{\xi}) = 1$. Then by Bayes theorem

$f_c(\lambda; \mathbf{y}_c, \boldsymbol{\xi}) \propto p(\mathbf{y}_c|\lambda, \boldsymbol{\xi}) = p(\mathbf{y}_c|\lambda, \boldsymbol{\xi})\pi(\lambda|\boldsymbol{\xi}) \propto \pi(\lambda|\mathbf{y}_c, \boldsymbol{\xi})$.

By (6) it therefore follows that

$g_c(\lambda) = \int f_c(\lambda; \mathbf{y}_c, \boldsymbol{\xi})\pi(\boldsymbol{\xi}|\mathbf{y}_c)d\boldsymbol{\xi} = \int \pi(\lambda|\mathbf{y}_c, \boldsymbol{\xi})\pi(\boldsymbol{\xi}|\mathbf{y}_c)d\boldsymbol{\xi} = \pi(\lambda|\mathbf{y}_c)$,

in accordance with (23).

Case 2, the parameter node case.

Next, suppose that $\boldsymbol{\beta}_c = (\boldsymbol{\gamma}, \boldsymbol{\xi})$, where $\boldsymbol{\gamma}$ consists of parameter child nodes of λ. The parameters in $\boldsymbol{\xi}$ are coparents with λ for $\boldsymbol{\gamma}$. Assume $\mathbf{Y}_c$ consists of data descendant nodes of $\boldsymbol{\gamma}$, formally

$$\mathbf{Y}_c \subseteq [\cup_{\gamma \in \mathrm{Ch}(\lambda) \cap \boldsymbol{\beta}_c} \mathrm{Desc}(\gamma)]. \tag{24}$$

Then λ is independent of $\mathbf{Y}_c$ given $\boldsymbol{\gamma}$. Assume as before that $\pi(\lambda) = \pi(\lambda|\boldsymbol{\xi}) = 1$. Then, by the same argument as in case 1 we have

$f_c(\lambda; \boldsymbol{\gamma}, \boldsymbol{\xi}) = \pi(\lambda|\boldsymbol{\gamma}, \boldsymbol{\xi})$.

By the independence between λ and $\mathbf{Y}_c$ given $\boldsymbol{\gamma}$ it follows that

$g_c(\lambda) = \int f_c(\lambda; \boldsymbol{\gamma}, \boldsymbol{\xi})\pi(\boldsymbol{\gamma}, \boldsymbol{\xi}|\mathbf{y}_c)d\boldsymbol{\gamma}d\boldsymbol{\xi} = \int \pi(\lambda|\boldsymbol{\gamma}, \boldsymbol{\xi})\pi(\boldsymbol{\gamma}, \boldsymbol{\xi}|\mathbf{y}_c)d\boldsymbol{\gamma}d\boldsymbol{\xi} =$

$\int \pi(\lambda|\boldsymbol{\gamma}, \boldsymbol{\xi}, \mathbf{y}_c)\pi(\boldsymbol{\gamma}, \boldsymbol{\xi}|\mathbf{y}_c)d\boldsymbol{\gamma}d\boldsymbol{\xi} = \pi(\lambda|\mathbf{y}_c)$,

in accordance with (23).

A2. The parent node information contribution

We will assume that λ has parent nodes. The case of λ being a top level parameter without parents, as in Section 5, can be dealt with in a similar way.

Case 1, the data node case.

Suppose first that the subvector $\mathbf{Y}_1$ of $\mathbf{Y}_p$ consists of child nodes of λ, the other components of $\mathbf{Y}_p$ being independent of λ given the parent nodes of λ. Denote by $\boldsymbol{\xi}$ the coparents with λ for $\mathbf{Y}_1$, and by $\boldsymbol{\mu}$ the parent nodes of λ. Assume λ and $\boldsymbol{\xi}$ are independent. Let $\boldsymbol{\beta}_p = (\mathbf{y}_1, \boldsymbol{\xi}, \boldsymbol{\mu})$. Then the parent node information contribution is

$f_p(\lambda; \mathbf{y}_1, \boldsymbol{\xi}, \boldsymbol{\mu}) \propto p(\mathbf{y}_1|\lambda, \boldsymbol{\xi}) p(\lambda|\boldsymbol{\mu})$.

By Bayes theorem

$\pi(\lambda|\mathbf{y}_p, \boldsymbol{\xi}, \boldsymbol{\mu}) = \pi(\lambda|\mathbf{y}_1, \boldsymbol{\xi}, \boldsymbol{\mu}) \propto p(\mathbf{y}_1|\lambda, \boldsymbol{\xi}, \boldsymbol{\mu}) \pi(\lambda|\boldsymbol{\xi}, \boldsymbol{\mu}) =$

$p(\mathbf{y}_1|\lambda, \boldsymbol{\xi}) p(\lambda|\boldsymbol{\mu}) \propto f_p(\lambda; \mathbf{y}_1, \boldsymbol{\xi}, \boldsymbol{\mu})$.

