Chapter 2

Node-Level Conflict Measures in Bayesian Hierarchical Models Based on Directed Acyclic Graphs

Jørund I. Gåsemyr and Bent Natvig

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70058

Abstract

Over the last decades, Bayesian hierarchical models defined by means of directed, acyclic graphs have become an essential and widely used methodology in the analysis of complex data. Simulation-based model criticism in such models can be based on conflict measures constructed by contrasting separate local information sources about each node in the graph. An initial suggestion of such a measure was not well calibrated. This shortcoming has, however, to a large extent been rectified by subsequently proposed alternative mutually similar tail probability-based measures, which have been proved to be uniformly distributed under the assumed model under various circumstances, and in particular, in quite general normal models with known covariance matrices. An advantage of this is that computationally costly precalibration schemes needed for some other suggested methods can be avoided. Another advantage is that noninformative prior distributions can be used when performing model criticism. In this chapter, we describe the basic framework and review the main uniformity results.

Keywords: cross-validation, data splitting, information contribution, MCMC, model criticism, pivotal quantity, preexperimental distribution, *p*-value

1. Introduction

Over the last decades, Bayesian hierarchical models have become an essential and widely used methodology in the analysis of complex data. Computational techniques such as Markow Chain Monte Carlo (MCMC) methods make it possible to treat very complex models and data structures. Analysis of such models gives intuitively appealing Bayesian inference based on posterior probability distributions for the parameters.

In the construction of such models, an understanding of the underlying structure of the problem can be represented by means of directed acyclic graphs (DAGs), with nodes in the graph

corresponding to data or parameters, and directed edges between parameters representing conditional distributions. However, a perfect understanding of the underlying structure is usually an unachievable goal, and there is always a danger of constructing inadequate models. Box [1] suggests a pattern for the model building process where an initial candidate model is assessed for adequacy, and if necessary modified and elaborated on, leading to a new candidate that again is checked for adequacy, and so on. As a tool in this model criticism process, Ref. [1] suggests using the prior predictive distribution of some checking function or test statistic as a reference for the observed value of this checking function, resulting in a prior predictive p-value. This requires an informative and realistic prior distribution, which is not always available or even desirable. Indeed, as pointed out in Ref. [2], in an early phase of the model building process, it is often convenient to use noninformative or even improper priors and thus avoid costly and time-consuming elicitation of prior information. Noninformative priors may be used also for the inference because relevant prior information is unavailable.

There exist many other methods for checking the overall fit of the model or an aspect of the model of special interest, based on locating a test statistic or a discrepancy measure in some kind of a reference distribution. The posterior predictive p-value (ppp) of Ref. [3] uses the posterior distribution as reference and does not require informative priors. But this method uses data twice and can as a result be very conservative [2, 4–6]. Hjort et al. [5] suggest remedying this by using the ppp value as a test statistic in a prior predictive test. The computation of the resulting calibrated cppp-value is, however, very computer intensive in the general case, and again realistic, informative priors are needed. A node-level discrepancy measure suggested in Ref. [7] is subject to the same limitations. The partial posterior predictive p-value of Ref. [4] avoids double use of data and allows noninformative priors but may be difficult to compute and interpret in hierarchical models.

Comparison with other candidate models through a technique for model comparison or model choice, such as predictive methods, maximum posterior probability, Bayes factors or an information criterion, can also serve as tools for checking model adequacy indirectly when alternative candidate models exist.

In this chapter, we will, however, focus on methods for criticizing models in the absence of any particular alternatives. We will review methods for checking the modeling assumptions at each node of the DAG. The aim is to identify parts or building blocks of the model that are in discordance with reality, which may be in need of adjustment or further elaboration. O'Hagan [8] regards any node in the graph as receiving information from two disjoint subsets of the neighboring nodes. This information is represented as a conditional probability density or a likelihood or as a combination of these two kinds of information sources. Adopting the same basic perspective, our aim is to check for inconsistency between such subsets. The suggestion in Ref. [8] is to normalize these information sources to have equal height 1 and to regard the height of the curves at the point of intersection as a measure of conflict. However, as shown in Ref. [2], this measure tends to be quite conservative. Dahl et al. [9] demonstrated that it is also poorly calibrated, with false warning probabilities that vary substantially between models. Dahl et al. [9] also identified the different sources of inaccuracy and modified the measure of Ref. [8] to an approximately χ^2-distributed quantity under the assumed model by

instead normalizing the information sources to probability densities. In Ref. [10], these densities were instead used to define tail probability-based conflict measures. Gåsemyr and Natvig [10] showed that these measures are uniformly distributed in quite general hierarchical normal models with fixed variances/covariances. In Ref. [11], such uniformity results were proved in various situations involving nonnormal and nonsymmetric distributions. These uniformity results indicate that the measures of Refs. [9] and [10] have comparable interpretations across different models. Therefore, they can be used without computationally costly precalibration schemes, such as the one suggested in Ref. [5]. Gåsemyr [12] focuses on some situations where the conflict measure approach can be directly compared to the calibration method of Ref. [5] and shows that the less computer-intensive conflict measure approach performs at least as well in these situations. Moreover, the conflict measure approach can be applied in models using noninformative prior distributions.

