TECHNICAL
REPORT

0350

ENHANCING THE PERFORMANCE OF A POSTERIOR PREDICTIVE CHECK

BERKHOF, J., VAN MECHELEN, I. and A. GELMAN

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Enhancing the performance of a posterior predictive check

Johannes Berkhof and Iven van Mechelen

Katholieke Universiteit Leuven, VU University Medical Center

Andrew Gelman

Columbia University

February, 2004

Please send all correspondence concerning this paper to Johannes Berkhof, Department of Clinical Epidemiology and Biostatistics, VU University Medical Center, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands. The research reported in this paper is supported by the Research Fund of the K.U. Leuven (grant OT/96/10) and the U.S. National Science Foundation. The authors are grateful to Xiao-Li Meng for helpful suggestions, to Doreen Arcus, Jerome Kagan, and Nancy Snidman for the infant temperament data, and to Marleen van Eck for the cortisol data.

Abstract

Posterior predictive checking is a method for detecting violations against a posited model. The basic idea is to compare the observed data to data sets generated from the posterior predictive distribution. Although comparable to classical goodness-of-fit testing, the test quantities or discrepancy measures used in posterior predictive checking are allowed to depend on unknown nuisance parameters. The posterior predictive check is computationally cheap but criticized for lack of power to detect model violations. We discuss the applicability of existing methods for enhancing the performance of a posterior predictive check. We also present new type of test quantities based on taking expectations, which have a similar interpretation as the discrepancy measure but yield a more powerful model check. Computational and distributional aspects of the modified model checks are discussed in theoretical and empirical examples.

Key words: posterior predictive check, discrepancy measure, posterior predictive p-value, frequentist properties.

1 Introduction

In any statistical analysis, it is important to check whether the data are adequately fitted by the model. Because statistical models are in general developed to yield accurate inferences for theoretically interesting model parameters, the fit of the model should be examined for theory-driven test quantities. Rubin (1984) considered comparing the observed data y to hypothetical replicated data sets $y_1^{\text{rep}}, \ldots, y_S^{\text{rep}}$ for test quantities $T(y)$ that can be defined by the user. The replicated data sets are generated from the posterior predictive distribution $p(y^{\text{rep}}|y)$. This approach, called posterior predictive checking, was introducted by Guttman (1967).

One may wish to consider statistics $T(y)$ that are functions of point estimates $\hat{\theta}_y$ of θ. In that case, it seems more natural to use discrepancy measures $D(y, \theta)$ that are functions of θ and therefore directly measure the discrepancy between model and data. For a detailed discussion of discrepancy measures, see Gelman, Meng, and Stern (1996) and Meng (1994). Using discrepancy measures, model violations are detected by comparing $D(y, \theta)$ to $D(y^{\text{rep}}, \theta)$ where (y^{rep}, θ) is drawn from the posterior predictive distribution $p(y^{\text{rep}}, \theta|y)$. The advantage of using discrepancy measures instead of statistics is not only conceptual but also computational, since in complex Bayesian modeling, point estimates $\hat{\theta}_y$ like the maximum likelihood estimate may be hard to obtain. Moreover, for performing predictive checks, $\hat{\theta}_{y^{\text{rep}}}$ would have to be computed for each simulation draw of y^{rep}, which would require nested looping if $\hat{\theta}_y$ is computed iteratively.

The exact distribution of discrepancy measures under the posited model is usually not available and a posterior predictive check then is attractive because it is computationally and analytically undemanding. Although easy to carry out, a posterior predictive check may fail to detect a model violation because the data are used twice: once to obtain the posterior and once to assess the fit of the model. Hence, if the posterior predictive check does not indicate that the model is violated, it remains to be examined whether the outcome of the model check is related to the double use of the data. This can be done by changing the replication distribution (Gelfand, Dey, and Chang, 1992;

Bayarri and Berger; 1999, 2000) or by modifying the test quantity (Robins, Van der Vaart, Ventura, 2000), usually at the cost of additional computational and/or analytical efforts.

In this paper, we discuss the applicability of existing methods for enhancing the performance of the model check. Besides, we present two modifications of $D(y, \theta)$ based on taking expectations but in slightly different ways. The idea is to reduce the dependence between $D(y^{\mathrm{rep}}, \theta)$ and θ by integrating out θ from $D(y, \theta)$. This yields test quantities that have a similar interpretation as the corresponding discrepancy measure $D(y, \theta)$ but have reference distributions with narrower tails. We also present efficient simulation techniques for computing the modified discrepancy measures.

The remainder of this paper is organized as follows. In Section 2, we review posterior predictive model checking and in Section 3, we discuss existing modifications of the model check. In Sections 4 and 5, we present the two modifications that are based on taking expectations and in Section 6, we offer some concluding remarks.

2 Model checking

This section contains a brief explanation of the concept of posterior predictive checking and a discussion of the Bayesian and frequentist properties of the posterior predictive check.

2.1 Plotting discrepancy measures

A posterior predictive check involves drawing $(\theta_s, y_s^{\mathrm{rep}})$ $(s = 1, \ldots, S)$ from the posited model $p(y^{\mathrm{rep}}, \theta|y)$ and comparing $D(y_s^{\mathrm{rep}}, \theta_s)$ to $D(y, \theta_s)$. The easiest way to draw from $p(y^{\mathrm{rep}}, \theta|y)$ is to draw θ from $p(\theta|y)$ and y^{rep} from $p(y^{\mathrm{rep}}|\theta)$. To judge the possible model violation, the draws can be presented in a two-dimensional scatterplot of $D(y_s^{\mathrm{rep}}, \theta_s)$ against $D(y, \theta_s)$ (Gelman *et al.*, 1996). Relatively large values of $D(y, \theta_s)$ compared to $D(y_s^{\mathrm{rep}}, \theta_s)$ contribute to the evidence against the posited model.

Instead of visualizing the fit of the model by a two-dimensional plot, one can also compute a discrepancy comparison $\delta(y^{\text{rep}}, y, \theta) = D(y^{\text{rep}}, \theta)$ - $D(y, \theta)$. The observed data y and the replicated data y^{rep} are equally close to the model if $\delta(y^{\text{rep}}, y, \theta)$ equals zero. Hence, a model violation can be detected from a histogram of $\delta(y^{\text{rep}}, y, \theta)$ compared to the value 0.

