Bayesian checking of hierarchical models *

M.J. Bayarri and M.E. Castellanos
University of Valencia and Rey Juan Carlos University

April 26, 2004

Abstract

Hierarchical models are increasingly used in many applications. Along with this increase use comes a desire to investigate whether the model is compatible with the observed data. Bayesian methods are well suited to eliminate the many (nuisance) parameters in these complicated models; in this paper we investigate Bayesian methods for model checking. Since we contemplate model checking as a preliminary, exploratory analysis, we concentrate in *objective Bayesian methods* in which careful specification of an informative prior distribution is avoided. Numerous examples are given and different proposals are investigated.

Key words and phrases: model checking; model criticism; objective Bayesian methods; p-values.

1 Introduction

With the availability of powerful numerical computations, use of hierarchical (or multilevel, or random effects) models have become very common in applications; They nicely generalize and extend classical one-level models to complicated situations, where the classical models would not apply. With their widespread use, comes along an increased need to check the adequacy of such models to the observed data. Recent Bayesian methods (Bayarri and Berger, 1999, 2000) have shown considerable promise in checking one-level models, specially in non-standard situations in which parameter-free testing statistics are not known. In this paper we show how these methods can be extended to checking hierarchical models; we also review other Bayesian proposals and critically compare them.

We approach model checking as a preliminary analysis in that if the data is compatible with the assumed model, then the full (and difficult) Bayesian process of model elaboration and model selection (or averaging) can be avoided. The role of Bayesian model checking versus

*This research was supported by the Spanish Ministry of Science and Technology, under Grant SAF2001-2931.

model selection has been discussed for example in Bayarri and Berger (1999, 2000) and O'Hagan (2003) and we will not repeat it here.

In general, in a model checking scenario, we relate observables $\mathbf{X}$ with parameters $\boldsymbol{\theta}$ through a parametric model $\mathbf{X} \,|\, \boldsymbol{\theta} \;\sim\; f(\mathbf{x} \,|\, \boldsymbol{\theta})$. We then observe data $\mathbf{x}_{obs}$ and wish to assess whether $\mathbf{x}_{obs}$ is compatible with the assumed (null) model $f(\mathbf{x} \,|\, \boldsymbol{\theta})$. Most of the existing methods for model checking (both Bayesian and frequentist) can be seen to correspond to particular choices of:

1. A diagnostic statistic T, to quantify incompatibility of the model with the observed data through $t_{obs} = T(\mathbf{x}_{obs})$.

2. A completely specified distribution for the statistic, $h(t)$, under the null model, in which to locate the observed t_{obs}.

3. A way to measure conflict between the observed statistic, and the null distribution, $h(t)$, for T. The most popular measures are tail areas (p-values) and relative height of the density $h(t)$ at t_{obs}.

In this paper, we concentrate on the optimal choice in 2, which basically reduces to choice of methods to eliminate the nuisance parameters $\boldsymbol{\theta}$ from the null model. Our recommendations then apply to *any* choices in 1 and 3. (We abuse notation and use the same $h(\cdot)$ to indicate both, the completely specified distribution for $\boldsymbol{X}$, after elimination of $\boldsymbol{\theta}$, and the corresponding distribution for T.) Of course, choice of 1 is very important; as a matter of fact, in some scenarios an optimal T can make choice of 2 nearly irrelevant. So our work will be most useful in complicated scenarios when such optimal T's are not known, or extremely difficult to implement (for an example of these, see Robins, Van der Vaart and Ventura, 2000). In these situations, T is often chosen casually, based on intuitive considerations, and hence we concentrate on these choices (with *no* implications whatsoever that these are our recommended choices for T; we simply do not address choice of T in this paper). As measures of conflict in 3, we explore the two best known *measures of surprise*, namely the p-value and the *relative predictive surprise, RPS* (see Berger, 1980/1985) used (with variants) by many authors. These two measures are defined as:

$$p \;=\; Pr^{h(\cdot)}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs})), \tag{1.1}$$

$$RPS \;=\; \frac{h(t(\mathbf{x}_{obs}))}{\sup_t h(t)} \;. \tag{1.2}$$

Note that *small* values of (1.1) and (1.2) denote incompatibility.

Frequentist and Bayesian choices for $h(\cdot)$ are discussed at length in Bayarri and Berger (2000), and we limit ourselves here to an extremely brief (and incomplete) mention of some of them.

The natural Bayesian choice for $h(\cdot)$ is the prior predictive distribution, in which the parameters get naturally integrated out with respect to the prior distribution (Box, 1980 pioneered use of p-values computed in the prior predictive for Bayesian model criticism). However, this requires a fairly informative prior distribution which might not be desirable for model checking for two reasons: *i)* it might not be desired to carefully quantify a prior in these earlier stages of the analysis, since the model might well not be appropriate and hence the effort is wasted; *ii)* most importantly, model checking with informative priors can not separate inadequacy of the prior from inadequacy of the model. In the sequel we will concentrate on *objective* Bayesian methods for model checking in which the priors are default, objective priors, usually improper. Note that this impropriety makes the prior predictive distribution undefined and hence not available for (objective) model checking.

Guttman (1967) and Rubin (1984) choice for $h(\cdot)$ is the *posterior* predictive distribution, resulting from integrating $\boldsymbol{\theta}$ out with respect to the posterior distribution instead of the prior. This allows use of improper priors, and hence of objective model checking. This proposal is very easy to implement by MCMC methods, and hence has become fairly popular in Bayesian model checking. However, its double use of the data can result in an extreme conservatism of the resulting p-values, unless the checking statistic is fairly ancillary (in which case the way to deal with the parameters is basically irrelevant). This conservatism is shown to hold asymptotically in Robins et al. (2000), and for finite n and several scenarios in, for example, Bayarri and Berger(1999, 2000), Bayarri and Castellanos (2001) and Bayarri and Morales (2003).

Alternative choices of $h(\cdot)$ for objective model checking are proposed in Bayarri and Berger (1997, 1999, 2000); Their asymptotic optimality is shown in Robins et all. (2000). In this paper we implement these choices in hierarchical model checking; we also compare the results with those obtained with posterior predictive distributions and several 'plug-in' choices for $h(\cdot)$.

In this paper we restrict attention to checking of a fairly simple normal-normal hierarchical model so as to best illustrate the different proposals and critically judge their behavior. However, the main ideas also apply to the checking of many other hierarchical models. In Section 2 we briefly review the different *measures of surprise* that we will derive and compare. In Section 3 we derive these measures for the hierarchical normal-normal model; we also study the sampling distribution of the different p-values, both when the null model is true, and when the data comes from alternative models.. In Section 4 we apply these measures to a particular simple test which allows easy and intuitive comparisons of the different proposals. In Section 5 we briefly summarize other methods for Bayesian checking of hierarchical models, namely those proposed by Dey et. al. (1998), O'Hagan (2003) and Marshall and Spiegelhalter (2001), comparing them with the previous proposals in an example.

2 Measures of surprise in the checking of hierarchical models

In this paper we will be dealing with the measures of surprise defined in equations (1.1) and (1.2). Their relative merits and drawbacks are discussed at length in Bayarri and Berger (1997, 1999) and will not be repeated here. In this section we derive these measures in the context of hierarchical models, and for some specific choices of the completely specified distribution $h(\cdot)$. We consider the general two-level model:

$$\begin{aligned} X_{ij} \,|\, \theta_i &\overset{ind.}{\sim} f(x_{ij} \,|\, \theta_i) && i = 1, \ldots, I; \quad j = 1, \ldots, n_i, \\ \boldsymbol{\theta} \,|\, \boldsymbol{\eta} &\overset{ind.}{\sim} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{\eta}) = \textstyle\prod_{i=1}^{I} \pi(\theta_i \,|\, \boldsymbol{\eta}), \\ \boldsymbol{\eta} &\sim \pi(\boldsymbol{\eta}) \,, \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)$.

To get a completely specified distribution $h(\cdot)$ for $\boldsymbol{X}$, we need to integrate $\boldsymbol{\theta}$ out from $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ with respect to some completely specified distribution for $\boldsymbol{\theta}$. We next consider three possibilities that have been proposed in the literature for such a distribution: empirical Bayes types (plug-in), posterior distribution, and partial posterior distribution, as they apply in the hierarchical scenario. Notice that, since we will be dealing with improper priors for $\boldsymbol{\eta}$, the natural (marginal) prior $\pi(\boldsymbol{\theta})$ is also improper and can not be used for this purpose (it would produce an improper $h(\cdot)$).

2.1 Empirical Bayes (plug-in) measures

This is the simplest proposal, very intuitive and frequently used in empirical Bayes analysis (see for example Carlin and Louis, 2000). It simply consists in replacing the unknown $\boldsymbol{\eta}$ in $\pi(\boldsymbol{\theta} \mid \boldsymbol{\eta})$ by an estimate (we use the MLE, but moment estimates are often used as well). In this proposal, $\boldsymbol{\theta}$ is integrated out with respect to

$$\pi^{EB}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \,|\, \widehat{\boldsymbol{\eta}}) = \pi(\boldsymbol{\theta} \,|\, \boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}), \tag{2.1}$$

where $\widehat{\boldsymbol{\eta}}$ maximizes the integrated likelihood:

$$f(\mathbf{x} \,|\, \boldsymbol{\eta}) = \int f(\mathbf{x} \,|\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \,|\, \boldsymbol{\eta}) d\boldsymbol{\theta}.$$

The corresponding proposal for a completely specified $h(\cdot)$ in which to define the measures of surprise is

$$m^{EB}_{prior}(t) = \int f(t \,|\, \boldsymbol{\theta}) \pi^{EB}(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{2.2}$$

The measures of surprise p^{EB}_{prior}, and RPS^{EB}_{prior} are now given by equations (1.1) and (1.2), respectively, in which $h(\cdot) = m^{EB}_{prior}(\cdot)$.

