Dept. of Math. University of Oslo
Statistical Research Report No 6
ISSN 0806–3842 July 2008

Extensions of a Conflict Measure of Inconsistencies in Bayesian Hierarchical Models

JØRUND GÅSEMYR and BENT NATVIG

July 3, 2008

ABSTRACT. In Dahl et al. (2007) we extended and refined some tools given in O'Hagan (2003) for criticism of Bayesian hierarchical models. Especially, avoiding double use of data by a data splitting approach was a main concern. Such tools can be applied at each node of the model, with a view to diagnosing problems of model fit at any point in the model structure. As in O'Hagan (2003) a Gaussian model of one-way analysis of variance was investigated. Through extensive MCMC simulations it was shown that our method detects model misspecification about as well as the one of O'Hagan, when this is properly calibrated, while retaining the desired false warning probability for data generated from the assumed model. In the present paper we suggest some new measures of conflict based on tail probabilities of the so-called integrated posterior distributions introduced in Dahl et al. (2007). These new measures are equivalent to the measure applied in the latter paper in simple Gaussian models, but seem more appropriately adjusted to deviations from normality and to conflicts not concerning location parameters. A general linear normal model with known covariance matrices is considered in detail.

Key words: double use of data, general linear model, integrated posterior distributions, Markov chain Monte Carlo simulations, model evaluation, tail probabilities

1. Introduction

O'Hagan (2003) introduces some tools for criticism of Bayesian hierarchical models. Such tools can be applied at each node of the model through analysis of what is called information contributions. The aim is to diagnose problems of model fit at any point in the model structure. His method relies on computing the posterior median of a conflict index, typically through MCMC simulations. In Dahl et al. (2007) we extended and refined the method of O'Hagan (2003), especially avoiding the double use of data by a data splitting approach. As in the latter paper a Gaussian model of one-way analysis of variance was investigated, and it was shown that O'Hagan's

Department of Mathematics, P.O. Box 1053 Blindern, N–0316 Oslo, Norway

approach gives unreliable false warning probabilities. Through extensive numerical experiments, accompanied by theoretical justifications from a non trivial special case, we showed that our method detects model misspecification about as well as the one of O'Hagan, when this is properly calibrated, while retaining the desired false warning probability for data generated from the assumed model. This also holds for Student-t and uniform distribution versions of the model.

In the present paper we suggest some new measures of conflict based on tail probabilities of the so-called integrated posterior distributions (ipd) introduced in Dahl et al. (2007). These new measures are equivalent to the measure applied in the latter paper in simple Gaussian models, but seem more appropriately adjusted to deviations from normality and to conflicts not concerning location parameters. This more simple case was treated in Dahl et al. (2007). In the present paper we also extend our notion of conflict to cover data nodes in addition to the parameter nodes considered in Dahl et al. (2007). This establishes a close link to the cross-validatory p-value discussed in Marshall & Spiegelhalter (2007). It is also shown that there is a close link between these new measures and the partial posterior predictive p-value introduced in Bayarri & Berger (2000). The latter is designed to avoid the double use of data by eliminating the influence of a chosen test statistic on the posterior distribution. For nodes that are parents to data nodes, our suggestions are closely related to the conflict measure introduced in Marshall & Spiegelhalter (2007). Our approach may hence serve as a unifying framework for these measures. A review of several Bayesian p-values along with related work is given in Dahl et al. (2007). In Bayarri & Castellanos (2007) an extensive numerical comparison of such p-values is given in a simple hierarchical model. As for these measures our aim for the new measures is that they are pre-experimentally close to be uniformly (0,1) distributed.

The paper is laid out as follows. In section 2 the new measures of conflict based on tail probabilities are introduced for data node conflicts. The link between these measures for data nodes and the partial posterior predictive p-value of Bayarri & Berger (2000) is discussed in section 3. In section 4 the new measures of conflict for parameter node conflicts are presented and the link to the conflict measure of Marshall & Spiegelhalter (2007) is discussed. Such conflicts are not considered by Bayarri & Berger (2000). A general linear normal model with known covariance matrices is considered in section 5 leaving the proofs of the theoretical results to an appendix. Some concluding remarks are given in section 6.

2. New measures of conflict for data nodes

The main purpose of this section is to introduce the new measures of conflict based on tail probabilities. It is most easy to motivate and discuss alternative variants of these measures in the context of data node conflicts. Such conflicts are not treated in Dahl et al. (2007).

In general a Bayesian hierarchical model can be supposed to be expressed as a directed acyclic graph. We define a child (c) node of a specific node as a node that can be reached by a directed edge from the specific node, including by definition the specific node itself. A parent (p) node of a specific node is a node that can reach the specific node by a directed edge.

As a start consider a splitting $(\boldsymbol{y}^p, y^c)$ of the data vector $\boldsymbol{y}$ with y^c scalar. We will consider a conflict at the node y^c. We assume $\boldsymbol{Y}^p$ and Y^c are independent given the vector of parent nodes $\boldsymbol{\beta}$ of y^c. The information contribution about y^c from $\boldsymbol{\beta}$ is the density for Y^c given $\boldsymbol{\beta}$. The corresponding integrated posterior distribution density g_p is defined by integrating the information contribution about y^c from $\boldsymbol{\beta}$ over the posterior distribution for $\boldsymbol{\beta}$ given $\boldsymbol{y}^p$. This is in the spirit of the predictive densities treated in Gelfand & Dey (1994). Hence,

$$g_p(y^c) = \int f_{Y^c}(y^c|\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\boldsymbol{y}^p)d\boldsymbol{\beta} = \pi_{Y^c}(y^c|\boldsymbol{y}^p), \tag{1}$$

the last equality following from the independence assumption. Denote the corresponding cumulative distribution function by G_p. Motivated by (6) of Dahl et al. (2007) a corresponding measure of conflict at the node y^c is given by

$$c_{y^c}^{2,\boldsymbol{y}^p,y_c} = \frac{(E^{g_p}(Y^c) - y^c)^2}{\mathrm{var}^{g_p}(Y^c)} = \frac{(E(Y^c|\boldsymbol{y}^p) - y^c)^2}{\mathrm{var}(Y^c|\boldsymbol{y}^p)}\,, \tag{2}$$

having applied (1).

Now the main idea is to measure the conflict at y_c in terms of the tail probabilities of G_p, rather than in terms of the mean and variance of this distribution as in (2). We will measure the conflict on a scale ranging from 0 to 1 in terms of 1 minus the tail probabilities of G_p, corresponding to the observed y^c. Hence, a value close to 1 of such a measure indicates an inconsistency in the model. Let $\bar{G}_p(\cdot) = 1 - G_p(\cdot)$. Our first proposal for a new conflict measure is

$$c_{y^c}^{3,\boldsymbol{y}^p,y^c} = 1 - 2\min(G_p(y^c), \bar{G}_p(y^c)) \tag{3}$$

To see the connection between (2) and (3), note that by (2) $c_{y^c}^{2,\boldsymbol{y}^p,y^c} = c$ iff $y^c = y_0^c(c) = E(Y^c|\boldsymbol{y}^p) - (c\,\mathrm{var}(Y^c|\boldsymbol{y}^p))^{1/2}$ or $y^c = y_1^c(c) = E(Y^c|\boldsymbol{y}^p) + (c\,\mathrm{var}(Y^c|\boldsymbol{y}^p))^{1/2}$. By letting $c_{y^c}^2$ be an abbreviation for $c_{y^c}^{2,\boldsymbol{y}^p,y^c}$, we consider the tail area $(-\infty, y_0^c(c_{y^c}^2)] \cup [y_1^c(c_{y^c}^2), \infty)$ corresponding to a measure of conflict at the node y^c at least equal to $c_{y^c}^2$. This leads by (2) to the conflict measure

$$\begin{aligned} P^{g_p}(c_{Y^c}^2 \leq c_{y^c}^2) &= G_p(y_1^c(c_{y^c}^2)) - G_p(y_0^c(c_{y^c}^2)) \\ &= G_p(E(Y^c|\boldsymbol{y}^p) + |E(Y^c|\boldsymbol{y}^p) - y^c|) - G_p(E(Y^c|\boldsymbol{y}^p) - |E(Y^c|\boldsymbol{y}^p) - y^c|) \end{aligned} \tag{4}$$

When g_p is symmetric around $E(Y^c|\boldsymbol{y}^p)$, we have $G_p(y_0^c(c_{y^c}^2)) = \bar{G}_p(y_1^c(c_{y^c}^2))$, and it follows that (4) equals the measure (3).