2.2 Posterior predictive p-value

The performance of the posterior predictive check under the posited model can be summarized by the p-value. The p-value as a measure of outlyingness has been disputed in the literature (Berger and Sellke, 1986; Kass and Raftery, 1995) although its merits have been recognized as well (Rubin, 1984; Meng, 1994; Gelman *et al.*, 1996; Bayarri and Berger, 2000; Robins *et al.*, 2000; Berger, 2003). The posterior predictive p-value is defined as

$$p_D \;=\; Pr\{D(y^{\text{rep}}, \theta) \geq D(y, \theta) | y\} \;=\; \iint I_{D(y^{\text{rep}}, \theta) \geq D(y, \theta)}\, p(y^{\text{rep}}, \theta | y)\, d\theta\, dy^{\text{rep}}\ . \tag{1}$$

If the posterior distribution of $D(y, \theta)$ does not depend on θ_0, the posterior predictive p-value coincides with the conditional p-value

$$p_{\text{cond}}(\theta) \;=\; Pr\{D(y^{\text{rep}}, \theta) \geq D(y, \theta) | \theta\}\ . \tag{2}$$

2.3 Sampling distribution of the posterior predictive p-value

The performance of the posterior predictive check under the posited model can be examined by making a frequency evaluation of the posterior predictive p-value (Rubin, 1984; Meng, 1994; Bayarri and Berger, 2000; Robins *et al.*, 2000). This in itself is of limited value as the posterior predictive p-value is a Bayesian posterior probability and not a frequentist p-value. Nevertheless, even within the Bayesian framework, it is desirable to have a p-value that has a uniform distribution over hypothetical observed data sets drawn from the true model (Rubin, 1984; Bayarri and Berger, 2000).

For the sampling distribution of hypothetical observed data sets, more than one choice is possible. Meng (1994) allowed the value of θ to vary across studies and assumed that the hypothetical data sets are drawn from the prior predictive distribution $p(y) = \int p(y|\theta)\,p(\theta)\,d\theta$. Let U denote a standard uniform variable. Under the assumption of a continuous sampling distribution for discrepancy $D(y,\theta)$, Meng (1994) showed that the prior predictive mean $E[p_D] = 1/2$ and that $E[h(p_D)] \leq E[h(U)]$ for all convex functions $h(.) : [0,1] \rightarrow \mathbb{R}$. The intuitive interpretation of the result is that the posterior predictive check tends to be conservative (i.e. the probability that the posterior predictive p-value is smaller than or equal to a nominal signficance level α is smaller than or equal to α). Robins *et al.* (2000) studied the asymptotic frequentist properties of the posterior predictive p-value under the assumption that the value of θ does not vary across studies but that there exists a true θ_0 and demonstrated that the posterior predictive check remains conservative when the sample size tends to infinity. The conservatism property makes the posterior predictive check justifiable from a frequentist point of view, but also indicates that the posterior predictive check may lack sensitivity to detect model violations.

2.4 Double use of the data

The conservatism of the posterior predictive check is strongly related to the double use of the data. We illustrate this by reformulating p_D as a posterior mean of conditional p-values (Meng, 1994):

$$p_D \;=\; \int p_{\mathrm{cond}}(\theta)\,p(\theta|y)\,d\theta \;=\; E_{\theta|y}[p_{\mathrm{cond}}(\theta)]\;. \tag{3}$$

An interesting aspect of formulation (3) is that it connects posterior predictive checking to classical hypothesis testing. In classical testing, the value of θ is known under the null $H_0 : \theta = \theta_0$ and the p-value then is $p_{\mathrm{cond}}(\theta_0)$. Classical testing is not applicable when the null distribution of $D(y,\theta)$ contains nuisance parameters that are not fixed. In posterior predictive checking, the nuisance parameters are dealt with by estimating the true unknown $p_{\mathrm{cond}}(\theta_0)$ by its posterior mean p_D as shown in (3). Therefore, the data are used both for computing conditional p-values and for averaging these p-values. It is the averaging part that leads to a conservative model check.

3 Remedies against conservatism

3.1 Partial posterior predictive check

For model checks that are based on statistic $T(y)$, Bayarri and Berger (2000) proposed the partial posterior predictive p-value, obtained by averaging the conditional p-values $p_{\text{cond}}(\theta)$ over draws θ from the partial posterior (instead of the posterior). The partial posterior distribution is obtained by conditioning on the part of y that is not contained in $T(y)$ and can be written as

$$p(\theta|y \setminus T(y)) \;\propto\; p(y|\theta)\, p(\theta) \,/\, p(T(y)|\theta) \;.$$

Because the information in y contained in $T(y)$ is not used in the partial posterior predictive distribution, double use of the data is avoided. Robins *et al.* (2000) showed that under asymptotic normality of $T(y)$, the partial posterior p-value, asymptotically, is a frequentist p-value. The idea of the partial posterior is similar to the one underlying case-deleted posteriors (Pettit and Smith, 1985; Geisser, 1993), that is, the data information used to produce the posterior should not be used when computing the test quantity. Bayarri and Berger (2000) also presented a conditional predictive p-value but it is more difficult to compute than the partial posterior predictive p-value and therefore is less useful for general consumption.

Draws from the partial posterior distribution can be obtained by weighing the posterior draws $\theta_1, \ldots, \theta_S$ proportional to $1/p(T(y)|\theta_1), \ldots, 1/p(T(y)|\theta_S)$. The partial posterior predictive p-value can be computed as

$$\hat{p}_{\text{ppost}} = \frac{\sum_{s=1}^{S} Pr\{T(y^{\text{rep}}) \;\geq\; T(y)|\theta_s\} \;/\; p(T(y)|\theta_s)}{\sum_{s=1}^{S} 1 \;/\; p(T(y)|\theta_s)} \;.$$

The estimator $\hat{p}_{\text{ppost}}$ performs fine because $p(T(y)|\theta)$ usually is less concentrated than $p(y|\theta)$ (DiCiccio et al., 1997). A problem may arise when the model is violated by the data because then $p(T(y)|\theta)$ is concentrated at a different θ than $p(y|T(y), \theta)$ and $p(y|\theta)$ becomes diffuse. Another problem is that $p(T(y)|\theta)$ often is not available in closed form in which case it has to be approximated by density-fitting techniques.

The partial posterior predictive check was developed only for model checking problems where the test quantity is a statistic $T(y)$. If the test quantity is a parameter-dependent discrepancy $D(y,\theta)$, double use of the data is not avoided because the test quantity directly depends on the posterior. Attempts to avoid double use of the data to some extent are not recommendable and may produce an extremely liberal model check as is illustrated in the following example.