Strictly for comparison purposes, we will be using later another empirical Bayes type of distribution in which the empirical Bayes prior (2.1) gets needlessly (and inappropriately) updated using again the same data. In this (wrong) proposal, $\boldsymbol{\theta}$ gets integrated out with respect to

$$\pi^{EB}(\boldsymbol{\theta} \,|\, \mathbf{x}_{obs}) \propto f(\mathbf{x}_{obs} \,|\, \boldsymbol{\theta})\pi^{EB}(\boldsymbol{\theta}), \tag{2.3}$$

resulting in

$$m^{EB}_{post}(t) = \int f(t \,|\, \boldsymbol{\theta})\pi^{EB}(\boldsymbol{\theta} \,|\, \mathbf{x}_{obs})d\boldsymbol{\theta}\,. \tag{2.4}$$

The corresponding measures of surprise p^{EB}_{post} and RPS^{EB}_{post} are again computed using (1.1) and (1.2), respectively, with $h(\cdot) = m^{EB}_{post}(t)$.

2.2 Posterior predictive measures

This proposal is also intuitive and seems to have a more Bayesian 'flavour' than the plug-in solution presented in the previous section. This along with its ease of implementation has made the method a popular one for objective Bayesian model checking. The idea is simple: since the prior $\pi(\boldsymbol{\theta})$ is improper (for improper $\pi(\boldsymbol{\eta})$), use the posterior instead to integrate $\boldsymbol{\theta}$ out. Thus, the proposal for $h(\cdot)$ is the posterior predictive distribution

$$m_{post}(t \,|\, \mathbf{x}_{obs}) = \int f(t \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \,|\, \mathbf{x}_{obs})d\boldsymbol{\theta}\,, \tag{2.5}$$

where $\pi(\boldsymbol{\theta} \,|\, \mathbf{x}_{obs})$ is the marginal from the joint posterior

$$\begin{aligned} \pi(\boldsymbol{\theta}, \boldsymbol{\eta} \,|\, \mathbf{x}_{obs}) &\propto f(\mathbf{x}_{obs} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ &= f(\mathbf{x}_{obs} \,|\, \boldsymbol{\theta}) \prod_{i=1}^{I} \pi(\theta_i \,|\, \boldsymbol{\eta})\pi(\boldsymbol{\eta}). \end{aligned}$$

The posterior p-value and the posterior RPS are denoted by p_{post} and RPS_{post}, and computed from (1.1) and (1.2), respectively, with $h(\cdot) = m_{post}(\cdot)$.

It is important to remark that, under regularity conditions, the empirical Bayes posterior $\pi^{EB}(\boldsymbol{\theta} \,|\, \mathbf{x}_{obs})$ given in (2.3) approximates the true posterior $\pi(\boldsymbol{\theta} \,|\, \mathbf{x}_{obs})$; Both are, in fact, asymptotically equivalent. Hence whatever inadequacy of $m^{EB}_{post}(t)$ in (2.4) for model checking is likely to apply as well to the posterior predictive $m_{post}(t \,|\, \mathbf{x}_{obs})$ in (2.5). We will see demonstration of the similar behavior of both predictive distributions in all the examples in this paper.

2.3 Partial Posterior predictive measures

Both, the empirical Bayes and posterior proposals presented in Sections 2.1 and 2.2 use the *same* data to *i)* 'train' the improper $\pi(\boldsymbol{\theta})$ into a proper distribution to compute a predictive distribution and *ii)* compute the observed t_{obs} to be located in this same predictive through the

measures of surprise. This can result in a severe conservatism incapable of detecting clearly inappropriate models. A natural way to avoid this double use of the data is to use part of the data for 'training' and the rest to compute the measures of surprise, as in cross-validation methods. The proposal in Bayarri and Berger (1999, 2000) is similar in spirit: since t_{obs} is used to compute the surprise measures, use the information in the data *not* in t_{obs} to 'train' the improper prior into a proper one, called a *partial posterior distribution*:

$$\pi_{ppp}(\boldsymbol{\theta} \,|\, \mathbf{x}_{obs} \setminus t_{obs}) \propto f(\mathbf{x}_{obs} \,|\, t_{obs}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \propto \frac{f(\mathbf{x}_{obs} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(t_{obs} \,|\, \boldsymbol{\theta})} \ .$$

The corresponding proposal for the completely specified $h(\cdot)$ is then the *partial posterior predictive distribution* computed as:

$$m_{ppp}(t \,|\, \mathbf{x}_{obs} \setminus t_{obs}) = \int f(t \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \,|\, \mathbf{x}_{obs} \setminus t_{obs}) d\boldsymbol{\theta} \ .$$

The *partial posterior predictive measures* of surprise will be denoted by p_{ppp} and RPS_{ppp} and, as before, computed using (1.1) and (1.2), respectively, with $h(\cdot) = m_{ppp}(\cdot)$.

Extensive discussions of the advantages and disadvantages of this proposal as compared with the previous ones can be found in Bayarri and Berger (2000) and Robins, van der Vaart and Ventura (2000). In this paper we demonstrate their performance in hierarchical models.

2.4 Computation of $\mathbf{p}_{\mathbf{h}(\cdot)}$ and $\mathbf{SPR}_{\mathbf{h}(\cdot)}$

Often, for a proposed $h(\cdot)$, the measures $p_{h(\cdot)}$ and $SPR_{h(\cdot)}$ can not be computed in close form; even more, the same $h(\cdot)$ is often not of close-form itself. In these cases we use Monte Carlo (MC), or Markov Chain Monte Carlo (MCMC) methods, to (approximately) compute them. If $\mathbf{x}^1, \ldots, \mathbf{x}^M$ is a sample of size M generated from $h(\mathbf{x})$, then $t_i = t(\mathbf{x}^i)$ is a sample from $h(t)$, and we approximate the measures of surprise as:

1. p-value
$$Pr^{h(\cdot)}(T \geq t_{obs}) = \frac{\#\text{ of } t_i \geq t_{obs}}{M}$$
2. Relative Predictive Surprise
$$SPR_{h(\cdot)} = \frac{\hat{h}(t_{obs})}{\sup\limits_{t} \hat{h}(t)} \ ,$$
where $\hat{h}(t)$ is an estimate (for instance, a kernel estimate) of the density h at t.

3 Checking hierarchical normal models

Consider the usual normal-normal two-level hierarchical (or random effects) model with I groups and n_i observations per group. The I means are assumed to be exchangeable. For simplicity,

we begin by assuming the variances σ_i^2 at the observation level to be known. The model is:

$$\begin{array}{lcll} X_{ij} \,|\, \theta_i & \overset{i}{\sim} & N(\theta_i, \sigma_i^2) & i = 1, \ldots, I, \quad j = 1, \ldots, n_i, \\ \pi(\boldsymbol{\theta} \,|\, \mu, \tau) & = & \prod_{i=1}^{I} N(\theta_i \,|\, \mu, \tau^2)\,, & \\ \pi(\mu, \tau^2) & = & \pi(\mu)\,\pi(\tau^2) \propto \frac{1}{\tau}\ . & \end{array} \tag{3.1}$$

In this paper we concentrate in checking the adequacy of the second level assumptions on the means θ_i. Of course, checking the normality of the observations is also important, but it will not be considered here; the techniques considered in this paper as applied to the checking of simple models have been explored in Bayarri and Castellanos (2001), Castellanos (1999) and Bayarri and Morales (2003).

Assume that choice of the departure statistic T is done in a rather casual manner, and that we are specially concerned about the upper tail of the distribution of the means. In this situation, a natural choice for T is

$$T = \max \quad \{\overline{X}_{1\cdot}, \ldots, \overline{X}_{I\cdot}\}, \tag{3.2}$$

where $\overline{X}_{i\cdot}$ denotes the usual sample mean for group i. This T is rather natural, but the analysis would be virtually identical with any other choice. Recall that if the statistic is fairly ancillary, then the answers from all methods are going to be rather similar, no matter how we integrate $\boldsymbol{\theta}$ out.

The distribution of the statistic (3.2) under the (null) model specified in (3.1) can be computed to be:

$$f_T(t \,|\, \boldsymbol{\theta}) = \sum_{k=1}^{I} N\left(t \,|\, \theta_k, \frac{\sigma_k^2}{n_k}\right) \prod_{\substack{l=1 \\ l \neq k}}^{I} F\left(t \,|\, \theta_l, \frac{\sigma_l^2}{n_l}\right)\ , \tag{3.3}$$

where $N(t \,|\, a, b)$ and $F(t \,|\, a, b)$ denote the density and distribution function, respectively, of a normal distribution with mean a and variance b evaluated at t.

We next integrate the unknown $\boldsymbol{\theta}$ from (3.3) using the techniques outlined in Section 2.

3.1 Empirical Bayes distributions

It is easy to see that the likelihood for μ and τ^2 is simply

$$f(\boldsymbol{x} \mid \mu, \tau^2) = \prod_{i=1}^{I} N\left(\bar{x}_i \mid \mu, \frac{\sigma_i^2}{n_i} + \tau^2\right), \tag{3.4}$$

from which $\hat{\mu}$ and $\hat{\tau}^2$ can be computed. Then (2.1) is given by

$$\pi^{EB}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \,|\, \hat{\mu}, \hat{\tau}^2) = \prod_{i=1}^{I} N(\theta_i \,|\, \hat{\mu}, \hat{\tau}^2),$$

which we use to integrated $\boldsymbol{\theta}$ out from (3.3). The resulting $m^{EB}_{prior}(\mathbf{x})$ does not have a close-form expression, but simulations can be obtained by simple MC methods: For $l = 1, \ldots, M$ simulate

$$\boldsymbol{\theta}_{(l)} = (\theta_{1(l)}, \ldots, \theta_{I(l)}) \sim \pi^{EB}(\boldsymbol{\theta}) = \prod_{i=1}^{I} \pi(\theta_i \,|\, \hat{\mu}, \hat{\tau}^2) \ ,$$

and for each $\boldsymbol{\theta}_{(l)},\ l = 1, \ldots, M$, simulate

$$\bar{\mathbf{x}}_{(l)} = (\overline{x}_{1\cdot(l)}, \ldots, \overline{x}_{I\cdot(l)}) \sim f(\bar{\mathbf{x}} \,|\, \boldsymbol{\theta}_{(l)}) = \prod_{i=1}^{I} f(\overline{x}_{i\cdot} \,|\, \theta_{i(l)}) \ .$$

