Posterior predictive model checking in hierarchical models

Article *in* Journal of Statistical Planning and Inference · February 2003

Impact Factor: 0.68 · DOI: 10.1016/S0378-3758(02)00303-8

2 authors, including:

Sandip Sinharay
Pacific Metrics

118 PUBLICATIONS **1,099** CITATIONS

SEE PROFILE

ELSEVIER

Journal of Statistical Planning and Inference 111 (2003) 209–221

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Posterior predictive model checking in hierarchical models

Sandip Sinharay[a,1], Hal S. Stern[b,*]

[a] *Educational Testing Service, Princeton, NJ, USA*
[b] *Department of Statistics, 4900H Berkeley Place, University of California, Irvine, CA 92697, USA*

Abstract

Model checking is a crucial part of any statistical analysis. Hierarchical models present special problems because assumptions made about the distribution of unobservable parameters are difficult to check. In this article, we review some approaches to model checking and apply posterior predictive model checking to a hierarchical normal–normal model analysis of data from educational testing experiments in eight schools. Then we carry out a simulation study to investigate the difficulties in model checking for hierarchical models. It turns out that it is very difficult to detect violations of the assumptions made about the population distribution of the parameters unless the extent of violation is huge or the observed data have small standard errors.

MSC: primary 62F15

Keywords: Discrepancy; Marginal model; Posterior predictive p-value; Random effects model

1. Introduction

Assessing the validity of model assumptions is a crucial part of any parametric statistical analysis. Hierarchical models, in which data $\boldsymbol{y}$ are modeled conditional on a collection of parameters $\boldsymbol{\theta}$ and these parameters are in turn described by a probability distribution with underlying parameters $\boldsymbol{\alpha}$, present special problems because assumptions made about the distribution of the unobservable parameters are difficult to check. As hierarchical models are increasingly popular in applications in the health sciences, animal breeding and many other fields (see, e.g., Gelman et. al., 1995; Carlin and Louis, 1996; Gilks et. al., 1996), the development of tools and techniques for model

* Corresponding author. Tel.: +949-824-1568.
E-mail address: sternh@uci.edu (H.S. Stern).
[1] Associate Research Scientist.

0378-3758/02/$ - see front matter
PII: S0378-3758(02)00303-8

checking is essential. In this paper, we review some methods and point out several interesting aspects of model checking in hierarchical models.

We view model checking as distinct from model selection where a number of candidate models are compared. Instead the question of interest is whether a particular specified model appears to provide an adequate fit. This situation can arise during the model building phase as we assess the fit of a proposed model in order to identify features of the data not addressed by the current model. It can also arise as a final step in a data analysis, checking that the final model has no major defects.

In the next section, we introduce the hierarchical normal–normal model that serves as a case study throughout the paper and the eight schools example. Section 3 surveys model checking methods. Posterior predictive model checks (Rubin, 1984, Gelman et. al., 1996) are described in this section and applied throughout. The posterior predictive model checking approach is applied to a hierarchical model analysis of data from educational testing experiments in eight schools in Section 4. A simulation study in Section 5 explores some difficulties in model checking for hierarchical models. It turns out that it is very difficult to detect violations of the assumptions made about the population distribution of the parameters $\boldsymbol{\theta}$ unless the extent of the violation is huge or the observed data have small standard errors.

2. Hierarchical normal–normal model

Hierarchical models are convenient and popular for modeling large data sets. One nice feature of such models is that they provide a natural method for addressing interrelationships among the units in a study. In addition they are computationally convenient. As an example of this type of model we describe the normal–normal hierarchical model and apply it to the analysis of an educational testing experiment.

2.1. Description of the model

Assume the data are measurements y_j, $j = 1, 2, \ldots, J$ that can be modeled as independent normal random variables conditional on the means θ_j, $j = 1, 2, \ldots, J$ and variances σ_j^2, $j = 1, 2, \ldots, J$. To keep things computationally simple we assume that the variances are known; this is often a reasonable approximation, for example, in the educational testing study considered later. There are a number of scenarios where this model is natural, e.g., the y_j's may be sample means from different subpopulations or estimates with approximately normal distributions. We assume that the parameters θ_j, $j = 1, 2, \ldots, J$ are drawn from a normal distribution with hyperparameters (μ, τ) so that

$$\theta_j | \mu, \tau \overset{\text{iid}}{\sim} N(\mu, \tau^2), \quad j = 1, 2, \ldots, J.$$

As the θ_j's are not directly observable, it can be difficult to identify a suitable distribution. The normal distribution is often chosen as an approximation when θ_j is continuous with reasonably large range. Determining when this assumption is plausible is one of our primary goals. To complete the model specification, we need a hyperprior distribution on the hyperparameters $\boldsymbol{\alpha} = (\mu, \tau)$. We assign a noninformative uniform hyperprior

distribution to μ given τ so that $p(\mu,\tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$. Although $p(\mu|\tau)$ is improper, it turns out to cause no problems for our model specification. The choice of $p(\tau)$ is discussed in the next section.

