Published in final edited form as:
Biometrics. 2012 March ; 68(1): 156–164. doi:10.1111/j.1541-0420.2011.01668.x.

Goodness-of-Fit Diagnostics for Bayesian Hierarchical Models

Ying Yuan and **Valen E. Johnson**[*]
Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A

Summary

This article proposes methodology for assessing goodness of fit in Bayesian hierarchical models. The methodology is based on comparing values of the pivotal discrepancy measures, computed using parameter values drawn from the posterior distribution, versus known reference distributions. Because the resulting diagnostics can be calculated from standard output of Markov chain Monte Carlo algorithms, their computational costs are minimal. Several simulation studies are provided, each of which suggests that diagnostics based on pivotal discrepancy measures have higher statistical power than comparable posterior-predictive diagnostic checks in detecting model departures. The proposed methodology is illustrated in a clinical application; an application to discrete data is described in supplementary material.

Keywords

Model checking; model criticism; model hierarchy; discrepancy measures; Markov chain Monte Carlo; posterior-predictive density

1 Introduction

In this article, we propose a class of model diagnostics that can be conveniently applied to assess model adequacy in Bayesian hierarchical models. Our method is based on the evaluation of pivotal discrepancy measures (PDMs), which are defined as functions of the data and model parameters that have an invariant distribution when evaluated at the data-generating (true) parameter value. A key result of this article is that the distributions of PDMs have the same invariant distribution when evaluated at parameter values drawn from the posterior distribution. This result holds if a PDM is a function of both data and parameters (Johnson, 2007), or parameters alone.

Diagnostics based on PDMs offer three important advantages over other commonly used model diagnostics. First, these diagnostics can be computed directly from the standard output of Markov chain Monte Carlo (MCMC) algorithms. Thus, posterior-predictive or prior-predictive sampling computations are not required for their use. Second, because PDMs can be defined as functions of model parameters only, they can be applied to diagnose model inadequacy at any level of a hierarchical model. Other model diagnostics, like those based on posterior-predictive sampling, do not share this property. Finally, because the marginal distribution of PDMs is known *a priori*, it is possible to calibrate their

[*] vejohnson@mdanderson.org.

Supplementary Material
The extension of our methodology to a discrete data set (referenced in Section 2.3) is provided in a Web Appendix available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

values exactly to obtain either uniformly distributed *p*-values or exact probabilistic bounds on distributions of order statistics of posterior samples of PDMs.

In current practice, most Bayesian model diagnostics are based on posterior-predictive methodology (e.g., Gelman, Meng and Stern, 1996). Like PDM diagnostics, posterior-predictive model diagnostics are based on the evaluation of discrepancy measures, which are chosen to detect specific departures from model assumptions. Under this approach, the adequacy of a model is assessed by comparing the value of the discrepancy measure evaluated at observed data to values of the discrepancy measure evaluated at data simulated from the posterior-predictive density. Although conceptually simple to implement, this approach suffers from a number of practical and theoretical shortcomings. From a practical perspective, posterior-predictive model assessment sacrifices statistical power to detect model inadequacy because observed data tend to be more consistent with the posterior distribution (which they have been used to compute) than are random values drawn from that posterior. This deficiency is related to another deficiency of posterior-predictive model diagnostics; namely, that posterior predictive *p*-values are not uniformly distributed under the assumed model. This makes the interpretation and comparison of posterior-predictive *p*-values difficult (e.g., Robins, van der Vaart, and Ventura, 2000).

Posterior-predictive model diagnostics are, by definition, also of limited value in their ability to detect model departures in hierarchical models. The evaluation of posterior-predictive model diagnostics implicitly depends on comparisons of the value of the discrepancy function evaluated at observed data to values of the discrepancy function evaluated at data simulated from the posterior distribution. As a consequence, posterior-predictive calibration of discrepancy measures can only occur at the data-generating level of a hierarchical model. This restriction severely limits the application of posterior-predictive methods for assessing the suitability of the assumptions made at levels in a model hierarchy in which the data do not appear.

As an alternative to posterior-predictive model checks, Dey et al. (1998) proposed what might be called prior-predictive model diagnostics. These diagnostics are based on comparisons of posterior distributions of a discrepancy measure evaluated at observed data to posterior distributions of data simulated from the prior-predictive density. Such prior-predictive assessment typically provides better statistical power than the posterior-predictive method, and *p*-values based on this method are uniformly distributed under the assumed model. However, prior-predictive model diagnostics are expensive to compute–generally increasing computation by an order of magnitude compared to that required to fit the original model.

Other approaches to assessment of Bayesian models include those of Bayarri and Berger (2000), Robins, van der Vaart, and Ventura (2000), and Hjort, Dahl and Steinbakk (2006), who proposed model diagnostics that can be calibrated to produce uniformly distributed *p*-values. Marshall and Spiegelhalter (2007) investigated a diagnostic test based on measuring the conflict between prior and likelihood replicates using cross-validation. Based on a measure of the local conflict between the prior and likelihood, Scheel, Green and Rougier (2010) proposed a graphical diagnostic for identifying influential model choices in Bayesian hierarchical models. Other methods for hierarchical model assessment include those of Hodges (1998), and Dahl, Gåsemyr and Natvig (2007).

2 Methodology

2.1 A motivating example

To motivate our method, we consider an example that stems from a clinical trial conducted at M.D. Anderson Cancer Center, which concerns radiation therapy in the treatment of esophageal cancer. Radiation therapy anywhere in the thorax potentially damages lung tissue, resulting in an inflammatory lung response known as radiation pneumonitis. Radiation pneumonitis can be extremely serious and even lethal in a subset of patients. Oncologists can estimate the level of radiation injury to the lungs by measuring the uptake of fluorescent-tagged glucose analogs in lung tissue following treatment. A preliminary statistical analysis of the uptake data collected at various locations in the patients' lung images suggested a linear relationship with the radiation dose. The slope and intercept of the line observed for each patient was further hypothesized to be related to the occurrence of radiation pneumonitis (Guerrero et al. 2007). The goal of our analysis is to evaluate the adequacy of a hierarchical linear model for describing the relation between the standardized uptake values (SUVs) of the glucose analog and the radiation dose.