Example 1. Consider $y = (y_1, \ldots, y_n)^t$ an $n \times 1$ vector of responses for which we assume $y \sim N(v\theta, I_n)$ where v is a known $n \times 1$ vector and $p(\theta) \propto 1$. We define test quantity $D_a(y,\theta) = (w'y - w'v\theta)/n$ with w a known $n \times 1$ vector. We further assume that $w'w = v'v = n$. A check based on $D_a(y,\theta)$ is equivalent to a check based on $T(y) = w'y/n$ because $D_a(y^{\mathrm{rep}},\theta) \geq D_a(y,\theta)$ if and only if $T(y^{\mathrm{rep}}) \geq T(y)$. Bayarri and Berger (2000) and Robins et *al.* (2000) showed that the partial posterior predictive p-value of this model check is a frequentist p-value. Here we outline this result using the discrepancy measure notation. The partial posterior distribution is

$$p(\theta|y \backslash T(y)) \;\propto\; p(y|\theta)p(\theta)/p(y|w'y) \;\propto\; N(\theta|\tilde{\theta}, \tilde{\sigma}^2) \; ,$$

where $\tilde{\theta} = v'Hy/v'Hv$ and $\tilde{\sigma}^2 = 1/v'Hv$ with $H = I_n - ww'/n$. If we condition on y except $T(y)$, we can evaluate $\sqrt{n}(D_a(y^{\mathrm{rep}},\theta) - D_a(y,\theta))$ as follows

$$\begin{aligned}
\sqrt{n}(D_a(y^{\mathrm{rep}},\theta) \;-\; D_a(y,\theta)) &= \frac{1}{\sqrt{n}}(w'y^{\mathrm{rep}} - w'v\theta) + \frac{1}{\sqrt{n}}(w'v\theta - w'v\tilde{\theta}) - \frac{1}{\sqrt{n}}(w'y - w'v\tilde{\theta}) \\
&= z_1 \;+\; z_2\,\frac{w'v(v'Hv)^{-1}v'w}{n} \;-\; \frac{1}{\sqrt{n}}(w'y - w'v\tilde{\theta}) \\
&= z\,\sqrt{w'v(v'Hv)^{-1}v'w/n \;+\; 1} \;-\; \frac{1}{\sqrt{n}}(w'y - w'v\tilde{\theta}) \; , \qquad (4)
\end{aligned}$$

were z_1, z_2, and z are standard normal variables. The partial posterior predictive p-value therefore is

$$p_{\mathrm{ppost}} = Pr\{z\,\sqrt{w'(v'Hv)^{-1}v'w/n \;+\; 1} \geq \frac{1}{\sqrt{n}}(w'y - w'v\tilde{\theta})\} \; . \qquad (5)$$

Because for any θ, it holds that $(w'y - w'v\tilde{\theta})\,/\,\sqrt{n} \mid \theta \;\sim\; N(0,\; w'v(v'Hv)^{-1}v'w/n \,+\, 1)$, the partial posterior predictive p-value is a frequentist p-value.

Next consider test quantity $D_b(y, \theta) = (w'y - w'v\theta)^2/n^{3/2}$. After conditioning on θ, $D_b(y, \theta)$ depends on y only via $w'y$. Therefore, it makes some intuitive sense to draw from the partial posterior distribution $p(\theta|y \setminus T(y))$. If we condition on y except $T(y)$, we have

$$\begin{aligned}\sqrt{n}(D_b(y^{\text{rep}}, \theta) - D_b(y, \theta)) &= \frac{1}{n}(w'y^{\text{rep}} - w'v\theta)^2 - \frac{1}{n}((w'y - w'v\tilde{\theta}) - (w'v\theta - w'v\tilde{\theta}))^2 \\ &= \chi_1^2 \;+\; \chi_1^2(\lambda(y))\,\frac{w'v(v'Hv)^{-1}v'w}{n}\;, \qquad (6)\end{aligned}$$

where $\chi_1^2(\lambda(y))$ is a chi-square variable with one degree of freedom and non-centrality parameter $\lambda(y) \;=\; (w'y - w'v\tilde{\theta})^2\,/\,w'v(v'Hv)^{-1}v'w$. The p-value is

$$\begin{aligned}p &= Pr\{D_b(y^{\text{rep}}, \theta) \geq D_b(y, \theta) \mid y,\, \setminus T(y)\} \\ &= Pr\{\chi_1^2 \;-\; \chi_1^2(\lambda(y))\,\frac{w'v(v'Hv)^{-1}v'w}{n} \geq 0\} \\ &= Pr\{\chi_1^2 \;-\; \chi^2(\lambda(y))\frac{\rho^2}{1-\rho^2} \;\geq\; 0\}\;,\end{aligned}$$

where $\rho = w'v/n$. If ρ approaches 1, the sampling distribution of $\lambda(y)|\theta$ converges to a χ_1^2 distribution and the p-value tends to zero. In that case, the model will always be rejected regardless the value of the observed discrepancy. The predictive check can perhaps be repaired at the expense of further analytical effort, but if the discrepancy measure is a function of θ, it is more natural to remain within the posterior predictive framework.

3.2 Modification of the discrepancy measure

The performance of the posterior predictive check can also be enhanced by modifying the test quantity such that the dependence between the test quantity and the nuisance parameters θ is reduced. This is a common idea in statistics and examples include the studentized residual statistic and the generalized test quantity (Tsui and Weerahandi, 1989). Robins *et al.* (2000) presented some generally applicable modifications that produce asymptotic frequentist tests. These include the centered statistic $T_{\mathrm{c}}(y) = T(y) - \mathrm{E}[T(y)|\theta]|_{\theta=\hat{\theta}}$, where $\hat{\theta}$ is the maximum likelihood estimator of θ, and the discrepancy measure

$$\tilde{D}(y,\theta) = D(y,\theta) - h(\theta)i(\theta)^{-1}n^{-1}S(\theta)$$

where $h(\theta) = \partial \lim_{n\to\infty} E[D(y,\theta)|\theta_0]/\partial\theta_0 \mid_{\theta_0=\theta}$, $i(\theta) = \lim_{n\to\infty} E[-\partial^2 p(y|\theta)/\partial\theta\partial\theta'|\theta]$, and $S(\theta)$ the score function (i.e. $\partial \log p(y|\theta)/\partial\theta$). Regarding the power, Robins *et al.* (2000) showed that the posterior predictive check based on $\tilde{D}(y,\theta)$ is a locally most powerful asymptotic test against Pitman alternatives if $D(y,\theta)$ is a score discrepancy. To derive the asymptotic frequentist properties, a number of conditions need to be fulfilled. An important condition that can be violated is that the sampling distribution of the test quantity is asymptotically normal with variance $O(n^{-1})$. We illustrate the use of $\tilde{D}(y,\theta)$ and the importance of the asymptotic normality condition for the example described in the previous subsection.