Focusing on the special problem of identifying outliers among the second-level parameters in a random-effects model, Ref. [13] defines similar conflict measures. In this setting, the group-specific means are the nodes of interest. In some models, there exist sufficient statistics for these means. Then, 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. In this context, the conflict measure approach can be viewed as an extension of cross-validation to situations where sufficient statistics do not exist. Also in Ref. [13] applications to the examination of exceptionally high hospital mortality rates and to results from a vaccination program are given. In Ref. [14], this methodology is used to check for inconsistency in multiple treatment comparison of randomized clinical trials. Presanis et al. [15] apply these conflict measures in complex cases of medical evidence synthesis.

2. Directed acyclic graphs and node-specific conflict

2.1. Directed acyclic graphs and Bayesian hierarchical models

An example of a DAG discussed extensively in Ref. [8] is the random-effects model with normal random effects and normal error terms defined by

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

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 appearing as bottom nodes and being the realizations of the random vector $\mathbf{Y}$. In the Bayesian model, the parameters, the components of $\boldsymbol{\theta}$, are also considered as random variables. 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. 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

nondescendants. Hence, writing $\boldsymbol{\theta} = (\boldsymbol{\nu}, \boldsymbol{\mu})$, with $\boldsymbol{\mu}$ representing the vector of top-level nodes, the joint density of $(\mathbf{Y}, \boldsymbol{\theta}) = (\mathbf{Y}, \boldsymbol{\nu}, \boldsymbol{\mu})$ is

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

where $\pi(\boldsymbol{\mu})$ is the prior distribution of $\boldsymbol{\mu}$. The posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ is the basis for the inference.

This setup can be generalized in various directions. The nodes may be allowed to represent vectors, at both the parameter and the data levels [10]. Instead of DAGs, one may consider chain graphs, as described in Ref. [16], with undirected edges representing mutual dependence as in Markov random fields. Scheel et al. [17] introduce a graphical diagnostic for model criticism in such models.

2.2. Information contributions

The representation of a Bayesian hierarchical model in terms of a DAG is often meant to reflect an understanding of the underlying 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 need adjustment.

The idea put forward in Ref. [8] is that for each node λ in a DAG one may in general think of each neighboring node as providing information about λ and that it is of interest to consider the possibility of conflict between different sources of information. For instance, one may want to contrast the local prior information provided by the factor $p(\lambda|\mathrm{Pa}(\lambda))$ with the likelihood information source formed by multiplying the factors $p(\gamma|\mathrm{Pa}(\gamma))$ for all child nodes $\gamma \in \mathrm{Ch}(\lambda)$. The full conditional distribution of λ given all the observed and unobserved variables in the DAG, i.e.,

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

is determined by these two types of factors. Here, $(\mathbf{y}, \boldsymbol{\theta})_{-\lambda}$ denotes the vector of all components of $(\mathbf{y}, \boldsymbol{\theta})$ except for λ.

Dahl et al. [9] normalize the product $\prod_{\gamma \in \mathrm{Ch}(\lambda)} p(\gamma|\mathrm{Pa}(\gamma))$ to a probability density function denoted by $f_c(\lambda)$, the likelihood or child node information contribution, whereas the local prior density is denoted by $f_p(\lambda)$ and called the prior or parent node information contribution. These information contributions are integrated with respect to posterior distributions for the unknown nuisance parameters to form integrated information contribution (iic) denoted by g_c and g_p. In this construction, a key to avoid the conservatism of the measure suggested in Ref. [8] 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, denoted by β_p the vector whose components are Pa(λ), and by β_c the vector whose components are*

$$\cup_{\gamma \in \mathrm{Ch}(\lambda)}(\{\gamma\} \cup Pa(\gamma)) - \{\lambda\} = \mathrm{Ch}(\lambda) \cup [Pa(Ch(\lambda)) - \{\lambda\}] \tag{4}$$

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 densities g_p, g_c *by*

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

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 Eq. (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. Definition 1 may also be extended to the case when λ is a vector, corresponding to a subset of parameter nodes.

