Posterior Predictive Assessment of Model Fitness via Realized Discrepancies

Andrew Gelman[1], Xiao-Li Meng[2], and Hal Stern[3]
University of California, Berkeley, The University of Chicago, and Iowa State University

September 5, 1995

Abstract

This paper considers the Bayesian counterparts of the classical tests for goodness of fit and their use in judging the fit of a single Bayesian model to the observed data. We focus on posterior predictive assessment, in a framework that also includes conditioning on ancillary statistics. The Bayesian formulation facilitates the construction and calculation of a meaningful reference distribution not only for any (classical) statistic, but also for any parameter-dependent "statistic" or *discrepancy variable*. The latter allows us to propose the *realized discrepancy assessment* of model fitness, which directly measures the true discrepancy between data and the posited model, for any aspect of the model which we want to explore. The computation required for the realized discrepancy assessment is a straightforward byproduct of the posterior simulation used for the original Bayesian analysis.

We illustrate with three applied examples. The first example, which serves mainly to motivate the work, illustrates the difficulty of classical tests in assessing the fitness of a Poisson model to a positron emission tomography image that is constrained to be nonnegative. The second and third examples illustrate the details of the posterior predictive approach in two problems: estimation in a model with inequality constraints on the parameters, and estimation in a mixture model. In all three examples, standard test statistics (either a χ^2 or a likelihood ratio) are not pivotal: the difficulty is not just how to compute the reference distribution for the test, but that in the classical framework no such distribution exists, independent of the unknown model parameters.

Keywords: Bayesian p-value, χ^2 test, discrepancy variable, graphical assessment, mixture model, model criticism, posterior predictive p-value, prior predictive p-value, realized discrepancy.

running title: Posterior Predictive Assessment of Model Fitness

[1] *Department of Statistics, University of California, Berkeley, CA 94720, U.S.A., gelman@stat.berkeley.edu*

[2] *Department of Statistics, University of Chicago, Chicago, IL 60637, U.S.A., meng@galton.uchicago.edu 312-702-0958 (O) 312-702-9810 (Fax)*

[3] *Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A., hstern@iastate.edu*

1 Introduction

1.1 Classical and Bayesian model assessment

Assessing the plausibility of a posited model (or assumptions in general) is always fundamental, especially in Bayesian data analyses. A Bayesian analysis, which conditions on the whole probability model, can be very misleading when the model is far from plausible. Any meaningful Bayesian analysis, therefore, should at least include a check to see if the posited model should be excluded because it fails to provide a reasonable summary of the data at hand. A standard classical approach for this kind of model-checking is to perform a goodness-of-fit test, which calculates a tail-area probability under the posited model to quantify the extremeness of the observed value of a selected discrepancy (e.g., differences between observations and predictions). The essence of the classical approach, as we see it, lies in comparing the observed value of the discrepancy to its reference (i.e., sampling) distribution derived under the posited model. The tail-area probability, or p-value, is only a computationally convenient way of locating the observed value in the reference distribution, in the days when graphical display could not be done routinely. Plotting the observed value against a reference distribution is more informative than merely reporting a p-value.

The main purpose of this paper is to extend the essence of the classical approach into the Bayesian framework, with the aim of providing pragmatic methods of assessing the fitness of a single model, especially in complex situations where classical frequentist calculation of a reference distribution of an informative statistic is infeasible. Our work is built upon several earlier papers, particularly Rubin (1984, §5), on assessing the fitness of a single Bayesian model. We use the term "assessment" instead of "testing" to highlight the fundamental difference between assessing the *discrepancies* between a model and data and testing the *correctness* of a model. We believe that there is a general concern that there has been too much emphasis on the latter problem and that, in the words of Tiao and Xu (1993), there is a need in practice for "... development of diagnostic tools with a greater emphasis on assessing the usefulness of an assumed model for specific purposes at hand rather than on whether the model is true."

For some problems, such as linear models, common goodness-of-fit tests serve as good diagnostics (when used and interpreted appropriately) and are easy to implement, because their reference distributions are known, at least approximately. Useful approximations to distributions of test statistics are possible for some problems (see, for example, Chernoff, 1954, concerning extensions of the linear model), but are not always available (see, for example, McCullagh, 1985, 1986, concerning the difficulty of finding distributions of classical goodness-of-fit tests in generalized linear models). The classical approach relying on known, or approximately known, reference distributions encounters difficulty in at least three kinds of models: models with severe restrictions on the parameters, such as positivity; models with probabilistic constraints, which may arise from a strong prior distribution; and unusual models that cannot be parameterized as generalized linear models. The following example illustrates the type of problem with which we are concerned, where it is difficult in practice and may be impossible in theory to construct a sensible classical test.