Example 1 (continued). Consider first the discrepancy $D_a(y,\theta) = \frac{1}{n}(w'y - w'v\theta)$. The mean $E[D_a(y,\theta)|\theta_0] = \rho(\theta_0 - \theta)$ and hence $h(\theta) = \rho$. Furthermore, $i(\theta) = 1/n$ and $S(\theta) = v'y - n\theta$ so $\tilde{D}_a(y,\theta) = (w'y - w'v\theta)/n - \rho(v'y - n\theta) = w'y - \rho v'y + (n-1)\rho\theta$, which is equivalent to the pivotal statistic $w'y - \rho v'y$. Hence the model check based on the modified discrepancy is a frequentist test. Next consider $D_b(y,\theta) = (w'y - w'v\theta)^2/n^{3/2}$. The asymptotic normality condition is not fulfilled ($\sqrt{n}D_b(y,\theta)$ is χ^2_1) and the modified discrepancy measure $\tilde{D}_b(y,\theta)$ coincides with $D_b(y,\theta)$ since $E[D_b(y,\theta)|\theta_0] = n + \theta_0 v'ww'v\theta_0 - 2\theta_0 v'ww'v\theta + \theta v'ww'v\theta$, and hence $h(\theta) = 0$. The posterior predictive check based on test quantity $D_b(y,\theta)$ is conservative (the predictive p-value tends to .5 if $\rho \to 1$) and therefore the posterior predictive check based on the modified discrepancy is also conservative.

In the following two sections, we present two modifications of $D(y,\theta)$ that are obtained by posterior averaging but in different ways. Posterior averaging does not only improve the asymptotic properties but can also improve the small-sample properties of an asymptotic frequentist test. Besides, it is conceptually appealing because the whole posterior of the test quantity is explored.

4 Mean discrepancy comparison

In Section 2.1, we formulated the test quantity as a discrepancy comparison $\delta(y^{\text{rep}}, y, \theta) = D(y^{\text{rep}}, \theta) - D(y, \theta)$. This measure compares the discrepancy between model and observed data to the discrepancy between model and replicated data and therefore is particularly attractive from a conceptual viewpoint since it formalizes in a most straightforward way what we care about in posterior predictive model checking, that is, a comparison of the closeness of y to the model with that of y^{rep}. The modification of the discrepancy comparison that we will present in this section stems from the symmetric nature of y and y^{rep} ($y|\theta$ and $y^{\text{rep}}|\theta$ have the same likelihood, by definition). Much classical and Bayesian literature on the topic is limited because it does not allow the simultaneous existence of y and y^{rep}, even thought this is the key comparison being made in model checking. From this perspective, conditioning on both y and y^{rep} is quite reasonable which is pursued in the following.

4.1 Definition

We present a Rao-Blackwellized version of $\delta(y^{\text{rep}}, y, \theta)$ by taking the mean conditional on y and y^{rep}:

$$\Delta(y^{\text{rep}}, y) \;=\; E\left\{\delta(y^{\text{rep}}, y, \theta)|y^{\text{rep}}, y\right\} \;=\; \int \delta(y^{\text{rep}}, y, \theta)\, p(\theta|y^{\text{rep}}, y)\, d\theta \;. \tag{7}$$

This yields a fair comparison of the discrepancies of y and y^{rep} as the conditional mean is antisymmetric with respect to y and y^{rep}. The mean discrepancy comparison $\Delta(y^{\text{rep}}, y)$ and $\delta(y^{\text{rep}}, y, \theta)$ have the same posterior mean and the posterior variance of $\delta(y^{\text{rep}}, y, \theta)$ is never exceeded by the one of $\Delta(y^{\text{rep}}, y)$. The latter immediately follows from the following decomposition:

$$\text{Var}[\,\delta(y^{\text{rep}}, y, \theta)|y\,] \;=\; \text{E}[\,\text{Var}[\,\delta(y^{\text{rep}}, y, \theta)|y^{\text{rep}}, y]\,|\,y] + \text{Var}[\,\text{E}[\,\Delta(y^{\text{rep}}, y)|y]\,|\,y]\;. \tag{8}$$

The first term at the right hand side of (8) measures the variability among the discrepancy comparisons that remains when conditioning on y^{rep}. This variability is removed when replacing $\delta(y^{\text{rep}}, y, \theta)$ by $\Delta(y^{\text{rep}}, y)$.

4.2 Frequentist properties

The posterior predictive check based on an anti-symmetric discrepancy comparison $\delta(y^{\text{rep}}, y, \theta)$ (which includes $\Delta(y^{\text{rep}}, y)$) has conservative operating characteristics. In the appendix, it is shown that under the assumption of a continuous sampling distribution, it holds that the prior predictive mean $\text{E}(p_\delta)$ equals $1/2$ and $\text{E}[p_\delta^q] \leq E[U^q]$ for $q \geq 1$ where U denotes a standard uniform variable. Hence $\text{Var}(p_\delta) \leq \text{Var}(U)$ which suggests that posterior predictive checks based on discrepancy comparisons are conservative. The result is less general than Theorem 1 of Meng (1994) where the inequality holds for all convex functions on $[0, 1]$.

The frequentist properties of the posterior predictive check are likely to improve when conditioning on y^{rep} and y because $\text{Var}[\Delta(y^{\text{rep}}, y)|y] \leq \text{Var}[\delta(y^{\text{rep}}, y, \theta)|y]$. A stronger result is obtained when it is assumed that $\delta(y^{\text{rep}}, y, \theta)$ and $\Delta(y^{\text{rep}}, y)$ are normal. Then it holds that $|p_\Delta - .5| \geq |p_\delta - .5|$ which implies that p_Δ has superior frequentist properties than p_δ.

4.3 Computation

The value of $\Delta(y^{\text{rep}}, y)$ is usually not obtainable in analytic form. We may approximate the value of $\Delta(y^{\text{rep}}, y)$ by averaging over draws from $p(\theta|y^{\text{rep}}, y)$ but this involves additional MCMC sampling for each y^{rep}. The computational burden can be reduced by importance sampling (Hammersley and Handscomb, 1964; Geweke, 1989), which yields a simulation consistent estimator for $\Delta(y^{\text{rep}}, y)$. Let $h(\theta)$ be an importance sampling density that approximates $p(\theta|y^{\text{rep}}, y)$. Then $\Delta(y^{\text{rep}}, y)$ can be estimated by

$$\hat{\Delta}(y^{\text{rep}}, y) = \frac{\sum_{s=1}^{S} w_s \delta(y^{\text{rep}}, y, \theta_q)}{\sum_{s=1}^{S} w_s} , \tag{9}$$

where θ_s is a draw from $h(\theta)$ and

$$w_s \propto \frac{p(\theta_s|y^{\text{rep}}, y)}{h(\theta_s)} \ .$$

An obvious choice for $h(\theta)$ is the posterior $p(\theta|y)$. This choice often gives satisfactory results because $p(\theta|y^{\text{rep}}, y)$ tends to be more concentrated than $p(\theta|y)$ (diCiccio *et al.*, 1997).