A total of I = 36 patients with esophageal cancer were treated and subsequently underwent positron emission tomography (PET) imaging to measure glucose uptake. SUVs derived from the PET images were then measured at various voxels within the affected anatomy and were matched to the radiation dose level delivered at those locations. We considered eight models to describe the relation between SUVs and radiation dose. The simplest was a three-level linear hierarchical model of the following form:

$$y_{ij} \sim N(\alpha_i+\beta_i x_j, \sigma_{ij}^2), \quad i=1,\ldots,I, \quad j=1,\ldots,J_i, \tag{1}$$

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left\{\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \boldsymbol{R}=\begin{pmatrix} \tau_1^2, 0 \\ 0, \tau_2^2 \end{pmatrix}\right\}, \quad \sigma_{ij}^2 \equiv \sigma^2 \sim IG(\gamma,\gamma), \tag{2}$$

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \delta^2, 0 \\ 0, \delta^2 \end{pmatrix}\right\}, \quad \tau_1^2, \tau_2^2 \sim IG(\nu,\nu). \tag{3}$$

Here, y_{ij} denotes an SUV measured in patient i at radiation dose level x_j, and α_i and β_i are patient-specific intercept and slope parameters. For these data, the same 50 dose levels were measured for each patient, so that $J_i \equiv 50$. Vague priors were assumed for $(\alpha, \beta)'$ and σ^2 by assuming a large value for δ and small values for γ, and for τ_1 and τ_2 by setting $\nu = 10^{-6}$. We chose a value of $\nu = 10^{-6}$ to reflect vague prior information regarding the parameters in the second stage variance matrix; we note that the empirical variances of the least squares estimates of the intercept and slope parameters were 0.026 and 0.552, respectively. For convenience, we refer to equations (1–3) as the first, second, and third levels of the model hierarchy, respectively. Seven other models were defined to have a similar structure in the first level of the model structure, but differed in their assumptions for the second and third levels. In the simplest of these models depicted above, σ^2 was assumed to be constant across all observations. In the more complicated models, the observational variance was allowed to differ according to dose or subject, which is why σ_{ij}^2 appears with an index in equations (1–2). A summary of all eight models is provided in Table 1 (discussed further in Section 4).

2.2 Pivotal discrepancy measures

Central to our methodology is the concept of a pivotal discrepancy measure, which is defined as a function of the data and model parameters, say $d(\mathbf{y}, \boldsymbol{\theta})$, that has a known and invariant sampling distribution when data $\boldsymbol{y}$ are generated from the model indexed by $\boldsymbol{\theta}$. For the purpose of this article, it is critical to distinguish between the values of the data-generating parameters and parameters sampled from the posterior distribution. We will denote the data-generating parameters with a 0 superscript, i.e., $\boldsymbol{\theta}^0$, and use $\tilde{\boldsymbol{\theta}}$ to indicate parameter values drawn from the posterior distribution.

Conceptually, the data generation scheme under a Bayesian hierarchical model can be equated to the following sampling process. First, a single value of the parameter defined at the highest level of a model hierarchy is generated. Based on this value, model parameters are generated at the second highest level of the hierarchy, and so on until the data in the final model stage are generated. From this perspective, it is easy to see that the quantity $(y_{ij} - \alpha_i^0 - \beta_i^0 x_j)/\sigma_{ij}^0$ in equations (1–3) is a pivotal quantity, being distributed as a $N(0, 1)$ random variable. Of course, this sampling interpretation of the hierarchical model is not necessary to obtain the results that follow; the same probability calculus applies if a more subjective Bayesian perspective is taken.

Johnson (2007) described the use of pivotal quantities that were functions of data $\mathbf{y}$ and $\boldsymbol{\theta}$ for model assessment. He showed that the sampling distribution of $d(\boldsymbol{y}, \boldsymbol{\theta}^0)$ was the same as the sampling distribution of $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ for arbitrary pivotal quantities derived from proper statistical models. His proof, however, required that the pivotal quantity be a function of the observed data $\boldsymbol{y}$. A more general result is required for the application of that methodology to hierarchical models, because many assumptions made in hierarchical models do not involve data. For example, Johnson's results do not immediately extend to the definition of pivotal quantities in the second level of the model, defined in equation (2). For this reason, we generalize the notion of a pivotal quantity and define a pivotal discrepancy measure (PDM) as a function of either data and parameters or of parameters alone, that has an invariant sampling distribution when evaluated at data and parameter values drawn from the assumed model. We denote a general PDM by $d(\mathbf{y}, \boldsymbol{\theta})$. In the hierarchical model of Section 2.1, $(\alpha_i^0 - \alpha^0)/\tau_1^0$ is a PDM that follows a $N(0, 1)$ distribution. The generalization of pivotal quantities to PDMs is important because it allows us to define model diagnostics at all levels within a hierarchical model. We note that diagnosing the fit of higher levels of a hierarchical model can be critical for model assessment when the hierarchical structure is important for inference, for example in predictive inference.

When $\tilde{\boldsymbol{\theta}}$ is a value drawn from the posterior distribution on $\boldsymbol{\theta}$ given $\mathbf{y}$, Johnson (2007) showed that the distribution of $d(\mathbf{y}, \boldsymbol{\theta}^0)$ is identical to the distribution of $d(\mathbf{y}, \tilde{\boldsymbol{\theta}})$ when the PDM depends on both $\mathbf{y}$ and $\boldsymbol{\theta}$. The following lemma extends this result to arbitrary PDMs.

Lemma 1—Suppose that $d(\mathbf{y}, \boldsymbol{\theta}^0)$ is a PDM distributed according to F. If $\tilde{\boldsymbol{\theta}}$ is drawn from the posterior distribution on $\boldsymbol{\theta}$ given $\mathbf{y}$, then $d(\mathbf{y}, \tilde{\boldsymbol{\theta}})$ is also distributed according to F.