1.2 A Motivating example

Gelman (1990, 1992) describes a positron emission tomography experiment whose goal is to estimate the density of a radioactive isotope in a cross-section of the brain. The two-dimensional image is estimated from gamma-ray counts in a ring of detectors around the head. Each count is classified in one of n bins, based on the positions of the detectors when the gamma rays are detected, and a typical experimental run has about 6,000,000 counts. The bin counts, y_i, are modeled as independent Poisson random variables with means θ_i that can be written as a linear function of the unknown image g, $\theta = Ag + r$, where $\theta = (\theta_1, \ldots, \theta_n)$, A is a known linear operator that maps the continuous g to a vector of length n, and r is a known vector of corrections. Both A and r, as well as the image, g, are nonnegative. In practice, g is discretized into "pixels" and becomes a long (length greater than n) nonnegative vector, and A becomes a matrix with nonnegative elements.

Were it not for the nonnegativity constraint, there would be no problem finding an image to fit the data; in fact, an infinite number of images g solve the linear equation, $y = Ag + r$. However, due to the Poisson noise, and perhaps to failures in the model, it often occurs in practice that no exact nonnegative solutions exist, and we must use an estimate (or family of estimates) $\hat{g}$ for which there is some discrepancy between the data, y, and their expectations, $\hat{\theta} = A\hat{g} + r$. The discrepancy between y and $\hat{\theta}$, however, should not be too large; given the truth of the model, it is limited by the variance in the independent Poisson distributions. For example, the χ^2 discrepancy,

$$X^2(y; \hat{\theta}) = \sum_{i=1}^{n} \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}, \tag{1}$$

should be no greater than could have arisen from a χ^2 distribution with n degrees of freedom ($y_i > 50$ for almost all the bins i, and so the χ^2 distribution, based on the normal approximation to the Poisson, is essentially exact). In fact, $X^2(y; \hat{\theta})$ should be considerably less, since $\hat{\theta}$ is the best fit to the data.

The posited model was fit to a real dataset, y, with $n = 22{,}464$. We would ultimately like to determine the posterior distribution, $P(\theta|y)$, given a reasonable prior distribution. As a first step, we need to determine if the model is invalidated by the observed data. For this dataset, the best-fit nonnegative image $\hat{g}$ was not an exact fit; the discrepancy between y and $\hat{\theta}$ was $X^2(y; \hat{\theta}) \approx 30{,}000$. This is unquestionably a rejection of the model, unexplainable by the Poisson variation. At this point, the model and data should be examined to find the causes of the lack of fit. Possible failures in the model include error in the specification of A and r, lack of independence or super-Poisson variance in the counts, and error from discretizing the continuous image, g.

There is no difficulty in deciding to reject the model here as long as $X^2(y; \hat{\theta})$ is greater than $n + 2\sqrt{2n} \approx 23{,}000$, for we can be almost certain that the model does not fit the data in this case. Suppose, however, that the experimental procedure is carefully examined, the model is made more accurate, and the new model is fit to the data, yielding a χ^2 discrepancy of 22,000, or 20,000. We should probably still be distrustful of the model, since a whole continuous image is being fit to the data. (In fact, as the total number of counts increases, the Poisson variances decrease proportionally, and it becomes increasingly likely that an exact fit image $\hat{g}$ will exist that solves $y = A\hat{g} + r$. Thus, conditional on the truth of the model, $X^2(y; \hat{\theta})$ must be zero, in the limit as the number of counts approaches infinity with a fixed number of bins, n. Due to massive near-collinearity, the positron emission tomography model is not near that asymptotic state even with 6,000,000 total counts.) If k linear parameters were fit, the χ^2 discrepancy defined in (1) would have a χ^2 distribution with $n - k$ degrees of freedom. It is difficult to determine how many degrees of freedom correspond to the fitting of a continuous image.

We have arrived at a practical problem: how to assess the quality of a model's fit to the observed data for complicated models in "close calls" for which the simple χ^2_n bound is too crude. The problem is important and common, because if we take modeling seriously, as we should with any Bayesian analysis, we will gradually improve models that clearly do not fit, and upgrade them into close calls.