4.4 Theoretical and empirical example

The performance and computation of the mean discrepancy comparison are considered in the following two examples.

Example 2. Suppose the linear regression model $y \sim N(X\theta, I_n)$ where X is an $N \times K$ design matrix. Assume $p(\theta) \propto 1$ so that $p(\theta|y) \sim N(\hat{\theta}, (X'X)^{-1})$ where $\hat{\theta} = (X'X)^{-1}X'y$. The fit of the model can be examined with a χ^2 goodness-of-fit measure $D(y, \theta) = (y - X\theta)'(y - X\theta)$. The properties of χ^2 goodness-of-fit measures in posterior predictive checking have been studied by Gelman *et al.* (1996). Here we examine the effect of replacing the discrepancy comparison by its mean. The discrepancy comparison $\delta(y^{\text{rep}}, y, \theta)$ can be written as

$$\begin{aligned} \delta(y^{\text{rep}}, y, \theta) &= (y^{\text{rep}} - X\theta)'(y^{\text{rep}} - X\theta) - (y - X\hat{\theta} - (X\theta - X\hat{\theta}))'(y - X\hat{\theta} - (X\theta - X\hat{\theta}) \\ &= \chi^2_n - \chi^2_k - D(X, \hat{\theta}) \ , \end{aligned}$$

and the mean discrepancy comparison $\Delta(y^{\text{rep}}, y)$ can be written as

$$\Delta(y^{\text{rep}}, y) \ = \ (y^{\text{rep}} - X\hat{\theta}^{\text{rep}})'(y^{\text{rep}} - X\hat{\theta}^{\text{rep}}) - (y - X\hat{\theta})'(y - X\hat{\theta}) \ = \ \chi^2_{n-k} - D(X, \hat{\theta}) \ . \qquad (10)$$

The posterior mean of $\delta(y^{\text{rep}}, y, \theta)$ (and $\Delta(y^{\text{rep}}, y)$) is $n - k - D(X, \hat{\theta})$ and the posterior variances of $\delta(y^{\text{rep}}, y, \theta)$ and $\Delta(y^{\text{rep}}, y)$ are $n + k$ and $n - k$, respectively. The ratio of variances is $(n+k)/(n-k)$ which indicates that the effect of modifying $\delta(y^{\text{rep}}, y, \theta)$ is substantial only when the fraction k/n is large. The posterior predictive check based on $\Delta(y^{\text{rep}}, y)$ is a frequentist test because the sampling distribution of the test statistic $D(y, \hat{\theta})$ does not depend on θ and is χ^2_{n-k}. Hence, by taking the

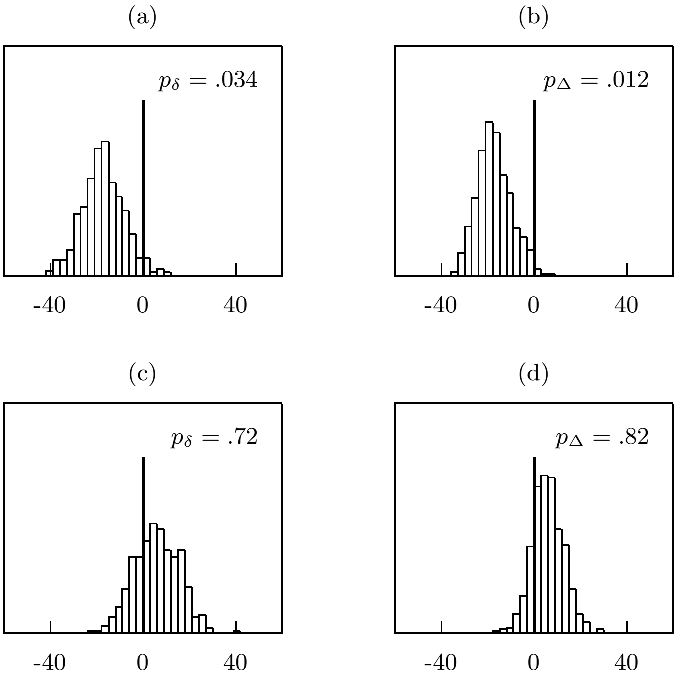

Figure 1: Infant temperament data: posterior predictive distributions of $\delta(y^{\text{rep}}, y, \theta)$ (a and c) and $\Delta(y^{\text{rep}}, y)$ (b and d) for the one class model (a and b) and the two class model (c and d), $\delta(y^{\text{rep}}, y, \theta)$ being the difference between likelihood ratio discrepancy measures.

mean of the discrepancy comparison, the conservatism of the goodness-of-fit measure is eliminated.

Example 3. Stern, Arcus, Kagan, Rubin, and Snidman (1994) and Rubin and Stern (1994) fitted latent class models to the scores from 93 infants on motor activity (M) at four months (scores 1, 2, 3, 4), crying intensity (C) at four months (scores 1, 2, 3), and fear intensity (F) at twelve months (scores 1, 2, 3). This model assumes that the infants can be divided in a number of classes; within each class, the probability vectors on the scores of M, C, and F are assumed to be independent and multinomial. Vaguely informative Dirichlet priors for the one and two class model were presented by Rubin and Stern (1994).

Gelman *et al.* (1996) checked the fit of the model using the likelihood ratio discrepancy measure (posited model versus saturated model): $D(y, \theta) = 2 \sum y_i \log[y_i/E(y_i|\theta)]$, where y_i is the frequency of a cell in the $4 \times 3 \times 3$ contingency table for combinations of scores on (M,C,F). Parameter-dependent discrepancy measures are useful in latent class modeling (Berkhof, van Mechelen, and Gelman, 2003) as they can address features of the different mixture components. We checked the fit of the one and two class model by examining the posterior predictive distributions of $\delta(y^{\text{rep}}, y, \theta)$ and $\Delta(y^{\text{rep}}, y, \theta)$. A posterior sample for the two class model was obtained via data augmentation. The MCMC settings were 5 chains of 5000 draws with independent starting values and a burnin period of 5000 draws per chain; convergence criterion $\sqrt{\hat{R}}$ (Gelman and Rubin, 1992) was smaller than 1.1 for all parameters. The value of $\Delta(y^{\text{rep}}, y, \theta)$ was approximated using 25000 posterior draws. Simulation from $p(\theta|y^{\text{rep}}, y)$ was avoided by using the the posterior distribution as an importance sampling function.

