Bayesian Analysis (2007) **2**, Number 4, pp. 719–734

Bayesian Model Assessment Using Pivotal Quantities

Valen E. Johnson[*]

Abstract. Suppose that $S(\mathbf{Y}, \boldsymbol{\theta})$ is a function of data $\mathbf{Y}$ and a model parameter $\boldsymbol{\theta}$, and suppose that the sampling distribution of $S(\mathbf{Y}, \boldsymbol{\theta})$ is invariant when evaluated at $\boldsymbol{\theta}_0$, the "true" (i.e., data-generating) value of $\boldsymbol{\theta}$. Then $S(\mathbf{Y}, \boldsymbol{\theta})$ is a pivotal quantity, and it follows from simple probability calculus that the distribution of $S(\mathbf{Y}, \boldsymbol{\theta}_0)$ is identical to the distribution of $S(\mathbf{Y}, \boldsymbol{\theta}_{\mathbf{Y}})$, where $\boldsymbol{\theta}_{\mathbf{Y}}$ is a value of $\boldsymbol{\theta}$ drawn from the posterior distribution given $\mathbf{Y}$. This fact makes it possible to define a large number of Bayesian model diagnostics having a known sampling distribution. It also facilitates the calibration of the joint sampling of model diagnostics based on pivotal quantities.

Keywords: prior-predictive density, posterior-predictive density, Bayesian model diagnostics, Bayesian chi-squared test.

1 Introduction

Within the Bayesian paradigm, a number of general approaches for model evaluation have been proposed. Although a number of informal measures have been proposed to assess model adequacy, many model assessment strategies have relied on the calculation of Bayesian analogs of frequentist p values. For example, Box (1980) proposed the calculation of Bayesian p values based on the prior-predictive distribution. In this approach, Bayesian p-values are calculated according to

$$\Pr[p(g(\mathbf{y})|A) < p(g(\mathbf{y}_d|A))],$$

where $\mathbf{y}$ denotes a value of the data generated from the joint distribution on model parameters and data, $\mathbf{y}_d$ denotes the value of the data actually observed, A refers to prior information available before the observation of $\mathbf{y}_d$, and $p(g(\mathbf{y})|A)$ represents the prior-predictive density of a checking function g.

Unfortunately, many Bayesians have been hesitant to utilize prior-predictive model diagnostics because the values of derived p values can depend critically on the model's prior density. Perhaps in response to this problem, Guttman (1967) and Rubin (1984) proposed model diagnostics based on posterior-predictive distributions, which were later extended to general discrepancy functions by Gelman, Meng, and Stern (1996) (see also Meng (1994) for accompanying theory). The essential idea of posterior-predictive model checks is to compare the observed value of a discrepancy function—that is, a function of the observed data and unobserved model parameters—to values of the discrepancy function evaluated at replicate observations simulated from the posterior-predictive density.

[*] University of Texas M.D. Anderson Cancer Center, Houston, TX, mailto:vejohnson@mdanderson.org

ba0008

Bayarri and Berger (2000), Robins, van der Vaart, and Ventura (2000) and others have noted a critical shortcoming of posterior-predictive p values: They are not (even asymptotically) uniformly distributed. That is, the presumed sampling distributions of discrepancy functions are not actually achieved in posterior-predictive simulations. Although this fact does not preclude the use of this methodology for performing case diagnostics, it severely limits its application for formal model assessment.

Cross-validation posterior-predictive densities can also be used to assess model fit (e.g., Gelfand (1996)). In this approach, a subset of observations, say $\mathbf{Y}_r$, is omitted from the complete data $\mathbf{Y}$, while the remaining data, $\mathbf{Y}_{-r}$, is used to estimate model parameters. Model assessment is then based on the conditional predictive ordinate (CPO) $f(\mathbf{y}_r|\mathbf{y}_{-r})$. Often, $\mathbf{Y}_r$ contains only a single observation. In this case, the product of CPOs,

$$\prod_{i=1}^{n} f(y_i|\mathbf{y}_{-r}),$$

is used as a summary measure of model fit. Because the distribution of this product is generally not available analytically, it must be evaluated numerically. The computations associated with this procedure can be expensive. Gelfand (1996) provides a discussion of numerical issues surrounding the use of the CPO and offers practical advice toward resolving them.

The purpose of this article is to explore a relation between the distribution of pivotal quantities evaluated at the true (i.e., data-generating) parameter value and the distribution of the same pivotal quantities evaluated at parameter values sampled from the posterior distribution. In general, these two distributions are identical, which makes it straightforward to define a large number of Bayesian model diagnostics. Such diagnostics can be tailored for the particular application at hand and are generally easy to compute using available MCMC output.

The remainder of this article is organized as follows. In the next section, I demonstrate that the distribution of a pivotal quantity evaluated at the data-generating parameter is the same as the distribution of a pivotal quantity evaluated at a posterior draw of the model parameter. Next, several approaches for evaluating the information contained in the joint posterior distribution of pivotal quantites based on a single, observed data vector are discussed. In Section 3, two examples that illustrate the use of pivotal quantities for assessing model adequacy are presented. A summary of findings and suggestions for future research appears in the concluding section.

2 Methodology

2.1 A Property of Pivotal Quantities

Let $\mathbf{Y}$ denote a random vector defined on a sample space $\mathcal{Y} \subset R^n$ and having a probability density function belonging to a parametric family $\{f(\mathbf{y}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subset R^s\}$. Similarly, let $\pi(\boldsymbol{\theta})$ denote the prior density function assigned to $\boldsymbol{\theta}$, and let $p(\boldsymbol{\theta}|\mathbf{y})$ denote the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}$. Also, suppose that $\mathbf{Y}$ has marginal density function $m(\mathbf{y})$ defined as

$$m(\mathbf{y}) = \int_\Theta f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

For simplicity, I assume that all densities are defined with respect to Lebesgue measure, although extensions to more general settings are straightforward.