1.3 A brief overview

The preceding example illustrates the need for model checking in complex situations. In Bayesian statistics, a model can be checked in at least three ways: (1) examining sensitivity of inferences to reasonable changes in the prior distribution and the likelihood; (2) checking that the posterior inferences are reasonable, given the substantive context of the model; and (3) checking that the model fits the data. We address the third of these concerns using the posterior predictive distribution for a *discrepancy variable*, an extension of classical test statistics to allow dependence on unknown (nuisance) parameters. The posterior predictive assessment was introduced by Guttman (1967), applied by Rubin (1981), and given a formal Bayesian definition by Rubin (1984). Our new methodological contribution is the adoption of more general discrepancy variables, which allows more direct assessment of the discrepancy between data and the posited model.

The recent rapid development of Bayesian computation allows us to fit more realistic and sophisticated models than previously possible, and thus there is a corresponding need for general methods to assess the fitness of these models when classical tests are not applicable, as our motivating example demonstrates. The approach we discuss here appears to be one such method. It is simple, both conceptually and computationally, and connects well to the classical goodness-of-fit methods that most researchers are familiar with. It is also very general, applicable for comparing observations with model predictions in any form. Our own applied work has benefited from the application of this method, as documented here and in many examples in Gelman, Carlin, Stern and Rubin (1995); also see Belin and Rubin (1995) and Upadhyay and Smith (1993). Meng (1994) discusses a similar method for testing parameter values within a model rather than for the entire model. West (1986) and Gelfand, Dey, and Chang (1992) also present posterior predictive approaches to model evaluation, in the context of sequential data and cross-validation of the existing data set, respectively, rather than comparing to hypothetical replications of the entire data set.

Our paper is organized as follows. Section 2 consists of definitions, a discussion of computational issues, and a theoretical illustration of the posterior predictive approach. Section 3 presents detailed illustration in two applied contexts. Section 4 provides discussion of various issues, including a brief comparison to the prior predictive approach of Box (1980).

2 Posterior predictive assessment of model fitness

2.1 Posterior predictive assessment using classical test statistics

We use the notation y for the observed data, H for the assumed model, and θ for the d-dimensional unknown model parameter (d can be infinite). We use T to denote a test statistic, a function from data space to the real numbers. In addition, we may specify a set of *ancillary statistics*, $A(y)$—functions of the data that are to be held constant in replications (for example, the sample size). In general, $A(y)$ may be a scalar, a vector, or the empty set.

To avoid confusion with the observed data, y, we define y^{rep} as the *replicated* data that could have been observed, or, to think predictively, as the data we would see if the experiment that produced y today were replicated tomorrow with the same model, H, the same (unknown) value of θ that produced y, and the same value of the ancillary statistics, $A(y)$. We denote the distribution of this replication by P_A:

$$P_A(y^{\text{rep}}|H,\theta) = P[y^{\text{rep}} \mid H, \theta, A(y^{\text{rep}}) = A(y)]. \tag{2}$$

In this notation, the classical p-value based on T is

$$p_c(y,\theta) = P_A[T(y^{\text{rep}}) \geq T(y) \mid H, \theta], \tag{3}$$

and a p-value very close to 0 indicates that the lack of fit in the direction of the test statistic, $T(y)$, is unlikely to have occurred under the model. In (3), y is fixed, with all the randomness coming from y^{rep}, and the value of p_c is obtainable only when it is free of the (nuisance) parameter θ. As we emphasized in Section 1.1, the key ingredient of the classical approach is not the p-value given in (3), but rather locating $T(y)$ in the distribution of $T(y^{\text{rep}})$ derived from (2).

In the Bayesian framework, the inference for θ is provided by its posterior distribution, $P(\theta|H,y)$, where the model H now also includes the prior distribution, $P(\theta)$. Correspondingly, the reference distribution of the future observation y^{rep}, given $A(y)$, is its *posterior predictive distribution*,

$$P_A(y^{\text{rep}}|H,y) = \int P_A(y^{\text{rep}}|H,\theta)P(\theta|H,y)d\theta. \tag{4}$$

The observed value of T, $T(y)$, is then plotted against the distribution of $T(y^{\text{rep}})$ induced by (4). The corresponding tail-area probability, analogous to (3), is

$$p_b(y) = P_A[T(y^{\text{rep}}) \geq T(y)|H,y] = \int p_c(y,\theta)P(\theta|H,y)d\theta, \tag{5}$$

which is the classical p-value of (3) averaged over the posterior distribution of θ. This is the p-value defined by Rubin (1984), which we term the *posterior predictive p-value* (also see Meng, 1994) to contrast it with the *prior predictive p-value* of Box (1980); see Section 4.1 for discussion.