Histograms of $\delta(y^{\text{rep}}, y, \theta)$ and $\Delta(y^{\text{rep}}, y)$ under the one and two class models are presented in Figure 1. The distribution of $\Delta(y^{\text{rep}}, y, \theta)$ has narrower tails both for the one and two class model. The effect on the p-value is small for the one-class model but even small changes can give rise to a different interpretation as a p-value of .012 is more conclusive evidence against the posited model than a p-value of .034 (see panels a and b). Regarding the computation of $\Delta(y^{\text{rep}}, y)$, the average simulation standard error of the approximation of $\Delta(y^{\text{rep}}, y)$ in the two class model was 1.2. We compared the estimate of $\Delta(y^{\text{rep}}, y)$ in the two class model with the one based on 5000 draws from the target density $p(\theta|y^{\text{rep}}, y)$ (instead of 25000 draws from the posterior). The average standard error then was equal to .10 and the estimated p-value p_Δ became .85. We see that the sample sizes are large enough for the importance sampling estimate to produce an accurate p-value.

5 Mean discrepancy measure

5.1 Definition and example

If the expectations of $D(y,\theta)$ and $D(y^{\mathrm{rep}},\theta)$ are taken over $p(\theta|y)$ and $p(\theta|y^{\mathrm{rep}})$ (instead of one distribution $p(\theta|y^{\mathrm{rep}},y)$), the test quantity of the posterior predictive check is

$$D^*(y) \;=\; E[D(y,\theta)|y] \;=\; \int D(y,\theta)p(\theta|y)d\theta\ . \tag{11}$$

This test quantity is equivalent to the discrepancy comparison

$$\Delta^*(y^{\mathrm{rep}},y) \;=\; D^*\left(y^{\mathrm{rep}}\right) - D^*\left(y\right)\ .$$

The posterior mean of $\Delta^*(y^{\mathrm{rep}},y)$ can be formulated as

$$\mathrm{E}[\Delta^*(y^{\mathrm{rep}},y)|y] \;=\; \mathrm{E}[\delta(y^{\mathrm{rep}},y,\theta)|y] \;+\; \mathrm{E}[D^*(y^{\mathrm{rep}})|y] - \mathrm{E}[D(y^{\mathrm{rep}},\theta)|y] \tag{12}$$

The difference between the latter two terms tends to zero for large sample size as can be shown by a first-order expansion. The underlying argument is that because the distributions $p(\theta|y^{\mathrm{rep}})$ and $p(\theta|y)$ concentrate around $\mathrm{E}[\theta|y]$, both $D^*(y^{\mathrm{rep}})$ and $D(y^{\mathrm{rep}},\theta)$ tend to the limit of $D(y^{\mathrm{rep}},\mathrm{E}[\theta|y])$ (it is assumed that $D(y^{\mathrm{rep}},\theta)$ converges to a constant and is a continuous function with bounded derivative in the neighborhood of $E[\theta|y]$). The asymptotic equality of the posterior means of $\Delta^*(y^{\mathrm{rep}},y)$ and $\delta(y^{\mathrm{rep}},y,\theta)$ suggests that what we showed for the mean discrepancy comparison $\Delta(y^{\mathrm{rep}},y)$ also holds for the mean discrepancy $D^*(y)$, that is, that the improvement of the model check is not attributable to a change in the posterior mean of the discrepancy measure but to a change in its posterior variance.

Using $D^*(y)$ instead of $D(y,\theta)$ does not always increase the sensitivity of the model check. The posterior variance of $\Delta^*(y^{\mathrm{rep}},y)$ can be larger than the posterior variance of $\delta(y^{\mathrm{rep}},y,\theta)$ depending on the choice of $D(y,\theta)$. We illustrate this for an outlier example.

Example 4. Consider $y=(y_1,\dots,y_n)^t$ an $n\times 1$ vector of responses for which we assume $y\sim N(\theta,I_n)$,

$p(\theta) \propto 1$, and hence $p(\theta|y) \sim N(\overline{y}, 1/n)$ where $\overline{y}$ is the sample mean. Let $D(y,\theta) = y_{\max} + a\theta$ where a is a known constant. The posterior variance of $\Delta^*(y^{\text{rep}}, y)$ is

$$\begin{aligned} \text{Var}[\Delta^*(y^{\text{rep}}, y)|y] \;=\; \text{Var}[y^{\text{rep}} + a\overline{y}^{\text{rep}}|y] \;&=\; \text{Var}[y^{\text{rep}}_{\max} - \overline{y}^{\text{rep}}|y] + (a+1)^2 \text{Var}[\overline{y}^{\text{rep}}|y] \;, \\ &=\; \text{Var}[\delta(y^{\text{rep}}, y, \theta)|y] + a(a+2)\, \text{Var}[\overline{y}^{\text{rep}}|y] \;. \end{aligned}$$

The posterior variance $\text{Var}(\Delta^*(y^{\text{rep}}, y)|y]$ is smaller than $\text{Var}(\delta(y^{\text{rep}}, y, \theta)|y]$ if and only if $|a+1| < 1$. The minimum of $\text{Var}[\Delta^*(y^{\text{rep}}, y)|y]$ is attained when a is set equal to -1. In that case, $D^*(y)$ is a centered test quantity obtained by subtracting $\overline{y}$ from $y_{\max}$.

5.2 Centering

As illustrated in Example 4, the choice of $D(y,\theta)$ is important in order to have a sensitive model check. In practical model checking, it often is advisable to transform the individual observations such that $D(y,\theta)|\theta$ is pivotal or nearly pivotal. If near-pivotality cannot be achieved by a simple transformation of the observations (like for instance standardization), then the dependence between $D(y,\theta)$ and θ can be reduced by centering the discrepancy measure around its sampling mean $E[D(y,\theta)|\theta]$, yielding

$$D_{\text{c}}\,(y,\theta) \;=\; D(y,\theta) - \text{E}[D(y,\theta)\,|\,\theta] \;.$$

The centering step was proposed by Robins *et al.* (2000) for asymptotically normal test statistics $T(y)$ with variance $O(n^{-1})$ and yields an asymptotic exact test. The asymptotic exactness follows on the asymptotic sampling mean of the centered statistic of zero. The same condition is fulfilled for the statistic $D^*_{\text{c}}(y) \;=\; E[D_{\text{c}}\,(y,\theta)\,|\,y]$ as can be shown by means of first-order expansions of $D(y,\theta)$ and $\text{E}[D(y,\theta)\,|\,\theta]$ around $D(y,\theta_0)$. However, there is no guarantee that the centered quantity performs well when the sample size is small or moderate. Centering is just one way to enhance the sensitivity of the model check and it usually does not change a bad test quantity into a good one. Asymptotic properties can be considered when developing test quantities for general consumption, but the formulation of test quantities should be guided by the posterior at hand.