2.2. The joint posterior distribution of the parameters

Combining the data distribution $p(\boldsymbol{y}|\boldsymbol{\theta})$, the prior distribution $p(\boldsymbol{\theta}|\mu,\tau)$ and the hyperprior distribution $p(\mu,\tau)$, the joint posterior distribution of $(\boldsymbol{\theta},\mu,\tau)$ is given up to a constant of proportionality as

$$p(\boldsymbol{\theta},\mu,\tau|\boldsymbol{y}) \propto p(\mu,\tau)p(\boldsymbol{\theta}|\mu,\tau)p(\boldsymbol{y}|\boldsymbol{\theta}) \propto p(\mu,\tau)\prod_{j=1}^{J} N(\theta_j|\mu,\tau^2)\prod_{j=1}^{J} N(y_j|\theta_j,\sigma_j^2), \tag{2.1}$$

where $N(x|a,b)$ is the normal density with argument x, mean a and variance b.

The $(J+2)$-dimensional posterior distribution is generally too complex to allow us to do any calculations directly. Instead it is common to use Markov chain Monte Carlo (MCMC) methods to obtain simulations from the posterior distribution. It is straightforward to develop a Gibbs sampling algorithm for the model. In fact, MCMC easily accommodates unknown observation variances as well. As we intend to carry out a simulation study in a subsequent section, we adopt an alternative faster method for simulation from the posterior distribution. As shown in Gelman et al. (1995, Section 5.4), the joint posterior distribution can be factored as the product

$$p(\boldsymbol{\theta},\mu,\tau|\boldsymbol{y}) \propto p(\tau|\boldsymbol{y})p(\mu|\tau,\boldsymbol{y})p(\boldsymbol{\theta}|\mu,\tau,\boldsymbol{y}),$$

where

$$p(\tau|\boldsymbol{y}) \propto p(\tau)V_\mu^{1/2}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}\exp\left(-\frac{(y_j-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right),$$

$$\mu|\tau,\boldsymbol{y} \sim N(\hat{\mu},V_\mu)$$

with

$$\hat{\mu} = \frac{\sum_{j=1}^{J} y_j/(\sigma_j^2+\tau^2)}{\sum_{j=1}^{J} 1/(\sigma_j^2+\tau^2)}, \quad V_\mu^{-1} = \sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}$$

and the elements of $\boldsymbol{\theta}$ are independent given $\mu,\tau,\boldsymbol{y}$ with

$$\theta_j|\mu,\tau,\boldsymbol{y} \sim N(\hat{\theta}_j,V_j),$$

$$\hat{\theta}_j = (y_j/\sigma_j^2+\mu/\tau^2)/(1/\sigma_j^2+1/\tau^2)$$

and

$$V_j = (1/\sigma_j^2+1/\tau^2)^{-1}.$$

Table 1
Observed effects and standard errors of SAT-V scores in eight schools

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28.39	14.9
B	7.94	10.2
C	−2.75	16.3
D	6.82	11.0
E	−0.64	9.4
F	0.63	11.4
G	18.01	10.4
H	12.16	17.6

Then simulations from the joint posterior distribution can be obtained using the following three steps:

1. generate τ from its marginal distribution $p(\tau|\boldsymbol{y})$;
2. generate μ from $p(\mu|\tau,\boldsymbol{y})$, using the value of τ generated in step 1;
3. generate θ_j from $p(\theta_j|\mu,\tau,\boldsymbol{y})$, $j=1,2,\ldots,J$, using the values of τ and μ generated in steps 1 and 2.

Note that the last 2 steps are trivial in that both require simulations from univariate normal distributions. The first step, simulation from the marginal distribution $p(\tau|\boldsymbol{y})$, involves a complicated but one-dimensional distribution. In the remainder of the article, we use a discrete grid approximation (using 10,000 equispaced points from 0 to 40) to simulate from this distribution.

The lone remaining issue concerns the hyperprior distribution $p(\tau)$. Note that the traditional prior distribution for a variance parameter, $p(\tau^2) \propto 1/\tau^2$, does not lead to a proper posterior distribution here. Instead, $p(\tau^2) \propto 1/\tau$ (equivalent to $p(\tau) \propto 1$) is often used as a noninformative prior distribution that yields a proper posterior distribution. A number of other choices for the hyperprior distribution ($p(\tau) \propto 1/\sqrt{\tau}$, $p(\tau) \propto \tau\sigma_c^2/(\tau^2+\sigma_c^2)^2$, $p(\tau) \propto \sigma_c/(\tau+\sigma_c)^2$, $p(\tau) \propto \tau/(\tau^2+\sigma_c^2)$, where σ_c^2 is a constant) were considered at various stages of this research. As the results obtained were not sensitive to the choice of hyperprior distribution, we present results only for the noninformative prior distribution specified above.