Clearly, the sampling and posterior predictive reference distributions of $T(y^{\mathrm{rep}})$ are identical when T is a pivotal quantity, conditional on $A(y)$, under the model H. For any posited Bayesian model, the posterior predictive approach has the virtue of constructing a well-defined reference distribution (with a corresponding tail-area probability), which is easy to simulate (see Section 2.3), for any test statistic.

Interestingly, the posterior predictive replication appears to be the replication that the classical approach intends to address. Figure 1a shows the posterior predictive reference set, which corresponds to repeating the experiment tomorrow with the same model, H, and the same (unknown) value of θ that produced today's data y. Because θ is unknown, but assumed to have the same value that generated the current data y, we simulate from its posterior distribution given y. Figure 1a is a conditional independence graph (see, e.g., Wermuth and Lauritzen, 1990) that displays the dependence relations between y, θ, and y^{rep}. Given θ, the data y and the replicate y^{rep} are independent, and both represent possible datasets resulting from the given value of θ and the model H. The fact that θ is displayed only once in the graph is *not* intended to imply that θ is known or "fixed," as in a frequentist analysis. (Figures 1b and 1c will be discussed in Section 4.1.)

2.2 Posterior predictive assessment using discrepancy variables

The Bayesian formulation not only handles the problem of nuisance parameters for classical test statistics but also naturally allows the use of test "statistics" that depend on θ. This generalization beyond Rubin's (1984) formulation is important because it allows us to compare directly the discrepancy between the observed data and the posited model, instead of between the data and the best fit of the model. It also, as we shall show in Section

2.3, greatly simplifies the computation of tail-area probabilities. Parameter-dependent test statistics have been suggested before by Tsui and Weerahandi (1989) and Meng (1994) for the case of testing parameters within a model. In particular, Tsui and Weerahandi (1989) termed such a statistic a "test variable" and Meng (1994) called it a "discrepancy variable"; we prefer the latter as it emphasizes that the focus here is to measure discrepancies between a model and the data, not to test whether a model is true.

For a selected discrepancy variable, $D(y;\theta)$, its reference distribution is derived from the joint posterior distribution of $y^{\rm rep}$ and θ,

$$P_A(y^{\rm rep}, \theta | H, y) = P_A(y^{\rm rep} | H, \theta) P(\theta | H, y). \tag{6}$$

This replication is a natural extension of (4), which is a marginal distribution of (6). Directly locating the *realized* value, $D(y;\theta)$, within its reference distribution, however, becomes infeasible when $D(y;\theta)$ depends on the unknown θ. The classical tail-area approach now becomes one possible technical device for measuring the location. Specifically, we can formally define a tail-area probability of D under its posterior reference distribution:

$$p_b(y) = P_A[D(y^{\rm rep};\theta) \geq D(y;\theta) | H, y], \tag{7}$$

which includes p_b of (5) as a special case. Interestingly, although the realized value itself is not observable, this posterior predictive p-value is well defined and calculable. The reference set for the discrepancy variable is the same as that in Figure 1a, except that it is now composed of pairs $(y^{\rm rep}, \theta)$ instead of just $y^{\rm rep}$. The term "realized" discrepancy is borrowed from Zellner (1975), where such discrepancies were used for conducting Bayesian analysis of regression error terms (i.e., no predictive replications were involved); also see Chaloner and Brant (1988) and Chaloner (1991) for related work.

By suitably eliminating its dependence on unknown θ, we can also construct classical test statistics from a discrepancy variable. For example, as illustrated in Section 2.4 for the χ^2 test, the classical statistics that arise from comparing data with the best fit of the models typically correspond to the *minimum discrepancy*:

$$D_{\text{min}}(y) = \min_{\theta} D(y;\theta).$$

Another possibility is the *average discrepancy* statistic,

$$D_{\text{avg}}(y) = E(D(y;\theta)|H,y) = \int D(y;\theta)P(\theta|H,y)d\theta.$$

The corresponding posterior predictive p-values are defined by (5) with T being replaced by D_{min} and D_{avg}, respectively.