If in addition g_p is unimodal, (3) is equal to our second proposal for a new conflict measure

$$c_{y^c}^{4,\boldsymbol{y}^p,y^c} = P^{g_p}(g_p(Y^c) \geq g_p(y^c)) \tag{5}$$

Now replace the fixed y^c on the right hand side of (2) by a random Y^c following the conditional distribution given by (1). If for instance G_p is normal, then $c_{Y^c}^{2,\boldsymbol{y}^p,y^c}$ is χ_1^2 distributed. Hence, due to the symmetry and unimodality of the normal density, we get from (4)

$$c_{y^c}^{3,\boldsymbol{y}^p,y^c} = c_{y^c}^{4,\boldsymbol{y}^p,y^c} = P^{g_p}(c_{Y^c}^2 \leq c_{y^c}^2) = \psi_1(c_{y^c}^{2,\boldsymbol{y}^p,y^c}),$$

where ψ_1 is the cumulative distribution function of the χ_1^2-distribution. Accordingly, $\psi_1(c_{Y^c}^{2,\boldsymbol{y}^p,y^c})$ is uniform on (0,1) under G_p. In particular the 3.85 level of $c_{Y^c}^{2,\boldsymbol{y}^p,y^c}$ corresponds to 0.95 for the new measures.

Although the new measures given by (3) and (5) equal the tail probability version of $c_{y^c}^{2,\boldsymbol{y}^p,y^c}$ given by (4) under special restrictions on G_p, our idea is to use these new measures without any restrictions on G_p. The $c_{y^c}^{4,\boldsymbol{y}^p,y^c}$ measure has the disadvantage compared to $c_{y^c}^{3,\boldsymbol{y}^p,y^c}$ that it is not invariant under non-linear transformations of the data. This is not considered a serious problem since there often is a natural scale for the data. One major advantage of the former measure is that it is particularly well suited to handle multimodality. Another major advantage is that it is readily extended to measure conflicts about vector nodes. Now g_p is a multidimensional distribution for $\boldsymbol{Y}^c$ defined parallel to (1). We will return to vector node conflicts in sections 4 and 5.

Returning to the scalar case, there may in some situations be of special interest to consider deviations of y^c from $E(Y^c|\boldsymbol{y}^p)$ in one of the two possible directions. Hence, it is natural to introduce

$$c_{y^c}^{3+,\boldsymbol{y}^p,y^c} = G_p(y^c), \qquad c_{y^c}^{3-,\boldsymbol{y}^p,y^c} = \bar{G}_p(y^c) \tag{6}$$

The first of these corresponds to the cross-validatory p-value discussed in Marshall & Spiegelhalter (2007), focusing on deviations towards the right tail of the cross-validatory predictive distribution for y^c, which coincides with G_p. Intuitively, all new measures should be well suited to handle skewness, an aspect of conflict analysis that may be problematic using the $c_{y^c}^{2,\boldsymbol{y}^p,y^c}$ measure. Another aspect that could be problematic with this measure, is the way uncertainty in spread parameters is treated in models based on separate parameters for location and spread. In the following simple normal example covering such a case, we show that our new measures are pre-experimentally uniformly (0,1) distributed as desired. Normal models that are much more complex, but with fixed variances and covariances are dealt with in section 5.

Example 1. Let $Y_1, \ldots, Y_k$ be independent $N(\mu, \tau^2)$, and let $\pi(\mu) = 1$, $\pi(\tau^2) = 1/\tau^2$ be improper, non-informative priors for the unknown parameters μ and τ^2. Introduce the data splitting $\boldsymbol{y}^p = (y_1, \ldots, y_{k-1}), y^c = y_k$, and consider the conflict measure $c_{y^c}^{3+,\boldsymbol{y}^p,y^c}$. Let ϕ be the standard normal density and $\gamma_{\text{inv}}(x; a, b) = (b^a/\Gamma(a))(1/x)^{a+1} \exp(-b/x)$ the inverse gamma density. Then

$$\begin{aligned}&\pi(\mu, \tau^2|\boldsymbol{y}^p) = \pi(\mu|\tau^2, \boldsymbol{y}^p)\pi(\tau^2|\boldsymbol{y}^p)\\ &= (\tau^2/(k-1))^{-1/2}\phi((\mu - \bar{y}^p)/(\tau^2/(k-1))^{1/2})\gamma_{\text{inv}}(\tau^2; (k-2)/2, s^2/2),\end{aligned}$$

where $s^2 = \sum_{i=1}^{k-1}(y_i - \bar{y}^p)^2$. By (1) it follows that

$$\begin{aligned}&g_p(y^c) = \int \tau^{-1}\phi((y^c - \mu)/\tau)\pi(\mu, \tau^2|\boldsymbol{y}^p)d\mu d\tau^2\\ &= \int (k\tau^2/(k-1))^{-1/2}\phi((y^c - \bar{y}^p)/(k\tau^2/(k-1))^{1/2})\gamma_{\text{inv}}(\tau^2; (k-2)/2, s^2/2)d\tau^2\\ &= (ks^2/((k-1)(k-2)))^{-1/2}h_{k-2}((y^c - \bar{y}^p)/(ks^2/((k-1)(k-2)))^{1/2}),\end{aligned}$$

where h_{k-2} denotes the t-distribution with $k-2$ degrees of freedom. If the variable $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ is distributed according to the assumed model, we therefore have

$$G_p(Y^c) = H_{k-2}(((k-1)/k)^{1/2}(Y^c - \bar{Y}^p)/(S^2/(k-2))^{1/2}),$$

where $S^2 = \sum_{i=1}^{k-1}(Y_i - \bar{Y}^p)^2$ and H_{k-2} is the cumulative distribution function corresponding to h_{k-2}. In the argument for H_{k-2}, the numerator, if scaled by the true standard deviation $\text{var}(Y_i)^{1/2} = \tau_0$, is standard normal and independent of the denominator, which similarly scaled is the square root of $1/(k-2)$ times a χ^2_{k-2} variable. Hence, this argument is t-distributed with $k-2$ degrees of freedom, and consequently $G_p(Y^c)$ is uniform on $[0, 1]$. Accordingly, by (6) $c_{y^c}^{3+,\boldsymbol{y}^p,y^c}$ is pre-experimentally uniform. By symmetry, this also applies to $c_{y^c}^{3-,\boldsymbol{y}^p,y^c}$. It also follows that $1 - 2\min(G_p(Y^c), \bar{G}_p(Y^c))$ is uniform on $[0, 1]$. Hence, by (4) $c_{y^c}^{3,\boldsymbol{y}^p,y^c}$ is pre-experimentally uniform. Since the c^3 and c^4 measures coincide in this case, due to symmetry and unimodality of g_p, this applies also to $c_{y^c}^{4,\boldsymbol{y}^p,y^c}$.

3. The link to the partial posterior predictive p-value

Bayarri & Berger (2000) introduces the partial posterior predictive p-value and the conditional predictive p-value. These Bayesian p-values are identical in most examples given in their paper. Furthermore, these values are based on a test statistic, T, which typically is a function of the entire data vector $\boldsymbol{Y}$. However, the definitions can equally well be applied to a scalar Y^c for which a conflict is considered, as discussed in the previous section. In this section we will demonstrate the link between the new measures presented in the previous section and the partial posterior predictive p-value denote by p_{ppost}.

Consider a splitting $(\boldsymbol{y}^p, y^c)$ of the data vector $\boldsymbol{y}$ with y^c scalar and let $\boldsymbol{\theta}$ be a parameter vector. Bayarri & Berger (2000) defines the partial posterior distribution for $\boldsymbol{\theta}$

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}\backslash y^c) \propto \frac{f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f_{Y^c}(y^c|\boldsymbol{\theta})} \tag{7}$$

The partial posterior predictive distribution for Y^c is then given by

$$m(y^c|\boldsymbol{y}\backslash y^c) = \int f_{Y^c}(y^c|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}\backslash y^c)d\boldsymbol{\theta} \tag{8}$$

Finally, p_{ppost} is given by

$$p_{\text{ppost}} = P^{m(y^c|\boldsymbol{y}\backslash y^c)}(Y^c \geq y^c) \tag{9}$$

We now have the following theorem

Theorem 1.
Let $(\boldsymbol{y}^p, y^c)$ be a data splitting with y^c scalar and such that $\boldsymbol{Y}^p$ and Y^c are independent given the parameter vector $\boldsymbol{\theta}$. Then

i) $p_{\text{ppost}} = 1 - c_{y^c}^{3+,\boldsymbol{y}^p,y^c}$

ii) *If g_p is nonincreasing,*
$p_{\text{ppost}} = 1 - c_{y^c}^{4,\boldsymbol{y}^p,y^c}$

iii) *If g_p is symmetric around $E^{g_p}(Y^c)$ and if $T = |Y^c - E^{g_p}(Y^c)|$ is the test statistic, then*
$p_{\text{ppost}} = 1 - c_{y^c}^{3,\boldsymbol{y}^p,y^c}$