2.3. An educational testing study

The data are from a study performed for the Educational Testing Service to analyze the effects of special coaching programs on standardized test scores and are given in Table 1. These data are described in more detail and analyzed by Rubin (1981); these are analyzed again by Gelman et al. (1995). Separate randomized experiments were performed to estimate the effects of coaching programs in each of eight high schools. In each school, the estimated coaching effect (in standardized test score points) and its standard error were obtained using an analysis of covariance regression model. The estimated coaching effects, which we label as y_j, are normally distributed under the usual regression assumptions with mean equal to the true coaching effect and sampling variances σ_j^2. The sampling variances can be treated as known for all practical purposes

Table 2
Summary of the posterior distribution of the parameters for the SAT coaching example

Parameters	Posterior quantiles				
	2.5%	25%	50%	75%	97.5%
θ_1	−2	7	10	16	31
θ_2	−5	3	8	12	23
θ_3	−11	2	7	11	19
θ_4	−7	4	8	11	21
θ_5	−9	1	5	10	18
θ_6	−7	2	6	10	28
θ_7	−1	7	10	15	26
θ_8	−6	3	8	13	33
μ	−3	5	8	11	18
τ	1	2	5	9	19

because the sample size in each of the eight experiments was relatively large. There was no prior reason to believe that any of the eight programs was more effective than any other or that some were more similar in effect to each other than to any other. Consequently, it seems reasonable to model the true coaching effects as independent draws from a population distribution. We apply the hierarchical normal model to this data. The assumption of a normal prior distribution for the coaching effects is made primarily for convenience. Our model checking focuses on this assumption later. Though analysis of these data is not of primary interest here, we provide numerical summaries of the posterior distribution in Table 2. Note that the variance τ^2 is estimated to be much smaller than the sampling variances. This leads to a lot of shrinkage of the coaching effect estimates towards a common value.

3. Review of model checking

It is very important for a complete data analysis to include some checks of the validity of the assumptions made. In addition, it makes sense to ask whether the model explains the observed variability in the data adequately or if there are features of the data not captured by the model. For example, before assuming that a sample is from a normal distribution, we might check if the normality assumption is valid for the data using plots (e.g., normal probability plot) or statistical tests (e.g., Shapiro–Wilk test or Anderson–Darling test). Perhaps the most widely known example of model checking is the use of regression diagnostics. This section briefly reviews some issues in model checking from the Bayesian perspective.

3.1. The marginal model

The basic approach to model checking is to compare the observed data (or some function of it) to values that would be expected under the suggested model (sometimes called fitted values). Standard residual plots are an example of this kind of approach where residuals are compared to their expected value if the model is true (zero).

For hierarchical models with data $\boldsymbol{y}$, parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{\alpha}$, model checking is often carried out on the marginal model

$$p(\boldsymbol{y}|\boldsymbol{\alpha}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\theta}.$$

The marginal model in the normal–normal hierarchical model is easily found as

$$y_j|\boldsymbol{\alpha} \sim N(\mu, \tau^2 + \sigma_j^2), \quad j = 1, 2, \ldots, J.$$

Then an obvious model checking diagnostic is a normal probability plot of the standardized residuals $z_j = (y_j - \mu)/\sqrt{\tau^2 + \sigma_j^2}$. An alternative is the weighted normal plot (Dempster and Ryan, 1985) where the z_j's are plotted against $\Phi^{-1}[F_n(z_j)]$ where $F_n(.)$ is the empirical cdf. These approaches require point estimates of μ and τ^2 but any point estimate ignores the posterior uncertainty about the values of these parameters that the Bayesian analysis addresses. An alternative is to conceptualize a series of normal probability plots or weighted normal plots, one for each posterior simulation of μ and τ^2.

3.2. The prior predictive method of Box

As remarked above, diagnostics must account for the uncertainty in the parameter estimates in the Bayesian approach. Let $\boldsymbol{y}^{\text{rep}}$ denote replicate data that we might observe if the experiment that generated $\boldsymbol{y}$ is replicated and let $\boldsymbol{\eta}$ denote all the parameters (both $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$). Box (1980) suggested checking Bayesian models using the marginal or prior predictive distribution $p(\boldsymbol{y}^{\text{rep}}) = \int p(\boldsymbol{y}^{\text{rep}}|\boldsymbol{\eta})\, p(\boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\eta}$ as a reference distribution for the observed data $\boldsymbol{y}$. In practice, diagnostic measures $D(\boldsymbol{y})$ are defined and the observed value $D(\boldsymbol{y})$ compared to the reference distribution of $D(\boldsymbol{y}^{\text{rep}})$ with any significant difference between them indicating a model failure.