Define a *pivotal quantity* to be a function $S : (\mathcal{Y}, \Theta) \rightarrow R^t$ for which the integral

$$G(\mathbf{s}) \equiv \int_\mathcal{Y} \mathcal{I}_{A(\boldsymbol{\theta},\mathbf{y})}\ f(\mathbf{y}|\boldsymbol{\theta})\,d\mathbf{y}, \qquad A(\boldsymbol{\theta},\mathbf{y}) = \{(\mathbf{y},\boldsymbol{\theta}) : S(\mathbf{y},\boldsymbol{\theta}) \leq \mathbf{s}\} \tag{1}$$

depends only on the value of the vector $\mathbf{s} \in R^t$ for all $\boldsymbol{\theta} \in \Theta$. In other words, the sampling distribution of S is invariant when the value of its second argument ($\boldsymbol{\theta}$) determines the sampling density of the first ($\mathbf{Y}$).

With these definitions, the following relation between pivotal quantities evaluated at the true parameter value and a value sampled from the posterior distribution holds:

Lemma: *Let $S(\mathbf{Y},\boldsymbol{\theta})$ denote a pivotal quantity, and suppose that $\boldsymbol{\theta}_0$ is a random vector drawn from density π. Given $\boldsymbol{\theta}_0$, let $\mathbf{Y}$ denote a random vector sampled from density $f(\mathbf{y}|\boldsymbol{\theta}_0)$, and let $\boldsymbol{\theta}_\mathbf{Y}$ denote a parameter vector drawn from the posterior distribution on $\boldsymbol{\theta}$ given $\mathbf{Y}$. Then $S(\mathbf{Y},\boldsymbol{\theta}_\mathbf{Y})$ and $S(\mathbf{Y},\boldsymbol{\theta}_0)$ are identically distributed.*

Demonstration of this lemma follows easily from properties of pivotal quantities and the specified joint sampling distribution on $(\boldsymbol{\theta},\mathbf{Y})$:

$$\begin{aligned}
\mathbf{Pr}\left[S(\mathbf{Y},\boldsymbol{\theta}_\mathbf{Y}) \leq \mathbf{s}\right] &= \int_\Theta\int_\mathcal{Y}\int_\Theta \mathcal{I}_{A(\boldsymbol{\theta}_\mathbf{y},\mathbf{y})}\, p(\boldsymbol{\theta}_\mathbf{y}|\mathbf{y}) f(\mathbf{y}|\boldsymbol{\theta}_0)\,\pi(\boldsymbol{\theta}_0)\,d\boldsymbol{\theta}_0\,d\mathbf{y}\,d\boldsymbol{\theta}_\mathbf{y}\\
&= \int_\Theta\int_\mathcal{Y}\int_\Theta \mathcal{I}_{A(\boldsymbol{\theta}_\mathbf{y},\mathbf{y})}\,\frac{f(\mathbf{y}|\boldsymbol{\theta}_\mathbf{y})\pi(\boldsymbol{\theta}_\mathbf{y})}{m(\mathbf{y})} f(\mathbf{y}|\boldsymbol{\theta}_0)\,\pi(\boldsymbol{\theta}_0)\,d\boldsymbol{\theta}_0\,d\mathbf{y}\,d\boldsymbol{\theta}_\mathbf{y}\\
&= \int_\Theta\int_\mathcal{Y}\int_\Theta \mathcal{I}_{A(\boldsymbol{\theta}_\mathbf{y},\mathbf{y})}\,\frac{f(\mathbf{y}|\boldsymbol{\theta}_0)\,\pi(\boldsymbol{\theta}_0)}{m(\mathbf{y})} f(\mathbf{y}|\boldsymbol{\theta}_\mathbf{y})\pi(\boldsymbol{\theta}_\mathbf{y})\,d\boldsymbol{\theta}_\mathbf{y}\,d\mathbf{y}\ d\boldsymbol{\theta}_0\\
&= \int_\Theta\int_\mathcal{Y} \mathcal{I}_{A(\boldsymbol{\theta}_\mathbf{y},\mathbf{y})}\, f(\mathbf{y}|\boldsymbol{\theta}_\mathbf{y})\,d\mathbf{y}\,\pi(\boldsymbol{\theta}_\mathbf{y})\,d\boldsymbol{\theta}_\mathbf{y}\\
&= \int_\Theta G(\mathbf{s})\,\pi(\boldsymbol{\theta}_\mathbf{y})\,d\boldsymbol{\theta}_\mathbf{y}\\
&= \mathbf{Pr}[S(\mathbf{Y},\boldsymbol{\theta}_0) \leq \mathbf{s}]
\end{aligned}$$

This result extends also to arbitrary functions $T(\boldsymbol{\theta},\mathbf{Y})$. To see why, note that samples of $\boldsymbol{\theta}$ drawn from a posterior distribution given $\mathbf{Y}$ have marginal density π whenever $\mathbf{Y}$ is drawn from its marginal density $m(\mathbf{y})$. In general, however, the distribution of $T(\boldsymbol{\theta}_0,\mathbf{Y})$ will not be available as a reference distribution for $T(\boldsymbol{\theta}_y,\mathbf{Y})$ when T is not pivotal, and so extensions to non-pivotal functions are not considered further here.

This property of pivotal quantities makes it possible to define model diagnostics in most Bayesian models. In particular, pivotal quantities can be defined in any model for which observations are assumed to have absolutely continuous density functions and to be conditionally independent given the value of the model parameter—in such cases the values of each observation's distribution function evaluated at a posterior draw of the parameter vector represent independent uniform random deviates. Extension to discrete models can be obtained by additional randomization over the probability mass assigned to individual observations.