Combining the set of information sources linked to a specific node in different ways 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. In this way, different types of conflict about the node may be revealed. 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 setup is illustrated in **Figure 1** in the case when the

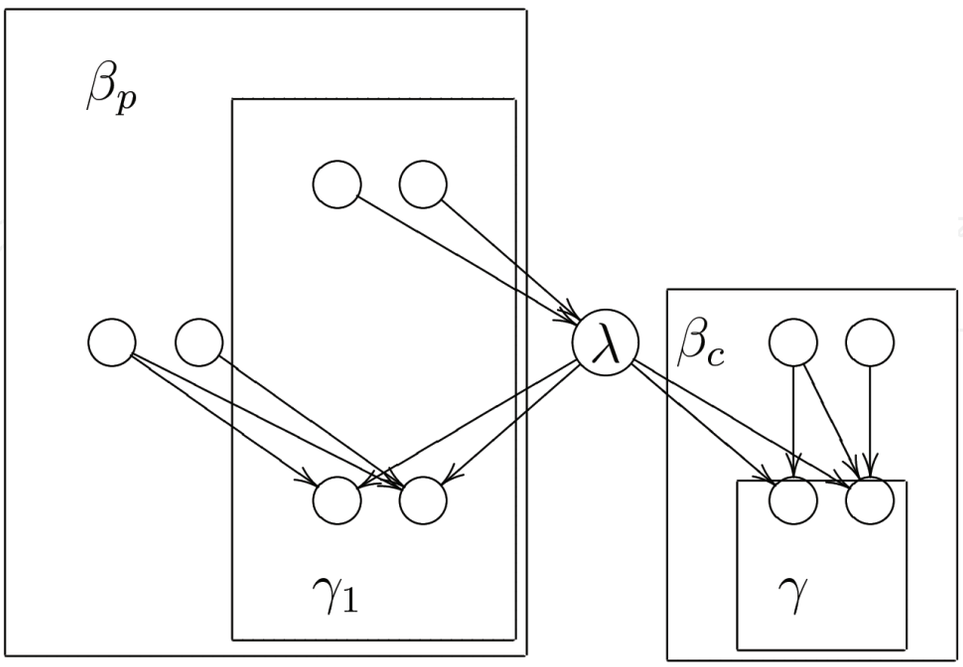

Figure 1. Part of a DAG showing information sources about λ.

set of common components, by abuse of notation denoted by $\boldsymbol{\beta}_p \cap \boldsymbol{\beta}_c$, is empty. For the general setup, Definition 1 is modified as follows.

Definition 2 *Let $\boldsymbol{\gamma}$ be a vector whose components are a subset of $Ch(\lambda)$, and define $\boldsymbol{\beta}_c$ as in Eq. (4). Denote by $\boldsymbol{\gamma}_1$ the rest of the child nodes of λ, and let $\boldsymbol{\beta}_p$ consist of $\boldsymbol{\gamma}_1$ together with its parent nodes in the same way as in Eq. (4), as well as $Pa(\lambda)$. The information contributions are then given by*

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

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

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

2.3. Node-specific conflict measures

The conflict measure c_λ^2 of Ref. [9] 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 reference distribution for this measure. For the conflict measures of Ref. [10], the uniform distribution on [0, 1] is the reference distribution. They focus on tail behavior but are based on the same iic distributions. The general distribution of information sources given in Definition 2 is also introduced in Ref. [10]. 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

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

and

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

The c_λ^{3+}-measure and the $P_\lambda^{\mathrm{conf}}$ measure of Ref. [13] are very similar. The latter measure is aimed at detecting outlying groups or units in a three-level hierarchical model, with the second-level parameters being location parameters for group-specific data. However, the measure is interpreted as a p value, with small values indicative of conflict. Gåsemyr and Natvig [10] also defines a measure based on defining a tail area in terms of the density g of G, namely

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

applicable also when λ is a vector.

Example 1. To illustrate the theory, consider the random-effects model 1, with the variance parameters σ^2, τ^2 assumed known, and with μ having the improper prior $\pi(\mu) = 1$. For simplicity, assume $n_i = n$ for all i. Suspecting the mth group of representing an outlier, let $\lambda=\lambda_m$ be the node of interest. Define the data splitting $\mathbf{Y}_p$, $\mathbf{Y}_c$ by letting $\mathbf{Y}_c = \mathbf{Y}_m = (Y_{m,1}, \ldots, Y_{m,n})$, and let $\boldsymbol{\beta}_c = \mathbf{y}_c$, $\boldsymbol{\beta}_p = \mu$. Denoting the normal density function by ϕ, it is easy to see that $g_c(\lambda) = f_c(\lambda) = \phi(\lambda; \overline{y}_c, \sigma^2/n)$. Furthermore, $f_p(\lambda; \mu) = \phi(\lambda; \mu, \tau^2)$. Given $\mathbf{y}_p$, μ has the density $\pi(\mu|\mathbf{y}_p) = \phi(\mu); \sum_{i=1}^{m-1} \overline{y}i/(m-1), (1/(m-1))\tau^2 + (1/(n(m-1)))\sigma^2)$. By a standard argument