Though this method is a natural one to use, the prior predictive distribution is undefined under improper priors (and can be quite sensitive to the prior distribution if vague prior distributions are used). This is a potentially significant disadvantage as many analyses with hierarchical models rely on improper or vague prior distributions.

3.3. Posterior predictive model checking

An alternative to the prior predictive approach of Box is to use the posterior distribution of $\boldsymbol{\eta}$ in place of its prior distribution in the Box definition. The replication $\boldsymbol{y}^{\text{rep}}$ then represents replicate data that we might observe if the experiment that generated $\boldsymbol{y}$ is replicated with the same value of $\boldsymbol{\eta}$ that generated the observed data. Note that here, as in our discussion of Box's approach, the definition of $\boldsymbol{\eta}$ includes all parameters. Since the value $\boldsymbol{\eta}$ that generated the observed data is unknown, we derive the posterior (given $\boldsymbol{y}$) predictive distribution of $\boldsymbol{y}^{\text{rep}}$ by averaging over the plausible values of $\boldsymbol{\eta}$, given by the posterior distribution $p(\boldsymbol{\eta}|\boldsymbol{y})$,

$$p(\boldsymbol{y}^{\text{rep}}|\boldsymbol{y}) = \int p(\boldsymbol{y}^{\text{rep}}|\boldsymbol{\eta})\, p(\boldsymbol{\eta}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{\eta}.$$

Guttman (1967) applies the posterior predictive distribution in a goodness of fit test. Rubin (1984) suggests simulating replicate data sets from the posterior predictive distribution for model checking. Any significant difference between the replications and the observed data indicates a possible failure of the model. Because the posterior distribution is used in this definition, we need not use a proper prior distribution.

There is flexibility in defining replications for different contexts, e.g., it is possible to compromise between the Box and Rubin approaches in hierarchical models by taking the posterior distribution for some parameters (e.g., $\boldsymbol{\alpha}$) and the prior distribution for others (e.g., $\boldsymbol{\theta}$). We can measure the discrepancy between the model and the data by defining one or more test statistics $D(\boldsymbol{y})$. Gelman et al. (1996) extend the posterior predictive approach to use discrepancies $D(\boldsymbol{y}, \boldsymbol{\eta})$ that depend on the data and the parameters. The divergence of the data from the posterior predictive distribution can be determined by comparing the posterior predictive distribution of $D(\boldsymbol{y}^{\text{rep}}, \boldsymbol{\eta})$ with the posterior distribution of $D(\boldsymbol{y}, \boldsymbol{\eta})$. The comparison can be carried out easily by simulation. We draw N simulations $\boldsymbol{\eta}^1, \boldsymbol{\eta}^2, \ldots, \boldsymbol{\eta}^N$, from the posterior distribution of $\boldsymbol{\eta}$ and then draw one $\boldsymbol{y}^{\text{rep}}$ from the predictive distribution $p(\boldsymbol{y}|\boldsymbol{\eta})$ using each simulated $\boldsymbol{\eta}$. We then have N draws from the joint posterior distribution $p(\boldsymbol{y}^{\text{rep}}, \boldsymbol{\eta}|\boldsymbol{y})$. The posterior predictive check boils down to comparing the values of the realized discrepancy $D(\boldsymbol{y}, \boldsymbol{\eta}^n)$ and the replicated discrepancy measures $D(\boldsymbol{y}^{\text{rep},n}, \boldsymbol{\eta}^n)$, $n = 1, 2, \ldots, N$ perhaps by plotting the pairs $(D(\boldsymbol{y}, \boldsymbol{\eta}^n), D(\boldsymbol{y}^{\text{rep},n}, \boldsymbol{\eta}^n))$ in a scatterplot. One popular summary of the comparison is the tail-area probability or p-value,

$$
\begin{aligned}
p_b &= P(D(\boldsymbol{y}^{\text{rep}}, \boldsymbol{\eta}) \geqslant D(\boldsymbol{y}, \boldsymbol{\eta})|\boldsymbol{y}) \\
&= \int\int I_{[D(\boldsymbol{y}^{\text{rep}}, \boldsymbol{\eta}) \geqslant D(\boldsymbol{y}, \boldsymbol{\eta})]}\, p(\boldsymbol{y}^{\text{rep}}|\boldsymbol{\eta})\, p(\boldsymbol{\eta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{y}^{\text{rep}}\, \mathrm{d}\boldsymbol{\eta},
\end{aligned}
$$

where $I_{[A]}$ is the indicator function for the event A. The p-value is estimated from the simulations as the proportion of the N replications for which $D(\boldsymbol{y}^{\text{rep},n}, \boldsymbol{\eta}^n) \geqslant D(\boldsymbol{y}, \boldsymbol{\eta}^n)$.