If in addition g_p is unimodal,

$$p_{\text{ppost}} = 1 - c_{y^c}^{4,\boldsymbol{y}^p,y^c}$$

Proof. Since $\boldsymbol{Y}^p$ and Y^c are independent given $\boldsymbol{\theta}$, the right hand side of (7) is proportional to $\pi(\boldsymbol{\theta}|\boldsymbol{y}^p)$. Hence, from (1), (8), (9) and (6)

$$p_{\text{ppost}} = P^{g_p}(Y^c \geq y^c) = \bar{G}_p(y^c) = 1 - c^{3+,\boldsymbol{y}^p,y^c}$$

The proof of ii) is very parallel to the one above using (5) instead of (6). For iii)

$$\begin{aligned} p_{\text{ppost}} &= P^{g_p}[|Y^c - E^{g_p}(Y^c)| \geq |y^c - E^{g_p}(Y^c)|] \\ &= 1 - P^{g_p}[(Y^c - E^{g_p}(Y^c))^2 \leq (y^c - E^{g_p}(Y^c))^2] \end{aligned}$$

iii) now follows by the argument leading to (3) and (5) just replacing $\text{var}^{g_p}(Y^c)$ in (2) by 1.

Note that the assumption that $\boldsymbol{Y}^p$ and Y^c are independent given $\boldsymbol{\theta}$ provides the link between the new measures and p_{ppost}. This assumption is not made in Bayarri & Berger (2000). However, they have only demonstrated the nice uniformity property of their p-values in examples where this assumption holds. Actually, in our opinion the interpretation of (7) and hence p_{ppost} in (9) is somewhat obscure without this assumption.

Moreover, in all the examples in Bayarri & Berger (2000) demonstrating the nice uniformity property of their p-values, one may transform the data in their paper. This can be done in such a way that the transformed data vector $\boldsymbol{Y}$ can be split into $(\boldsymbol{Y}^p, Y^c)$, where $\boldsymbol{Y}^p$ and Y^c are independent given $\boldsymbol{\theta}$ and Y^c is their test statistic. A conflict analysis as described in the previous section can then be based on this splitting.

The following example shows how a data transformation can be used to make the tools of our data node conflict analysis available as an alternative to the posterior predictive p-value analysis.

Example 2. This is Example 1 in Bayarri & Berger (2000). Let $X_1, \ldots, X_n$ be independent $N(0, \sigma^2)$, where σ^2 has the improper prior $\pi(\sigma^2) = 1/\sigma^2$. Let $T = t(X) = |\bar{X}|$. Bayarri & Berger (2000) shows that p_{ppost} is pre-experimentally uniformly distributed. In order to analyse this in terms of data node conflict, we define $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ as an orthogonal transformation of $\boldsymbol{X}$ for which $Y_n = n^{1/2}\bar{X}$. With the data splitting $\boldsymbol{y}^p = (y_1, \ldots, y_{n-1}), y^c = y_n$, we have that $\boldsymbol{Y}^p$ and Y^c are independent given σ^2. Moreover, $f(y^c|\sigma^2)$ is symmetric around 0 for every σ^2. Hence, g_p is also symmetric around 0. It follows from iii) of theorem 1 and the pre-experimental uniformity of the partial posterior predictive p-value that $c_{y^c}^{3,\boldsymbol{y}^p,y^c}$ is pre-experimentally uniform. The same applies to $c_{y^c}^{4,\boldsymbol{y}^p,y^c}$ due to unimodality.

4. New measures of conflict for parameter nodes

Let as a start λ be any scalar parameter of interest given by an interior node of the network. Let $\boldsymbol{\gamma}$ be the vector of neighbouring nodes of λ, possibly containing data nodes. Let $(\boldsymbol{\gamma}^p, \boldsymbol{\gamma}^c)$ be a decomposition of $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}^c$ contains some or all of the child nodes of λ and $\boldsymbol{\gamma}^p$ contains all the parent nodes of λ as well as child nodes not present in $\boldsymbol{\gamma}^c$. Let $\boldsymbol{\beta}^c$ be the vector consisting of $\boldsymbol{\gamma}^c$ as well as nodes that are coparents with λ for the child nodes in $\boldsymbol{\gamma}^c$. Let $\boldsymbol{\beta}^p$ be the vector consisting of $\boldsymbol{\gamma}^p$ and the coparents with λ for child nodes in $\boldsymbol{\gamma}^p$. Such coparents could e.g. be a variance parameter, if λ is a location parameter for the components of $\boldsymbol{\gamma}$ that are child nodes of λ, or other regression parameters, if λ is a coefficient in a vector of regression parameters. Suppose it is of interest to contrast information contributions about λ from $\boldsymbol{\beta}^p$ and $\boldsymbol{\beta}^c$. This set up is more general than the one described in Dahl et al. (2007) since we now do not necessarily assume that $\boldsymbol{\beta}^p$ consists only of the parent nodes of λ.

The information contribution about λ from $\boldsymbol{\beta}^p(\boldsymbol{\beta}^c)$ is the density proportional to all the likelihood factors expressing components of $\boldsymbol{\beta}^p(\boldsymbol{\beta}^c)$ as function of λ, and is denoted $f(\lambda; \boldsymbol{\beta}^p)(f(\lambda; \boldsymbol{\beta}^c))$. Note that O'Hagan (2003) based on geometric intuition normalises these functions by scaling them to have equal height 1. Our idea of instead normalising them to densities is also used in Scheel et al. (2008) as a basis for a graphical technique for diagnosing conflicts in hierarchical models. Define a data splitting $(\boldsymbol{y}^p, \boldsymbol{y}^c)$. Parallel to (1) the following integrated posterior distribution densities are defined

$$g_p(\lambda) = \int f(\lambda; \boldsymbol{\beta}^p)\pi(\boldsymbol{\beta}^p|\boldsymbol{y}^p)d\boldsymbol{\beta}^p, \qquad g_c(\lambda) = \int f(\lambda; \boldsymbol{\beta}^c)\pi(\boldsymbol{\beta}^c|\boldsymbol{y}^c)d\boldsymbol{\beta}^c \quad (10)$$

Denote the corresponding cumulative distribution functions by G_p and G_c. Let $(\lambda_p^*, \lambda_c^*)$ be a pair of independent samples from G_p and G_c respectively. Define $\delta = \lambda_p^* - \lambda_c^*$ and let G and g be the cumulative distribution function and density of δ. Parallel to (6), (4) and (5), replacing y^c by 0, we suggest the following conflict measures

$$c_\lambda^{3+,\boldsymbol{y}^p,\boldsymbol{y}^c} = G(0) \qquad c_\lambda^{3-,\boldsymbol{y}^p,\boldsymbol{y}^c} = \bar{G}(0) \quad (11)$$

$$c_\lambda^{3,\boldsymbol{y}^p,\boldsymbol{y}^c} = 1 - 2\min(G(0), \bar{G}(0)) \quad (12)$$

$$c_\lambda^{4,\boldsymbol{y}^p,\boldsymbol{y}^c} = P^g(g(\delta) \geq g(0)) \quad (13)$$

We will now show that these measures coincide with (6), (3) and (5) respectively in the data node case considered in section 2. In this case λ_p^* corresponds to Y^c, 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 $\bar{G}(0) = \bar{G}_p(y^c)$. It follows that (6) and (3) are special cases of (11) and (12). Moreover,

$$P^g(g(X) \geq g(0)) = P^{g_p}(g_p(Y^c) \geq g_p(y^c)),$$

showing that (5) is a special case of (13).