2.2 Joint distributions of pivotal quantities

An important limitation of the property delineated in the previous section is that it applies to the marginal distribution of pivotal quantities defined using a single, independent draw of $(\mathbf{Y}, \boldsymbol{\theta}_{\mathbf{Y}})$ from its joint distribution. It does not describe the joint distribution of pivotal quantities based on multiple draws of $(\mathbf{Y}, \boldsymbol{\theta}^i_{\mathbf{Y}})$ from the same posterior distribution, where $\{\boldsymbol{\theta}^i_{\mathbf{Y}}\}$ denote posterior samples of $\boldsymbol{\theta}$ based on the same data vector $\mathbf{Y}$. This difficulty was discussed in Johnson (2004) for the special case of pivotal statistics based on Pearson's chi-squared goodness-of-fit test. However, several strategies for combining information from multiple values of a pivotal quantity defined from the same posterior distribution may be considered. The most direct is numerical evaluation of the joint sampling distribution of the pivotal quantity using prior-predictive-posterior (PPP) simulation (e.g., Dey, Gelfand, Swartz, Vlachos (2003), Hjort, Dahl, Steinbakk (2006)).

Defining $A_i(\mathbf{Y}) \equiv \{(\mathbf{Y}, \boldsymbol{\theta}^i_{\mathbf{Y}}) : S(\mathbf{Y}, \boldsymbol{\theta}^i_{\mathbf{Y}}) \leq \mathbf{s_i}\}$, $i = 1, \ldots, c$, the joint sampling distribution of $\{S(\mathbf{Y}, \boldsymbol{\theta}^i_{\mathbf{Y}})\}$ can be expressed

$$\mathbf{Pr}\left[S(\mathbf{Y}, \boldsymbol{\theta}^1_{\mathbf{Y}}) \leq \mathbf{s_1}, \ldots, S(\mathbf{Y}, \boldsymbol{\theta}^c_{\mathbf{Y}}) \leq \mathbf{s_c}\right] = \tag{2}$$
$$\int_\Theta \cdots \int_\Theta \int_{\mathcal{Y}} \int_\Theta I_{A_1} \cdots I_{A_c} f(\mathbf{y}|\boldsymbol{\theta}_0)\ \pi(\boldsymbol{\theta}_0)\ \lambda(d\boldsymbol{\theta}_0)\ \mu(d\mathbf{y})\ \prod_i p(\boldsymbol{\theta}^i_{\mathbf{y}}|\mathbf{y})\lambda(d\boldsymbol{\theta}^i_{\mathbf{y}}).$$

This expression corresponds to Box's (1980) prior predictive density for the joint distribution on the pivotal quantity S and can be estimated by repeatedly drawing values of $\mathbf{Y}$ from its marginal distribution $m(\mathbf{y})$.

It is interesting to compare the joint distribution specified in (2) to that based on posterior-predictive simulations. Posterior-predictive simulations can be implemented in at least two ways. In one, the discrepancy function is evaluated at pairs of posterior parameter draws and posterior-predictive data values drawn from the sampling distribution given the *same* parameter value. In the case of pivotal quantities, this procedure hardly makes sense (though it is frequently implemented) because the distribution of the discrepancy function for such pairs is known: it is the distribution of the pivotal quantity. Furthermore, draws of the pivotal quantity generated in this way are independent whenever draws from the sampling distribution are independent.

Alternatively, posterior-predictive approximations to the joint distribution of pivotal quantities might be based on the joint distribution of pivotal quantities evaluated at posterior-predictive data vectors $\mathbf{Z}$ and values of $\boldsymbol{\theta}_Y^j$ not used in the generation of $\mathbf{Z}$. Such a procedure leads to a joint posterior distribution that can be expressed

$$\mathbf{Pr}\left[S(\mathbf{Z},\boldsymbol{\theta}_{\mathbf{Y}}^1) \le \mathbf{s_1},\ldots,S(\mathbf{Z},\boldsymbol{\theta}_{\mathbf{Y}}^c) \le \mathbf{s_c}\right] = \tag{3}$$
$$\int_\Theta \cdots \int_\Theta \int_{\mathcal{Z}} \int_\Theta \mathcal{I}_{A_1}\cdots\mathcal{I}_{A_c} f(\mathbf{z}|\boldsymbol{\theta}_{\mathbf{y}}^0)\ p(\boldsymbol{\theta}_{\mathbf{y}}^0|\mathbf{y})\,d\boldsymbol{\theta}_{\mathbf{y}}^0\ d\mathbf{z}\prod_{i=1}^{c} p(\boldsymbol{\theta}_{\mathbf{y}}^i|\mathbf{y})d\boldsymbol{\theta}_{\mathbf{y}}^i.$$

The precise nature of the relation between (3) and (2) is difficult to study analytically, although Meng (1994) describes theoretical properties of psuedo-Bayesian p values based on posterior-predictive schemes. In general, however, it is clear that the marginal distribution of $S(\mathbf{Z},\boldsymbol{\theta}_{\mathbf{Y}})$ is not the same as the nominal distribution of $S(\mathbf{Y},\boldsymbol{\theta}_{\mathbf{Y}})$.

The primary disadvantage of PPP simulation is its high computational burden. In realistic statistical models applied to complex data, implementing a MCMC algorithm for the observed data can be computationally expensive; running the same algorithm 10,000 or more times for data simulated from the prior model will often not be feasible. Such computations are particularly problematic in the model refinement stages of data analysis. To avoid PPP simulations, several strategies can be considered.

One approach toward avoiding PPP simulation is to base model assessments on informal graphical comparisons of the posterior distribution of a pivotal quantity (or several pivotal quantities) to their nominal sampling distribution. In many cases, simple graphical diagnostics based on pivotal quantities provide a clear indication of model inadequacy. If the posterior distribution of the pivotal quantity does suggest model lack-of-fit and a p value is required, then PPP simulation can be implemented to more formally assess the model deviation.