Posterior predictive checks have been criticized for being conservative. To quote Dey et al. (1997), "the observed data, through the posterior, suggests which values of the parameter are likely under the model. Then to assess adequacy, the observed data is checked against data generated using such parameter values apparently making it difficult to criticize the model". In addition, recent work by Robins et al. (2000) show that the posterior predictive p-value is not uniformly distributed under the null hypothesis, not even asymptotically. Bayarri and Berger (2000) propose alternative model checking strategies that are less conservative and produce p-values having asymptotic uniform distribution under the null hypothesis. Their approaches however are more difficult to carry out and interpret.

In the remainder, we use posterior predictive model checks to address the fit of hierarchical models. The posterior predictive checks are easy to carry out and interpret. They are especially useful if we think of the current model as a plausible ending point with modifications to be made only if substantial lack of fit is found, for examples, see Glickman and Stern (1998) and Belin and Rubin (1995).

4. Posterior predictive model checking in the normal–normal model

4.1. Model checking in the SAT coaching example

We illustrate the posterior predictive approach using the SAT coaching study data. Simulations from the posterior distribution were obtained using the algorithm of Section 2.2. One thousand draws from the posterior distribution of $(\boldsymbol{\theta}, \mu, \tau^2)$ were used for model checking. For each of the 1000 draws, a replicate data set $\boldsymbol{y}^{\text{rep}}$ was drawn from the posterior predictive distribution and values of a number of test statistics were calculated for each replication. We initially consider the following simple test statistics: the largest of the eight observed outcomes, $\max_j(y_j)$, the smallest of the eight observed outcomes, $\min_j(y_j)$, the average, $\text{mean}_j(y_j)$, and the sample standard deviation, $\text{sd}_j(y_j)$. These measures are merely intended to illustrate the approach and identify gross failures of the model. Histograms showing the posterior predictive distribution of the test statistics are provided in Fig. 1; the corresponding observed values (from the data) are also given along with the posterior predictive p-values. As none of the p-values are very high or low, they suggest that the model generates replicate data which are similar in these respects to the observed values in the study. So, using posterior predictive model checking with these measures, we find no failure of the model. We take a closer look at the measures themselves in the next section.

4.2. Choice of discrepancy measures

In posterior predictive model checking, one very important issue concerns the choice of discrepancy measures. It is difficult to discuss the choice of discrepancy measures in general terms because the most informative measures typically depend on the context.

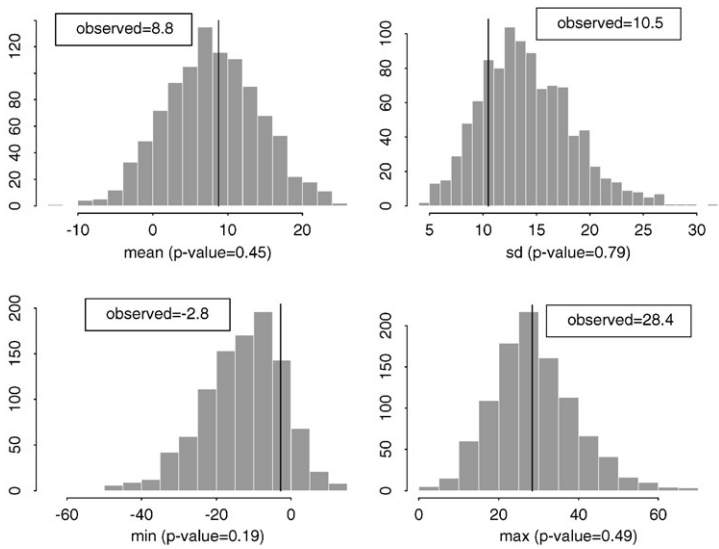

Fig. 1. Histograms of the posterior predictive distribution of the discrepancy measures for the SAT example.

They should, like test statistics in traditional statistical tests, reflect aspects of the model that are relevant to scientific purposes to which the inference will be applied. Here we discuss some issues concerning the choice of discrepancy measures or test statistics that are suggested by our example.

If we take a sufficient statistic for a model parameter as a discrepancy measure, then that quantity will automatically be reproduced in replicate data sets by the parameter in question. This is why the sample mean in the normal hierarchical model is not a good discrepancy measure—the posterior predictive distribution will be centered around the sample mean regardless of whether the model fits the data. Regarding the other measures shown in Fig. 1, the observed minimum and maximum do not appear terribly unusual. This too is not surprising because the posterior distribution of μ and τ are implicitly adjusted to fit the data as well as possible. Hence, even if there is an outlying value of $\boldsymbol{y}$, the posterior distribution of τ (and μ) are likely to be inflated enough to generate replicate values similar to the outlying value. The observed minimum appears to be a bit bigger than expected, but not troublingly so.