Furthermore, this correspondance between the data node conflict measures of section 2 and the parameter node conflict measures of the present section 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 towards 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 since λ is in reality not known, we take the expectation of this conflict with respect to the distribution G_c, which reflects the uncertainty about λ when influence from data $\mathbf{y}^p$ are removed. Hence, we are lead to consider

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

If for instance G_p and G_c are normal, then also G is normal and it follows that $(\delta - E^g(\delta))^2/\mathrm{var}^g(\delta)$ is χ^2_1-distributed under G. We then have

$$\begin{aligned}
c_\lambda^{4,\boldsymbol{y}^p,\boldsymbol{y}^c} &= P^g(g(\delta) \geq g(0)) \\
&= P^g\Big[\frac{(\delta - E^g(\delta))^2}{\mathrm{var}^g(\delta)} \leq \frac{(E^g(\delta))^2}{\mathrm{var}^g(\delta)}\Big] \\
&= \psi_1\Big(\frac{(E^g(\delta))^2}{\mathrm{var}^g(\delta)}\Big) = \psi_1\Big[\frac{(E^{g_p}(\lambda) - E^{g_c}(\lambda))^2}{\mathrm{var}^{g_p}(\lambda) + \mathrm{var}^{g_c}(\lambda)}\Big] \\
&= \psi_1(c_\lambda^{2,\boldsymbol{y}^p,\boldsymbol{y}^c}),
\end{aligned}$$

having recalled the definition of $c_\lambda^{2,\boldsymbol{y}^p,\boldsymbol{y}^c}$ in (6) of Dahl et al. (2007). Hence, calibrating the latter measure against the cumulative distribution function of the χ^2_1-distribution is equivalent to calibrating the c_λ^{4,y^p,y^c} measure against the uniform distribution on $[0,1]$ in this case. Since in this case g is symmetric and unimodal, $c_\lambda^{3,\boldsymbol{y}^p,\boldsymbol{y}^c} = c_\lambda^{4,\boldsymbol{y}^p,\boldsymbol{y}^c}$. Accordingly, the same applies to the $c_\lambda^{3,\boldsymbol{y}^p,\boldsymbol{y}^c}$ measure. If $c_\lambda^{2,\boldsymbol{y}^p,\boldsymbol{y}^c}$ is in fact pre-experimentally χ^2_1-distributed, as in corollary 1 of Dahl et al. (2007), it follows that the new measures are pre-experimentally uniformly distributed.

Since these new measures take the functional form of g_p and g_c into account, it is our intuition that they reflect the level of conflict in a way that is better than the $c_\lambda^{2,\boldsymbol{y}^p,\boldsymbol{y}^c}$ measure depending only on expectations and variances. Computationally, sample based estimation of $c_\lambda^{3,\boldsymbol{y}^p,\boldsymbol{y}^c}$, $c_\lambda^{3+,\boldsymbol{y}^p,\boldsymbol{y}^c}$ and $c_\lambda^{3-,\boldsymbol{y}^p,\boldsymbol{y}^c}$ should be straightforward. Estimating $c_\lambda^{4,\boldsymbol{y}^p,\boldsymbol{y}^c}$ seems to require a kernel estimate of g, and is hence somewhat more demanding.

In the special case when λ is a location parameter which is a parent node for one or more data nodes y^c, Marshall & Spiegelhalter (2007) in their equation (10) defines a conflict measure which is very closely related to c_λ^{3+}. A distribution similar to G is constructed, based on sample differences for variables generated from G_p and a distribution identical to or similar to G_c. While the prior distributions for $\boldsymbol{\beta}^c$ used in (10) above to compute G_c are derived from the hierarchical model, Marshall & Spiegelhalter (2007) use specific reference priors for the same purpose. They also mention the possibility of using measures very close to c_λ^{3-} and c_λ^{3}.

The $c_\lambda^{4,\boldsymbol{y}^p,\boldsymbol{y}^c}$ measure is very attractive since it can be applied to collections of nodes. Indeed (13) extends to this case by interpreting λ and δ as vectors. Parallel to the results for scalar λ and δ we have the following theorem in the multinormal case

Theorem 2.
Assume $g_p(\boldsymbol{\lambda})$ and $g_c(\boldsymbol{\lambda})$ and hence also $g(\boldsymbol{\delta}) = \int g_p(\boldsymbol{\delta} + \boldsymbol{\lambda})g_c(\boldsymbol{\lambda})d\boldsymbol{\lambda}$ are multinormal densities of dimension n. Let ψ_n be the cumulative distribution function of the χ^2_n-distribution. We then have

$$c_{\boldsymbol{\lambda}}^{4,\boldsymbol{y}^p,\boldsymbol{y}^c} = \psi_n((E^g(\boldsymbol{\delta}))^T(\mathrm{cov}^g(\boldsymbol{\delta}))^{-1}E^g(\boldsymbol{\delta})),$$

where

$$E^g(\boldsymbol{\delta}) = E^{g_p}(\boldsymbol{\lambda}) - E^{g_c}(\boldsymbol{\lambda})$$
$$\mathrm{cov}^g(\boldsymbol{\delta}) = \mathrm{cov}^{g_p}(\boldsymbol{\lambda}) + \mathrm{cov}^{g_c}(\boldsymbol{\lambda})$$

If $E^g(\boldsymbol{\delta})$ *is pre-experimentally multinormal with mean* $\mathbf{0}$ *and covariance matrix* $\mathrm{cov}^g(\boldsymbol{\delta})$*, which is assumed non random, then* $c_{\boldsymbol{\lambda}}^{4,\boldsymbol{y}^p,\boldsymbol{y}^c}$ *is pre-experimentally uniformly distributed.*

In the appendix we will prove that this last condition is satisfied for very general linear normal models as long as the covariance matrices involved are known and the improper prior $\mathbf{1}$ for the top vector location parameter applied. This generalizes corollary 1 of Dahl et al. (2007). Marshall & Spiegelhalter (2007) consider vectors of location parameters being parents to data y^c and define the conflict for such vectors in terms of a χ^2_n-type expression as the one appearing in theorem 2. This seems to be meant to be used universally, though with caution outside Gaussian models. This corresponds to what would be the natural generalization of the c^2_λ measure of Dahl et al. (2007) to the vector case.

Our set up can be adjusted to the case where it is of interest to examine whether a prior specification of a top vector node $\boldsymbol{\theta}$ at a certain fixed value $\boldsymbol{\theta}_0$, is supported by the data. Let $\boldsymbol{\beta}^c$ be the vector consisting of all the child nodes of $\boldsymbol{\theta}$, and let parallel to (10)

$$g_c(\boldsymbol{\theta}) = \int f(\boldsymbol{\theta}; \boldsymbol{\beta}^c)\pi(\boldsymbol{\beta}^c|\boldsymbol{y})d\boldsymbol{\beta}^c\,.$$

Here $\pi(\boldsymbol{\beta}^c|\boldsymbol{y})$ is the posterior distribution for $\boldsymbol{\beta}^c$ given $\boldsymbol{y}$ obtained by replacing the fixed value $\boldsymbol{\theta}_0$ by a non-informative, possibly improper, prior for the parent node $\boldsymbol{\theta}$ of $\boldsymbol{\beta}^c$. Hence, all data is used in the formation of the density $g_c(\boldsymbol{\theta})$. Analogous to (5), by a top-down view rather than a bottom-up one, this leads e.g. to the conflict measure

$$c^4_{\boldsymbol{\theta}} = P^{g_c}(g_c(\boldsymbol{\theta}) \geq g_c(\boldsymbol{\theta}_0))$$

Example 3. This set up can e.g. be applied to an extension of example 2, with $E(X_i) = \theta, i = 1, \ldots, n$. Using the improper prior $\pi(\theta) = 1$, a calculation similar to the one used in example 1 shows that $g_c(\theta) = (s^2/(n(n-1)))^{-1/2}h_{n-1}((\theta - \bar{x})/(s^2/(n(n-1)))^{1/2})$, where $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$. Arguing as in example 1, it follows that all measures $c^{3+}_\theta, c^{3-}_\theta, c^3_\theta$ and c^4_θ are pre-experimentally uniformly distributed.

5. The general linear normal model

In this section we consider a general linear normal model described by a directed acyclic graph with a tree structure. All nodes except the bottom ones represent vectors of location parameters. Each node is multinormal given its parent node with expectation equal to a linear function of this node.

We assume that the matrix representing this function has full rank. The covariance matrix is assumed known. In this model, a vector of regression parameters is generally considered as a single node. Since also all covariance matrices are assumed known, the vector γ of the neighbouring nodes of $\boldsymbol{\lambda}$ coincides with $\boldsymbol{\beta}$, as defined in the general description in the beginning of section 4. The top node is equipped with the improper prior $\mathbf{1}$. Hence, marginally every node has the improper prior $\mathbf{1}$. The bottom nodes represent the data $\boldsymbol{y}$. Many models of practical importance are special cases, as for instance some Bayesian dynamic models. Under a suitable data splitting $(\boldsymbol{y}^p, \boldsymbol{y}^c)$ we will show that $c_{\boldsymbol{\lambda}}^{4,\boldsymbol{y}^p,\boldsymbol{y}^c}$ for any vector node $\boldsymbol{\lambda}$ is pre-experimentally uniformly distributed.