For comparison purposes, we will also consider integrating $\boldsymbol{\theta}$ w.r.t. the (inappropriate) empirical-Bayes posterior distribution. The resulting $m^{EB}_{post}(\mathbf{x})$ is also trivial to simulate from using a similar MC scheme, except that θ is now simulated from :

$$\boldsymbol{\theta}_{(l)} = (\theta_{1(l)}, \ldots, \theta_{I(l)}) \sim \pi^{EB}(\boldsymbol{\theta} \,|\, \mathbf{x}_{obs}) = \prod_{i=1}^{I} N(\widehat{E}_i, \widehat{V}_i)$$

where

$$\widehat{E}_i = \frac{n_i \overline{x}_{i\cdot}/\sigma_i^2 + \hat{\mu}/\hat{\tau}^2}{n_i/\sigma_i^2 + 1/\hat{\tau}^2} \quad \text{and} \quad \widehat{V}_i = \frac{1}{n_i/\sigma_i^2 + 1/\hat{\tau}^2} \ .$$

3.2 Posterior Predictive distribution

This proposal integrates $\boldsymbol{\theta}$ out from (3.3) w.r.t. its posterior distribution. For the non-informative prior $\pi(\mu, \tau^2) \propto 1/\tau$, the joint posterior is

$$\begin{aligned} \pi_{post}(\boldsymbol{\theta}, \mu, \tau^2 | \mathbf{x}_{obs}) &\propto f(\mathbf{x} \,|\, \boldsymbol{\theta}, \mu, \tau^2) \pi(\boldsymbol{\theta} \,|\, \mu, \tau^2) \pi(\mu, \tau^2) \\ &= \prod_{i=1}^{I} N\left(\overline{x}_{i\cdot} \,|\, \theta_i, \frac{\sigma_i^2}{n_i}\right) \prod_{i=1}^{I} N(\theta_i \,|\, \mu, \tau^2) \, \frac{1}{\tau} \ . \end{aligned} \tag{3.5}$$

To simulate from the resulting posterior predictive distribution $m_{post}(\mathbf{x} \,|\, \mathbf{x}_{obs})$, we first simulate from $\pi_{post}(\boldsymbol{\theta}, \mu, \tau^2 | \mathbf{x}_{obs})$ and for each simulated $\boldsymbol{\theta}$, we simulate $\boldsymbol{x}$ from $f(\boldsymbol{x} \mid \boldsymbol{\theta})$. To simulate from the joint posterior (3.5) we use an easy Gibbs sampler defined by the full conditionals:

$$\mu \,|\, \boldsymbol{\theta}, \tau^2, \mathbf{x}_{obs} \sim N(E_\mu, V_\mu) \quad \text{with } E_\mu = \frac{\sum_{i=1}^{I} \theta_i}{I} \quad \text{and} \quad V_\mu = \frac{\tau^2}{I} \ , \tag{3.6}$$

$$\tau^2 \,|\, \boldsymbol{\theta}, \mu, \mathbf{x}_{obs} \sim \chi^{-2}(I-1, \widetilde{\tau}^2) \ \text{ where } \ \widetilde{\tau}^2 = \frac{\sum_{i=1}^{I} (\theta_i - \mu)^2}{I-1} \ , \tag{3.7}$$

$$\theta_i \,|\, \mu, \tau^2, \mathbf{x}_{obs} \sim N(E_i, V_i), \quad \text{where} \tag{3.8}$$

$$E_i = \frac{n_i \overline{x}_{i\cdot}/\sigma_i^2 + \mu/\tau^2}{n_i/\sigma_i^2 + 1/\tau^2} \quad \text{and} \quad V_i = \frac{1}{n_i/\sigma_i^2 + 1/\tau^2}$$

All the full conditionals are standard distributions, trivial to simulate from; $\chi^{-2}(\nu, a)$ refers to an *scaled* inverse Chi-square distribution: it is the distribution of $(\nu\, a)/Y$ where $Y \sim \chi^2(\nu)$.

3.3 Partial Posterior distribution

To simulate from the partial posterior predictive distribution, m_{ppp}, we proceed similarly to Section 3.2, except that simulations for the parameters are generated from the partial posterior distribution:

$$\pi_{ppp}(\boldsymbol{\theta}, \mu, \tau^2 \,|\, \mathbf{x}_{obs} \setminus t_{obs}) \propto \frac{\pi_{post}(\boldsymbol{\theta}, \mu, \tau^2 \,|\, \mathbf{x}_{obs})}{f(t_{obs} \,|\, \boldsymbol{\theta})} \,,$$

where $\pi_{post}(\boldsymbol{\theta}, \mu, \tau^2 \,|\, \mathbf{x}_{obs})$ is given in (3.5). The full conditionals for the Gibbs sampler are:

$$\mu \,|\, \boldsymbol{\theta}, \tau^2, \mathbf{x}_{obs} \setminus t_{obs} \propto \frac{\pi(\mu \,|\, \boldsymbol{\theta}, \tau^2, \mathbf{x}_{obs})}{f(t_{obs} \,|\, \boldsymbol{\theta})} \propto \pi(\mu \,|\, \boldsymbol{\theta}, \tau^2, \mathbf{x}_{obs}) \tag{3.9}$$

$$\tau^2 \,|\, \boldsymbol{\theta}, \mu, \mathbf{x}_{obs} \setminus t_{obs} \propto \frac{\pi(\tau^2 \,|\, \boldsymbol{\theta}, \mu, \mathbf{x}_{obs})}{f(t_{obs} \,|\, \boldsymbol{\theta})} \propto \pi(\tau^2 \,|\, \boldsymbol{\theta}, \mu, \mathbf{x}_{obs}) \tag{3.10}$$

$$\boldsymbol{\theta} \,|\, \mu, \tau^2, \mathbf{x}_{obs} \setminus t_{obs} \propto \frac{\pi(\boldsymbol{\theta} \,|\, \mu, \tau^2, \mathbf{x}_{obs})}{f(t_{obs} \,|\, \boldsymbol{\theta})} \,. \tag{3.11}$$

The full conditionals (3.9) and (3.10) are identical to (3.6) and (3.7), respectively, and hence they are easy to simulate from. On the other hand, (3.11) is not of close form, and we use Metropolis-Hastings within Gibbs for the full conditional of each θ_i

$$\pi_{ppp}(\theta_i \,|\, \mu, \tau, \boldsymbol{\theta}_{-i}, \mathbf{x}_{obs} \setminus t_{obs}) \propto \frac{\pi_{post}(\theta_i \,|\, \mu, \tau^2, \mathbf{x}_{obs})}{f(t_{obs} \,|\, \boldsymbol{\theta})} \propto \frac{N(\theta_i \,|\, E_i, V_i)}{f(t_{obs} \,|\, \boldsymbol{\theta})} \,, \tag{3.12}$$

where E_i, V_i are given in (3.8). Next we need to find a good proposal to simulate from (3.12). An obvious proposal would simply be the posterior $\pi_{post}(\theta_i \,|\, \mu, \tau^2, \mathbf{x}_{obs})$, but this can be a very bad proposal when the data is indeed 'surprising' for the entertained model. In particular, the posterior distribution centers around the MLE $\widehat{\boldsymbol{\theta}}$ while the partial posterior centers around the *conditional* MLE, $\widehat{\boldsymbol{\theta}}_{cMLE}$, that is,

$$\widehat{\boldsymbol{\theta}}_{cMLE} = \arg\max f(\mathbf{x}_{obs} \,|\, t_{obs}, \boldsymbol{\theta}) = \arg\max \frac{f(\mathbf{x}_{obs} \,|\, \boldsymbol{\theta})}{f(t_{obs} \,|\, \boldsymbol{\theta})}.$$

It is intuitively obvious that, when the data is not 'surprising', that is, when t_{obs} comes from the 'null' model, then $f(\mathbf{x}_{obs} \,|\, t_{obs}, \boldsymbol{\theta})$ would be similar to $f(\mathbf{x}_{obs} \,|\, \boldsymbol{\theta})$ and the partial and posterior distributions would also be similar. However, when data is 'surprising' and t_{obs} is not a 'typical' value, then the 'null" model and the conditional model can be considerably different, as well as the corresponding MLE's. For Metropolis proposals, Bayarri and Berger (2000) then suggest generating from the posterior distribution but then 'moving' the generated values closer to the mode of the target distribution (the partial posterior) by adding

$$\widehat{\boldsymbol{\theta}}_{cMLE,i} - \widehat{\boldsymbol{\theta}}_{MLE,i},$$

possibly multiplied by a Uniform(0,1) random generation.

To avoid computation of $\widehat{\boldsymbol{\theta}}_{cMLE}$, which can be rather time consuming, we use instead an estimate $\widetilde{\boldsymbol{\theta}}_c$ which we expect to be close enough (for our purposes) to $\widehat{\boldsymbol{\theta}}_{cMLE}$ for this model and this T (see Bayarri and Morales, 2003). In particular, we take all components to be equal and given by

$$\widetilde{\theta}_c = \frac{\sum_{l=1}^{I-1} \overline{X}_{(l\cdot)}}{I-1} \ ,$$

where $(\overline{X}_{(1\cdot)}, \ldots, \overline{X}_{(I\cdot)})$ denote the group means sorted in ascendent order. That is, we simple remove the largest sample mean and then average (we could have also used a weighted average if the sample sizes were very different).

Then, the resulting algorithm to simulate from (3.12) at stage k, given the (simulated) values $(\boldsymbol{\theta}^k_{-i}, \theta^k_i, \mu^k, \tau^{2(k)})$, is:

1. Simulate $\theta^*_i \sim N(\theta_i \,|\, E_i, V_i)$.
2. Move the simulation θ^*_i to
$$\widetilde{\theta}^*_i = \theta^*_i + U \cdot (\widetilde{\theta}_c - \widetilde{\theta}_{MLE,i})$$
where U is random number in (0,1).
3. Accept candidate $\widetilde{\theta}^*_i$ with probability:

$$\alpha = \min \left\{ 1, \frac{N(\widetilde{\theta}^*_i \,|\, E_i, V_i) N(\theta^k_i \,|\, E_i, V_i) f(t_{obs} \,|\, \boldsymbol{\theta}^k_{-i}, \widetilde{\theta}^k_i)}{N(\widetilde{\theta}^k_i \,|\, E_i, V_i) N(\theta^*_i \,|\, E_i, V_i) f(t_{obs} \,|\, \boldsymbol{\theta}^k_{-i}, \widetilde{\theta}^*_i)} \right\}$$

3.4 Examples

For illustration, we now compute the measures of surprise, that is, the p-values and the Relative Predictive Surprise indexes for the different proposals. We use a couple of data sets with 5 groups and 8 observations in each group. In both of them the null model is not the model generating the data; in *Example 1* one of the means comes from a different normal with a larger mean, whereas in *Example 2* the means come from a Gamma distribution. Recall that the null model (3.1) had the group means i.i.d normal.