In performing graphical assessments, it is important to consider the complexity of the model being assessed. Simple models with low dimensional parameters generally produce posterior samples of pivotal quantities that are highly correlated. In contrast, models containing large numbers of parameters usually produce posterior samples of pivotal quantities that are less highly correlated. Because the realized pivotal quantities in more complex models are less dependent, the posterior distribution of pivotal quantities from such models should be expected to more closely match their marginal distribution. Thus, even small deviations from the nominal distribution may be associated with extreme PPP p values when the posterior distribution of the pivotal quantity has approximately the same dispersion as the nominal distribution of a single value.

With these considerations in mind, it is often useful to summarize evidence contained in the posterior distribution of a pivotal quantity by the proportion of posterior pivotal quantities that result in the rejection of the null hypothesis of model adequacy at a specified level of significance. For example, one might report that 50% of the pivotal quantities generated from the posterior distribution result in rejection of the null hypotheses at the 5% level of significance. Such summaries do not, however, account for the dispersion of the posterior distribution of the pivotal quantity or the complexity of the fitted model.

Beyond informal summaries of model assessment, it is also possible to obtain probabilistic bounds on PPP p values using bounds on order statistics of dependent samples of random variables. This procedure is discussed in more detail in the next section. Other approaches for approximating PPP p values are discussed in the final section of this article and are currently the topic of active investigation.

2.3 Probabilistic bounds on PPP p values

Gascuel and Caraux (1992) and Rychlik (1992) derived bounds on the distribution of order statistics from identically distributed, dependent random variables and provided constructions in which these bounds were obtained. Letting $X_{(1)}, \ldots, X_{(n)}$ denote order statistics from a dependent sample of random variables each having distribution function G, and letting $P_{m:n}$ denote the distribution function for the m'th order statistic out of n, one such bound can be expressed

$$P_{m:n}(t) \geq \max\left\{0, \frac{nG(t) - m + 1}{n - m + 1}\right\}. \tag{4}$$

To apply this bound to a posterior sample of pivotal quantities, let $S_{(m)}$ denote the m'th order statistic from n values of sampled pivotal quantities, and assume that large values of the pivotal quantity are associated with model misspecification. If G now denotes the sampling distribution of a single value of the pivotal quantity S, then it follows that

$$\mathbf{Pr}(S_{(m)} > t) \leq 1 - \max\left(0, \frac{nG(t) - m + 1}{n - m + 1}\right). \tag{5}$$

For example, if t represents the 0.99 quantile from G, it follows for large n that the probability that the $S_{(.8n)} > t$ is less than or equal to 0.05. In other words, a finding that the .8 quantile from the posterior distribution of pivotal quantities exceeds the .99 quantile from the nominal distribution implies that the PPP p value is less than .05.

In applying this bound, two facts should be considered. First, to avoid producing a bound that is dependent on the particular posterior sample of pivotal quantities generated, the value of m/n should be bounded below 1 and should be selected so that uncertainty in the m/n'th quantile of the posterior distribution of pivotal quantities is small. Second, if the distribution of the pivotal quantity is not exact (e.g., Pearson's χ^2 statistic), extreme values of m/n should be regarded with caution since in that case the marginal distribution of the pivotal quantity in the extreme tails of the distribution will also not be exact.

3 Examples

In this section, the use of pivotal quantities to define Bayesian model diagnostics is explored in two examples. The purpose of the first example is to illustrate differences between model diagnostics defined using prior-predictive methodology and diagnostics obtained using posterior-predictive methodology. The second is intended to provide insight regarding the accuracy of the bounds and approximations to the joint posterior distribution on pivotal quantites described in Sections 2.2 and 2.3, and to illustrate the definition of graphical diagnostics based on pivotal quantities.

3.1 Mortality tables

To facilitate comparison with posterior-predictive model diagnostics, consider the mortality data presented in Broffitt (1988). These data were subsequently analyzed by Carlin (1992) and Gelman, Meng, and Stern (1996). The data consist of observed deaths, by age, for participants in a health insurance policy. Letting y_t denote the number of deaths out of N_t participants of age t, Gelman, Meng, and Stern (1996) assume a Poisson model of the form

$$P(\mathbf{y}|\boldsymbol{\theta}) \propto \prod_{\mathbf{t=35}}^{\mathbf{64}} \theta_{\mathbf{t}}^{\mathbf{y_t}} \exp(-\mathbf{N_t}\theta_{\mathbf{t}}) \tag{6}$$

for the age-specific mortality rates $\boldsymbol{\theta} = \{\theta_t\}$. The prior distribution for $\boldsymbol{\theta}$ is assumed only to be an increasing convex function of $\boldsymbol{\theta}_t$ with respect to age. In order to obtain a proper prior distribution on $\boldsymbol{\theta}$, I impose the additional constraint that the components of $\boldsymbol{\theta}$ be bounded in the interval (0,0.05) (the highest observed mortality rate was less than 0.02; Gelman, Meng and Stern also assume a uniform distribution for $\boldsymbol{\theta}$ but do not specify the corresponding interval). A plot of y_t/N_t appears in Figure 1; the possibility that mortality rates decrease at the higher ages may suggest that $\boldsymbol{\theta}$ may not, in fact, be convex.

The pivotal quantity chosen by Gelman, Meng and Stern for assessing the fit of the Poisson model to the mortality data was the Pearson-type discrepancy measure

$$S(\mathbf{y}, \boldsymbol{\theta}) = \sum_t \frac{(y_t - N_t\theta_t)^2}{N_t\theta_t}.$$

Gelman, Meng and Stern provide a thorough analysis of the distribution of $S(\boldsymbol{\theta}, \mathbf{y})$ under a variety of distributional assumptions for various estimates of $\boldsymbol{\theta}$ and for various assumptions regarding the number of degrees of freedom that should be associated with the distribution on S. They report a posterior-predictive p-value of 10% for the minimum χ^2 statistic, and a posterior-predictive p-value of 6.3% for the realized discrepancy.