5.3 Computation

The computation of $\Delta^*(y)$ is problematic when direct sampling from the posterior is not possible. Then computation involves additional MCMC sampling for each y^{rep}. An intermediate approach is to split θ into $\theta = (\theta_1', \theta_2')'$ and to average the discrepancies using draws from $p(\theta_1|y, \theta_2)$ rather than draws from $p(\theta|y)$. In many Bayesian models, conditional sampling is straightforward and has motivated the development of Gibbs sampling schemes. The resulting test quantity is a discrepancy measure (namely a function of θ_2) and is given by

$$D^*(y, \theta_2) = \int D(y, \theta)\, p(\theta_1|y, \theta_2)\, d\theta_1 \ .$$

The discrepancy $D^*(y, \theta_2)$ may be regarded as an intermediate between $D(y, \theta)$ and $D^*(y)$ as both $D(y, \theta)$ and $D^*(y, \theta_2)$ have posterior mean $D^*(y)$.

Conditioning on θ_2 is effective only if the distribution of $D(y, \theta)|y, \theta_2$ is diffuse. Therefore, splitting θ into θ_1 and θ_2 should be done such that the posterior of $D(y, \theta)|y, \theta_2$ is relatively concentrated in comparison to the posterior of $D(y, \theta)|y, \theta_1$. If $D(y, \theta)|y, \theta_1$ is very concentrated, the performance loss from using $D^*(y, \theta_2)$ instead of $D^*(y)$ is almost negligible.

5.4 Empirical example

Example 5. The data y_{ij} are saliva cortisol concentrations collected at time points $t_{1j}, \ldots, t_{n_j,j}$ for 87 male white-collar workers. Van Eck *et al.* (1996) collected and analyzed these data using a mixed effects model in which the random intercepts β_{0j} and slopes β_{1j} are related to individual trait characteristics. We fitted a random slopes model to the data and captured the diurnal pattern by a fourth-degree polynomial. The likelihood of $\ln(y_{ij})$ is $\ln(y_{ij})|\beta_j \sim N(\beta_{0j} + \beta_{1j}t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 +$

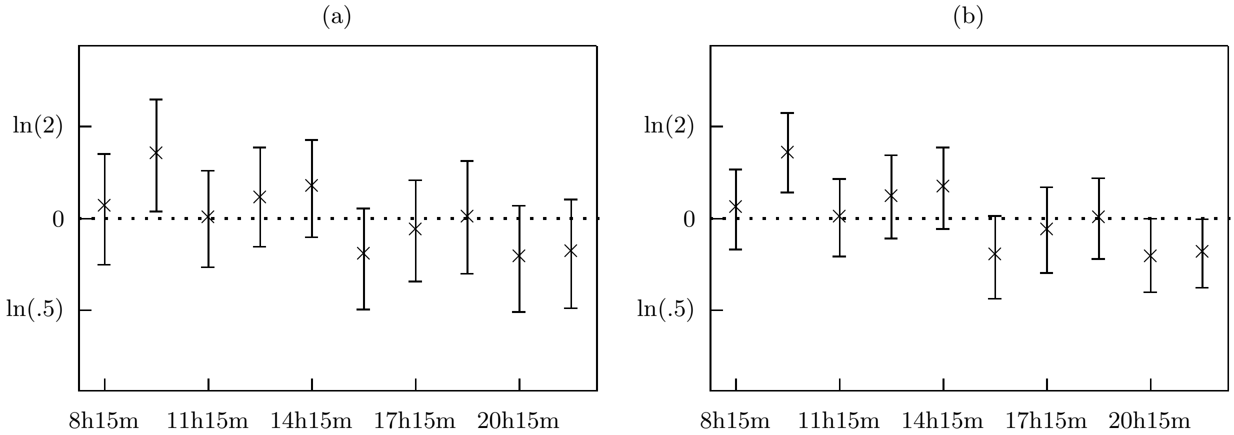

Figure 2: 95% posterior predictive intervals of (a) $D_{(k)}(y^{\mathrm{rep}},\theta)-D_{(k)}(y,\theta)$ against time; (b) $D^*_{(k)}(y^{\mathrm{rep}},\theta_2)-D^*_{(k)}(y,\theta_2)$ against time.

$\beta_4 t_{ij}^4,\sigma^2)$ and $(\beta_{0j},\beta_{1j})' \sim N((\beta_0,\beta_1)',W)$. The chosen priors are $p(\sigma^2)\propto 1/\sigma^2$, $p(\beta_k)\propto 1$, and $W^{-1}\sim Wish(.001E_2,.001)$, where E_2 is a 2×2 identity matrix. The model was estimated by Gibbs sampling (burnin period 2000, the subsequent 2000 draws were stored).

We examined whether the homoskedasticity of the within-subject errors (after log-transformation of the cortisol values) is violated by the data. Because the data are concentrated around 10 equally spaced time points, we formulated 10 intervals of one-and-a-half hour denoted by $A_{(1)},\ldots,A_{(10)}$ and carried out a posterior predictive check per interval. The k-th discrepancy measure is defined as

$$D_{(k)}(y,\theta)=\ln\,(\sum_{j=1}^{87}\sum_{i=1}^{n_j} e'_{ij}e_{ij}I_{\{t_{ij}\epsilon A_{(k)}\}})\ ; \qquad e_{ij}=\ln\,(y_{ij})-\beta_{0j}-\beta_{1j}t_{ij}-\beta_2t_{ij}^2-\beta_3t_{ij}^3-\beta_4t_{ij}^4\ .$$

In panel a of Figure 2, 95% posterior predictive intervals of the discrepancy comparisons are drawn (constructed from 1000 posterior predictive draws). It can be read that the homoskedasticity assumption is possibly violated at the second interval but the posterior predictive check is not conclusive. As a following step, the performance of the check was increased. Because direct sampling from the posterior is not possible, θ was split into $\theta_{1j}=(\beta_{0j},\beta_{1j})$ and $\theta_2=(\beta_2,\beta_3,\beta_4,\sigma^2,W)$ and the discrepancy measures were averaged using draws from $p(\theta_{1j}|y, \theta_2)$. The number of draws of θ_{1j} was set at 500 and was sufficient for accurate approximation of $D^*_{(k)}(y, \theta_2)$. The results are presented in panel b. The 95% intervals have shrinked substantially and the assumption of homoskedasticity is violated at the second interval and possibly violated at the seventh, ninth, and tenth interval, which seems sufficient evidence for reconsidering the posited model. To conclude, in this example a check on the assumption of homoskedasticity could be enhanced by a simple modification that does not require demanding computations inherent in nested MCMC schemes.