Example 1. The group means are 1.560, 0.641, 1.982, 0.014, 6.964, simulated from:

$$\begin{aligned} X_{ij} &\sim N(\theta_i, 4) \quad i = 1, \ldots, 5, \ j = 1, \ldots, 8 \\ \theta_i &\sim N(1, 1) \quad j = 1, \ldots, 4 \\ \theta_5 &\sim N(5, 1) \end{aligned}$$

Example 2. The group means are: 0.749, 0.769, 5.770, 1.856, 0.753, simulated from:

$$\begin{aligned} X_{ij} &\sim N(\theta_i, 4) \qquad i = 1, \ldots, 5,\ j = 1, \ldots, 8 \\ \theta_i &\sim Ga(0.6, 0.2) \quad i = 1, \ldots, 5 \end{aligned}$$

	p^{EB}_{prior}	SPR^{EB}_{prior}	p^{EB}_{post}	SPR^{EB}_{post}	p_{post}	SPR_{post}	p_{ppp}	SPR_{ppp}
Example 1	0.130	0.456	0.347	0.941	0.409	0.951	0.010	0.025
Example 2	0.121	0.426	0.301	0.886	0.381	0.954	0.011	0.036

Table 1: p-values and SPR for Examples 1 and 2.

In Table 1 we show all measures of surprise for the two examples. The partial posterior measures clearly detect the inadequate models, with very small p-values and RPS. On the other hand, none of the other predictive distributions work well for this purpose, no matter how we choose to locate the observed t_{obs} in them (with p-values or RPS). The prior empirical Bayes are conservative, with p and RPS an order of magnitude larger than the ones produced by the partial posterior predictive distribution. Both, the posterior empirical Bayes and predictive posterior measures are extremely conservative, indicating almost perfect agreement of the observed data with the quite obviously wrong null models. Besides, it can be seen that EB posterior and posterior predictive measures are very similar to each other. This is not a specific feature of these examples, but occurs very often. We further explore it in a rather simple null model in Section 4.

We next study the behavior of the different p-values, when considered as a function of $\boldsymbol{X}$, under the null and under some alternatives.

3.5 Null sampling distribution of the p-values

A crucial characteristic of p values is that, when considered as random variables, $p(\boldsymbol{X})$ have $U(0,1)$ distributions under the null models (at least asymptotically). This endorses p-values with a very desirable property, namely having the same interpretation across models. The uniformity of p-values has often taken as their defining characteristic (more discussion and references can be found in Bayarri and Berger, 2000).

In this section we simulate the null sampling distribution of $p^{EB}_{prior}(\mathbf{X})$, $p_{post}(\mathbf{X})$ y $p_{ppp}(\mathbf{X})$, when $\mathbf{X}$ comes from a hierarchical normal-normal model as defined in (3.1). (We do not show the behavior of $p^{EB}_{post}(\mathbf{X})$ because it is bassically identical to that of $p_{post}(\mathbf{X})$.)

In particular, we have simulated 1000 samples from the following model:

$$\begin{aligned} X_{ij} &\sim N(\theta_i, 4) \qquad i = 1, \ldots, I,\ j = 1, \ldots, 8\ , \\ \theta_i &\sim N(0, 1) \qquad i = 1, \ldots, I\ . \end{aligned}$$

We have considered three different 'group sizes': $I = 5$, 15 and 25. (Since here we are checking the distribution of the means, the adequate "asymptotics" is in the number of groups.)

We compute the different p-values for 1000 simulated samples. Figure 1 shows the resulting histograms. As we can see, $p_{ppp}(\mathbf{X})$ has already a (nearly) uniform distribution even for I (number of groups) as small as 5. On the other hand, the distributions of both $p^{EB}_{prior}(\mathbf{X})$ and $p_{post}(\mathbf{X})$ are quite far from uniformity, the distribution of $p_{post}(\mathbf{X})$ being the farthest. Moreover, the deviation from the $U(0, 1)$ is in the direction of more conservatism (given little probability to small p values, and concentrating around 0.5), as it is the case in simpler models. Notice that conservatism usually results in lack of power (and thus in not being able to detect data coming from wrong models). Particularly worrisome is the behavior of $p_{post}(\mathbf{X})$ for small number of groups. We have also performed similar simulations for larger I's (number of groups) to investigate whether the distribution of these p-values approaches uniformity as I grows. In Figure 2 we show the histograms for $I = 100$ and $I = 200$ of p-values $p_{post}(\mathbf{X})$ and $p^{EB}_{prior}(\mathbf{X})$ (we do not show $p_{ppp}(\boldsymbol{X})$ as it is virtually uniform). The distributions of these p-values do not seem to change much as I is doubled from $I = 100$ to $I = 200$, and they are still quite far from uniformity, still showing a tendency to concentrate around middle values for p. We do not know whether these p-values are asymptotically $U(0, 1)$.

3.6 Distribution of p-values under some alternatives

In this section we study the behavior of $p^{EB}_{prior}(\mathbf{X})$, $p_{post}(\mathbf{X})$ y $p_{ppp}(\mathbf{X})$, when the 'null' normal-normal model is wrong. In particular, we focus on violations of normality at the second level.

Specifically, we simulate data sets from three different models. In all the three, we take the distribution at the first level to be the same and in agreement with the first level in the null model (3.1):

$$X_{ij} \sim N(\theta_i, \sigma^2 = 4) \quad i = 1, \ldots, I, j = 1, \ldots, 8$$

We use three different distributions for the group means (remember, under the null model, the θ_i's were normal):

1. Exponential distribution: $\theta_i \sim Exp(1), \quad i = 1, \ldots, I.$
2. Gumbel distribution: $\theta_i \sim Gumbel(0, 2), \quad i = 1, \ldots, I$, where the Gumbel($\alpha$, β) density is
$$f(x \,|\, \alpha, \beta) = \frac{1}{\beta} exp\left(-\frac{x-\alpha}{\beta}\right) exp\left(-exp\left(-\frac{x-\alpha}{\beta}\right)\right) \quad \text{for} \quad -\infty < x < \infty$$

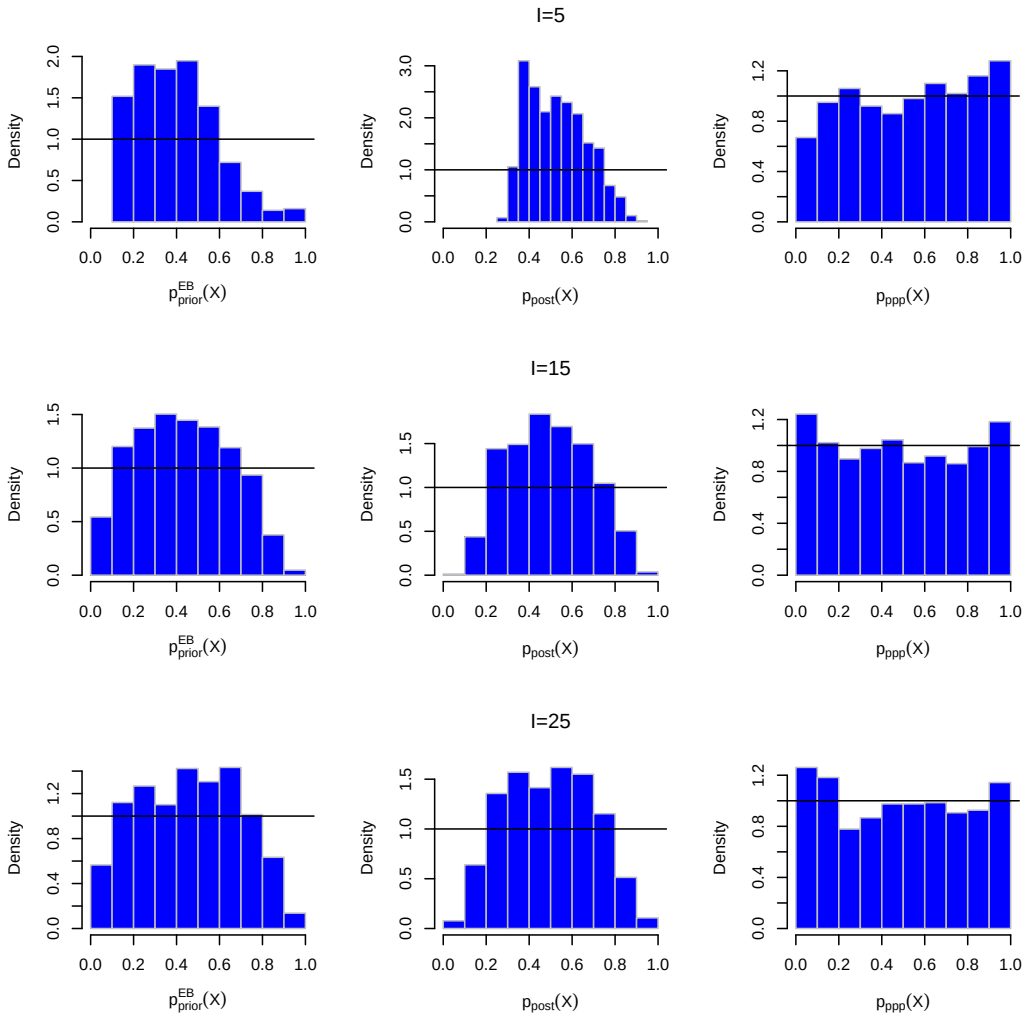

Figure 1: Null distribution of $p^{EB}_{prior}(\mathbf{X})$ (first column), $p_{post}(\mathbf{X})$ (second column) and $p_{ppp}(\mathbf{X})$ (third column) for $I = 5$ (first row), 10 (second row) and 15 (third row).