In particular, this lemma applies even if $d(\boldsymbol{y}, \boldsymbol{\theta}^0) \equiv d(\boldsymbol{\theta}^0)$; its proof appears in the Appendix.

With this result, our general strategy for hierarchical model assessment can be described in two steps. First, we identify a pivotal discrepancy measure $d(\boldsymbol{y}, \boldsymbol{\theta})$ with a known sampling distribution F that targets a specific departure from the model assumptions. Second, we assess the adequacy of the model by determining if $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ can be regarded as a draw from F. That is, we treat $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ as a standard test statistic.

2.3 Construction of pivotal discrepancy measures

In the general process of model assessment, it is useful to examine both global goodness-of-fit diagnostics, as well as diagnostics targeted at specific features of a model. Targeted diagnostics are useful when a particular aspect of a model is questioned and are especially important when the number of parameters involved in the model assumption under question is small. Omnibus tests are useful for examining overall model fit and can be used to screen models to determine if more specific assumptions need to be tested. One omnibus PDM that we find useful is based on standardized residuals. For ease of exposition, we define this statistic for residuals obtained in the first level of a normal hierarchical model, although the extension to other exponential family models and other levels within a model hierarchy is straightforward. We assume throughout that observations are conditionally independent, given the value of the model parameter, in the first stage of the model. The idea behind this discrepancy measure is illustrated in Figure 1 and can be described as follows.

1. Let $\mu_i(\tilde{\boldsymbol{\theta}})$ and $V_i(\tilde{\boldsymbol{\theta}})$ denote the mean and variance of observation y_i based on a value of $\tilde{\boldsymbol{\theta}}$ drawn from the posterior distribution. Define standardized residuals according to $w_i = V_i(\tilde{\theta})^{-\frac{1}{2}}\{y_i - \mu_i(\tilde{\theta})\}$.

2. Partition $\{w_i\}$ into K groups according to the values of $\mu_i(\tilde{\boldsymbol{\theta}})$. The motivation for this partitioning procedure is that the residuals in the same group tend to deviate from 0 in the same direction when the mean structure is not correctly specified. Consequently, grouping residuals according to their expectations yields higher power for detecting model misspecification. In Figure 1, the thick vertical lines indicate a partition of hypothetical residuals from a simple linear regression model into three groups.

3. Within the kth group of residuals, $k = 1, \ldots, K$, construct a PDM $d_k(\boldsymbol{y}, \boldsymbol{\theta})$ that has a χ^2 distribution under the assumed model by binning residuals into L cells based on the value of their cumulative distribution function at the sampled parameter $\tilde{\boldsymbol{\theta}}$, using predetermined bin boundaries $0 = b_0 < \cdots < b_L = 1$. For example, in the first-level model, equation (1), residuals are placed into bin l if the value of $\Phi\{(y_{ij} - \alpha_i^0 - \beta_i^0 x_j)/\sigma_{ij}^0\}$ falls in the interval $(b_{l-1}, b_l]$, where $\Phi(\cdot)$ denotes the standard normal distribution function. Cell boundaries determined in this way are indicated by the horizontal lines in Figure 1. Letting n_k denote the total number of residuals in group k, O_l the observed number of residuals in bin l and $p_l = \Phi(b_l) - \Phi(b_{l-1})$, a χ^2 statistic $d_k(\boldsymbol{y}, \boldsymbol{\theta})$ can be defined by

$$d_k(\boldsymbol{y}, \theta) = \sum_{l=1}^{L} \left(\frac{O_l - n_k p_l}{\sqrt{n_k p_l}} \right)^2.$$

   It is important to note that the degrees of freedom of this χ^2 statistic is $L - 1$, because the distribution of the PDM evaluated at a posterior sample is the same as it is under the data-generating value of the parameter. Thus, no degree-of-freedom adjustments for model complexity are required.

4. Sum the χ^2 statistics defined for each of the K groups to obtain a global PDM. The resulting discrepancy measure takes the form

$$d(\boldsymbol{y}, \theta) = \sum_{k=1}^{K} d_k(\boldsymbol{y}, \theta). \quad (4)$$

When the assumed model is correct and the sample size is large, $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ approximately follows a $\chi^2_{K(L-1)}$ distribution.

For discrete data, it is often useful to define discrepancy measures by randomizing probability-transformed values to bins. Letting g and G denote the probability mass function and cumulative density function of a discrete random variable $\boldsymbol{y}$, uniform deviates can be obtained according to

$$z_i = G(y_i^- | \theta) + u g(y_i | \theta), \tag{5}$$

where u is a random uniform deviate and y_i^- denotes the largest value in the sample space that is smaller than y_i. At both the true parameter and a parameter value drawn from the posterior distribution given $\boldsymbol{y}$, $\{z_i\}$ are independent and follow a uniform distribution. The vector $\mathbf{z} = \{z_i\}$ can thus be used to define discrepancy measures similar to those described above for continuous data. An example of the use PDMs to assess the fit of models for binary data is provided in the supplemental material.

We note, of course, that our purpose in performing this randomization is not the same as that used in performing a randomized test in the Neyman-Pearson framework. Our randomization procedure is performed over many draws from an MCMC sampling algorithm; thus, no single value of u dominates the procedure.

2.4 Assessing the joint distribution of PDMs

We assess model adequacy by treating $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ as a test statistic drawn from the reference distribution F. Let $\{\tilde{\boldsymbol{\theta}}^j, j = 1, \ldots, J\}$ denote samples from the posterior distribution of $\boldsymbol{\theta}$ given a fixed value $\boldsymbol{y}$ obtained from the output of an MCMC algorithm. Although the marginal distribution of each element of $\{d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}}^j)\}$ is known (i.e., F), the values of the PDM evaluated at draws from the same posterior are not independent because they are based on the same value of $\boldsymbol{y}$ (Johnson, 2004). This dependence makes summarizing evidence against a model based on a posterior sample $\{\tilde{\boldsymbol{\theta}}^j\}$ more complicated. Johnson (2007) proposed several schemes for circumventing this difficulty for pivotal quantities dependent on data.