We define a descendent node of a specific node as a node that can be reached by a directed path from the specific node, including by definition the specific node itself. An ancestor node of a specific node is a node that can reach the specific node by a directed path. We now have the following theorem

Theorem 3.

i) *Let $\boldsymbol{\lambda}$ be the vector of parent nodes of the data vector $\boldsymbol{y}$ and let $(\boldsymbol{y}^p, \boldsymbol{y}^c)$ be a data splitting such that either of the following two conditions are satisfied*

   a) $\boldsymbol{Y}^p$ and $\boldsymbol{Y}^c$ are independent given $\boldsymbol{\lambda}$ and the linear mapping $\boldsymbol{\lambda} \rightarrow E(\boldsymbol{Y}^p|\boldsymbol{\lambda})$ has full rank.

   b) $\boldsymbol{\lambda}$ can be decomposed as $(\boldsymbol{\lambda}^p, \boldsymbol{\lambda}^c)$, where $\boldsymbol{y}^c$ consists of all child nodes of components of $\boldsymbol{\lambda}^c$, and where $\boldsymbol{\lambda}^p$ and $\boldsymbol{\lambda}^c$ are independent given an ancestor parameter node $\boldsymbol{\beta}$ of the components of $\boldsymbol{\lambda}^c$.

   Then the conflict measure $c_{\boldsymbol{y}^c}^{4,\boldsymbol{y}^p,\boldsymbol{y}^c}$ comparing the information contributions from $\boldsymbol{y}^c$ on the one side and $\boldsymbol{\lambda}$ or $\boldsymbol{\lambda}^c$ on the other about $\boldsymbol{y}^c$ is pre-experimentally uniformly distributed.

ii) *Let $\boldsymbol{\lambda}$ be a parameter node where the vector of neighbouring nodes $\boldsymbol{\beta}$ can be decomposed as $(\boldsymbol{\beta}^p, \boldsymbol{\beta}^c)$. Here $\boldsymbol{\beta}^c$ consists of child nodes of components of $\boldsymbol{\lambda}$. Furthermore, let $(\boldsymbol{y}^p, \boldsymbol{y}^c)$ be a data splitting such that either of the following two conditions are satisfied*

   a) $\boldsymbol{y}^c$ consists of all the data decendent nodes of $\boldsymbol{\beta}^c$.

   b) $\boldsymbol{y}^c = (\boldsymbol{y}_1^c, \boldsymbol{y}_2^c)$ with $\boldsymbol{y}_1^c = \boldsymbol{\beta}^c$.

   Then the conflict measure $c_{\boldsymbol{\lambda}}^{4,\boldsymbol{y}^p,\boldsymbol{y}^c}$ comparing the information contributions from $\boldsymbol{\beta}^p$ and $\boldsymbol{\beta}^c$ about $\boldsymbol{\lambda}$ is pre-experimentally uniformly distributed.

It is tacitly assumed in i) of theorem 3 that the graph is manipulated such that $\boldsymbol{\lambda}$ or $\boldsymbol{\lambda}^c$ is a single node contributing information about $\boldsymbol{y}^c$. This gives great flexibility in the choice of data splitting for data node conflicts.

Such manipulations are also allowed in ii), but are often less relevant since conflicts about parameter nodes are primarily of interest in relation to the original formulation of the model.

If $\boldsymbol{y}^c$ of i) of theorem 3 is a scalar y^c, it follows from the proof that the density g is normal with expectation $E(Y^c|\boldsymbol{y}^p) - y^c$ and variance $\mathrm{var}(Y^c|\boldsymbol{y}^p)$. Hence,

$$c_{y^c}^{3+,\boldsymbol{y}_p,y_c} = G(0) = \varphi(-(E(Y^c|\boldsymbol{y}^p) - y^c)/(\mathrm{var}(Y^c|\boldsymbol{y}^p))^{1/2})$$

It also follows from this proof that $E(Y^c|\boldsymbol{Y}^p) - Y^c$ is normal with expectation 0 and variance $\mathrm{var}(Y^c|\boldsymbol{y}^p)$. Hence, also $c_{y^c}^{3+,\boldsymbol{y}^p,y^c}$ is pre-experimentally uniformly distributed.

Example 4. We can apply this to an example considered in section 3.3 of Bayarri & Berger (2000) where $Y^c = T = \boldsymbol{w}^T\boldsymbol{X}$. Here $\boldsymbol{X}$ arises from a standard normal regression model with i.i.d. noise terms. We transform $\boldsymbol{X}$ to $\boldsymbol{Y} = W\boldsymbol{X}$, where $\boldsymbol{w}^T$ is the last row vector of an orthogonal matrix W, assuming without loss of generality that $\boldsymbol{w}$ has norm 1. By i) of theorem 1 it follows that the partial posterior predictive p-value of T is pre-experimentally uniformly distributed. Hence, the result of the mentioned section of Bayarri & Berger (2000) can be obtained as a special case of theorem 3.

Let us consider the one-way analysis of variance example in Dahl et al. (2007) as a special case of the general linear normal model. We assume

$$\begin{aligned} &y_{ij}|\boldsymbol{\lambda}, \sigma^2 \sim^{\text{ind}} N(\lambda_i, \sigma^2), \qquad i = 1, \ldots, k; \quad j = 1, \ldots, n_i \\ &\lambda_i|\mu, \tau^2 \sim^{\text{ind}} N(\mu, \tau^2), \qquad i = 1, \ldots, k, \end{aligned} \tag{14}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^T$. σ^2 and τ^2 are assumed fixed and known and we choose an improper prior 1 for μ. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, $i = 1, \ldots, k$. Two data splittings are considered. The horizontal splitting is given by

$$\begin{aligned} \boldsymbol{y}^p &= (y_{11}, \ldots, y_{1m_1}, \ldots, y_{k1}, \ldots, y_{km_k})^T \\ \boldsymbol{y}^c &= (y_{1m_1+1}, \ldots, y_{1,n_1}, \ldots, y_{km_k+1}, \ldots, y_{kn_k})^T, \end{aligned} \tag{15}$$

where $1 \leq m_i < n_i$ for $i = 1, \ldots, k$. The vertical splitting is for $1 \leq \ell < k$ given by

$$\boldsymbol{y}^p = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_\ell^T)^T \qquad \boldsymbol{y}^c = (\boldsymbol{y}_{\ell+1}^T, \ldots, \boldsymbol{y}_k^T)^T \tag{16}$$

By manipulating the graph such that $\boldsymbol{\lambda}$ is a single node, (15) is allowed by a) of i) of theorem 3. By letting $\boldsymbol{\lambda}^c = (\lambda_{\ell+1}, \ldots, \lambda_k)^T$ be considered as a single node, (16) is allowed by b) of i) of theorem 3 noting that $\boldsymbol{\beta} = \mu$. Now consider ii) of theorem 3. In Dahl et al. (2007) the conflict between the information contributions about λ_k from $\beta^p = \mu$ and $\boldsymbol{\beta}^c = \boldsymbol{y}_k$ is assessed. The horizontal splitting (15) is not allowed neither by a) nor b), since $\boldsymbol{y}^c$ does not contain all the data descendent nodes $\boldsymbol{y}_k$. a) is violated by the vertical splitting (16) unless $\ell = k - 1$. However, (16) is allowed by b) by choosing $\boldsymbol{y}_1^c = \boldsymbol{y}_k$. Letting $\boldsymbol{y}^c = (\boldsymbol{y}_1^c, \boldsymbol{y}_2^c)$ enables one to analyse conflicts about several nodes by a simple data splitting. This saves computational efforts, but at the expense of detection power. If instead the conflict between the information contributions about the single node $\boldsymbol{\lambda}^c$ from $\beta^p = \mu$ and $\boldsymbol{\beta}^c = \boldsymbol{y}^c$ is considered, (16) is allowed by a).

Example 5. Consider a Bayesian dynamic model of the form $\lambda_t = A_t\lambda_{t-1} + \epsilon_t$, $Y_t = B_t\lambda_t + \eta_t$, where $\epsilon_t, \eta_t, t = 1, \dots, T$ are independent, normally distributed noise terms with known variances. Let $t > 1$ be arbitrary. The conditions in b) of part i) of theorem 3 are satisfied by choosing

$\boldsymbol{y}^p = (y_1, \dots, y_{t-1})$, $\boldsymbol{y}^c = (y_t, \dots, y_T)$, $\boldsymbol{\lambda}_p = (\lambda_1, \dots, \lambda_{t-1})$, $\boldsymbol{\lambda}_c = (\lambda_t, \dots, \lambda_T)$, $\beta = \lambda_{t-1}$.