3. Log-Normal distribution: $\theta_i \sim LogNormal(0, 1) \quad i = 1, \ldots, I.$

We have considered $I = 5$ and $I = 10$, simulated 1000 samples from each model and computed the different p-values for each sample. In Table 2 we show $Pr(p \leq \alpha)$ for the three $p-values$ and some values of α. p_{ppp} seems to show adequate power (lower for the exponential alternative, and largest for the log-normal); both p^{EB}_{prior} and p_{post} show considerable lack of power in comparison. In particular, notice the extreme low power of p_{post} in all instances, producing basically no p-values smaller than 0.2.

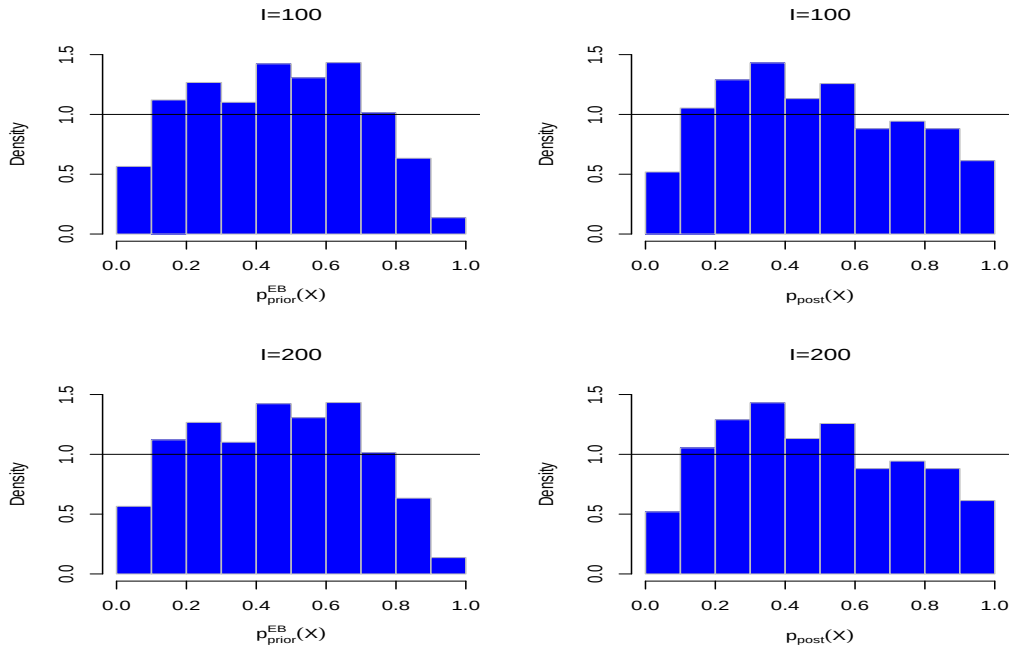

Figure 2: Null distribution of $p^{EB}_{prior}(\mathbf{X})$ and $p_{post}(\mathbf{X})$ when $I = 100$ (first row) and $I = 200$ (second row).

4 Testing $\mu = \mu_0$

As we have seen in Section 3, the specified predictive distributions for T (empirical Bayes, posterior and partial posterior) used to located the observed t_{obs} had to be dealt with by MC and MCMC methods. To gain understanding in the behavior of the different proposals to 'get rid' of the unknown parameters, we address here a simpler "null model" which results in simpler expressions and allows for easier comparisons.

Suppose that we have the normal-normal hierarchical model (3.1) (with σ_i^2 known) but that we are interested in testing:

$$H_0 : \mu = \mu_0.$$

A natural T to consider to investigate this H_0 is the grand mean:

$$T = \frac{\sum_{i=1}^{I} n_i \overline{X}_{i\cdot}}{\sum_{i=1}^{I} n_i},$$

where $\overline{X}_{i\cdot}$, $i = 1, \ldots, I$ are the sample means for the I groups. The (null) sampling distribution of T is:

$$T \mid \boldsymbol{\theta} \sim N(\mu_T, V_T), \quad \text{with} \quad \mu_T = \frac{\sum_{i=1}^{I} n_i \theta_i/\sigma_i^2}{\sum_{i=1}^{I} n_i/\sigma_i^2}, \quad V_T = \frac{1}{\sum_{i=1}^{I} n_i/\sigma_i^2} \tag{4.1}$$

α	0.02	0.05	0.1	0.2	0.02	0.05	0.1	0.2
Normal-Exponencial								
	I=5				I=10			
p_{ppp}	0.040	0.083	0.148	0.243	0.124	0.196	0.288	0.419
p_{post}	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.046
p^{EB}_{prior}	0.000	0.000	0.000	0.235	0.002	0.063	0.176	0.369
Normal-Gumbel								
	I=5				I=10			
p_{ppp}	0.124	0.219	0.322	0.462	0.208	0.314	0.425	0.550
p_{post}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003
p^{EB}_{prior}	0.000	0.000	0.000	0.268	0.001	0.067	0.187	0.383
Normal-Lognormal								
	I=5				I=10			
p_{ppp}	0.160	0.220	0.310	0.415	0.319	0.416	0.500	0.611
p_{post}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.020
p^{EB}_{prior}	0.000	0.001	0.00 2	0.235	0.013	0.064	0.126	0.235

Table 2: $Pr(\mathrm{p} \leq \alpha)$ for p_{ppp}, p_{post}, y p^{EB}_{prior}, for different values of I and the three alternative models.

Again we will integrate $\boldsymbol{\theta}$ out from (4.1) with the previous proposals and compare the resulting predictive distributions for T, $h(t)$, and the corresponding measures of surprise (which we take relative to μ_0):

$$p = Pr^{h(\cdot)}(|t(\mathbf{X}) - \mu_0| \geq |t(\mathbf{x}_{obs}) - \mu_0|), \tag{4.2}$$

$$RPS = \frac{h(t(\mathbf{x}_{obs}))/h(\mu_0)}{\sup_t h(t)/h(\mu_0)} . \tag{4.3}$$

4.1 Empirical Bayes Distributions

In this case the integrated likelihood for τ^2 is simply given by (3.4) with μ replaced by μ_0, from which $\hat{\tau}^2$ the m.l.e. of τ^2 can be computed. For this null model, it is possible to derive close-form expressions for the prior and posterior empirical Bayes distributions given in (2.2) and (2.4) respectively.

Indeed, the joint Empirical Bayes prior predictive for $\bar{\boldsymbol{X}} = (\bar{X}_{1\cdot}, \ldots, \bar{X}_{I\cdot})$ is

$$m^{EB}_{prior}(\bar{\boldsymbol{x}}) = \prod_{i=1}^{I} N\left(\bar{x}_{i\cdot} \,|\, \mu_0, \frac{\sigma_i^2}{n_i} + \hat{\tau}^2\right) ,$$

so that the corresponding distribution for T, $m^{EB}_{prior}(t)$, is normal with mean and variance given by

$$E^{EB}_{prior} = \mu_0, \qquad V^{EB}_{prior} = \frac{1}{(\sum_{i=1}^{I} n_i)^2} \sum_{i=1}^{I} n_i^2 \left(\frac{\sigma_i^2}{n_i} + \hat{\tau}^2 \right) . \tag{4.4}$$

The Empirical Bayes *posterior* predictive distribution $m^{EB}_{post}(\bar{\boldsymbol{x}})$ can be derived in a similar manner resulting also in a normal $m^{EB}_{post}(t)$ with mean and variance:

$$E^{EB}_{post} = \frac{\sum_{i=1}^{I} n_i \widetilde{E}_i}{\sum_{i=1}^{I} n_i}, \qquad V^{EB}_{post} = \frac{1}{(\sum_{i=1}^{I} n_i)^2} \sum_{i=1}^{I} n_i^2 \left(\frac{\sigma_i^2}{n_i} + \widetilde{V}_i \right), \tag{4.5}$$

where

$$\widetilde{E}_i = \frac{n_i \overline{x}_{i\cdot}/\sigma_i^2 + \mu_0/\hat{\tau}^2}{n_i/\sigma_i^2 + 1/\hat{\tau}^2} \quad \text{and} \quad \widetilde{V}_i = \frac{1}{n_i/\sigma_i^2 + 1/\hat{\tau}^2}$$

The measures of surprise (4.2) and (4.3) can also be computed in close form. The (prior) Empirical Bayes measures are

$$p^{EB}_{prior} = 2 \cdot \left(1 - \Phi \left(\frac{|t_{obs} - \mu_0|}{\sqrt{V^{EB}_{prior}}} \right) \right) , \qquad RPS^{EB}_{prior} = exp \left\{ - \frac{(t_{obs} - \mu_0)^2}{2V^{EB}_{prior}} \right\} ,$$

where Φ denotes the standard normal distribution function. The *posterior* Empirical Bayes measures can similarly be derived in close-form, but they are of much less interest and we do not produce them here (see Castellanos, 2002).

The inadequacies of m^{EB}_{post} for testing the null model can already be seen in the above formulae, but they are more evident in the particular homocedastic, balanced case: $\sigma_i^2 = \sigma^2$ and $n_i = n \; \forall \; i, \; i = 1, \ldots, I$. In this case the distribution of T simplifies to:

$$T \; \sim \; N \left(\frac{\sum_{i=1}^{I} \theta_i}{I}, \frac{\sigma^2}{In} \right).$$

Also, there is a closed-form expression for the m.l.e. of τ^2:

$$\hat{\tau}^2 = \max \left\{ 0, \frac{\sum_{i=1}^{I} (\overline{x}_{i\cdot} - \mu_0)^2}{I} - \frac{\sigma^2}{n} \right\}.$$

Then, the mean and variance of m^{EB}_{prior}, as given in (4.4) are

$$E^{EB}_{prior} = \mu_0, \qquad V^{EB}_{prior} = \frac{\frac{\sigma^2}{n} + \hat{\tau}^2}{I}.$$

Similarly, the mean and variance of m^{EB}_{post}, given in (4.5) reduce to

$$E^{EB}_{post} = \frac{n t_{obs}/\sigma^2 + \mu_0/\hat{\tau}^2}{n/\sigma^2 + 1/\hat{\tau}^2}, \qquad V^{EB}_{post} = \frac{2n\sigma^2\hat{\tau}^2 + \sigma^4}{nI(n\hat{\tau}^2 + \sigma^2)} .$$

For a given μ_0 (and fixed τ), it is now easy to investigate the behavior of m^{EB}_{prior} and m^{EB}_{post} as $t_{obs} \to \infty$, indicating flagrant incompatibility between the data and H_0. m^{EB}_{prior} centers at μ_0, which in principle allows for declaring incompatible a very large value t_{obs}; however, the variance also grows to ∞ as t_{obs} grows, thus alleviating the incompatibility, and maybe 'missing' some surprisingly large t_{obs}. Thus, the behavior of m^{EB}_{prior} is reasonable, but might be conservative. On the other hand, the behavior of m^{EB}_{post} is completely inadequate. Indeed, for very large t_{obs}, it centers precisely at t_{obs} thus precluding detecting as unusual *any* value t_{obs}, no matter how large!. Moreover, the variance is seen to go to $(2\sigma^2)/(n\,I)$, a finite constant. It is immediate to see that m^{EB}_{post} should not be used to check this particular (and admittedly simple) model; as a matter of fact, for $t_{obs} \to \infty$ (extremely inadequate models) we expect p-values of around 0.5. We remark that the previous argument does not belong to any particular measure of surprise, rather it reflects the inadequacy of m^{EB}_{post} for model checking, whatever measure of surprise we use. Note also that we expect similar inadequacies to occur with the posterior predictive distribution.