Hence, from (6)

$g_p(\lambda) = \int f_p(\lambda; \mathbf{y}_1, \boldsymbol{\xi}, \boldsymbol{\mu}) \pi(\boldsymbol{\xi}, \boldsymbol{\mu}|\mathbf{y}_p) d\boldsymbol{\xi} d\boldsymbol{\mu} =$

$\int \pi(\lambda|\mathbf{y}_p, \boldsymbol{\xi}, \boldsymbol{\mu}) \pi(\boldsymbol{\xi}, \boldsymbol{\mu}|\mathbf{y}_p) d\boldsymbol{\xi} d\boldsymbol{\mu} = \pi(\lambda|\mathbf{y}_p)$,

in accordance with (22). Hence, if data child nodes appearing in $\mathbf{Y}_p$ constitute the child node part of $\boldsymbol{\beta}_p$, then G_p is a posterior distribution given $\mathbf{y}_p$. This also agrees with the condition of A1, case 1 that $\mathbf{Y}_c$ consists of the child nodes of λ in $\boldsymbol{\beta}_c$. Moreover, recalling that any node is a descendant of itself, both conditions agree with (13), and consequently Theorem 1 is proved in the data node case. In this case (13) essentially says that the splitting corresponds to the partition of $\mathrm{Ch}(\lambda)$ into child and parent node information contributions.

Case 2, the parameter node case.

Next, suppose $\boldsymbol{\beta}_p$ is of the form $\boldsymbol{\gamma}_1, \boldsymbol{\xi}, \boldsymbol{\mu}$, where $\boldsymbol{\gamma}_1 = \mathrm{Ch}(\lambda) \cap \boldsymbol{\beta}_p$ is a subvector of $\mathrm{Ch}(\lambda)$ consisting of parameter nodes, $\boldsymbol{\xi}$ are coparents with λ for $\boldsymbol{\gamma}_1$, assumed independent of λ, and as before $\boldsymbol{\mu}$ are the parent nodes of λ. Supppose that

$$\mathbf{Y}_p \subseteq \mathbf{Y} - \mathbf{Y} \cap [\cup_{\gamma \in \mathrm{Ch}(\lambda) \cap \boldsymbol{\beta}_c} \mathrm{Desc}(\gamma)]. \tag{25}$$

Let the subvector $\mathbf{Y}_1$ of $\mathbf{Y}_p$ consist of descendant nodes of $\boldsymbol{\gamma}_1$. By (25), the other components of $\mathbf{Y}_p$ are independent of λ given $\boldsymbol{\mu}$. Hence, λ and $\mathbf{Y}_p$ are independent given $\boldsymbol{\beta}_p$, and by this conditional independence and Bayes theorem we have

$\pi(\lambda|\boldsymbol{\gamma}_1, \boldsymbol{\xi}, \boldsymbol{\mu}, \mathbf{y}_p) = \pi(\lambda|\boldsymbol{\gamma}_1, \boldsymbol{\xi}, \boldsymbol{\mu}) \propto p(\boldsymbol{\gamma}_1|\lambda, \boldsymbol{\xi}, \boldsymbol{\mu})\pi(\lambda|\boldsymbol{\xi}, \boldsymbol{\mu}) =$

$p(\boldsymbol{\gamma}_1|\lambda, \boldsymbol{\xi})p(\lambda|\boldsymbol{\mu}) \propto f_p(\lambda; \boldsymbol{\gamma}_1, \boldsymbol{\xi}, \boldsymbol{\mu})$.

Hence, from (6)

$g_p(\lambda) = \int f_p(\lambda; \boldsymbol{\gamma}_1, \boldsymbol{\xi}, \boldsymbol{\mu})\pi(\boldsymbol{\gamma}_1, \boldsymbol{\xi}, \boldsymbol{\mu}|\mathbf{y}_p)d\boldsymbol{\gamma}_1 d\boldsymbol{\xi} d\boldsymbol{\mu} =$

$\int \pi(\lambda|\boldsymbol{\gamma}_1, \boldsymbol{\xi}, \boldsymbol{\mu}, \mathbf{y}_p)\pi(\boldsymbol{\gamma}_1, \boldsymbol{\xi}, \boldsymbol{\mu}|\mathbf{y}_p)d\boldsymbol{\gamma}_1 d\boldsymbol{\xi} d\boldsymbol{\mu} = \pi(\lambda|\mathbf{y}_p)$,

in accordance with (22). Hence, if parameter child nodes $\boldsymbol{\gamma}_1$ appear in the f_p information contribution, then G_p is a posterior distribution given $\mathbf{y}_p$ as long as all data descendant nodes of λ included in $\mathbf{Y}_p$ are also descendants of $\boldsymbol{\gamma}_1$, which is a consequence of (25). By letting $\mathbf{Y}_p$ contain all such data descendant nodes of $\boldsymbol{\gamma}_1$, and hence letting $\mathbf{Y}_c$ consist of all data descendant nodes of $\boldsymbol{\gamma} = \mathrm{Ch}(\lambda) \cap \boldsymbol{\beta}_c$, (24) is also satisfied, and G_c is a posterior distribution given $\mathbf{y}_c$. Combining (24) and (25) precisely leads to the splitting (13), proving Theorem 1 in the parameter node case.

Specializing to the standard set up of Definition 1, where $\mathrm{Ch}(\lambda) \subseteq \boldsymbol{\beta}_c$, we see that the requirement for (13) and hence Theorem 1 to hold in both cases is that $\mathbf{Y}_c$ consists of all data descendant nodes of λ, while $\mathbf{Y}_p$ consists of the rest of the data nodes.