The analysis based on the recognition that S is approximately a pivotal quantity is more direct. By evaluating this quantity at a draw from the posterior distribution $\boldsymbol{\theta}_{\mathbf{y}}^i$, it follows that the marginal distribution of $S(\mathbf{y}, \boldsymbol{\theta}_{\mathbf{y}}^i)$ is (for large N_t) approximately that of a χ^2_{30} random variable.

A plot of the posterior distribution of $S(\mathbf{y}, \boldsymbol{\theta}_{\mathbf{y}}^i)$ against its reference χ^2_{30} distribution appears in Figure 2. For these data, the probability that an individual value of $S(\mathbf{y}, \boldsymbol{\theta}_{\mathbf{y}}^i)$ exceeds a randomly drawn value from a χ^2_{30} distribution is 95.2%, and the posterior probability that a value of $S(\mathbf{y}, \boldsymbol{\theta}_{\mathbf{y}}^i)$ exceeds the 0.95 quantile from a χ^2_{30} distribution is

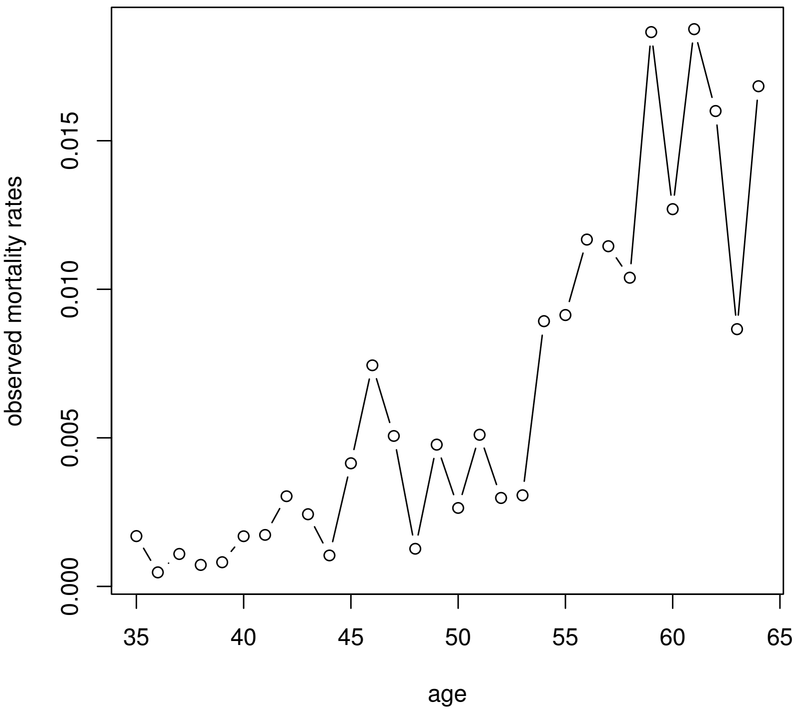

Figure 1: Mortality data. This plot depicts the ratio of the number of deaths to number enrolled in an insurance plan versus age.

approximately 0.61. Both summaries provide substantial evidence that this model does not fit these data. In many applications, plots similar to Figure 2 and the summary statistics cited above will provide an adequate basis for model assessment. However, it is sometimes necessary to perform more formal assessment of model adequacy using a single, summary diagnostic derived from the joint posterior distribution of pivotal values. As stated above, such assessments can be obtained through PPP simulation or by the approach described at the end of Section 2.

To illustrate the more formal PPP scheme for these data, define the prior-predictive summary statistic $V(\mathbf{Z}_j)$ according to

$$V(\mathbf{Z}_j) = \Pr\left[S(\boldsymbol{\theta}_{\mathbf{Z}_j}, \mathbf{Z}_j) > X^2\right], \tag{7}$$

where $\mathbf{Z}_j$ denotes an observation drawn from the prior-predictive distribution, X^2 denotes an independent χ^2_{30} random variable, and probability is computed with respect to the posterior distribution given $\mathbf{Z}_j$. An overall Bayesian p-value for model adequacy can then be defined as the proportion of $V(\mathbf{Z}_j)$ values that exceed $V(\mathbf{Y})$, the value of (7) evaluated at the observed data.

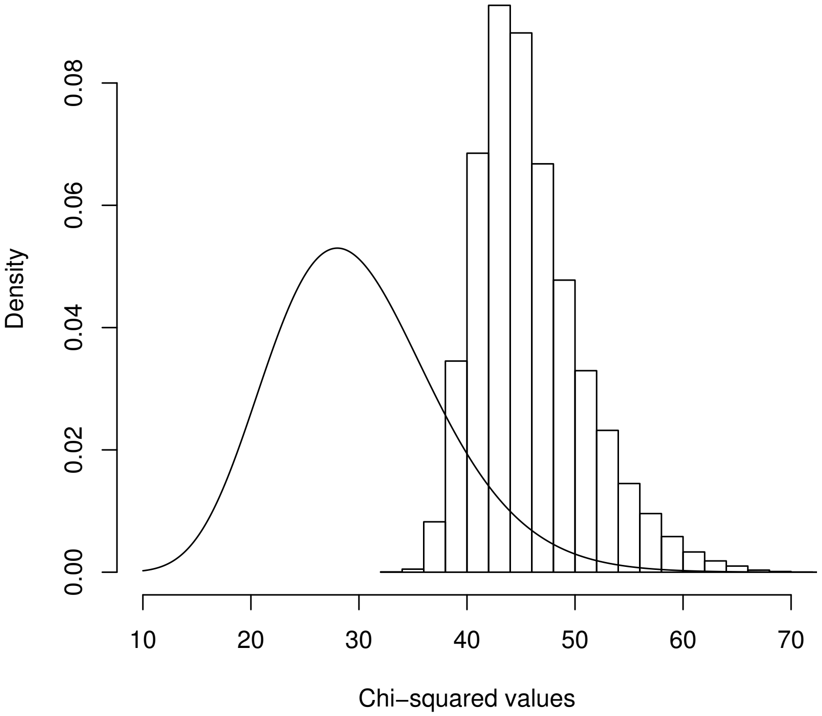

Figure 2: Histogram estimate of the posterior distribution of $S(\boldsymbol{\theta}^i_{\mathbf{y}}, \mathbf{y})$ for the mortality data. For comparison, the marginal χ^2_{30} distribution for a singe value of $S(\boldsymbol{\theta}^i_{\mathbf{y}}, \mathbf{y})$ is displayed as a solid line in the plot.