4.2 Posterior Distribution

No major simplifications occur for this specific H_0. The posterior distribution is not of close-form (nor even for the homocedastic, balanced case), and hence neither is the posterior predictive distribution. We can however easily generate from it with virtually the same Gibbs sampler used in Section 3.2: if suffices to (obviously) ignore the full conditional for μ and replace μ with the value μ_0 in the other two full conditionals (3.7) and (3.8), which were standard distributions.

4.3 Partial Posterior Distribution

There is no close-form expression for the partial posterior distribution either, but considerably simplification occurs since the Metropolis-within-Gibbs step is no longer needed and a straight Gibbs sampler suffices. The full conditional for τ^2 is as given in (3.10) with μ replaced by μ_0; the full conditional of each θ_i is here also normal:

$$\pi(\theta_i \,|\, \tau^2, \boldsymbol{\theta}_{-i}, \mathbf{x}_{obs} \setminus t_{obs}) = N(\theta_i \mid E_i^0, V_i^0)$$

where

$$E_i^0 = \frac{1}{V_i^0}\left[\frac{n_i}{\sigma_i^2}\left(\overline{x}_{i\cdot} - t_{obs} + \frac{\sum_{l\neq i} n_l\,\theta_l/\sigma_l^2}{\sum_j n_j/\sigma_j^2}\right) + \frac{1}{\tau^2}\,\mu_0\right], \tag{4.6}$$

$$\frac{1}{V_i^0} = \frac{n_i}{\sigma_i^2} + \frac{1}{\tau^2} - \frac{n_i^2}{\sigma_i^4 \sum_j n_j/\sigma_j^2}\,. \tag{4.7}$$

Details of the derivations appear in the Appendix.

4.4 Some examples

We next consider four examples in which we carry out the testing $H_0: \mu = 0$. We consider $I = 8$ groups, with $n = 12$ observations per group, and $\sigma^2 = 4$. In one of the examples (*Example 1*) H_0 is true and the means θ_i are generated from a $N(0, 1)$. In the remaining three examples the null H_0 is wrong, with $\theta_i \sim N(1.5, 1)$ in *Example 1*, $\theta_i \sim N(2.5, 1)$ in *Example 2*, and $\theta_i \sim N(2.5, 3)$ in *Example 3*. The simulated sample means are:

Example 1: $\overline{\mathbf{x}}$ = (-2.18, -1.47, -0.87, -0.38, 0.05, 0.29, 0.96, 2.74).

Example 2: $\overline{\mathbf{x}}$ = (-0.05, 0.66, 1.37, 1.70, 1.72, 2.14, 2.73, 3.68).

Example 3: $\overline{\mathbf{x}}$ = (1.53, 1.65, 1.71, 1.75, 1.87, 2.16, 2.47, 3.68).

Example 4: $\overline{\mathbf{x}}$ = (0.50, 1.52, 1.59, 2.73, 2.88, 3.54, 4.21, 5.86).

In Figure 3 we show the predictive distributions for all proposals in the 4 examples. A quite remarkable feature is that in every occasion, m^{EB}_{post} basically coincides with m_{post}, so much that they can hardly be told apart; We were expecting them to be close, but not so close. Also, when the null is true (Example 1), all distributions rightly concentrate around the null and, as expected, the most concentrated is m^{EB}_{post} (and m_{post}), and the least the m_{ppp} (m^{EB}_{prior} ignores the uncertainty in the estimation of τ^2). When the null model is wrong, however, even though both m_{ppp} and m^{EB}_{prior} have the right location, m_{ppp} is more concentrated than m^{EB}_{prior}, thus indicating more promise in detecting extreme t_{obs}; Notice the hopeless (and identical) behavior of m^{EB}_{post} and m_{post}: both concentrate around t_{obs}, no matter how extreme; that is, there is no hope that it can detect incompatibility of a very large t_{obs} with the hypothetical value of 0.

In Table 3 we show the different measures of surprise for the four Examples. All behave well when the null is true, but only the ppp and the prior empirical Bayes measures detect the wrong models (ppp more clearly, thus anticipating greater power). On the other hand m^{EB}_{post} and m_{post} produce very similar measures and both are incapable of detecting clearly unappropriate null models. Notice that the conservatism of the posterior predictive measures (and the posterior empirical Bayes ones) is extreme.

5 A Comparison with other Bayesian methods for Model Checking

In this section we retake the main goal of checking the adequacy of the hierarchical model:

$$\begin{aligned} X_{ij} \,|\, \theta_i &\overset{i}{\sim} N(\theta_i, \sigma^2) \qquad\qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i, \\ \pi(\boldsymbol{\theta} \,|\, \mu, \tau) &= \prod_{i=1}^{I} N(\theta_i \,|\, \mu, \tau^2), \end{aligned}$$

	Example 1		Example 2		Example 3		Example 4	
	p	SPR	p	SPR	p	SPR	p	SPR
ppp	0.859	0.966	0.008	0.016	0.000	0.000	0.005	0.009
EB prior	0.831	0.977	0.016	0.056	0.007	0.026	0.013	0.047
EB post	0.711	0.998	0.313	0.888	0.305	0.879	0.378	0.953
post	0.712	0.991	0.333	0.945	0.325	0.919	0.392	0.995

Table 3: p-values and RPS for testing $\mu = 0$ in the four examples.

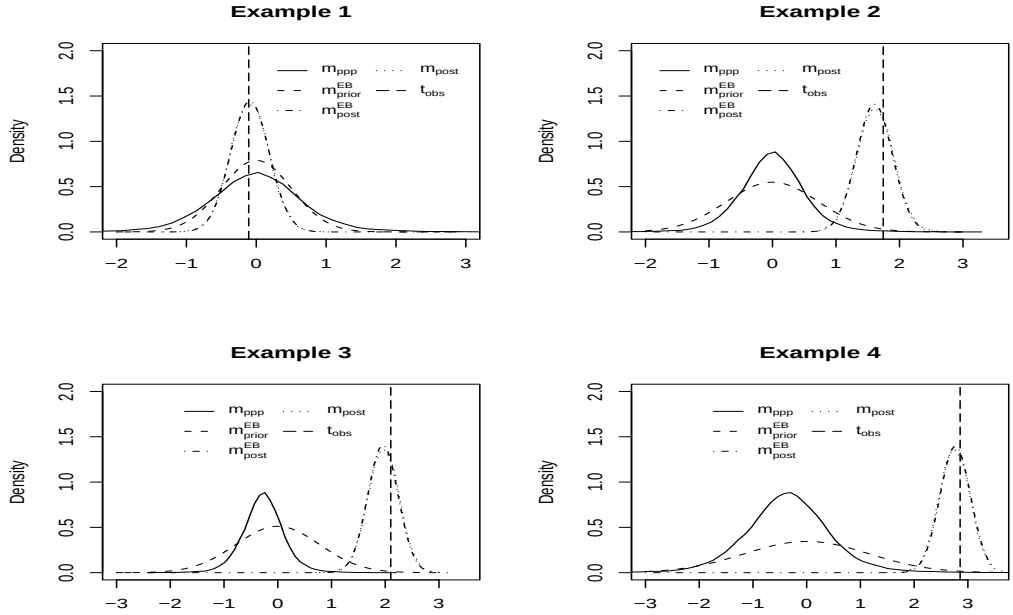

Figure 3: Different predictive distribution for T in each example.

with σ^2 unknown, as well as μ, τ^2. We first provide some few details needed to derive the MS used so far when σ^2 is unknown, then we briefly review three recent methods for Bayesian checking of hierarchical models, proposed in Dey et al.(1998), O'Hagan(2003) and Marshall and Spiegelhalter (2001). We do not specifically address here (because the philosophy is somewhat different) the much earlier, likelihood/Empirical Bayes proposal of Lange and Ryan (1989), which basically consists in checking the normality of some standardized version of estimated residuals. We apply the four methods considered so far and the three new methods to a data set proposed by O'Hagan (2003).

O'Hagan (2003) Example:

In the general scenario of checking the normal-normal hierarchical model, O'Hagan (2003) uses the following data set:

Group 1	2.73,	0.56,	0.87,	0.90,	2.27,	0.82.	$\overline{x}_{1\cdot}= 1.36$
Group 2	1.60,	2.17,	1.78,	1.84,	1.83,	0.80.	$\overline{x}_{2\cdot}= 1.67$
Group 3	1.62,	0.19,	4.10,	0.65,	1.98,	0.86.	$\overline{x}_{3\cdot}= 1.57$
Group 4	0.96,	1.92,	0.96,	1.83,	0.94,	1.42.	$\overline{x}_{4\cdot}= 1.34$
Group 5	6.32,	3.66,	4.51,	3.29,	5.61,	3.27.	$\overline{x}_{5\cdot}= 4.44$

Note that $\overline{x}_{5\cdot}$ is considerably far from the other 4 sample means. □

5.1 Methods used so far.

The empirical Bayes methods (both the prior and the posterior) have an easy generalization to the unknown σ^2 case. It suffices to substitute σ^2 by its usual m.l.e. estimate and apply the procedures in Section 3 for σ^2 known.