2.3 Computation for posterior predictive assessment

Computation of reference distributions of discrepancy variables and the corresponding tail-area probabilities can be performed analytically for some simple problems (e.g., Meng, 1994), but in complicated models, such as arise in practical applications, it is more easily accomplished via Monte Carlo simulation. This typically is not an extra computational burden, because simulation is a standard tool for Bayesian analysis with complex models. In other words, the required computation is typically a byproduct of the usual Bayesian simulation that provides a set of draws of θ from the posterior distribution, $P(\theta|H,y)$.

Specifically, consider the computation required for comparing the realized discrepancy $D(y;\theta)$ to its reference distribution under (6). Given a set of (possibly dependent) draws, $\theta^j, j = 1,\ldots,J$, we need only perform the following two steps for each j:

1. Given θ^j, draw a simulated replicated data set, $y^{\text{rep}\,j}$, from the sampling distribution, $P_A(y^{\text{rep}}|H,\theta^j)$;

2. Calculate $D(y;\theta^j)$ and $D(y^{\text{rep}\,j};\theta^j)$.

Having obtained $\{(D(y;\theta^j), D(y^{\text{rep}\,j};\theta^j))\,, j = 1,\ldots,J\}$, we can make a scatterplot (see Figures 7 and 9) to make a graphical assessment, and estimate p_b of (7) by the proportion of the J pairs for which $D(y^{\text{rep}\,j};\theta^j)$ exceeds $D(y;\theta^j)$.

For many problems, step 1 is easy (step 2 is almost always straightforward). In addition, once the replicates have been drawn, the same draws $\{(\theta^j, y^{\text{rep}\,j}), j = 1, \ldots, J\}$ can be used for as many realized discrepancy variables as one wishes. This is particularly convenient if we are interested in measuring the discrepancy of various aspects of the model (e.g., mean-squared errors, quantiles, patterns of residuals, etc.). In cases where the classical p-value based on $D(y;\theta)$ (i.e., treating θ as known) is easy to calculate analytically, one can simulate p_b more efficiently by directly averaging the classical p-values, $p_c(y, \theta^j), j = 1, \ldots J$. In any case, we recommend making the scatterplot whenever feasible because it tells us the typical magnitudes of $D(y;\theta)$ and $D(y^{\text{rep}};\theta)$. We also note that "double parametric bootstrap" or various Bayesian bootstrap methods can sometimes be used to obtain approximations to posterior predictive distributions (e.g., Rubin, 1987, Ch. 4; Tsay, 1992).

Simulating reference distributions for D_{min} and D_{avg} is more complicated because one must minimize or average over θ when evaluating their values. To compute D_{min}, one needs to determine, for each j, the value θ for which $D(y^{\text{rep}\,j};\theta)$ is minimized; this may not be an easy computation. The computation for $D_{\text{avg}}(y)$ requires a potentially even more difficult integration. Thus, while the minimum and average discrepancies are interesting theoretically and for comparison to classical methods, they are typically much harder to compute than the realized discrepancy, particular with complicated models.

2.4 A theoretical example: χ^2 discrepancies

As a theoretical illustration, consider the χ^2 discrepancy, by which we simply mean a sum of squares of standardized residuals of the data with respect to their expectations under a posited model. For simplicity, assume that the data are expressed as a vector of n independent observations (not necessarily identically distributed), $y = (y_1, \ldots, y_n)$, given the parameter vector θ, and no ancillary statistics. The χ^2 discrepancy is then

$$X^2(y;\theta) = \sum_{i=1}^{n} \frac{(y_i - E(y_i|\theta))^2}{\text{var}(y_i|\theta)}. \tag{8}$$

For example, the discrepancy in equation (1) in Section 1.2 is the above formula for the Poisson distribution, evaluated at the estimate $\hat{\theta}$. In this section, we will assume that, given θ, expression (8) has an approximate χ_n^2 distribution.

Now suppose we are interested in assessing the fitness of a model, H, that constrains θ to lie in a subspace of $\Re^n$, and for the purpose of theoretical comparisons we focus on calculating posterior predictive p-values. Given a prior distribution, $P(\theta)$, on the subspace, we can calculate the posterior predictive p-value based on X^2 as

$$p_b(y) = \int P(\chi_n^2 \geq X^2(y;\theta))P(\theta|H,y)d\theta, \tag{9}$$

where χ_n^2 represents a chi-squared random variable with n degrees of freedom. This computation is straightforward, once draws from $P(\theta|H,y)$ are obtained, because the tail-area probability function for χ_n^2 can be found in any standard statistical software.