$$\begin{aligned} g_p(\lambda) &= \int f_p(\lambda; \mu)\pi(\mu|\mathbf{y}_p)d\mu \\ &= \phi(\lambda; \sum_{i=1}^{m-1} \overline{y}_i/(m-1), (1 + 1/(m-1))\tau^2 + (1/(n(m-1)))\sigma^2). \end{aligned}$$

It follows that $g(\delta) = \phi(\delta); \sum_{i=1}^{m=1} \overline{y}_i/(m-1) - \overline{\mathbf{y}}_c, (m/(m-1))(\tau^2 + \sigma^2/n)$. The conflict measures (Eqs. (9), (10), (11), and (12)) can hence be calculated analytically, with no simulation needed in this case.

In a simulation study of the c_λ^2-measure in Ref. [9] using a warning level equal to the 95% quantile of the χ_1^2-distribution, a false warning probability of close to 5% is obtained for a normal random-effects model with unknown variance parameters as in Eq. (1) and also in similar random-effects models with heavy-tailed t- and uniformly distributed random effects. Also with respect to detection power, this measure performs well when compared to a calibrated version of the measure given in Ref. [8], if an optimal data splitting is used. Refs. [10] and [11] prove preexperimental uniformity of the conflict measures in various situations, 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 parameter. Another way of stating this is that we obtain a proper p-value by subtracting these measures from 1. These results are reviewed in Section 5 of the present chapter.

2.4. Integrated information contributions as posterior distributions

In most cases, the conflict measures of Refs. [9] and [10] are based on simulated samples from G_p and G_c. Definitions 1 and 2 suggest obtaining such samples by running an MCMC algorithm to generate posterior samples of the unknown parameters in $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_c$ and then generate samples λ_p^* and λ_c^* from the respective information contributions for each such parameter sample. If the information contributions are standard probability densities, this procedure is straightforward. If not, one may instead often use the fact that, under certain conditions on the data splitting, the distributions G_p and G_c are 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*

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

the latter expression by abuse of notation meaning the components of $\mathbf{Y}$ *not present in* $\mathbf{Y}_c$. *Assume* λ *and the coparents* $Pa\left(Ch(\lambda) \cap \beta_p\right) - \lambda$ *are independent. We then have*

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

and, specifying as prior density

$$\begin{aligned} \pi(\lambda|Pa(Ch(\lambda) \cap \boldsymbol{\beta}_c) - \lambda) = 1, \\ g_c(\lambda) = \pi(\lambda|\mathbf{y}_c). \end{aligned} \tag{14}$$

The proof is given in Appendix A in the online supporting information for Ref. [11]. Specializing to the standard setup of Definition 1, where $\mathrm{Ch}(\lambda) \subseteq \boldsymbol{\beta}_c$, we see that the requirement for Eq. (13) to hold is that $\mathbf{Y}_c$ consists of all data descendant nodes of λ. In Ref. [9], this splitting was compared with two other splittings for c_λ^2 and found to be optimal with respect to detection power. This measure was also found to be a well-calibrated measure under this splitting.

3. Noninvariance and reparametrizations

The iic distributions and the corresponding conflict measures are parametrization dependent. Based on experience so far, the conflict measures seem to be fairly robust to changes in parametrization. However, this noninvariance can be handled in a theoretically satisfactory way under certain circumstances.