For both, the posterior predictive and the partial posterior predictive measures, we need to specify a new (non-informative) joint prior. Since we can use the standard non-informative prior for σ^2, we take:

$$\pi(\mu, \sigma^2, \tau^2) \;\propto\; \frac{1}{\sigma^2}\,\frac{1}{\tau}. \tag{5.1}$$

To simulate from the posterior distribution, we again use Gibbs sampling. The full conditionals for $\boldsymbol{\theta}$, μ and τ^2 are the same as for the known σ^2 and they are given in (3.8), (3.6) and (3.7), respectively. The full conditional for the new parameter, σ^2, is:

$$\sigma^2 \,|\, \boldsymbol{\theta}, \mu, \tau^2, \mathbf{x}_{obs} \sim \chi^{-2}(m, \widetilde{\sigma}^2) \;,$$

where

$$m = \sum_{i=1}^{I} n_i \;, \qquad \widetilde{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(x_{ij} - \theta_i)^2 \;.$$

The (joint) partial posterior distribution is

$$\pi_{ppp}(\boldsymbol{\theta}, \sigma^2, \mu, \tau^2 \,|\, \mathbf{x}_{obs} \setminus t_{obs}) \propto \frac{\pi(\boldsymbol{\theta}, \sigma^2, \mu, \tau^2 | \mathbf{x}_{obs})}{f(t_{obs} \,|\, \boldsymbol{\theta}, \sigma^2)},$$

and again we use the same general procedure as for the σ^2 known scenario (see Section 3). We only need to specify how to simulate from the full conditional of σ^2:

$$\pi_{ppp}(\sigma^2 \,|\, \boldsymbol{\theta}, \mu, \tau^2, \mathbf{x}_{obs} \setminus t_{obs}) \propto \frac{\chi^{-2}(m, \widetilde{\sigma}^2)}{f(t_{obs} \,|\, \boldsymbol{\theta}, \sigma^2)} \;.$$

We use Metropolis-Hastings with $\chi^{-2}(m, \widetilde{\sigma}^2)$ as proposal distribution. The acceptance probability (at stage k) of candidate $\sigma^{2\,*}$, given the simulated values $(\boldsymbol{\theta}^{(k)}, \sigma^{2\,(k)}, \mu^{(k)}, \tau^{2\,(k)})$ is:

$$\alpha = min\left\{1, \frac{f(t_{obs}|\boldsymbol{\theta}^{(k)}, \sigma^{2\,(k)})}{f(t_{obs}|\boldsymbol{\theta}^{(k)}, \sigma^{2\,*})}\right\} .$$

We next derive the different measures of surprise for O'Hagan data.

O'Hagan (2003) Example (cont.) :

The empirical Bayes, posterior predictive and partial posterior predictive measures of surprise applied to this data set are

p^{EB}_{prior}	SPR^{EB}_{prior}	p^{EB}_{post}	SPR^{EB}_{post}	p_{post}	SPR_{post}	p_{ppp}	SPR_{ppp}
0.195	0.625	0.371	0.972	0.405	0.977	0.015	0.036

Table 4: MS (σ^2 unknown) for O'Hagan data set.

We again observe the same behavior as the one repeatedly observed in previous examples: in spite of such an 'obvious' data set, only the partial posterior measures detect the incompatibility between data and model; the empirical Bayes prior measures are too conservative, and the posterior predictive measures (and their very much alike empirical Bayes posterior ones) are completely hopeless. □

5.2 Simulation-based model checking

This method is proposed in Dey et al. (1998), as a computationally intense method for model checking. This methods works not only with checking statistics T, but more generally, with discrepancy measures d, that is, with functions of the parameters and the data; this feature also applies to the posterior predictive checks that we have been considering all along. In essence, the method consists in comparing the posterior distribution $d\,|\,\mathbf{x}_{obs}$ with R posterior distributions of d given R data sets $\mathbf{x}^r$, for $r = 1, \ldots, R$, generated from the (null) *predictive* model; note that the method *requires* proper priors. Comparison is carried out via *Monte Carlo Tests* (Besag and Clifford, 1989).

Letting $\mathbf{x}^r$, for $r = 0$ denote the observed data $\mathbf{x}_{obs}$, their algorithm is as follows:

- For each posterior distribution of d given $\mathbf{x}^r$, $r = 0, \ldots R$, compute the vector of quantiles $\mathbf{q}^{(r)} = (q^{(r)}_{.05}, q^{(r)}_{.25}, q^{(r)}_{.5}, q^{(r)}_{.75}, q^{(r)}_{.95})$.

- Compute the vector $\overline{\mathbf{q}}$ of averages, over r, of these quantiles: $\overline{\mathbf{q}} = (\overline{q}_{.05}, \overline{q}_{.25}, \overline{q}_{.5}, \overline{q}_{.75}, \overline{q}_{.95})$.

- Compute the $r+1$ Euclidean distances between $\mathbf{q}^{(r)}$, $r = 0, 1, \ldots, R$ and $\overline{\mathbf{q}}$.
- Perform a 0.05 one-sided, upper tail Monte Carlo test, that is, check whether or not the distance corresponding to the original data is smaller than the 95% percentile of the $r+1$ distances.

In reality, this method is not a competitor of the ones we have been considering previously, since it *requires* proper priors, and hence is not available for objective model checking. We, however, apply it also to O'Hagan data.

O'Hagan (2003) Example (cont.) :

In order to perform the *Simulation-based model checking*, we need proper priors. We use the ones proposed in O'Hagan (2001):

$$\mu \sim N(2, 10), \quad \sigma^2 \sim 22W, \quad \tau^2 \sim 6W, \quad \text{where } W \sim \chi^{-2}_{20}. \tag{5.2}$$

Along with the statistic used so far, we have also considered a measure of discrepancy which in this case is just a function of the parameters:

$$T_1 = \max \overline{X}_{i\cdot} \ , \qquad T_2 = \max |\theta_i - \mu| \ .$$

With 1000 simulated data sets from the null, the results are shown in Table 5. It can be seen that, with the given prior, incompatibility is detected with T_2, but not with T_1. We do not know whether T_2 would detect incompatibility with other priors (see related results in Section 5.3).

	$\|\mathbf{q}^{(0)} - \overline{\mathbf{q}}\|$	.95 quantile
T_1	2.313	13.463
T_2	1.823	0.808

Table 5: Euclidean distance between $\mathbf{q}^{(0)}$ and $\overline{\mathbf{q}}$ and the 0.95 quantile of all distances.

□

5.3 O'Hagan method

O'Hagan (2003) proposes a general method to investigate adequacy of graphical models at each node. We will not describe his method in full generality, but only in how it applies to checking the second level of our normal-normal hierarchical model.

To investigate conflict between the data and the normal assumption for each of the group means, the proposal investigates conflict between the likelihood for θ_i

$$\prod_{i=1}^{n_i} f(x_{ij} \,|\, \theta_i, \sigma^2),$$

and the (null) density for θ_i:

$$\pi(\theta_i \,|\, \mu, \tau^2).$$

To check conflict between two known univariate densities/likelihoods, O'Hagan proposes a "measure of conflict" based on their relative heights at an 'intermediate' value. Specifically, the likelihoods/densities are first normalized so that their maximum height is 1 (notice that this is equivalent to dividing by their respective maximum, as in RPS before). Then the (common) density height, z, at the intersection point between the two modes is computed. The proposed measure of conflict is $c = -2 \ln z$. For the particular case of comparing two normal distributions, $N(\omega_i, \gamma_i^2)$, for $i = 1, 2$, this measure is:

$$c = \left(\frac{\omega_1 - \omega_2}{\sqrt{\gamma_1} + \sqrt{\gamma_2}} \right)^2 . \tag{5.3}$$

O'Hagan indicates that a conflict measure smaller than 1 should be taken as indicative of no conflict, whereas values of 4 or larger would indicate clear conflict. No indication is given for values lying between 1 and 4.

When, as usual, the distributions involved depend on unknown parameters, the measures of conflict (in particular (5.3)), can not be computed. O'Hagan proposal is then to use the median of their *posterior* distribution. Notice that this is closely related to computing a relative height on the posterior predictive distribution and, hence, the concern exists that it can be too conservative for useful model checking. In fact this conservatism was highlighted in the discussions by Bayarri(2003) and Gelfand(2003).

Interestingly enough, O'Hagan defends use of *proper* priors for the unknown parameters, so neither posterior predictive nor posterior distributions are needed for implementation of his proposal (since the prior predictives and priors are proper). Alternatively, if one wishes to insist on using posterior distributions (instead of the, more natural, prior distributions), then proper priors are no longer needed, and the method can thus be generalized. Accordingly, we also apply his proposal with the non-informative prior (5.1).

O'Hagan (2003) Example (cont.):

We compute the measure (5.3) for the data set proposed by O'Hagan (2003). To derive the posterior distributions, we use both, the proper priors proposed by O'Hagan for this example, given in (5.2), and the non informative prior (5.1). The posterior medians for c are shown in Table 6. It can be seen that the results are very dependent on the prior used: the spurious group 5 is detected with the specific proper prior used , but not with the non-informative priors (thus suffering from the expected conservatism). We recall that data was clearly indicating an anomalous group 5.

□

	θ_1	θ_2	θ_3	θ_4	θ_5
O'Hagan prior	0.433	0.143	0.218	0.458	4.809
Non informative Prior	0.158	0.092	0.110	0.165	1.363

Table 6: Posterior medians of c_i, $i = 1, \ldots, 5$, for O'Hagan data set.

5.4 'Conflict' p-value

Marshall and Spiegelhalter (2001) proposed this approach based on, and generalizing, cross-validation methods (see Gelfand et al., 1992; Bernardo and Smith, 1994).

In cross-validation, to check adequacy of group i, data in group i, $\mathbf{X}_i$, is used to compute the 'surprise' statistic (or diagnostic measure), whereas the rest of the data, $\mathbf{X}_{-i}$, is used to train the improper prior. A *mixed* p-value is accordingly computed as:

$$p_{i,mix} = Pr^{m_{cross}(\cdot \,|\, \mathbf{X}_{-i})} \left(T_i \geq T_i^{obs}\right) , \tag{5.4}$$

where the completely specified distribution used to computed the i-th p-value is

$$m_{cross}(t_i \,|\, \mathbf{X}_{-i}) = \int f(t_i \,|\, \theta_i, \sigma^2)\, \pi(\theta_i \,|\, \mu, \tau^2)\, \pi(\mu, \tau^2, \sigma^2 \,|\, \mathbf{X}_{-i})\, d\boldsymbol{\theta},$$

and thus there is no double use of the data.