The simplest way to evaluate the distribution of the $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}}^j)$ values is to graphically compare a histogram of their values to the reference distribution F. If model fit is reasonable, then the histogram should fall within the central region of F.

A more formal comparison between the posterior distribution of $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ and F can be based on bounds on order statistics of dependent samples of random variables (Caraux and Gascuel, 1992; Rychlik, 1992). Let $d_{(r)}$ denote the rth order statistic of the $\{d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}}^j)\}$ values. Then it follows that

$$P(d_{(r)} > t) \le min\left[1, \frac{J\{1 - F(t)\}}{J - r + 1}\right], \tag{6}$$

where J is the posterior sample size. For example, if t equals the 0.99 quantile of F, then for large J the probability that $d_{(0.8J)} > t$ is less than or equal to 0.05. In other words, a finding that 20% of $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ values exceeds the .99 quantile from the nominal distribution implies a p-value less than or equal to 0.05. To eliminate the necessity of choosing r, we extend Johnson's approach (2007) and perform a numerical search over the order statistics to identify the minimum value of the resulting p-value, say $p_{\min}$. The numerical search of $p_{\min}$

is computationally trivial, involving only simple evaluation of the right side of equation (6) at the marginal distribution of each order statistic.

Several points should be noted with regard to identifying $p_{\min}$. First, in order to avoid defining a value of $p_{\min}$ that is dependent on the particular posterior sample $\{d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})\}$ obtained from the MCMC algorithm, J should be chosen large enough so that uncertainty in the r/Jth quantile of the distribution of $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ is small. Second, in the case in which the distribution of $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ is not exact, the extreme tails of the distribution of $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ should be excluded from the search. In our simulation and examples, we excluded the upper 0.5% percentile of the PDMs. Finally, the bounds determined by equation (6) are typically quite conservative. With these comments in mind, we propose the following rule of thumb for evaluating model adequacy based on $p_{\min}$.

1. If $p_{\min} < 0.05$, there is strong evidence of model inadequacy.
2. If $0.05 < p_{\min} < 0.25$, there is some evidence of inadequacy, and the posterior distribution of PDMs warrants more precise evaluation using prior predictive posterior simulations (e.g., Dey et al 1998; Johnson 2007).
3. If $p_{\min} > 0.25$, the diagnostic does not provide evidence of lack of fit.

For case 2, the use of prior-predictive assessment of the PDM essentially makes our proposal equivalent to the method of Dey et al. (1998). The computational burden associated with obtaining a formal p-vaue in this case will often be prohibitive, and so we do not generally recommend formal calibration of the PDM's p-value with prior-predictive methods. Instead, we recommend that values of $p_{\min} < 0.25$ simply be regarded as suggesting evidence of model lack of fit.

Finally, we note that this rule of thumb can be applied for PDMs defined at any level of the model hierarchy, including PDMs that do not depend on first-level data.

3 Simulation studies

To evaluate the utility of model diagnostics based on PDMs, we performed a simulation study to investigate their performance. In this study, we fit simulated data to the three-level hierarchical model in equations (1–3). In each set of simulated data, $I = 30$ subjects were observed at $J_i = 50$ doses. To test the ability of PDMs and our rule of thumb to detect model departures, we simulated data under schemes that violated both first- and second-level model assumptions.

Three types of inadequacy in the first level of the model were considered:

1. Misspecified mean structure, in which $y_{ij} \sim N(\alpha_i+\beta_i x_j+\gamma x_j^2, \sigma^2)$ for $\gamma = 1.6$, 2.2, and 3.0 (i.e., a quadratic effect was added).
2. Heterogeneous variance, in which $y_{ij} \sim N(\alpha_i+\beta_i x_j, x_j^c \sigma^2)$ for $c = 0.28$, 0.35, and 0.5.
3. Non-normality of errors, in which $y_{ij} \sim t_f(\alpha_i + \beta_i x_j, \sigma^2)$ where $t_f(a, b)$ denotes a t-distribution with a mean a, variance b and degrees of freedom $f = 4$, 5, and 6.

In addition, we also simulated two types of violations in the second level of the model:

1. Non-exchangeability of regression parameters, in which $\beta_i \sim N(\beta, \tau_2^2)$ for $i = 1, \cdots, 15$, and $\beta_i \sim N(\beta+c, \tau_2^2)$ for $i = 16, \cdots, 30$, with $c = -2, -2.5$, and -3.

2. Non-normality of the second-level distribution, in which $\beta_i \sim t_f(\beta, \tau_2^2)$ for $f = 2$, 3 and 5.

Data generating values of other parameters were simulated from the following prior distributions: $\tau_1^2 \sim IG(10, 11)$, $\tau_2^2 \sim IG(10, 11)$, $\sigma^2 \sim IG(10, 11)$, $\alpha \sim N(2, 5^2)$ and $\beta \sim N(2, 5^2)$. We set $x_j = (j - 1)/50$ for $j = 1, \cdots, 50$. The vague prior densities described in Section 2.1 were used to compute the posterior distributions for each simulated data set.

Model assessment was based on 10,000 posterior draws of parameters from a Gibbs sampler. We used the PDM in equation (4) as the first-level pivotal discrepancy measure. To construct this discrepancy measure, a total of 50 residual groups were formed according to each value of x. Residuals for each x value were partitioned into 4 equal probability cells according to the sampled value of the normal variance. To facilitate comparisons with Dey et al. (1998), we defined the PDM for assessing the second stage model to be

$$d_\beta = \sum_{i=1}^{30} (\beta_i - \beta)^2 / \tau_\beta^2.$$

The rule of thumb described in Section 2.4 was used to determine the adequacy of each simulated model. We compared our method to the prior-predictive method of Dey et al. (1998) and the posterior-predictive method of Gelman et al. (1996) using the same discrepancy functions for each method. For all methods, the null hypothesis of model adequacy was rejected when the p-value was less than 0.05. In the prior-predictive approach, the empirical distribution of the discrepancy measures was estimated from 1,000 prior-predictive replicates. Similarly, 1,000 posterior-predictive data values were drawn from each posterior distribution to perform posterior-predictive diagnosis. The posterior-predictive method could not be calibrated for d_β because d_β is not a function of y; thus, posterior-predictive methodology was not used to assess second-level model assumptions.