In this example, the prior-predictive probability that $V(\mathbf{Z}_j)$ exceeds $V(\mathbf{Y})$ was estimated in the following way. First, an MCMC algorithm was used to generate $5,000$ draws from the prior distribution of $\boldsymbol{\theta}$. For each of these values, a prior-predictive data value $\mathbf{Z}_j$ was drawn according to (6). Samples from the posterior distribution of $S(\mathbf{Z}_j, \boldsymbol{\theta}_{\mathbf{Z}_j})$ were then used to estimate the value of $V(\mathbf{Z}_j)$. This resulted in a sample of 5,000 $V(\mathbf{Z}_j)$ values; these values provided a reference distribution for $V(\mathbf{Y})$. In this case, $V(\mathbf{Y}) = 0.952$, and of the 5,000 values of $V(\mathbf{Z}_j)$ generated by this procedure, only 156 exceeded 0.952. It follows that the Bayesian p value for lack-of-fit is approximately 0.031.

Next, bounds based on order statistics were obtained by sampling 10^5 values of the pivotal quantity S from the posterior distribution. For this sample, $S_{(96,000)}$ (the 0.96 quantile of the empirical distribution of the sampled pivotal quantities) was 56.39. A χ^2_{30} distribution function evaluated at 56.39 is .99754; applying (5) thus implies the

Group	Observations	Sample mean
1	2.73, 0.56, 0.87, 0.90, 2.27, 0.82	1.36
2	1.60, 2.17, 1.78, 1.84, 1.83, 0.80	1.67
3	1.62, 0.19, 4.10, 0.65, 1.98 ,0.86	1.57
4	0.96, 1.92, 0.96, 1.83, 0.94, 1.42	1.34
5	6.32, 3.66, 4.51, 3.29, 5.61, 3.27	4.44

Table 1: Hypothetical data presented by O'Hagan (2003).

probability of seeing a value of $S_{(96,000)}$ that exceeds 56.39 is

$$\mathbf{Pr}(S_{(96,000)} > 56.39) < 1 - \max\left(0, \frac{99754 - 96000 + 1}{100000 - 96000 + 1}\right) = .062.$$

Similarly, bounds of $p < 0.07$ were achieved for all quantiles of order statistics between 0.7 and 0.98. Note that obtaining these bounds required essentially no additional computation above that required to obtain the posterior sample for $\boldsymbol{\theta}$, and that this bound compares favorably to the the value obtained through PPP methodology.

3.2 Hierarchical Linear Models

O'Hagan (2003) provides an analysis of a simple hierarchical linear model and compares his model diagnostic approach—based on conflict measures—to diagnostics proposed by Chaloner (1994). The particular model he considers may be written

$$y_{ij}|\lambda_i, \sigma^2 \sim N(\lambda_i, \sigma^2), \qquad i = 1, \ldots, 5, \qquad j = 1, \ldots, 6,$$
$$\lambda_i|\mu, \tau^2 \sim N(\mu, \tau^2), \qquad i = 1, \ldots, 5,$$
$$\mu \sim N(2, 10), \qquad \sigma^2 \sim 22\chi^{-2}_{20}, \qquad \tau^2 \sim 6\chi^{-2}_{20},$$

where all parameters and data $\mathbf{y} = \{y_{i,j}\}$ are assumed to be conditionally independent. Data to which the model were applied appear in Table 1. A cursory examination of this table suggests that the fifth group has a mean that is not compatible with the means of the other groups.

The simple form of the linear hierarchical model assumed for these data makes model assessment based on pivotal quantities straightforward. Defining

$$\epsilon_{ij} = (y_{ij} - \lambda_i)/\sigma \qquad \text{and} \qquad \epsilon_i = (\lambda_i - \mu)/\tau,$$

it follows that the vectors $\mathbf{e}_i = \{\epsilon_{ij}\}$ have independent, standard normal sampling distributions when evaluated at a draw of (λ_i, σ^2) from the posterior. So do the quantities ϵ_i when evaluated at a posterior sample of (λ_i, μ, τ). Quantile-quantile plots for each of these quantities appear in Figure 3.

Several interesting patterns are immediately apparent from Figure 3. For instance, the points in the quantile-quantile plot for Group 5 all fall above the diagonal, clearly indicating that the sampled value of the mean λ_5 for this group was too *small*. In the quantile-quantile plot for ϵ_i, the largest value corresponds to ϵ_5, which suggests that the value of λ_5 was too *large*. Furthermore, the slopes of the dashed lines fall below the 45° line in all groups except 5, suggesting that the observational variance was overestimated in these groups. These patterns manifest themselves with high probability as different samples from the posterior are drawn, and taken together clearly indicate a problem with the fifth group's mean parameter. These facts suggest that the posterior distribution of λ_5 is too small given the data in this group, but too large to have come from the same population as the other λ's. There is also evidence of a suspicious observation in the third group, corresponding to y_{33}, and this pattern also appears with high probability as additional samples are drawn from the posterior.