Let ϕ be the parameter, in a standard parametrization, corresponding to a specific node in the DAG. 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 function $\lambda(\phi)$, 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. Such a parametrization will be considered as a canonical or reference parametrization if it exists, as opposed to the standard parametrization involving ϕ. Accordingly, the conflict measures given in Eqs. (9)–(12) are preferably based on this reference 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 Eq. (13) and the prior for λ satisfies Eq. (14), i.e., equals 1. According to an argument given in Section 1.3 of Ref. [18], such a prior expresses noninformativity for likelihoods of the form (Eq. (15)). Computationally, we may, however, use the standard parametrization. When generating $\phi^*{}_c$ as posterior samples from $\pi(\phi|\mathbf{Y}_c)$, the prior density $|d\lambda/d\phi|$ for ϕ must be used. Then, we may calculate $\lambda_c^* = \lambda(\phi^*{}_c)$. To represent the iic distribution $G_p(\lambda)$, we may calculate $\lambda_p^* = \lambda(\phi^*{}_p)$for samples $\phi^*{}_p$ from $\pi(\phi|\mathbf{y}_p)$ according to the given model. Now, the c_λ^4-measure can be estimated from (Eq. (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 (Eq. (11)) and its one-sided versions (Eq. (10)), we may use the samples from $\pi(\phi|\mathbf{y}_c)$ and $\pi(\phi|\mathbf{y}_p)$ directly. To see this, note that the condition $\lambda_p^* \geq \lambda_c^*$ is equivalent to the condition $\phi^*{}_p \geq \phi^*{}_c$ (assuming that λ is increasing as a function of ϕ). 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. Extensions to deterministic nodes: Relation to cross-validation, prediction and hypothesis testing

4.1. Cross-validation and data node conflict

The model variables $\mathbf{Y}$ are represented by the bottom nodes in the DAG describing the hierarchical model. The framework can be extended to also cover conflict concerning these nodes. In this way, cross-validation can be viewed as a special case of the conflict measure approach.

Let Y_c be an element in the vector $\mathbf{Y}$ of observable random variables. We define the prior iic density $g_p(y_c)$ exactly as in Eq. (6), with λ replaced by y_c. The Dirac measure at the observed value y_c represents a degenerate iic information contribution about Y_c. This leads to the following definitions:

$$c_{y_c}^{3+} = G_p(y_c), \quad c_{y_c}^{3-} = \overline{G}_p(y_c), \tag{16}$$

$$c_{y_c}^{3} = 1 - 2\min(G_p(y_c), \overline{G}_p(y_c)), \tag{17}$$

$$c_{y_c}^{4} = P^{g_p}(g_p(Y_c) \geq g_p(y_c)). \tag{18}$$

The measures (Eqs. (16)–(18)) are called data node conflict measures. To see that these definitions are consistent with Eqs. (10)–(12), note that λ_p^* corresponds to Y_c, and λ_c^* is deterministic and corresponds to y_c. We define $X = Y_c - y_c$, corresponding to δ. We then have $g(x) = g_p(x + y_c)$. Hence,

$$G(0) = \int_{-\infty}^{0} g(x)dx = \int_{-\infty}^{y_c} g_p(y)dy = G_p(y_c),$$

and accordingly, $\overline{G}(0) = \overline{G}_p(y_c)$. It follows that Eqs. (16) and (17) are special cases of Eqs. (10) and (11). Moreover,

$$P^g(g(X) \geq g(0)) = P^{g_p}(g_p(Y_c) \geq g_p(y_c)),$$

showing that Eq. (18) is a special case of Eq. (12).

Furthermore, this correspondence between the data node conflict measures (Eqs. (16) and (17)) and the parameter node conflict measures (Eqs. (10) and (11)) can be used to motivate these latter measures. We will treat the c^{3+} measure as an example. Consider again a parameter node

λ. If λ were actually observable and known to take the value λ_c, the data node version of the c^{3+} measure could be used to measure deviations toward the right tail of G_p as

$$G_p(\lambda_c) = \int_{-\infty}^{\lambda_c} g_p(\lambda)d\lambda = \int_{-\infty}^{0} g_p(\delta + \lambda_c)d\delta.$$

Now λ is in reality not known, but we can take the expectation of this conflict with respect to the distribution G_c, which reflects the uncertainty about λ when influence from data $\mathbf{y}_p$ is removed. The result is the following theorem:

Theorem 2

$$E^{G_c}(G_p(\lambda) = c_\lambda^{3+}.$$

Proof:

$$E^{G_c}(G_p(\lambda) = \int_{-\infty}^{\infty} g_c(\lambda)\left(\int_{-\infty}^{0} g_p(\delta + \lambda)d\delta\right)d\lambda = \int_{-\infty}^{0}\left(\int_{-\infty}^{\infty} g_p(\delta + \lambda)g_c(\lambda)d\lambda\right)d\delta$$

$$= \int_{-\infty}^{0} g(\delta)d\delta = G(0) = c_\lambda^{3+}$$

by Eq. (10).

4.2. Cross-validation and sufficient statistics

Suppose the node λ of interest is the parent of the subvector $\mathbf{Y}_c$ of $\mathbf{Y}$. Suppose also that Y_c is a sufficient statistic for $\mathbf{Y}_c$. Evidently then, the measures c_λ^{3+} and $c_{Y_c}^{3+}$ address the same kind of possible conflict in the model. The following theorem, proved in Ref. [11], states that the two measures agree under certain conditions. This is a generalization of a result in Ref. [13], which also unnecessarily assumed symmetry for the conditional density of Y_c.