Computations for posterior predictive p-values for minimum χ^2, X^2_{min}, and average χ^2, X^2_{avg}, are more complicated in general, as discussed in Section 2.3. However, when H is a linear model (i.e., θ is constrained to lie on a hyperplane of dimension k), the minimum χ^2 discrepancy is essentially equivalent to the classical goodness-of-fit test statistic. (The classical χ^2 test is sometimes evaluated at the maximum likelihood estimate (MLE) and sometimes at the minimum-χ^2 estimate, a distinction of some controversy (see, e.g., Berkson, 1980); we consider the minimum χ^2 in our presentation, but similar results could be obtained using the MLE.) Thus $X^2_{\text{min}}(y)$ is approximately pivotal with a χ^2_{n-k} distribution (e.g., Cochran, 1952). Consequently, the posterior predictive p-value can be approximated by $P(\chi^2_{n-k} \geq X^2_{\text{min}}(y))$.

Furthermore, if θ is given a diffuse uniform prior distribution in the subspace defined by a linear model H, then the posterior predictive distributions of $X^2(y;\theta)$ and $X^2_{\text{avg}}(y)$ are closely related to that of $X^2_{\text{min}}(y)$. With the diffuse prior distribution, the posterior distribution of $X^2(y;\theta) - X^2_{\text{min}}(y)$ is approximately χ_k^2. Then we can decompose the average

χ^2 statistic as follows:

$$X^2_{\text{avg}}(y) = E[X^2_{\text{min}}(y) + (X^2(y;\theta) - X^2_{\text{min}}(y))|y] \approx X^2_{\text{min}}(y) + k,$$

and thus the average χ^2 discrepancy is approximately equivalent to the minimum χ^2 discrepancy, just shifted by a constant, k.

For the realized discrepancy, $X^2(y;\theta)$, the same decomposition can be applied to (9), which yields

$$\begin{aligned} p_b(y) &= \int P[\chi^2_n \geq X^2_{\text{min}}(y) + (X^2(y;\theta) - X^2_{\text{min}})]P(\theta|H,y)d\theta \\ &= P(\chi^2_n \geq X^2_{\text{min}}(y) + \chi^2_k) = P(\chi^2_n - \chi^2_k \geq X^2_{\text{min}}(y)), \end{aligned}$$

where χ^2_n and χ^2_k are independent random variables. In other words, assessing a linear model using $X^2(y;\theta)$ is equivalent to using $X^2_{\text{min}}(y)$ but with a different reference distribution: instead of a χ^2_{n-k}, the reference distribution is the difference between two independent χ^2 random variables, $\chi^2_n - \chi^2_k$. This implies that the posterior predictive p-value for the realized discrepancy, $X^2(y;\theta)$, is larger than that from $X^2_{\text{min}}(y)$; the reference distribution of the former has a larger variance, $2(n+k)$ versus $2(n-k)$. Suppose, for example, $n = 250$, $k = 200$, and data y are observed for which $X^2_{\text{min}}(y) = 80$. Under $X^2_{\text{min}}(y)$, this is three standard deviations away from the mean of the χ^2_{50} reference distribution—an indication of lack of fit. The corresponding reference distribution for $X^2(y;\theta)$ is $\chi^2_{250} - \chi^2_{200}$, which has the same mean of 50 but with a larger standard deviation of 30, and thus the data do not appear to be a surprise at all.

How do we interpret this difference? The lack of fit under X^2_{min} shows that the data are not as close to the best fitting model, in terms of the sum of the standardized residuals, as would be expected from a model with a large number of parameters. However, it is possible that this lack of fit will not adversely affect practical inferences from the data. In the example considered here, the realized discrepancy indicates that the data are reasonably close to what could be expected in replications under the hypothesized model. The extra

30 by which the minimum discrepancy exceeds its expectation seems large compared to 50 degrees of freedom but small when examined in the context of the 250-dimensional space of y. In the next section, we indicate the possibility of using the difference between the posterior predictive assessments from these two discrepancies to detect whether the lack of fit is due to the likelihood or due to the prior specification.