Table 2 displays the operating characteristics of the three diagnostics under each of the five data generation schemes. For first-level model assessment, the PDM method had substantially higher power to detect model inadequacy than the posterior-predictive method, even when based on the conservative bounds in equation (6). These gains are quite remarkable since they are also accompanied by a drastic reduction in computational effort.

As expected, the PDM method based on the exact probabilistic bounds was less powerful than exact prior-predictive simulations. Nonetheless, the power of the PDM method was comparable to or better than that of the prior-predictive method when models were flagged whenever $p_{min} < 0.25$ (i.e., the numbers in parentheses in Table 1). In terms of computational expense, the PDM method offers an enormous advantage over the prior-predictive method. The PDM diagnostics were computed directly from standard MCMC output, whereas the prior-predictive diagnostics required the simulation of 1,000 additional chains for each data value generated. Similar comments apply to the second-level model assessment. The PDM method was again slightly less powerful than the prior-predictive method, but was dramatically faster in terms of the computing time.

It is also interesting to note that the first-level model diagnostics had little power in detecting second-level model deviations, and that the second-level diagnostics had little power to detect first-level violations. Clearly, targeting of diagnostics to a given level succeeds at isolating the effect of a deficiency at that level.

4 An application to dose-uptake data

We return to the motivating example described in Section 2.1. Eight candidate models were applied to the dose-uptake data. The models differed according to their assumptions regarding the variance components in the second and third levels of the model (see Table 1). In each model, we used diffuse priors for σ^2 (or σ_i^2) of the form $\pi(\sigma^2) \propto 1/\sigma^2$, and vague priors for $\boldsymbol{R}$. When $\boldsymbol{R}$ was assumed to be of the form $\boldsymbol{R}=\mathrm{diag}(\tau_1^2, \tau_2^2)$, we assumed $\tau_1^2 \sim IG(10^{-6}, 10^{-6})$ and $\tau_2^2 \sim IG(10^{-6}, 10^{-6})$. When $\boldsymbol{R}$ had the more general structure

$$\begin{pmatrix} \tau_1^2, \tau_{1,2} \\ \tau_{1,2}, \tau_2^2 \end{pmatrix},$$

its prior was assumed to be an inverse-Wishart distribution on two degrees of freedom and a scale matrix diag$(10^{-6}, 10^{-6})$. A noninformative prior was assigned to α and β, i.e., we assumed $\pi(\alpha, \beta) \propto$ constant.

Because the number of observations was large (1,800), we applied the PDM in equation (4) to assess the adequacy of the first-level model. We formed 50 residual groups according to the values of the tissue exposure, x_j, and within each group we used normal quantiles to partition the residuals into 4 cells. The resulting pivotal discrepancy measure nominally follows a χ^2_{150} under the assumed model.

The second-level model was examined by investigating the distribution of PDMs given by $\varepsilon_{\alpha_i} = (\alpha_i - \alpha)/\tau_1$ and $\varepsilon_{\beta_i} = (\beta_i - \beta)/\tau_2$. Figure 2 displays a normal scores plot for a single posterior draw of these standardized residuals from the same MCMC iteration under model 1 in Table 1. According to the lemma, these residuals are marginally distributed as independent standard normal deviates when evaluated at a single draw from the posterior distribution when the second-level model is correct. Figure 2 does not suggest a lack of fit based on the residuals ε_{α_i}. However, it is clear that the residuals ε_{β_i} are not normally distributed. Similar patterns were observed for the residuals obtained under all the other models listed in Table 1.

These observations prompted us to use the Shapiro-Wilks test statistic as the PDM to conduct a more formal assessment of the second-level model. We assessed the adequacy of the model specification of α_i and β_i separately. Specifically, we applied the Shapiro-Wilks test to $\{\tilde{\alpha}_i\}$ and $\{\tilde{\beta}_i\}$ to obtain bounds on the p-values.

Table 1 summarizes the diagnostics for eight models using our method, each based on 10,000 posterior draws of the parameters. From these model diagnostics, we see that models 1–6 clearly failed to provide an adequate fit to the data in the first level of the hierarchy. Under each of these models, all 10,000 sampled values of the PDM exceeded the 0.95 quantile of the reference χ^2_{150} distribution, and $p_{\min}$ was always less than 10^{-4}. A lack of fit was also detected by the posterior-predictive method (p-value < 10^{-4}). Model 7 provided a better fit than models 1 to 6; however, it still led to 56.8% of sampled PDM values exceeding the 0.95 quantile of the χ^2_{150} distribution; $p_{\min}$ was 0.02, suggesting a significant lack of fit for this model as well. The posterior-predictive diagnostic failed to detect this lack of fit; its p-value was 0.07. For model 8, only 9% of the PDM values from model 8 were larger than the 0.95 quantile of χ^2_{150}, and $p_{\min}$ was 0.43, suggesting an adequate model fit. This finding was consistent with the posterior-predictive method (p-value = 0.26).

In the second-level model assessment, no model lack of fit was detected for the intercept terms. Except for Model 5, the percentage of PDM values that exceeded the 95% quantile of the reference distribution was less than 10%, and $p_{\min} > 0.35$. Although Model 5 demonstrated more evidence of lack of fit, it was not decisive based on our rule of thumb. Our diagnostic did indicate a strong lack of fit for the distribution of the β_i parameters in all models; $p_{\min}$ was always less than 10^{-4}.

The distribution of the discrepancy measures can also be explored graphically. Figure 3 shows a plot of the distribution of the first-level model discrepancy values based on 10,000 posterior samples of $\tilde{\boldsymbol{\theta}}$ under each model. To facilitate interpretation, the empirical distributions of $d(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ for models 1 to 7 are smoothed, and only the empirical distribution of model 8 discrepancy values is presented as a histogram. As before, this plot clearly reveals a lack of fit of the first-level hierarchy in models 1 to 7, and suggests adequate fit of the first-level assumptions for model 8.