Theorem 3 *Suppose the conditional density for the scalar variable Y_c given the parameter λ is of the form $f_{Y_c}(y|\lambda) = f_{c,0}^2(y - \lambda)$. Then,*

$$c_{Y_c}^{3+} = c_\lambda^{3+}.$$

When a sufficient statistic exists, the cross-validatory p-value is considered by Ref. [13] as the gold standard, and the aim of their construction is to provide a measure which is generally applicable and matches cross-validation when a sufficient statistic exists.

4.3. Prediction

As mentioned in Section 2, the c^4 measure can be used to assess conflict concerning vectors of nodes. Applying this at the data node level, we may assess the quality of predictions of a subvector $\mathbf{Y}_c$ of $\mathbf{Y}$ based on a complementary subvector y_p of observations. The relevant

measure is given by Eq. (18), with Y_c replaced by the vector $\mathbf{Y}_c$. This is particularly well suited to models where data accumulate as time evolves. Such a conflict measure can be used to assess the overall quality of the model. It can also be used as a tool for model comparison and model choice.

4.4. Hypothesis testing

Suppose the top-level nodes $\boldsymbol{\mu}$ appearing in Eq. (2) are assumed fixed and known according to the model, so that $\pi(\boldsymbol{\mu})$ is a Dirac measure at these fixed values of the components of $\boldsymbol{\mu}$. Hence, the DAG has deterministic nodes both at the top and at the bottom, namely the vectors $\boldsymbol{\mu}$ and $\mathbf{y}$, respectively. We may then check for a conflict concerning a component λ of $\boldsymbol{\mu}$ by introducing a random version $\tilde{\lambda}$ of λ and contrast the corresponding $g_c(\tilde{\lambda})$ with the fixed value λ. The random $\tilde{\lambda}$ has the same children and coparents as λ, and the vector $\boldsymbol{\beta}_c$, the information contribution $f_c(\tilde{\lambda}; \beta_c)$ and the iic density g_c are defined as in Eqs. (4), (5) and (6). The respective conflict measures are defined as in Eqs. (16)–(18) with y_c replaced by λ and G_p and g_p replaced by G_c and g_c. If the model is rejected when the conflict exceeds a certain predefined warning level, this corresponds to a formal Bayesian test of the hypothesis $\tilde{\lambda} = \lambda$. Using the conflict measure (Eq. (18)), we may put the whole vector $\boldsymbol{\mu}$ to test in this way.

5. Preexperimental uniformity of the conflict measures

In this section, we review some results concerning the distribution of the conflict measures. If c is one of the measures (Eqs. (10), (11), (12), (16), (17) or (18)), then preexperimentally, i.e., prior to observing the data $\mathbf{y}$, c is a random variable taking a value in [0, 1]. A large value of c indicates a possible conflict in the model, and uniformity of c corresponds to 1 – c being a proper p-value. This does not mean that we propose a formal hypothesis testing procedure for model criticism, possibly even adjusted for multiple testing, nor that we think that a fixed significance level represents an appropriate criterion signaling the need for changing the model. A relatively large value of c may be accepted if there are convincing arguments for believing in a particular modeling aspect, while a less extreme value of c may indicate a need for adjustments in modeling aspects that are considered questionable for other reasons. But the terms "relatively large" and "less extreme" must refer to a meaningful common scale. In our view, uniformity of the conflict measure under all sources of uncertainty is the natural ideal criterion for being a well-calibrated conflict measure, the fulfillment of which ensures comparable assessment of the level of conflict across models. This means that we aim for preexperimental uniformity in cases where the prior distribution is highly noninformative, and also, as discussed in the following subsection, in cases where an informative prior represents part of the randomness in the data-generating process (aleatory uncertainty) rather than subjective (epistemic) uncertainty about the location of a fixed but unknown λ. In this chapter, we limit attention to situations where exact uniformity is achieved. The pivotality condition (Eq. (15)) turns out to be a key assumption needed to obtain such exact results. Refs. [10] and [12] provide some examples where exact uniformity is achieved in other cases.

5.1. Data-prior conflict

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 aleatory rather than epistemic uncertainty. 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 conflict between the prior and likelihood information contributions, we consider this a data-prior conflict. The following theorem, proved in Ref. [11], deals with this kind of conflict. Note that in this situation, the $\mathbf{Y}_p$ part of the data splitting is empty.