To avoid the issue of defining the statistic or discrepancy measure $T_i = T(\mathbf{X}_i)$ (which can be difficult for non-normal generalized linear models) Marshall and Spiegelhalter (2001) aim to preserving the cross-validation spirit while avoiding choice of a particular statistic or discrepancy measure $T_i = T(\mathbf{X}_i)$. Specifically, they propose use of *conflict* p-values for each group i, computed as follows:

- Simulate θ_i^{rep} from the posterior $\theta_i \,|\, \mathbf{X}_{-i}$.
- Simulate θ_i^{fix} from the posterior $\theta_i \,|\, \mathbf{X}_i$.
- Compute $\theta_i^{diff} = \theta_i^{rep} - \theta_i^{fix}$.
- Compute the 'conflict' p-value for group i, $i = 1, \ldots, I$ as

$$p_{i,\,con} = Pr(\theta^{diff} \leq 0 \,|\, \boldsymbol{x}) \ . \tag{5.5}$$

Marshall and Spiegelhalter (2001)) show that for location parameters θ_i, the conflict p-value (5.5) is equal to the cross-validation p-value (5.4) based on statistics $\hat{\theta}_i$ with symmetric likelihoods and using uniform priors in the derivation of θ_i^{fix}.

A clear disadvantage of this approach (as well as with the cross-validation mixed p-values) is that we have as many p-values as groups, and multiplicity might be an issue. (O'hagan measures

might suffer from it too.) Since we are dealing with p-values, adjustment is most likely done by classical methods (controlling either the family-wise error rate, as the Bonferroni method, or the false discovery rate and related methods, as the Benjamini and Hochberg, 1995, method). None of these methods is full proof and the danger exists that they also result in a lack of power.

O'Hagan (2003) Example (cont.):

We compute the *conflict* p-values for O'Hagan data set. We again use both, O'Hagan priors and non-informative priors. The results are shown in Table 7. Taken at face value, these p-values behave nicely and detect the outlying group. We have not explored any 'corrections' for multiplicity.

	Group 1	Group 2	Group 3	Group 4	Group 5
Non Informative priors	0.662	0.592	0.610	0.677	0.003
O'Hagan priors	0.842	0.737	0.733	0.879	0.003

Table 7: Conflict p-values for the O'Hagan data set using Non Informative Priors and O'Hagan Priors.

□

6 Conclusions

In this paper we have investigated the checking of hierarchical models from an objective Bayesian point of view (that is, introducing only the information in the data and model). We have explored several ways of eliminating the unknown parameters to derive 'reference' distributions. We have also explored different ways of characterizing 'incompatibility'. We propose use of the *partial posterior predictive measures*, which we compare with many other proposals. Some of our findings are:

- MS_{ppp} behave considerably better than the alternative MS^{EB}_{prior}, MS^{EB}_{post} and MS_{post}. The behavior of MS_{post} can be particularly bad with casually chosen T's, failing to reject clearly wrong models (but notice that the T we use is also proposed in Gelman et al. (1995, pages 181-182). As a matter of fact, the measures MS_{post} are very similar to the clearly inappropriate MS^{EB}_{post}.

- In our (limited) simulation study, the null sampling distribution of p_{ppp} is found to be approximately uniform, while these of p^{EB}_{prior} and p_{post} are far from uniformity. Also, p_{ppp} is the most powerful for the considered alternatives.

- The simulation-based model checking seems to work well in detecting the incompatibility between the model and the data, but it requires proper priors.

– O'Hagan method is highly sensitive to the prior chosen, and in fact it seems to be conservative with non-informative priors.

– The conflict p-values $p_{i,con}$ seems to work well, but they produce as many p-values as number of groups and multiplicity might be an issue.

7 Appendix

Derivation of the full conditional of θ_i in Section 4.3

The full condirional partial posterior density for θ_i:

$$\pi(\theta_i \,|\, \tau^2, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_I, \mathbf{x}_{obs} \setminus t_{obs}) \propto$$

$$\propto \quad \frac{\pi_{post}(\theta_i \,|\, \tau^2, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_I, \mathbf{x}_{obs})}{f(t_{obs} \,|\, \theta_1, \ldots, \theta_i, \ldots, \theta_I)}$$

$$\propto \quad exp\left\{-\frac{1}{2}\left(\frac{n_i}{\sigma_i^2}+\frac{1}{\tau^2}\right)\left(\theta_i - \frac{n_i\overline{x}_{i\cdot}/\sigma_i^2 + \mu_0/\tau^2}{n_i/\sigma_i^2 + 1/\tau^2}\right)^2\right\} exp\left\{\frac{1}{2}\underbrace{\sum_{l=1}^{I}\frac{n_l}{\sigma_l^2}}_{s}\left(t_{obs} - \frac{\sum_{j=1}^{I} n_j\theta_j/\sigma_j^2}{\sum_{j=1}^{I} n_j/\sigma_j^2}\right)^2\right\}$$

$$\propto \quad exp\left\{-\frac{1}{2}\left(\theta_i^2\left(\frac{n_i}{\sigma_i^2}+\frac{1}{\tau^2}\right) - 2\theta_i\left(\frac{n_i}{\sigma_i^2}\overline{x}_{i\cdot} + \frac{1}{\tau^2}\mu_0\right)\right)\right\} exp\left\{\frac{1}{2s}\left(st_{obs} - \frac{n_i}{\sigma_i^2}\theta_i - \sum_{l\neq i}\frac{n_l}{\sigma_l^2}\theta_l\right)^2\right\}$$

$$\propto \quad exp\left\{-\frac{1}{2}\left(\theta_i^2\left(\left(\frac{n_i}{\sigma_i^2}+\frac{1}{\tau^2}\right) - \frac{n_i^2}{\sigma_i^4 s}\right) - 2\theta_i\left(\frac{n_i}{\sigma_i^2}\overline{x}_{i\cdot} + \frac{1}{\tau^2}\mu_0 - \frac{n_i\left(\sum_{l\neq i} n_l\theta_l/\sigma_l^2 - st_{obs}\right)}{\sigma_i^2 s}\right)\right)\right\},$$

which, after some algebra, reduces to

$$\pi(\theta_i \,|\, \tau^2, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_I, \mathbf{x}_{obs} \setminus t_{obs}) \propto exp\left\{-\frac{1}{2V_i^0}(\theta_i - E_i^0)^2\right\},$$

whith E_i^0 and V_i^0 given in (4.6) and (4.7) respectively. The result then follows if V_i^0 can be shown to be greater than 0, which is equivalent to showing that

$$\begin{aligned}(V_i^0)^{-1} &= \frac{n_i}{\sigma_i^2} + \frac{1}{\tau^2} - \frac{n_i^2}{\sigma_i^4\sum_{j=1}^{I} n_j/\sigma_j^2} > 0 \\ &\Leftrightarrow \frac{n_i}{\sigma_i^2}\left(1 - \frac{n_i}{\sigma_i^2\sum_{j=1}^{I} n_j/\sigma_j^2}\right) + \frac{1}{\tau^2} > 0,\end{aligned}$$

which is true because $1 - \frac{n_i/\sigma_i^2}{\sum_{j=1}^{I} n_j/\sigma_j^2} > 0$.

References

Bayarri, M.J. and Berger, J.O. (1997). Measures of surprise in Bayesian analysis. *ISDS Discussion Paper* 97-46, Duke University.

Bayarri, M.J. and Berger, J.O. (1999). Quantifying surprise in the data and model verification. In *Bayesian Statistics 6*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, London: Oxford University Press, 53-82.

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

Bayarri, M.J. and Castellanos, M.E. (2001). A comparison between p-values for goodness-of-fit checking. In *Monographs of Official Statistics. Bayesian Methods with applications to science, policy and official statistics*, 1-10.

Bayarri, M.J. and Morales, J. (2003). Bayesian measures of surprise for outlier detection. *Journal Statistical Planning and Inference*, **111**, 3-22.

Bayarri, M.J. (2003). Discussion on HSSS model criticism. In *Highly Structured Stochastic Systems*, eds. P. J. Green, N. L. Hjort and S. T. Richardson. Oxford University Press.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach. *Journal of the Royal Statistical Society, Series B*, **57**, 289-300.

Berger, J.O. (1980/85). *Statistical Decision Theory and Bayesian Analysis. Second Edition.* New York: Springer-Verlag.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* Wiley, Chichester.

Besag, J. and Clifford, P. (1989). Generalized Monte Carlo Significance Tests *Biometrika*, **76, 4**, 633-642.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A*, **143**, 383-430.

Carlin, B.P. and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Second Edition. Chapman & Hall.

Castellanos M.E. (1999), *Medidas de sorpresa para bondad de ajuste.* Master Thesis. Dpto. Estadística i I.O. Universitat de València.

Castellanos M.E. (2002), *Diagnósticos Bayesianos de Modelos.* PhD. Dissertation. Dpto. Estadística y Mat. Aplicada. Universidad Miguel Hernández.

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

Gelfand, A.E., Dey, D.K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith), 147-167. Oxford University Press, Oxford.

Gelfand, A.E. (2003). Discussion to HSSS model criticism. In *Highly Structured Stochastic Systems*, eds. P. J. Green, N. L. Hjort and S. T. Richardson. Oxford University Press.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D. (1995) *Bayesian Data Analysis*. Chapman & Hall/CRC.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B*, **29**, 83-100.

Lange, N. and Ryan L. (1989). Assessing Normality in Random Effects Models. *The Annals of Statistics*, **17**, 624-642

Marshall, E.C. and Spiegelhalter D.J. (2001). Simulation-based tests for divergent behavior in hierarchical models. *Technical Report in http://www.med.ic.ac.uk/divisions/60/biostat/criticism.ps*

O'Hagan, A. (2003). HSSS model criticism (with discussion). In *Highly Structured Stochastic Systems*, eds. P. J. Green, N. L. Hjort and S. T. Richardson. 423-453. Oxford University Press.

Rubin, D.B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151-1172.

Robins, J.M., van der Vaart, A., and Ventura, V. (2000). The asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association*, **95**, 1143-1156.