6 Concluding remarks

We discussed several methods for improving the sensitivity of a predictive model check. By means of examples, we showed that the applicability of the different methods varies with the testing problem under consideration. When the test quantity is a statistic, the model check can be improved by using the partial posterior predictive distribution instead of the posterior predictive distribution. This approach handles double use of the data in a natural way but involves evaluation of the likelihood of $T(y)$ at each posterior draw. This likelihood often is not available in closed form in which case one has to make an approximation. This can be done but it is not appealing and an alternative strategy may be adopted where the test quantity is modified rather than the replication distribution. We discussed some modifications suitable when the test quantity has a normal sampling distribution.

In addition to the existing modifications, we presented the mean discrepancy comparison and the mean discrepancy measure. The two methods operate similarly in that, approximately, they do not lead to a change in the posterior predictive mean but to a reduction in the posterior predictive variance. A difference between the two methods is that the mean discrepancy comparison always improves the performance of the check whereas this may not be the case for the mean discrepancy measure: If the formulation of the discrepancy measure is poor, computation of the posterior mean may produce an inferior model check. Therefore it is important to center the observations and some-times center the discrepancy measure itself. With regard to the computation, the mean discrepancy comparison can be computed by importance sampling which is a fast method when the sample sizes are small. A mean discrepancy measure can only be efficiently computed when direct sampling of the posterior is possible after conditioning on a subset of parameters. The examples showed that both methods may lead to a substantial improvement in the sensitivity of the check while analytical calculations and demanding computations are avoided.

Appendix

The prior predictive mean of $p^q_{\mathrm{cond}}(\theta)$ is smaller than or equal to the q-th moment of a standard uniform variable. To show this, the prior predictive mean of $p^q_{\mathrm{cond}}(\theta)$ is written as

$$\begin{aligned} \mathrm{E}[p^q_{\mathrm{cond}}(\theta)|\theta] &= \mathrm{E}[\Pr\{\delta(y^{\mathrm{rep}}_1, y, \theta) \geq 0 \cap \cdots \cap \delta(y^{\mathrm{rep}}_q, y, \theta) \geq 0|\theta, y\}] \\ &= \Pr\{\delta(y^{\mathrm{rep}}_1, y, \theta) \geq 0 \cap \cdots \cap \delta(y^{\mathrm{rep}}_q, y, \theta) \geq 0|\theta\}\ , \end{aligned} \tag{13}$$

where $y^{\mathrm{rep}}_1, \ldots, y^{\mathrm{rep}}_q$ are replicated data sets. Under the prior predictive sampling distribution, y, $y^{\mathrm{rep}}_1, \ldots, y^{\mathrm{rep}}_q$ are independent and identically distributed so that interchanging y and y^{rep}_j does not affect probability (13). Besides, as a continuous sampling distribution for $\delta(y^{\mathrm{rep}}_1, y, \theta)$ is assumed, the sum of probability (13) and the q probabilities obtained by interchanging y and y^{rep}_j $(j = 1, \ldots, q)$ is not larger than 1. Combining the two statements gives $\mathrm{E}[p^q_{\mathrm{cond}}(\theta)|\theta] \leq 1/(q+1) = \mathrm{E}[U_q]$ for $q \geq 1$ where U denotes a standard uniform variable. The equality is attained for $q = 1$. The rest of the proof is analogous to the proof of Theorem 1 of Meng (1994): by Jensen's inequality, $\mathrm{E}[p^q_\delta|\theta] \leq \mathrm{E}[p^q_{\mathrm{cond}}(\theta)|\theta]$ for all $q \geq 1$ and $\mathrm{E}[p_\delta|\theta] = \mathrm{E}[p_{\mathrm{cond}}(\theta)|\theta]$. Combining the two results concludes the proof.

References

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

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

Berger, J.O. (2003). Could Fisher, Jeffreys, and Newman have agreed on testing (with Discussion)? *Statistical Science*, **18**, 1–32.

Berger, J.O., and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence (with discussion). *Journal of the American Statistical Association*, **82**, 112–122.

Berkhof, J., van Mechelen, I., and Gelman, A. (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, **13**, 423–442.

DiCiccio, T.J., Kass, R.E., Raftery, A., and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903–915.

Geisser, S. (1993). *Predictive inference: an introduction.* London: Chapman and Hall.

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

Gelman, A., and Rubin, D.B. (1992). Inferences from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econo-*

metrica, **57**, 1317–1340.

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.

Hammersley, J.M., and Handscomb, D.C. (1964). *Monte Carlo Methods.* New York: Wiley.

Kass, R.E., and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 377–395.

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

Pettit, L.I., and Smith, A.F.M. (1985). Outliers and influential observations in linear models (with discussion). In *Bayesian Statistics 2* (Editors J.M. Bernardo, M. DeGroot, D. Lindley, and A.F.M. Smith), 473–494. Amsterdam: North-Holland.

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

Rubin, D.B., and Stern, H.S. (1994). Testing in latent class models using a posterior predictive check distribution. In *Latent Variable Analysis: Applications for Developmental Research*, (Editors E. von Eye and C. Clogg), 420–438.

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**, 1143–1156.

Stern, H., Arcus, D., Kagan, J., Rubin, D.B., and Snidman, N. (1994). Statistical choices in infant temperament research. *Behaviormetrika*, **21**, 1–17.

Tsui, K.-W., and Weerahandi, S. (1989). Generalized P values in significance testing of hypothesis in the presence of nuisance parameters. *Journal of the American Statistical Association*, **84**, 602–607.

Van Eck, M., Berkhof, H., Nicolson, N., and Sulon, J. (1996). The effects of perceived stress, traits, mood states, and stressful daily events on salivary cortisol. *Psychosomatic Medicine*, **58**, 447–458.