The final measure shown in Fig. 1 is the standard deviation. We might expect the result here to be like that of the sample mean. After all, the sample variance is associated with the variance parameter τ^2 (and σ_j^2). It is a bit surprising that the posterior predictive p-value for the standard deviation (or sample variance) is noticeably larger than 0.5. The posterior predictive p-value is not high enough for us to reject the model, however it does suggest that the observed data are less variable than expected under the model. This same result has also been found in other studies, e.g., Stern and Cressie (1995). In the next section, a simulation study addresses this issue. We also consider other measures that are better able to assess the fit of the normal–normal hierarchical model.

5. Simulation studies

We next use simulation to assess the performance of posterior predictive model checking in the hierarchical normal–normal model. We focus on model checking for the assumed prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$. Clearly it is also important to check the assumptions embodied in the data distribution $p(\boldsymbol{y}|\boldsymbol{\theta})$ and other, often implicit, assumptions like exchangability. Assessing the validity of data distribution assumptions is common using a variety of standard tools (e.g., residual analysis in regression). By contrast, testing the fit of the prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$ to the data is difficult because neither $\boldsymbol{\theta}$ nor $\boldsymbol{\alpha}$ is directly observed.

5.1. The simulation approach

We simulate data sets from the normal–normal hierarchical model, using the following steps.

1. Generate $\boldsymbol{\theta}$ from $\theta_j|\mu,\tau \overset{\text{iid}}{\sim} N(\mu,\tau^2)$, $j=1,2,\ldots,J$ where we fix $\mu=7.9$ and $\tau=7.9$ (the method of moments estimates of these parameters for the SAT data). We consider $J=8$ (as in the study), 16, 50 (for some scenarios) and 100;

Table 3
Quantiles (2.5%, 50%, 97.5%) of the posterior predictive p-values for four measures when the model used to analyze the data (normal–normal) is the correct model. Results are based on 1000 simulated data sets from the normal–normal model

Discrepancy measure	Number of schools	σ^2		
		1	20	168
Mean	8	(0.47,0.50,0.53)	(0.47,0.50,0.53)	(0.47,0.50,0.53)
	16	(0.47,0.50,0.53)	(0.47,0.50,0.53)	(0.47,0.50,0.53)
	100	(0.47,0.50,0.53)	(0.47,0.50,0.53)	(0.47,0.50,0.53)
S.D.	8	(0.47,0.50,0.53)	(0.45,0.49,0.61)	(0.44,0.53,0.96)
	16	(0.47,0.50,0.53)	(0.45,0.49,0.52)	(0.43,0.49,0.94)
	100	(0.47,0.50,0.53)	(0.46,0.49,0.52)	(0.40,0.45,0.63)
Minimum	8	(0.28,0.53,0.60)	(0.22,0.53,0.76)	(0.13,0.43,0.77)
	16	(0.24,0.54,0.61)	(0.17,0.53,0.81)	(0.11,0.47,0.86)
	100	(0.19,0.55,0.63)	(0.11,0.50,0.88)	(0.08,0.50,0.96)
Maximum	8	(0.40,0.47,0.75)	(0.24,0.48,0.80)	(0.21,0.57,0.87)
	16	(0.39,0.46,0.75)	(0.19,0.49,0.84)	(0.16,0.54,0.90)
	100	(0.37,0.46,0.80)	(0.13,0.50,0.91)	(0.04,0.48,0.92)

2. Generate $\boldsymbol{y}$ from $y_j|\theta_j,\sigma^2 \overset{\text{ind}}{\sim} N(\theta_j,\sigma^2)$, $j=1,2,\ldots,J$ using the values of θ_j from step 1 and a specified value of σ^2. We use $\sigma^2=1$, 20, 168, the latter representing the average of the σ_j^2's in the SAT study.

We then perform posterior predictive model checking for each simulated data set and calculate the posterior predictive p-value based on 1000 posterior simulations for a given discrepancy.

As as alternative model, to assess the power of the posterior predictive approach, we use a normal data model with an exponential prior distribution. We simulate data using the same basic procedure as for the normal–normal model, except that in step 1, we generated θ_j's using $\theta_j|\mu \overset{\text{iid}}{\sim} \text{Exp}(\mu)$, where we fixed $\mu=7.9$ to make the prior mean of the θ_j's conform to that in the first simulation. The exponential distribution is used because it is not symmetric and thus might represent a serious failure of the normal–normal model.