The residual values displayed in Figure 3 can also be used to generate chi-squared random variables for each value of the parameter vector $(\lambda_i, \mu, \sigma, \tau)$ drawn from the posterior. Figure 4 displays the posterior distributions of the chi-squared statistics obtained by taking the sum-of-squared residuals for each panel in Figure 3 for each of 1,000 draws from the posterior on $(\lambda_i, \mu, \sigma, \tau)$. In each panel, a vertical line indicates the expected value of the chi-squared statistics. From the plot, it is clear that the chi-squared values obtained from Group 5 (i.e., $\sum_j \epsilon_{5,j}^2$) and for the group means (i.e., $\sum_i \epsilon_i^2$) are too large, while the chi-squared variables for Groups 2 and 4 are unexpectedly small.

The chi-squared values depicted in Figure 4 can be used to obtain Bayesian p values through prior-predictive simulations similar to those illustrated for the mortality data. Defining the summary statistic V according to (7) for each subset of residuals, prior-predictive simulations yield p values of $(0.62, 0.99, 0.09, 0.96, 0.07)$ for the five residual groups, and a p value of 0.07 for the group means. Except for the first residual group, these values tend to be either close to 0 and 1, again indicating model lack-of-fit.

4 Discussion

Results reported in this article are derived from the simple fact that the distribution of a pivotal quantity $S(\mathbf{Y}, \boldsymbol{\theta}_{\mathbf{Y}})$ equals the distribution $S(\mathbf{Y}, \boldsymbol{\theta}_0)$ whenever $\boldsymbol{\theta}_{\mathbf{Y}}$ denotes a parameter value drawn randomly from the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Y}$ and $\boldsymbol{\theta}_0$ denotes the parameter value underlying the generation of $\mathbf{Y}$. The simplicity of this result makes it possible to analytically define reference distributions for many Bayesian model diagnostics.

A potential difficulty that arises in the use of pivotal quantities to define formal model diagnostics involves the requirement to perform PPP simulations to calibrate the distribution of the joint sampling distribution on pivotal quantities corresponding to a single observation vector. Bounds based on order statistics provide a useful mechanism for avoiding this problem. Other strategies to avoid PPP simulation are currently under investigation and include approaches based on modeling the marginal distribution of pivotal quantities as mixtures of low-dimensional parametric densities and computing "average" Bayes factors from the pivotal quantities themselves (e.g., Johnson (2005)).

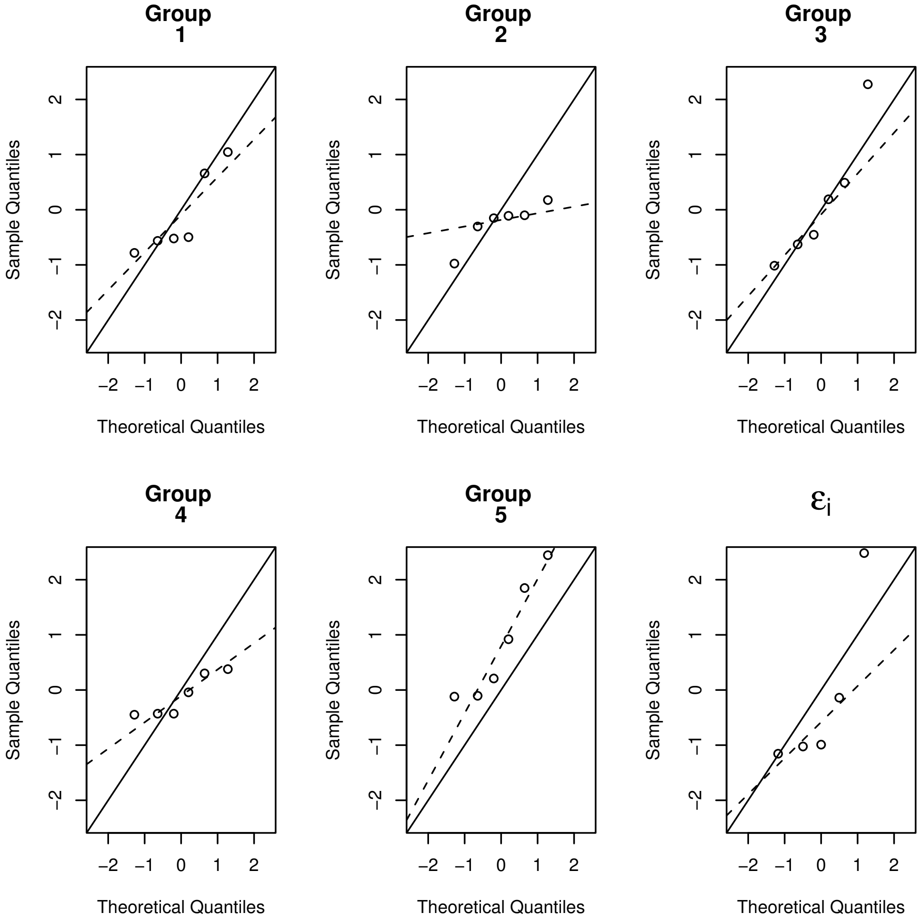

Figure 3: Quantile-quantile plots of residuals $\mathbf{e}_i$ and ϵ_i for a randomly selected parameter value drawn from the posterior distribution. The dashed lines represent the R defaults for the function qqline, which pass through the upper and lower quartiles of the empirical distribution function. The 45° line through the origin is displayed for reference.