Theorem 4 *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 preexperimentally uniformly distributed for both the c_λ^3- and c_λ^4-measures.*

The theorem obviously applies to the location parameter of normal and *t*-distributions with fixed variance parameters, as well as the location parameter in the skew normal distribution [19]. If the vector $\mathbf{Y}$ consists of IID normal variables, the theorem also applies to the location parameter, using as scalar variable the sufficient statistic $\overline{\mathbf{Y}}$. If the n components of $\mathbf{Y}$ are IID exponentially distributed with failure rate λ, their sum is a sufficient statistic that is gamma distributed with shape parameter n and scale parameter λ. We may then use the fact that for a variable Y which is gamma distributed with known shape parameter and unknown scale parameter λ, the quantity $\log(Y) - \log(\lambda)$ is a pivotal statistic, and uniformity is obtained by combining Theorem 4 with the approach of Section 3. In the standard parametrization, the appropriate prior distribution is $\pi(\lambda) = 1/\lambda$). Details are given in Ref. [11], which also deals with the gamma, inverse gamma, Weibull and lognormal distributions in a similar way.

5.2. 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 Eqs. (7) and (8), we contrast the information in $f_c(\lambda; \mathbf{Y}_c)$ with that in $f_p(\lambda; \mathbf{Y}_p)$. We use the term data-data conflict in this context, since there is no prior information incorporated in f_p, and the two information contributions play symmetric roles. However, as a particular application, one may think of $\mathbf{Y}_c$ as a scalar variable representing a possible outlier.

The following theorem is proved in Ref. [11].

Theorem 5 *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 preexperimentally uniformly distributed for both the c^3_λ- and c^4_λ-measures.

Theorem 5 can be applied if the components of $\mathbf{Y}_c$ and $\mathbf{Y}_p$ are normally or lognormally distributed with known variance parameter, exponentially distributed, or gamma, inverse gamma or Weibull with known shape parameter, since pivotal quantities based on sufficient statistics exist for these distributions.

5.3. Normal hierarchical models with fixed covariance matrices

Allowing for each y and v appearing in Eq. (2) to be interpreted as vectors of nodes, we now assume that each conditional distribution in the decomposition (Eq. (2)) is multinormal with fixed and known covariance matrices. The random-effects model (Eq. (1)) is a simple example of this. We also assume that the top-level parameter vector $\boldsymbol{\mu}$ has the improper prior 1 and that each linear mapping $\mathrm{Pa}(v) \to E(v|\mathrm{Pa}(v))$ has full rank.

Now let λ be any node in the model description. It is standard to verify that, regardless of how the vector of neighboring and coparent nodes $\boldsymbol{\beta}$ is decomposed into $\boldsymbol{\beta}_p$, containing $\mathrm{Pa}(\lambda)$, and $\boldsymbol{\beta}_c$, the densities $f_p(\lambda; \beta_p)$ and $f_c(\lambda; \beta_c)$ of Eqs. (5) and (8) are multinormal with fixed covariance matrices. Furthermore, this is true also for the iic densities g_p and g_c of Eq. (6), regardless of the data splitting. It follows that the density g of the difference δ between independent samples from g_p and g_c is multinormal with expectation $E^G(\delta) = E^{G_p}(\lambda) - E^{G_c}(\lambda)$ and covariance matrix $\mathrm{cov}^G(\delta) = \mathrm{cov}^{G_p}(\lambda) + E^{G_c}(\lambda)$. It follows that $\left(\delta - E^G(\delta)\right)^t \mathrm{cov}^G(\delta)^{-1} \left(\delta - E^G(\delta)\right)$ is χ^2-distributed with $n = \dim(\lambda)$ degrees of freedom, and the probability under G that $g(\delta) < g(0)$ is easily seen to be $\Psi_n\left(E^G(\delta)^t \mathrm{cov}^G(\delta)^{-1} E^G(\delta)\right)$, where Ψ_n is the cumulative distribution function for the χ^2_n-distribution. The preexperimental uniformity of this quantity is proved in Ref. [10].

Theorem 6 *Consider a hierarchical normal model as described above.*

i. *Let λ be an arbitrary scalar or vector parameter node. If the data splitting satisfies Eq. (13), then c^4_λ is uniformly distributed preexperimentally.*

ii. *Suppose the data splitting $(\mathbf{Y}_p, \mathbf{Y}_c)$ satisfies $Ch\left(Pa(\mathbf{Y}_c)\right) = \mathbf{Y}_c$. Then, $c^4_{\mathbf{Y}_c}$ is uniformly distributed preexperimentally.*

If λ in (i) or Y_c in (ii) are one dimensional, then G is symmetric and unimodal, and therefore, the respective c^3-measures are defined and coincide with the c^4-measures. Gåsemyr et al. [10] also show that in that case the c^{3+}- and c^{3-}-measures are uniformly distributed preexperimentally.