5 Discussion

This article proposes a simple diagnostic method to assess the adequacy of Bayesian hierarchical models. Our method is based on comparisons of PDMs computed at draws from the posterior distribution to their known reference distributions. Large discrepancies between these distributions indicate model inadequacy. Compared to other diagnostic methods for Bayesian models, which are often simulation-intensive, our diagnostics are defined using only posterior samples of parameters obtained as part of standard MCMC output.

One limitation of the proposed PDM diagnostics is that informal assessment of the joint distribution of PDMs is based on a general bound on the order statistics of the dependent variables, which applies to any form of dependency between deviates. Consequently, the *p*-value obtained from this bound can be quite conservative.

Another limitation of our method is that it requires the use of proper prior distributions on all model parameters. Although limiting arguments can be used to extend this result for particular cases and stages of hierarchical models, a more general formulation in which guidelines for defining the extent of permissible impropriety would be useful.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to thank James Hodges, the Associate Editor, and two referees for numerous comments that improved the presentation of results in this article. Ying Yuan and Valen Johnson acknowledge support from NIH grants R01 CA154591 and R01 CA158113, respectively.

References

Bayarri MJ, Berger JO. P values for composite null models. Journal of the American Statistical Association. 2000; 95:1127–1142.

Caraux G, Gascuel O. Bounds on distribution functions of order statistics for dependent variates. Statistics and Probability Letters. 1992; 14:103–105.

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

Dey DK, Gelfand AE, Swartz TB, Vlachos PK. A simulation-intensive approach for checking hierarchical models. Test. 1998; 7:325–346.

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

Guerrero T, Johnson VE, Hart J, Pan T, Khan M, Luo D, Liao Z, Ajani J, Stevens C, Komaki R. Radiation pneumoitis: local dose versus fluorodeoxyglucose uptake response in irradiated lung. International Journal of Radiation Oncology, Biology, and Physics. 2007; 68:1030–1035.

Hodges JS. Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). Journal of the Royal Statistical Society, Series B. 1998; 60:497–536.

Johnson VE. A Bayesian χ^2 test for goodness-of-fit. Annals of Statistics. 2004; 32:2361–2384.

Johnson VE. Bayesian model assessment using pivotal quantities Bayesian Analysis. 2007; 2:719–734.

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

Marshall EC, Spiegelhalter DJ. Identifying outliers in Bayesian hierarchical models: a simulation-based approach. Bayesian Analysis. 2007; 2:409444.

Robins JM, van der Vaart A, Ventura V. Asymptotic distribution of P values in composite null models. Journal of the American Statistical Association. 2000; 95:1143–1159.

Rychlik T. Stochastically extremal distributions of order statistics for dependent samples. Statistics and Probability Letters. 1992; 13:337–341.

Scheel I, Green PJ, Rougier JC. A graphical diagnostic for identifying influential model choices in Bayesian hierarchical models. Scandinavian Journal of Statistics. 2010 online early view. 10.1111/j.1467-9469.2010.00717.x

Appendix: Proof of lemma

Proof

When the PDM depends on **y**, a proof of this lemma follows from Theorem 1 in Johnson (2007). We focus here on the case in which the PDM depends only on $\boldsymbol{\theta}$. Let $\pi(\boldsymbol{\theta})$ denote the prior distribution on $\boldsymbol{\theta}$, $f(\boldsymbol{y}|\boldsymbol{\theta})$ denote the sampling distribution of $\boldsymbol{y}$ given $\boldsymbol{\theta}$ for $\mathbf{y} \in \mathbf{Y}$, and $m(\boldsymbol{y})$ denote the marginal distribution of y. Let Θ denote the parameter space, and let A denote any subset of Θ. Then it follows that

$$\begin{aligned}
\Pr(\tilde{\theta} \in A) &= \int_{\Theta}\int_{\boldsymbol{Y}}\int_{\Theta} I_{A(\tilde{\theta})}\, p(\tilde{\theta}|\boldsymbol{y}) f(\boldsymbol{y}|\theta)\pi(\theta)\mathrm{d}\theta \mathrm{d}\boldsymbol{y}\mathrm{d}\tilde{\theta} \\
&= \int_{\Theta}\int_{\boldsymbol{Y}}\int_{\Theta} I_{A(\tilde{\theta})} \frac{f(\boldsymbol{y}|\tilde{\theta})\pi(\tilde{\theta})}{m(y)} f(\boldsymbol{y}|\theta)\pi(\theta)\mathrm{d}\theta \mathrm{d}\boldsymbol{y}\mathrm{d}\tilde{\theta} \\
&= \int_{\Theta}\int_{\boldsymbol{Y}}\int_{\Theta} I_{A(\tilde{\theta})} f(\boldsymbol{y}|\tilde{\theta})\pi(\tilde{\theta}) \frac{f(\boldsymbol{y}|\theta)\pi(\theta)}{m(y)} \mathrm{d}\theta \mathrm{d}\boldsymbol{y}\mathrm{d}\tilde{\theta} \\
&= \int_{\Theta}\int_{\boldsymbol{Y}}\int_{\Theta} I_{A(\tilde{\theta})} f(\boldsymbol{y}|\tilde{\theta})\pi(\tilde{\theta}) p(\theta|\boldsymbol{y})\mathrm{d}\theta \mathrm{d}\boldsymbol{y}\mathrm{d}\tilde{\theta} \\
&= \int_{\Theta}\int_{\boldsymbol{Y}} I_{A(\tilde{\theta})} f(\boldsymbol{y}|\tilde{\theta})\pi(\tilde{\theta})\mathrm{d}\boldsymbol{y}\mathrm{d}\tilde{\theta} \\
&= \int_{\Theta} I_{A(\tilde{\theta})} \pi(\tilde{\theta})\mathrm{d}\tilde{\theta} \\
&= \Pr(\theta^0 \in A)
\end{aligned}$$

Since the distribution of $\tilde{\boldsymbol{\theta}}$ is the same as $\boldsymbol{\theta}^0$, so is the distribution of PDMs based on $\boldsymbol{\theta}^0$.