Hence, the data node conflict $c_{\boldsymbol{y}^c}^{4,\boldsymbol{y}^p,\boldsymbol{y}^c}$ is pre-experimentally uniformly distributed by theorem 3. To exemplify part (ii), we may let $\boldsymbol{y}^p, \boldsymbol{y}^c$ be as above, and $\lambda = \lambda_t$, $\beta^p = \lambda_{t-1}$, $\boldsymbol{\beta}^c = (y_t, \lambda_{t+1})$. We may then conclude that $c_{\lambda_t}^{4,\boldsymbol{y}^p,\boldsymbol{y}^c}$ is pre-experimentally uniform.

6. Concluding remarks

In this paper we have adopted the idea of O'Hagan (2003) of measuring the conflict at any node of a graph, representing a Bayesian hierarchical model, by contrasting local information contributions from neighbouring nodes. At the same time we have been insisting on avoiding double use of data, and aiming for correct pre-experimental probabilites for false warnings. Accordingly, we have developed some new measures of internal inconsistencies in such models. Through the χ^2-type measure treated in Dahl et al. (2007), well suited when the local information contributions about the node are symmetric and unimodal, this has lead to the tail probability based measures considered in the present paper. It turns out that other measures with the same ambition of pre-experimental correctness and avoidance of double use of data, such as the cross validatory p-value, the conflict p-value of Marshall and Spiegelhalter (2007) and the partial posterior predictive p-value of Bayarri and Berger (2000), to a large extent can be seen as special cases of our measures. Hence, our methodology may serve as a unifying framework for these measures.

We have shown theoretically that our conflict measures are pre-experimentally uniformly distributed under the assumed model in some cases. The empirical results of Dahl et al. (2007) are also quite promising with respect to having approximate pre-experimental uniformity more generally. However, further empirical studies are needed, and we plan to return to this computationally demanding exercise in a future paper.

The main reason for pursuing a correct pre-experimental probability of false warning, is not to use the conflict measures in a traditional, frequentist hypothesis testing framework, although one may use the conflict measures in conjunction with Bonferroni-like adjustments to the significance level to control the overall false alarm probabilities. Rather, the main purpose is to standardize the measures for conflict in hierarchical models in such a way that they can be interpreted in the same way at different levels of the hierarchy, and across different models and distribution types.

With this focus of coming as close as possible to pre-experimental uniformity of the conflict measures, we are in this paper not concerned with the quite heavy computational burden that a comprehensive implementation of our conflict analysis within a large network would imply. In very complex models with many nodes, it may be necessary to develop approximate methods that can reduce the computational effort.

Acknowledgements

This work has benefitted from the "Evaluation of Bayesian Hierarchical Models" programme, supported by the Research Council of Norway.

References

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

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

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

Gelfand, A. E. & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc. Ser. B* **56**, 501–514.

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

O'Hagan, A. (2003). HSSS model criticism (with discussion). In *Highly Strucured Stochastic Systems* (eds P.J. Green, N.L. Hjort & S. Richardson), 423–453. Oxford University Press, Oxford.

Scheel, I., Green, P.J. & Rougier, J.C. (2008). Identifying influential model choices in Bayesian hierarchical models. Technical Report, Department of Mathematics, University of Bristol.

Appendix

Proof of Theorem 3. The proof is based on the assumptions linked to the general linear model, and is built up through the proofs of four lemmas and two propositions.

Lemma 1 *Let $\boldsymbol{\lambda}$ be a parameter node, and suppose $\boldsymbol{z}$ is a subvector of $\boldsymbol{y}$ consisting of descendant nodes of $\boldsymbol{\lambda}$. Then $\boldsymbol{Z}$ can be written in the form*

$\boldsymbol{Z} = A\boldsymbol{\lambda} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is multinormal with mean $\mathbf{0}$ and some covariance matrix Σ, and where $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\lambda}$.

Proof. If $\boldsymbol{z}$ is a child node of $\boldsymbol{\lambda}$, i.e. there is exactly one edge between $\boldsymbol{\lambda}$ and $\boldsymbol{z}$, the assertion follows by the assumptions. Note that $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\lambda}$ since the covariance matrix is assumed known. If there is a single path with exactly two edges, and hence one node $\boldsymbol{\beta}$, between $\boldsymbol{\lambda}$ and $\boldsymbol{z}$, we can write the relation as $\boldsymbol{\beta} = A_1\boldsymbol{\lambda}+\boldsymbol{\epsilon}_1$, $\boldsymbol{Z} = A_2\boldsymbol{\beta}+\boldsymbol{\epsilon}_2$, and hence $\boldsymbol{Z} = A_2A_1\boldsymbol{\lambda}+(A_2\boldsymbol{\epsilon}_1+\boldsymbol{\epsilon}_2)$, which is of the given form. The lemma follows by induction in the case when there is a single path from $\boldsymbol{\lambda}$ to $\boldsymbol{z}$. The general case follows by breaking $\boldsymbol{z}$ down into components such that each component is arrived at through one path starting in $\boldsymbol{\lambda}$.

Lemma 2 *Let $\boldsymbol{\theta}$ be the top node. Then $\boldsymbol{\theta}$ given $\boldsymbol{y}$ as well as $E(\boldsymbol{\theta}|\boldsymbol{Y})$ are multinormal, and we have*

$$\mathrm{cov}(\boldsymbol{\theta}|\boldsymbol{y}) = \mathrm{cov}(E(\boldsymbol{\theta}|\boldsymbol{Y}))$$

Proof. By lemma 1 it follows that $\boldsymbol{Y}$ can be written in the form $\boldsymbol{Y} = V\boldsymbol{\theta}+\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is multinormal with expectation $\mathbf{0}$ and some covariance matrix Σ. Since the conditional expectation of each node given its parent node is a full rank linear function, the matrix V must have full rank equal to $\dim(\boldsymbol{\theta})$. There exists a matrix R, the product of a diagonal matrix of scaling factors and an orthogonal matrix diagonalizing Σ, such that $R\Sigma R^T = I$. This implies that the transformed data vector $\boldsymbol{X} = R\boldsymbol{Y}$ has I as covariance matrix. Obviously, $\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \pi(\boldsymbol{\theta}|\boldsymbol{x})$. Hence, we may assume that $\Sigma = \mathrm{cov}(\boldsymbol{\epsilon}) = \mathrm{cov}(\boldsymbol{Y} - V\boldsymbol{\theta}) = I$. We have

$$\boldsymbol{Y} - V\boldsymbol{\theta} = (I - V(V^TV)^{-1}V^T)\boldsymbol{Y} + (V(V^TV)^{-1}V^T\boldsymbol{Y} - V\boldsymbol{\theta})$$

The product of the transposed of the first summand and the second summand is 0. Hence, due to the improper prior for $\boldsymbol{\theta}$, it follows that

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \exp(-(1/2)(\boldsymbol{\theta} - (V^TV)^{-1}V^T\boldsymbol{y})^TV^TV(\boldsymbol{\theta} - (V^TV)^{-1}V^T\boldsymbol{y})),$$

and hence $\boldsymbol{\theta}$ is multinormal given $\boldsymbol{y}$, $E(\boldsymbol{\theta}|\boldsymbol{y}) = (V^TV)^{-1}V^T\boldsymbol{y}$ and $\mathrm{cov}(\boldsymbol{\theta}|\boldsymbol{y}) = (V^TV)^{-1}$. It also follows that $E(\boldsymbol{\theta}|\boldsymbol{Y})$ is multinormal with expectation $\boldsymbol{\theta}$ and covariance matrix $(V^TV)^{-1}$.