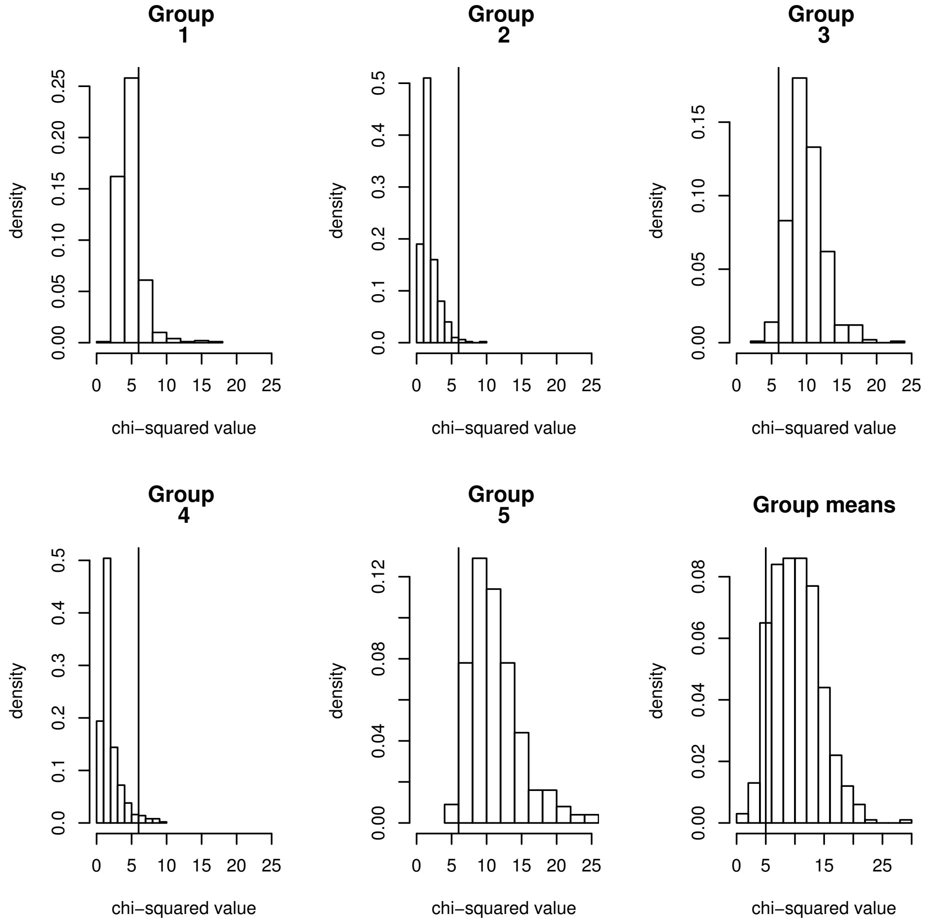

Figure 4: Histogram estimate of the joint posterior distribution residual chi-squared values. The first five panels depict values of $\sum_{j=1}^{6} \epsilon_{ij}^2$, while the last panel displays the posterior distribution of $\sum \epsilon_i^2$. Vertical lines in each plot indicate marginal expectations.

References

Bayarri, M.J. and Berger, J.O. (2000). "*P* Values for composite null models." *Journal of the American Statistical Association*, 95:1127–1142. 719

Box, G. (1980). "Sampling and Bayes' inference in scientific modelling and robustness (with discussion)." *Journal of the Royal Statistical Society, Series A*, 143:383–430. 719

Broffitt, J.D. (1988). "Increasing and increasing convex Bayesian graduation." *Transactions of the Society of Actuaries*, 40:115–148. 725

Carlin, B. P. (1992). "A simple Monte Carlo approach to Bayesian graduation." *Transactions of the Society of Actuaries*, 44:55–76. 725

Chaloner, K. (1994). "Residual analysis and outliers in Bayesian hierarchical models." In Freeman, P.R. and Smith, A.F.M. (eds.), *Aspects of Uncertainty: a Tribute to D.V. Lindley*, 149–157. Chicester: Wiley. 728

Chaloner, K. and Brant, R. (1988). "A Bayesian approach to outlier detection and residual analysis." *Biometrika*, 75:651–9.

Dey, D.K., Gelfand, A.E., Swartz, T.B. and Vlachos, P.K. (2003). "A simulation-intensive approach for checking hierarchical models." *Test*, 7:325–346. 722

Gascuel, O. and Caruax, G. (1992). "Bounds on the expectation of order statistics for dependent variates." *Statistics & Probability Letters*, 15:143–148. 724

Gelfand, A.E. (1996). "Model determination using sampling-based methods." In Gilks, W., Richardson, S. and Spiegelhalter, J. (eds.) *Markov Chain Monte Carlo in Practice*, 145–162, London:Chapman & Hall. 720

Gelman, A., Meng, X.-L., and Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies (with discussion)." *Statistica Sinica*, 6:733–807. 719, 725

Guttman, I. (1967). "The use of the concept of a future observation in goodness-of-fit problems." *Journal of the Royal Statistical Society, Series B*, 29:83–100. 719

Hjort, N., Dahl, F.A., Steinbakk, G.H. (2006). "Post-processing posterior predictive p values." *Journal of the American Statistical Association*, 101:1157–1174. 722

Johnson, V.E. (2004). "A Bayesian χ^2 test for goodness-of-fit." *Annals of Statistics*, 32:2361–2384. 722

Johnson, V.E. (2005). "Bayes factors based on test statistics." *Journal of the Royal Statistical Society, Series B*, 67:689–701. 729

Meng, X.-L. (1994). "Posterior predictive p-values." *Annals of Statistics*, 22:1142–1160. 719, 723

O'Hagan, A. (2003). "HSSS model criticism." In Green, P., Hjort, N., Richardson, S. (eds.) *Highly Structured Stochastic Systems*, 423–444, Oxford: Oxford University Press. 728

Robins, J.M., van der Vaart, A., and Ventura, V. (2000). "Asymptotic distribution of P values in composite null models." *Journal of the American Statistical Association*, 95:1143–1159. 720

Rubin, D. (1984). "Bayesianly justifiable and relevant frequency calculations for the applied statistician." *The Annals of Statistics*, 12:1151–1172. 719

Rychlik, T. (1992). "Stochastically extremal distributions of order statistics for dependent samples." *Statistics and Probability Letters*, 13:337–341. 724

Acknowledgments

I would like to thank the editor, associate editor and an anonymouse referee for several suggestions that improved the presentation of material in this article.