5.2. Study of discrepancy measures for detecting nonnormality of random effects

One thousand data sets were simulated from the normal–normal model as described in the previous section, and the posterior predictive p-value is recorded for each data set for each of the four discrepancy measures used in Section 4.1. The distribution of the p-values for each discrepancy measure is given in Table 3. The number of schools ($J=8$, 16 and 100) and the value of σ^2 are varied. Note that the normal assumption is correct in this model. Table 3 indicates that the simulated p-values are typically distributed evenly around 0.5. It is noteworthy that for large variance and

Table 4
Quantiles (2.5%, 50%, 97.5%) of the posterior predictive p-values for four measures when the model used to analyze the data (normal–normal) is not the correct model. Results are based on 1000 simulated data sets from the normal–exponential model

Discrepancy measure	Number of schools	σ^2		
		1	20	168
Mean	8	(0.47,0.50,0.53)	(0.47,0.50,0.53)	(0.47,0.50,0.53)
	16	(0.47,0.50,0.53)	(0.47,0.50,0.53)	(0.47,0.50,0.53)
	100	(0.47,0.50,0.53)	(0.47,0.50,0.53)	(0.47,0.50,0.53)
S.D.	8	(0.47,0.50,0.53)	(0.45,0.49,0.79)	(0.44,0.55,0.97)
	16	(0.47,0.50,0.53)	(0.44,0.49,0.53)	(0.43,0.51,0.93)
	100	(0.47,0.50,0.53)	(0.46,0.49,0.52)	(0.41,0.45,0.67)
Minimum	8	(0.16,0.41,0.60)	(0.14,0.42,0.72)	(0.12,0.41,0.75)
	16	(0.09,0.30,0.54)	(0.07,0.32,0.74)	(0.09,0.41,0.83)
	100	(0.01,0.06,0.35)	(0.01,0.09,0.54)	(0.05,0.39,0.90)
Maximum	8	(0.36,0.45,0.69)	(0.22,0.40,0.78)	(0.19,0.57,0.89)
	16	(0.34,0.42,0.67)	(0.13,0.31,0.70)	(0.11,0.51,0.89)
	100	(0.29,0.36,0.60)	(0.02,0.12,0.54)	(0.01,0.26,0.84)

small number of schools (the values most like those in the SAT study), the p-value for the standard deviation appears to be significantly shifted towards higher values. This is consistent with the observed value in the SAT study. Some theoretical work for the normal–normal model in Sinharay (1998) suggests that this pattern is in fact to be expected.

We carry out a second set of simulations using the normal–exponential hierarchical model. Once again one thousand data sets were simulated. The results are shown in Table 4. Here, the fitted normal–normal model is not correct and we hope that the posterior predictive p-values for (at least) some measures will tend towards extreme values. The results support our discussion of Section 4.2. The sample mean is not a useful measure, the posterior predictive p-values are always near 0.5. If the number of schools is large and the sampling variance not too large, then the maximum and especially the minimum appear to be effective diagnostics. The standard deviation is not an effective measure. Interestingly, the distribution of p-values for the standard deviation is closer to 0.5 in this case (when the model is false) than when the model is true.

5.3. A new measure

The simulation results of Section 5.2 indicate that our basic measures are not very useful for detecting failure of the normal assumption for $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$. As an alternative, we consider a measure defined exclusively on the θ_j's,

$$D_{\boldsymbol{\theta}} = |\theta_{\max} - \theta_{\text{med}}| - |\theta_{\min} - \theta_{\text{med}}|,$$

Table 5
Quantiles (2.5%, 50%, 97.5%) of $E(D_{\boldsymbol{\theta}}|\boldsymbol{y})$ when the data are simulated from the normal–exponential model and the normal–normal model is fit. Results are based on 1000 simulated data sets from the normal–exponential model

Number of schools	σ^2		
	1	20	168
8	(−2.08,7.65,32.47)	(−2.68,1.67,25.59)	(−4.12,0.11,10.35)
16	(0.75,12.40,37.17)	(−1.27,4.03,27.30)	(−2.45,0.18,10.97)
50	(8.37,20.62,47.29)	(0.67,8.83,30.86)	(−1.04,0.32,6.57)
00	(12.82,25.44,53.91)	(2.57,11.00,33.18)	(−0.60,0.38,5.48)

where $\theta_{\max}$, θ_{med} and $\theta_{\min}$ are respectively the maximum, median and minimum of the θ_j's. This measure would be expected to be sensitive to skewness in the distribution of θ. It is unusual in that it relies only on parameters, but seems natural given our interest in the distribution of $\boldsymbol{\theta}$. For simulations where the normal assumption is true, $D_{\boldsymbol{\theta}}$ should be distributed around zero, but if the population distribution of the θ_j's is actually skewed, then it will be distributed away from zero. One thousand data sets were simulated from the normal–normal model using the algorithm described in Section 5.1 with $\sigma^2 = 168$ (motivated by the SAT example). For each data set, we generated 1000 simulations from the posterior distribution of $(\boldsymbol{\theta}, \mu, \tau)$ under the hierarchical normal–normal model and calculated $E(D_{\boldsymbol{\theta}}|\boldsymbol{y})$. The posterior distribution of the $E(D_{\boldsymbol{\theta}}|\boldsymbol{y})$'s was centered very close to zero as expected (95% of the values are in the interval $(-6.47, 5.90)$). The analysis was repeated to show the posterior distribution of $E(D_{\boldsymbol{\theta}}|\boldsymbol{y})$ over 1000 simulated data sets from the normal–exponential model. The surprising result is that the distribution was still centered near zero (95% of the values are in the interval $(-4.18, 10.43)$) even though the random effects were drawn from an exponential distribution, which is skewed and asymmetric.