Lemma 3 *For any parameter node $\boldsymbol{\lambda}$ and an arbitrary data vector $\boldsymbol{y}$ we have that $\boldsymbol{\lambda}$ given $\boldsymbol{y}$ as well as $E(\boldsymbol{\lambda}|\boldsymbol{Y}) - \boldsymbol{\lambda}$ are multinormal, and that*

$$\mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}) = \mathrm{cov}(E(\boldsymbol{\lambda}|\boldsymbol{Y}) - \boldsymbol{\lambda})$$

Proof. Suppose first that $\boldsymbol{\lambda} = \boldsymbol{\theta}$. Since $\boldsymbol{\theta}$ is deterministic, we get

$$\mathrm{cov}(E(\boldsymbol{\theta}|\boldsymbol{Y}) - \boldsymbol{\theta}) = \mathrm{cov}(E(\boldsymbol{\theta}|\boldsymbol{Y})) = \mathrm{cov}(\boldsymbol{\theta}|\boldsymbol{y})$$

by lemma 2. Also, $\boldsymbol{\theta}$ given $\boldsymbol{y}$ as well as $E(\boldsymbol{\theta}|\boldsymbol{Y}) - \boldsymbol{\theta}$ are multinormal. Now let $\boldsymbol{\lambda}$ be any other node. Let $\boldsymbol{\nu}$ be the parent node of $\boldsymbol{\lambda}$, and write $\boldsymbol{\lambda}$ as

$\boldsymbol{\lambda} = V\boldsymbol{\nu} + \boldsymbol{\eta}$, where $E(\boldsymbol{\eta}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{\eta}) = R$, and $\boldsymbol{\eta}$ is independent of $\boldsymbol{\nu}$ by the assumptions. We prove the lemma by induction on the number of edges between $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. Suppose that, with respect to arbitrary data, the lemma is true for the node $\boldsymbol{\nu}$, which is one edge closer to $\boldsymbol{\theta}$ than $\boldsymbol{\lambda}$ is. Decompose $\boldsymbol{y}$ as $(\boldsymbol{x}, \boldsymbol{z})$, where $\boldsymbol{x}$ consists of the descendant data nodes of $\boldsymbol{\lambda}$. Assume first that both $\boldsymbol{x}$ and $\boldsymbol{z}$ are non-empty. Applying lemma 1, we have $\boldsymbol{X} = A\boldsymbol{\lambda} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{\epsilon}) = \Sigma$, and $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\lambda}$. By induction $\boldsymbol{\nu}$ given $\boldsymbol{z}$ is multinormal with $\mathrm{cov}(\boldsymbol{\nu}|\boldsymbol{z}) = \mathrm{cov}(E(\boldsymbol{\nu}|\boldsymbol{Z}) - \boldsymbol{\nu}) = K$. Then $\boldsymbol{\lambda}$ given $\boldsymbol{z}$ is multinormal with expectation $E(\boldsymbol{\lambda}|\boldsymbol{z}) = VE(\boldsymbol{\nu}|\boldsymbol{z})$ and covariance matrix $Q = VKV^T + R$. It follows that

$$\begin{aligned}\pi(\boldsymbol{\lambda}|\boldsymbol{y}) &\propto f(\boldsymbol{x}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda}|\boldsymbol{z}) \\ &\propto \exp((-1/2)([(\boldsymbol{x} - A\boldsymbol{\lambda})^T\Sigma^{-1}(\boldsymbol{x} - A\boldsymbol{\lambda})] \\ &\qquad + [(\boldsymbol{\lambda} - VE(\boldsymbol{\nu}|\boldsymbol{z}))^T Q^{-1}(\boldsymbol{\lambda} - VE(\boldsymbol{\nu}|\boldsymbol{z})])) \end{aligned} \tag{17}$$

Hence, $\boldsymbol{\lambda}$ is multinormal given $\boldsymbol{y}$. Collecting the quadratic terms in (17) we find that the precision of $\boldsymbol{\lambda}$ given $\boldsymbol{y}$ is

$$C^{-1} = \mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y})^{-1} = A^T\Sigma^{-1}A + Q^{-1} \tag{18}$$

Collecting the linear terms we obtain

$$E(\boldsymbol{\lambda}|\boldsymbol{y}) = C[A^T\Sigma^{-1}\boldsymbol{x} + Q^{-1}VE(\boldsymbol{\nu}|\boldsymbol{z})] \tag{19}$$

Since by induction $E(\boldsymbol{\nu}|\boldsymbol{Z}) - \boldsymbol{\nu}$ is multinormal, it follows by (19) that $E(\boldsymbol{\lambda}|\boldsymbol{Y})$ and hence also $E(\boldsymbol{\lambda}|\boldsymbol{Y}) - \boldsymbol{\lambda}$ are multinormal. By (19) we get

$\mathrm{cov}(E(\boldsymbol{\lambda}|\boldsymbol{Y}) - \boldsymbol{\lambda}) = \mathrm{cov}(\boldsymbol{\lambda} - CA^T\Sigma^{-1}\boldsymbol{X} - CQ^{-1}VE(\boldsymbol{\nu}|\boldsymbol{Z})) =$

$\mathrm{cov}(CC^{-1}\boldsymbol{\lambda} - CA^T\Sigma^{-1}(A\boldsymbol{\lambda} + \boldsymbol{\epsilon}) - CQ^{-1}VE(\boldsymbol{\nu}|\boldsymbol{Z}))$

By (18) and the independence of the noise terms $\boldsymbol{\epsilon}$ of $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ of $\boldsymbol{\nu}$, this equals

$\mathrm{cov}(C[Q^{-1}\boldsymbol{\lambda} - A^T\Sigma^{-1}\boldsymbol{\epsilon} - Q^{-1}VE(\boldsymbol{\nu}|\boldsymbol{Z})]) =$

$CA^T\Sigma^{-1}AC + \mathrm{cov}(C[Q^{-1}(V\boldsymbol{\nu} + \boldsymbol{\eta}) - Q^{-1}VE(\boldsymbol{\nu}|\boldsymbol{Z})]) =$

$CA^T\Sigma^{-1}AC + \mathrm{cov}(CQ^{-1}\boldsymbol{\eta} + CQ^{-1}V(\boldsymbol{\nu} - E(\boldsymbol{\nu}|\boldsymbol{Z}))) =$

$CA^T\Sigma^{-1}AC + CQ^{-1}RQ^{-1}C + CQ^{-1}VKV^TQ^{-1}C =$

$C[A^T\Sigma^{-1}A + Q^{-1}]C = C = \mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y})$.

If $\boldsymbol{x}$ is empty, the contribution from $\boldsymbol{x}$ simply vanishes, and the proof just simplifies. If $\boldsymbol{z}$ is empty, the full rank assumption assures that $A^T\Sigma^{-1}A$ is invertible, and the proof is valid also in this case.

Lemma 4 *Let $(\boldsymbol{y}^p, \boldsymbol{y}^c)$ be a data splitting, and suppose the parameter node $\boldsymbol{\lambda}$ is such that $\boldsymbol{y}^c$ consists of descendant nodes of $\boldsymbol{\lambda}$, and such that $\boldsymbol{Y}^p$ and $\boldsymbol{Y}^c$ are independent given $\boldsymbol{\lambda}$. Then $E(\boldsymbol{Y}^c|\boldsymbol{Y}^p) - \boldsymbol{Y}^c$ is multinormal, and*

$$\mathrm{cov}(\boldsymbol{Y}^c|\boldsymbol{y}^p) = \mathrm{cov}(E(\boldsymbol{Y}^c|\boldsymbol{Y}^p) - \boldsymbol{Y}^c)$$

Proof. Using lemma 1, we write $\boldsymbol{Y}^c$ in the form $\boldsymbol{Y}^c = A\boldsymbol{\lambda} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\lambda}$ and of $\boldsymbol{Y}^p$ and has covariance matrix Σ. Then

$\mathrm{cov}(\boldsymbol{Y}^c|\boldsymbol{y}^p) = \Sigma + A\mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^p)A^T$,

$\mathrm{cov}(E(\boldsymbol{Y}^c|\boldsymbol{Y}^p) – \boldsymbol{Y}^c) = \mathrm{cov}(AE(\boldsymbol{\lambda}|\boldsymbol{Y}^p) – A\boldsymbol{\lambda} – \boldsymbol{\epsilon}) = A\mathrm{cov}(E(\boldsymbol{\lambda}|\boldsymbol{Y}^p) – \boldsymbol{\lambda})A^T + \Sigma$

The lemma therefore follows from lemma 3, with $\boldsymbol{y}^p$ in place of $\boldsymbol{y}$.