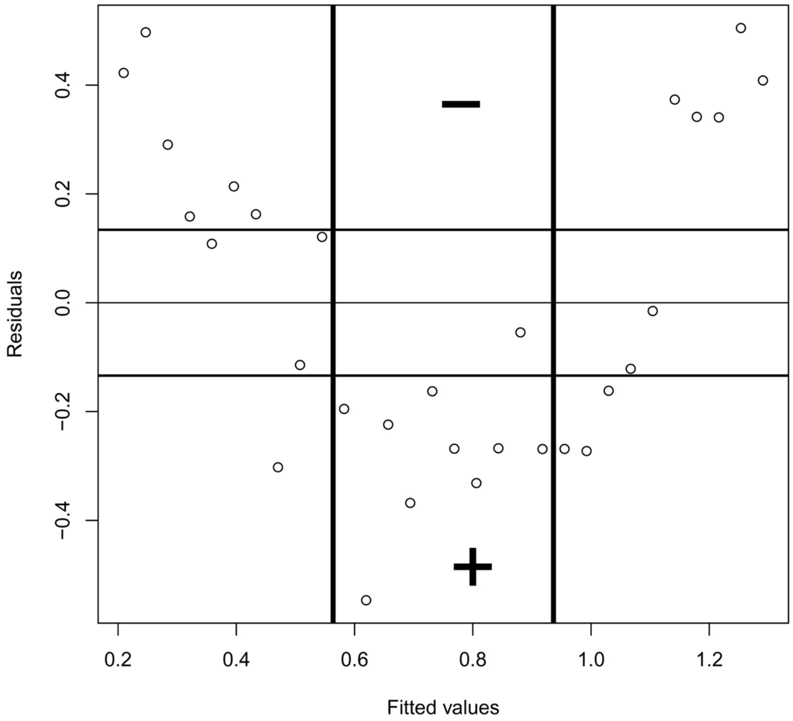

Figure 1.

Construction of a chi-squared discrepancy measure. Standardized residuals are first split into three groups using the two thick vertical lines, and then partitioned into chi-squared cells according to the two thick horizontal lines. These residuals were obtained from a simple linear model that failed to incorporate a quadratic effect. Note that by partitioning the observations into K=3 groups, the deviation from linearity is more clearly detected in each cell. For example, the two cells marked with "−" and "+" contribute $(0 - 3.33)^2/3.33 = 3.33$ and $(9 - 3.33)^2/3.33 = 9.63$, respectively, to the second χ_2^2 deviate, and these large values would be missed if all 30 residuals in the plot were collapsed to form a single χ_2^2 random variable based only on the horizontal partitions.

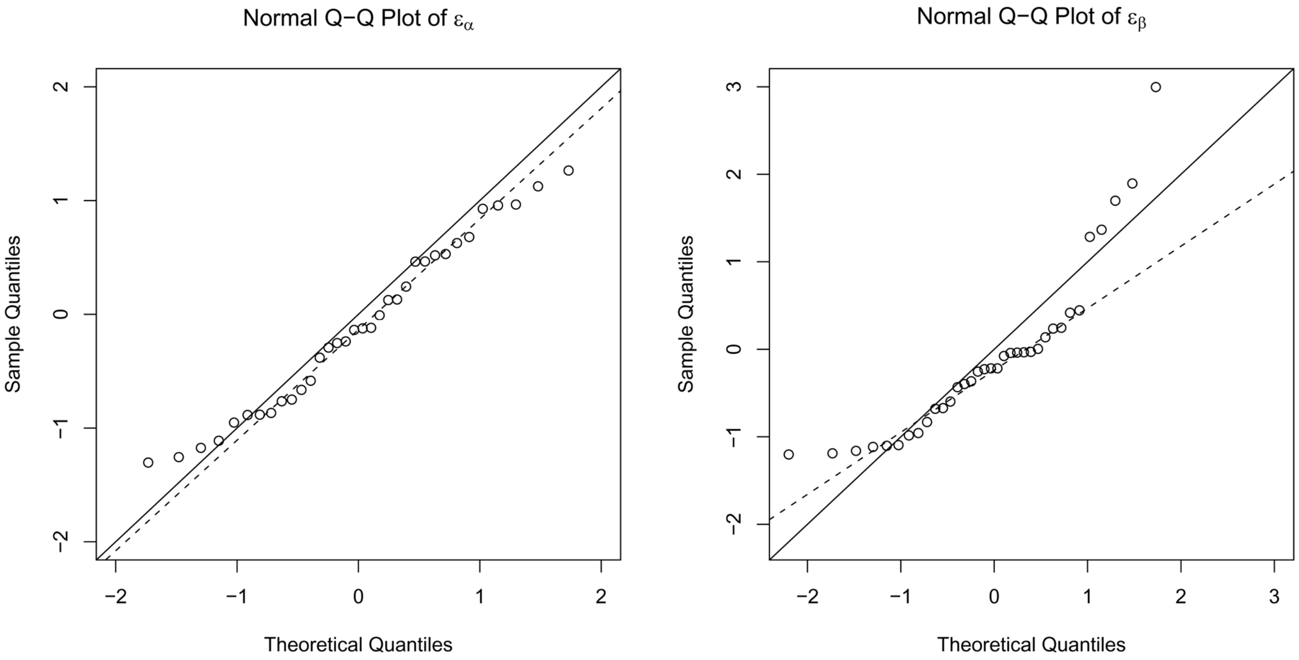

Figure 2.

Quantile-quantile plots of residuals ε_α and ε_β for a single, randomly selected parameter drawn from the posterior distribution obtained within one update in the MCMC algorithm. The dashed lines represent the defaults for the function qqline (from the R soft-ware package), which pass through the upper and lower quartiles of the empirical distribution function. The 45-degree line through the origin is displayed for reference.