This unexpected result suggests that even when the θ_j's are not normally distributed, the data $\boldsymbol{y}$ may still be well described by the normal–normal model. Of course, in the limit of small sampling variance ($\sigma_j^2 \approx 0$), the $\boldsymbol{y}$'s are approximately equal to the θ_j's and it is possible to directly verify assumptions about the distribution of the θ_j's using standard methods. Table 5 provides simulation results for a variety of choices for the number of schools and sampling variance. Simulations verify that for $\sigma_j^2 \approx 0$, we can detect failures of the normal hierarchical assumption easily—in that case, the histogram showing the distribution of $E(D_{\boldsymbol{\theta}}|\boldsymbol{y})$ over many normal–exponential simulations has very little or no mass below 0. If the number of schools is eight, then it is quite difficult to detect the nonnormality unless the σ_j^2's are very close to 0. But as the number of schools is increased, we can detect the nonnormality for larger values of σ_j^2's. When the number of schools is 100, the nonnormality of the θ_j's can be detected for values of σ_j^2's up to 20 or so (recall the variance of the random effects was fixed at $7.9^2 = 62.4$). So the measure $D_{\boldsymbol{\theta}}$ has the ability to detect the model failure regarding the random effects, but only if the sampling variances are small.

6. Conclusion

The results presented here indicate that it is difficult to detect violations of the assumption about the random effects distribution using posterior predictive model checks. The simulations suggest that we can only detect a model failure using the posterior predictive method if the random effects are nonnormal and the data sampling variances are small.

It should be emphasized that it may not be a problem that model violations are not easily detected. After all, we should only worry about the choice of the distribution of $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$ if it represents such a poor choice that it seriously impacts the conclusion of the data analysis. In fact, the difficulty detecting failures of the assumed normal prior distribution in the normal–normal hierarchical model can be thought of as a form of model robustness in that the posterior distribution of the normal–normal model appears to adequately describe the data even when the population distribution is not normal.

References

Bayarri, S., Berger, J., 2000. P-values for composite null models. J. Amer. Statist. Assoc. 95, 1127–1142.

Belin, T.R., Rubin, D.B., 1995. The analysis of repeated-measures data on schizophrenic reaction times using mixture models. Statist. Med. 14, 747–768.

Box, G.E.P., 1980. Sampling and Bayes' inference in scientific modelling and robustness. J. Roy. Statist. Soc. A 143, 383–430.

Carlin, B.P., Louis, T.A., 1996. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall, London.

Dempster, A.P., Ryan, L.M., 1985. Weighted normal plots. J. Amer. Statist. Assoc. 80, 845–850.

Dey, D.K., Gelfand, A.E., Vlachos, P.K., Swartz, T.B., 1997. A simulation-intensive approach for checking hierarchical models. Comput. Sci. Statist. 28, 162–171.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & Hall, London.

Gelman, A., Meng, X.L., Stern, H.S., 1996. Posterior predictive assessment of model fitness via realized discrepancies (with discussion) Statistica Sinica 6, 733–807.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

Glickman, M.E., Stern, H.S., 1998. A state-space model for National Football League scores. J. Amer. Statist. Assoc. 93, 25–35.

Guttman, I., 1967. The use of the concept of a future observation in goodness-of-fit problems. J. Roy. Statist. Soc. B 29, 83–100.

Robins, J.M., van der Vaart, A., Ventura, V., 2000. The asymptotic distribution of p-values in composite null models. J. Amer. Statist. Assoc. 95, 1143–1172.

Rubin, D.B., 1981. Estimation in parallel randomized experiments. J. Ed. Statist. 6, 377–401.

Rubin, D.B., 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Statist. 12, 1151–1172.

Sinharay, S., 1998. A look at the application of posterior predictive model checking in hierarchical Bayesian models. Unpublished Technical Report, Department of Statistics, Iowa State University, Ames, IA.

Stern, H.S., Cressie, N., 1995. Bayesian and constrained Bayesian inference for extremes in epidemiology. 1995 Proceedings of the Epidemiology Section, American Statistical Association, pp. 11–20.