Example 2. Consider the following DAG model, a regression model with randomly varying regression coefficients.

$$Y_{i,j} \sim N(\mathbf{X}^t_{i,j}\xi_i, \sigma^2),\ \xi_i \sim N(\xi, \Omega),\ j = 1, \ldots, n,\ i = 1, \ldots, m,\ \pi(\xi) \propto 1. \tag{19}$$

The m units could be groups of individuals, with $y_{i,j}$ the measurement for a group member with individual covariate vector $\mathbf{X}_{i,j}$, or individuals with the successive $y_{i,j}$ representing

repeated measurements over time. In this model, we could check for a possible exceptional behavior of the mth unit by means of the conflict measure $c^4_{\xi_m}$. With a data splitting for which $\mathbf{Y}_c = \mathbf{Y}_m = (Y_{m,1}, \ldots, Y_{m,n})$ the conditions for Theorem 6, part (i), are satisfied if $\dim(\xi) \leq n$, and the measure is preexperimentally uniformly distributed.

6. Concluding remarks

The assumption of fixed covariance matrices in the previous subsection is admittedly quite restrictive. In general, the presence of unknown nuisance parameters, such as parameters describing the covariance matrices in a normal model, makes the derivation of exact uniformity at least difficult and often impossible. Promising approximate results are reported in Ref. [9] for the closely related c^2_λ measure. Further empirical studies are needed in order to examine to what extent the conflict measures are approximately uniformly distributed in other situations. As an informal tool to be used in conjunction with subject matter insight, the conflict measure approach does not require exact uniformity in order to be useful.

Author details

Jørund I. Gåsemyr* and Bent Natvig

*Address all correspondence to: gaasemyr@math.uio.no

University of Oslo, Norway

References

[1] Box GEP. Sampling and Bayes' inference in scientific modelling and robustness (with discussion and rejoinder). Journal of the Royal Statistical Society. Series A. 1980;**143**:383-430

[2] Bayarri MJ, Castellanos ME. Bayesian checking of the second levels of hierarchical models. Statistical Science. 2007;**22**:322-343

[3] Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies (with discussion and rejoinder). Statistica Sinica. 1996;**6**:733-807

[4] Bayarri MJ, Berger JO. P values in composite null models (with discussion). The Journal of the American Statistical Association. 2000;**95**:1127-1142

[5] Hjort NL, Dahl FA, Steinbakk GH. Post-processing posterior predictive p-values. The Journal of the American Statistical Association. 2006;**101**:1157-1174

[6] Dahl FA. On the conservativeness of posterior predictive p-values. Statistics and Probability Letters. 2006;**76**:1170-1174

[7] Dey D, Gelfand A, Swartz T, Vlachos P. A simulation-intensive approach for checking hierarchical models. Test. 1998;**7**:325-346

[8] O'Hagan A. HSSS model criticism (with discussion). In: Green PJ, Hjort NL, Richardson S, editors. Highly Structured Stochastic Systems. Oxford: Oxford University Press; 2003. pp. 423-444

[9] Dahl FA, Gåsemyr J, Natvig B. A robust conflict measure of inconsistencies in Bayesian hierarchical models. Scandinavian Journal of Statistics. 2007;**34**:816-828

[10] Gåsemyr J, Natvig B. Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models. Scandinavian Journal of Statistics. 2009;**36**:822-838

[11] Gåsemyr J. Uniformity of node level conflict measures in Bayesian hierarchical models based on directed acyclic graphs. Scandinavian Journal of Statistics. 2016;**43**:20-34

[12] Gåsemyr J. Alternatives to post-processing posterior predictive *p*-values. Submitted 2017

[13] Marshall EC, Spiegelhalter DJ. Identifying outliers in Bayesian hierarchical models. A simulation based approach. Bayesian Analysis. 2007;**2**:409-444

[14] Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Statistics in Medicine. 2010;**29**:932-944

[15] Presanis AM, Ohlssen D, Spiegelhalter D, De Angelis D. Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. Statistical Science. 2013;**28**:376-397

[16] Lauritzen SL. Graphical Models. Oxford: Oxford University Press; 1996

[17] Scheel I, Green P, Rougier JC. A graphical diagnostic to identifying influential model choices in Bayesian hierarchical models. Scandinavian Journal of Statistics.2011;**38**:529-550

[18] Box GEP, Tiao GC. Bayesian Inference in Statistical Analysis. New York: Wiley; 1992

[19] Azzalini A. A class of distributions which include the normal ones. Scandinavian Journal of Statistics. 1985;**12**:171-178

IntechOpen