Note that it is not needed in the proof that the linear mapping $\boldsymbol{\lambda} \to E(\boldsymbol{Y}^c|\boldsymbol{\lambda})$ has full rank. Hence, the lemma is also valid with $\boldsymbol{Y}^c$ replaced by a dimension reducing linear transformation of $\boldsymbol{Y}^c$. In particular, if $T = t(\boldsymbol{Y}^c)$ is a linear statistic, we obtain

$$\mathrm{var}(T|\boldsymbol{y}^p) = \mathrm{var}(E(T|\boldsymbol{Y}^p) - T)$$

Proposition 1 *Make the assumptions of i) of Theorem 3. Then* $E(\boldsymbol{Y}^c|\boldsymbol{Y}^p) - \boldsymbol{Y}^c$ *is multinormal and*

$$\mathrm{cov}(\boldsymbol{Y}^c|\boldsymbol{y}^p) = \mathrm{cov}(E(\boldsymbol{Y}^c|\boldsymbol{Y}^p) - \boldsymbol{Y}^c)$$

Proof. Consider first a). Applying a slightly extended version of lemma 1 twice, with the pair $(\boldsymbol{\lambda}, \boldsymbol{z})$ of the lemma replaced by respectively $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and $(\boldsymbol{\lambda}, \boldsymbol{y}^p)$, we obtain an alternative description of the submodel for data $\boldsymbol{y}^p$, only consisting of the three nodes $\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{y}^p$, and with only two edges. Since $\boldsymbol{\lambda} \to E(\boldsymbol{Y}^p|\boldsymbol{\lambda})$ has full rank, this description is allowed. If now also the mapping $\boldsymbol{\lambda} \to E(\boldsymbol{Y}^c|\boldsymbol{\lambda})$ has full rank, using the independence of $\boldsymbol{Y}^p$ and $\boldsymbol{Y}^c$ given $\boldsymbol{\lambda}$ we can extend this to a description of the entire model, with $\boldsymbol{y}^c$ as an extra node, and with one extra edge from $\boldsymbol{\lambda}$ to $\boldsymbol{y}^c$. The assertion then follows directly from lemma 4. If not, the assertion follows nevertheless in view of the remark after lemma 4. To prove b) we also apply the slightly extended version of lemma 1 twice, now with $(\boldsymbol{\lambda}, \boldsymbol{z})$ replaced by respectively $(\boldsymbol{\beta}, \boldsymbol{\lambda}^c)$ and $(\boldsymbol{\lambda}^c, \boldsymbol{y}^c)$. Due to the independence of $\boldsymbol{\lambda}^p$ and $\boldsymbol{\lambda}^c$ given $\boldsymbol{\beta}$, this transformation of the graph does not affect the submodel for $\boldsymbol{\lambda}^p$ and $\boldsymbol{y}^p$. Part b) now follows from lemma 4, with $\boldsymbol{\lambda}^c$ in place of $\boldsymbol{\lambda}$.

Proposition 2 *Let* $\boldsymbol{\lambda}$ *be a parameter node, and suppose that the data splitting* $(\boldsymbol{y}^p, \boldsymbol{y}^c)$ *is such that* $\boldsymbol{y}^c$ *consists of descendant nodes of* $\boldsymbol{\lambda}$*, and such that* $\boldsymbol{Y}^p$ *and* $\boldsymbol{Y}^c$ *are independent given* $\boldsymbol{\lambda}$*. Then* $E(\boldsymbol{\lambda}|\boldsymbol{Y}^p) - E(\boldsymbol{\lambda}|\boldsymbol{Y}^c)$ *is multinormal, and*

$$\mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^p) + \mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^c) = \mathrm{cov}(E(\boldsymbol{\lambda}|\boldsymbol{Y}^p) - E(\boldsymbol{\lambda}|\boldsymbol{Y}^c))$$

Proof. Lemma 3 assures the multinormality of $E(\boldsymbol{\lambda}|\boldsymbol{Y}^p) - E(\boldsymbol{\lambda}|\boldsymbol{Y}^c)$. Again we write $\boldsymbol{Y}^c$ in the form $\boldsymbol{Y}^c = A\boldsymbol{\lambda} + \boldsymbol{\epsilon}$, using lemma 1. By our standard assumptions, A has full rank. Since $\pi(\boldsymbol{\lambda}|\boldsymbol{y}^c)$ is invariant under linear transformations of $\boldsymbol{y}^c$, we may assume as in the proof of lemma 2 that $\mathrm{cov}(\boldsymbol{\epsilon}) = I$. Due to the improper prior for $\boldsymbol{\lambda}$, it follows as in that proof that $E(\boldsymbol{\lambda}|\boldsymbol{y}^c) = (A^TA)^{-1}A^T\boldsymbol{y}^c$, and that $\mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^c) = (A^TA)^{-1}$. Using these facts, and also the independence given $\boldsymbol{\lambda}$ of $\boldsymbol{Y}^p$ and $\boldsymbol{Y}^c$, and hence of $\boldsymbol{Y}^p$ and $\boldsymbol{\epsilon}$, in addition to the independence of $\boldsymbol{\lambda}$ and $\boldsymbol{\epsilon}$, as well as lemma 3 with $\boldsymbol{y}^p$ in place of $\boldsymbol{y}$, we get

$$\mathrm{cov}(E(\boldsymbol{\lambda}|\boldsymbol{Y}^p) - E(\boldsymbol{\lambda}|\boldsymbol{Y}^c)) = \mathrm{cov}(E(\boldsymbol{\lambda}|\boldsymbol{Y}^p) - (A^TA)^{-1}A^T(A\boldsymbol{\lambda}+\boldsymbol{\epsilon})) =$$
$$\mathrm{cov}((E(\boldsymbol{\lambda}|\boldsymbol{Y}^p) - \boldsymbol{\lambda}) - (A^TA)^{-1}A^T\boldsymbol{\epsilon}) = \mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^p) + (A^TA)^{-1} =$$
$$\mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^p) + \mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^c)$$

To prove the first part of theorem 3, we define $\boldsymbol{\delta} = \boldsymbol{Y}^c - \boldsymbol{y}^c$, where $\boldsymbol{Y}^c$ is g_p-distributed. By a multidimensional version of (1) we then have

$$E^g(\boldsymbol{\delta}) = E^{g_p}(\boldsymbol{Y}^c) - \boldsymbol{y}^c = E(\boldsymbol{Y}^c|\boldsymbol{y}^p) - \boldsymbol{y}^c,$$

which by proposition 1 is pre-experimentally multinormal with expectation $E(E(\boldsymbol{Y}^c|\boldsymbol{Y}^p) - \boldsymbol{Y}^c) = 0$. Also

$$\mathrm{cov}^g(\boldsymbol{\delta}) = \mathrm{cov}^{g_p}(\boldsymbol{Y}^c) = \mathrm{cov}(\boldsymbol{Y}^c|\boldsymbol{y}^p)$$

Using proposition 1 again it follows that the pre-experimental covariance matrix of $E^g(\boldsymbol{\delta})$ equals $\mathrm{cov}(E(\boldsymbol{Y}^c|\boldsymbol{Y}^p) - \boldsymbol{Y}^c) = \mathrm{cov}^g(\boldsymbol{\delta})$. Hence, part i) follows from a data node version of theorem 2.

The conditions in a) in the last part of the theorem imply that also the conditions of proposition 2 are satisfied. Note that under the given conditions $\boldsymbol{Y}^c$ is independent of $\boldsymbol{\lambda}$ given $\boldsymbol{\beta}^c$, and that $f(\boldsymbol{\lambda};\boldsymbol{\beta}^c) = \pi(\boldsymbol{\lambda}|\boldsymbol{\beta}^c)$ due to the improper prior for $\boldsymbol{\lambda}$. Hence, from a $\boldsymbol{\lambda}$ vector version of (10) $g_c(\boldsymbol{\lambda}) = \pi(\boldsymbol{\lambda}|\boldsymbol{y}^c)$. Accordingly, g_c is multinormal with $E^{g_c}(\boldsymbol{\lambda}) = E(\boldsymbol{\lambda}|\boldsymbol{y}^c)$ and $\mathrm{cov}^{g_c}(\boldsymbol{\lambda}) = \mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^c)$. A corresponding result is valid for g_p. With $\boldsymbol{\delta}$ as in theorem 2, it follows that the pre-experimental expectation of $E^g(\boldsymbol{\delta})$ is $E(E(\boldsymbol{\lambda}|\boldsymbol{Y}^p) - E(\boldsymbol{\lambda}|\boldsymbol{Y}^c)) = 0$. By proposition 2 the pre-experimental covariance matrix is

$$\mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^p) + \mathrm{cov}(\boldsymbol{\lambda}|\boldsymbol{y}^c) = \mathrm{cov}^{g_p}(\boldsymbol{\lambda}) + \mathrm{cov}^{g_c}(\boldsymbol{\lambda})$$

By proposition 2 it also follows that $E^g(\boldsymbol{\delta})$ is pre-experimentally multinormal. The last part for the case a) then follows from theorem 2.

For the last part for the case b) consider the reduced data set $(\boldsymbol{y}^p, \boldsymbol{y}_1^c)$, also denoting the splitting. Then the conclusion follows from case a). However, the full data set with the splitting $(\boldsymbol{y}^p, \boldsymbol{y}^c)$ gives rise to the same g_p and g_c and the proof is completed.