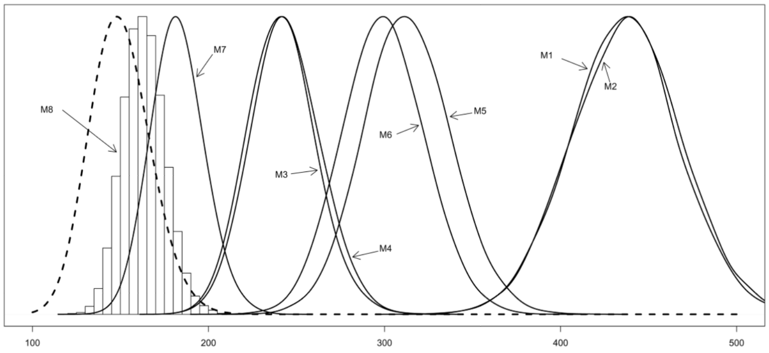

Figure 3.

Histogram estimates of the posterior distribution of the first-level discrepancy measure $d(\boldsymbol{y}, \boldsymbol{\theta})$ under eight models (M1 to M8). The histogram estimates under models 1 to 7 have been smoothed. The reference χ^2_{150} distribution is displayed as a broken line.

Table 1

Summaries of assessment of eight models for dose-uptake data. Models differ according to their assumptions regarding σ^2_{ij} and $\boldsymbol{R}$. For both first-level and second-level assessments, the first row in each section shows the percentage of discrepancy measure values larger than the 95% quantile of the corresponding reference distributions. The second row in each section presents the probabilistic upper bound on the *p*-values. For the first-level model assessment, the third row shows the posterior-predictive *p*-values. Posterior-predictive *p*-values cannot be computed for the second-level model and so are not represented in the table.

Model		M1	M2	M3	M4	M5	M6	M7	M8
σ^2_{ij}		$\boldsymbol{\sigma^2}$	$\boldsymbol{\sigma^2}$	σ^2_i	σ^2_i	$\boldsymbol{x_j\sigma^2}$	$\boldsymbol{x_j\sigma^2}$	$x_j\sigma^2_i$	$x_j\sigma^2_i$
$\boldsymbol{R}$		$\begin{pmatrix}\tau^2_1, 0\\ 0, \tau^2_2\end{pmatrix}$	$\begin{pmatrix}\tau^2_1, \tau_{12}\\ \tau_{12}, \tau^2_2\end{pmatrix}$	$\begin{pmatrix}\tau^2_1, 0\\ 0, \tau^2_2\end{pmatrix}$	$\begin{pmatrix}\tau^2_1, \tau_{12}\\ \tau_{12}, \tau^2_2\end{pmatrix}$	$\begin{pmatrix}\tau^2_1, 0\\ 0, \tau^2_2\end{pmatrix}$	$\begin{pmatrix}\tau^2_1, \tau_{12}\\ \tau_{12}, \tau^2_2\end{pmatrix}$	$\begin{pmatrix}\tau^2_1, 0\\ 0, \tau^2_2\end{pmatrix}$	$\begin{pmatrix}\tau^2_1, \tau_{12}\\ \tau_{12}, \tau^2_2\end{pmatrix}$
First-level assessment		100	100	100	100	100	100	56.8	9.0
		0	0	0	0	0	0	0.02	0.43
		0	0	0	0	0	0	0.07	0.26
Second-level assessment	α_i	5.6	6.6	5.2	4.4	12.9	9.6	5.7	5.4
		0.67	0.57	0.69	1.0	0.20	0.38	0.85	0.91
	β_i	100	100	99.9	99.9	100	100	100	100
		0	0	0	0	0	0	0	0

Table 2

Percentage of rejection (%) of model adequacy in 1,000 simulations for prior-predictive, posterior-predictive and PDM methods under a hierarchical model. For the PDM method, the rule of thumb is used to determine model adequacy; numbers in parentheses are the percentage of simulations in which either strong or some evidence of inadequacy was detected (i.e., $p_{min} < 0.25$).

		First-level assessment			**Second-level assessment**	
Inadequacy	**Parameter**	**Prior predictive**	**Posterior predictive**	**Pivotal**	**Prior predictive**	**Pivotal**
			First level of the model			
Mean	$\gamma = 1.6$	32.4	9.6	13.9 (39.1)	4.9	0.4 (7.4)
	$\gamma = 2.2$	64.2	35.7	41.3 (69.5)	4.8	0.4 (7.4)
	$\gamma = 3.0$	91.7	77.8	80.7 (93.6)	4.8	0.4 (7.2)
Variance	$c = 0.28$	28.0	5.9	11.5 (36.6)	5.6	0.4 (6.6)
	$c = 0.35$	47.8	16.1	29.6 (58.3)	5.6	0.4 (6.7)
	$c = 0.50$	92.8	67.6	83.2 (96.6)	5.6	0.4 (6.7)
Distribution	$f = 6$	27.7	6.5	17.3 (42.0)	4.4	0.4 (6.0)
	$f = 5$	51.1	20.2	39.5 (67.0)	5.0	0.8 (7.2)
	$f = 4$	88.7	65.2	85.0 (94.7)	4.8	0.7 (7.7)
			Second level of the model			
Exchangeability	$c = -2$	5.6	1.0	1.4 (7.5)	42.6	8.3 (51.1)
	$c = -2.5$	5.6	0.9	1.3 (7.7)	77.2	24.3 (82.0)
	$c = -3$	5.7	0.9	1.3 (7.6)	94.7	56.9 (96.4)
Distribution	$f = 5$	3.9	0.4	0.8 (5.8)	32.3	12.6 (36.4)
	$f = 3$	4.7	0.8	1.4 (6.6)	57.5	34.7 (61.7)
	$f = 2$	4.5	0.6	0.9 (6.2)	82.5	66.4 (84.5)
None		5.0	0.4	1.0 (7.1)	5.0	0.4 (7.4)