If prior knowledge of θ is added, as expressed by a nonuniform prior distribution, the posterior predictive p-value for $X^2_{\min}(y)$ is unchanged, since $X^2_{\min}(y)$ is still a pivotal quantity in the linear model case, but the tests based on $X^2_{\text{avg}}(y)$ and $X^2(y;\theta)$ now change, as they are now measuring discrepancy from the prior model as well as the likelihood. Sensitivity to the prior distribution is discussed in Section 3 in the context of our applied examples, and further discussed in Section 4.2.

3 Illustration with two applied examples

3.1 Fitting an increasing, convex mortality rate function

For a simple real-life example, we reanalyze the data of Broffitt (1988), who presents a problem in the estimation of mortality rates (Carlin, 1992, provides another Bayesian analysis of these data). For each age, t, from 35 to 64 years, inclusive, Table 1 gives N_t, the number of people insured under a certain policy and y_t, the number of insured who died. (People who joined or left the policy in the middle of the year are counted as half.) We wish to estimate the mortality rate (probability of death) at each age, under the assumption that the rate is increasing and convex over the observed range. The observed mortality rates are shown in Figure 2 as a solid line. The observed deaths at each age, y_t, are assumed to follow independent binomial distributions, with rates equal to the unknown mortality rates, θ_t, and known population sizes, N_t (equivalently, we could consider the values N_t as random but treat them as ancillary statistics). Because the population for each age was in the hundreds or thousands, and the rates were so low, we use the Poisson approximation for mathematical convenience: $P(y|\theta) \propto \prod_t \theta_t^{y_t} e^{-N_t\theta_t}$. An optimization routine was used to maximize this likelihood, under the constraint that the mortality rate be increasing and convex. The maximum likelihood fit is displayed as the dotted line in Figure 2. Having obtained an estimate, we would like to check its fit to the data. The obvious possible flaws of the model are the Poisson distribution and the assumed convexity.

The χ^2 discrepancy between the data and the maximum likelihood estimate is 30.0, and the minimum χ^2 discrepancy (using minimum χ^2 estimates in place of the MLE) is 29.3. These are based on 30 data points, with 30 parameters being fit. There are obviously, however, less than 30 free parameters, because of the constraints implied by the assumption of increasing, convex rates. In fact, the MLE lies on the boundary of constraint space; the solution is essentially characterized by only four parameters, corresponding to the two endpoints and the two points of inflection of the best-fit increasing, convex curve in Figure 2. So perhaps a χ^2_{26} distribution is a reasonable approximation to the reference distribution for the minimum χ^2 statistic?

As a direct check, we can simulate the sampling distribution of the minimum χ^2 statistic, assuming $\theta = \hat{\theta}$, the MLE. The resulting distribution of $X^2_{\text{min}}(y^{\text{rep}})$ is shown in Figure 3; it has a mean of 23.0 and a variance of 43.4 (by comparison, the mean and variance of a χ^2_{26} distribution are 26 and 52, respectively). The observed value, $X^2_{\text{min}}(y) = 29.3$, is plotted as a vertical line in Figure 3; it corresponds to a tail-area probability of 16%. The distribution of Figure 3 is only an approximation, however, as the value of θ that generates the current data is unknown. In particular, we do not expect the true θ to lie exactly on the boundary of the constrained parameter space. Moving θ into the interior would lead to simulated data that would fit the constraints better, and thus the distribution of Figure 3 provides a conservative p-value for the minimum χ^2 discrepancy.

To conduct a Bayesian inference, we need to define a prior distribution for θ. Since we were willing to use the MLE, we use a uniform prior distribution, under the constraint of increasing convexity. (The uniform distribution is also chosen here for simplicity; Broffitt, 1988, and Carlin, 1992, apply various forms of the gamma prior distribution.) Samples from the posterior distribution are generated by simulating a random walk through the space of permissible values of θ, using the Metropolis algorithm. Nine parallel sequences were simulated, three starting at the MLE and three at each of two crude extreme estimates of θ—one a linear function, the other a quadratic, chosen to loosely fit the raw data. Convergence of the simulations was monitored using the method of Gelman and Rubin (1992), with the iterations stopped after the within-sequence and total variances were roughly equal for all components of θ. Nine draws from the posterior distribution for θ, one from each of the simulated sequences, are plotted as dotted lines in Figure 4, with the MLE from Figure 2 displayed as a solid line for comparison.

3.2 Posterior predictive assessment for the mortality rate example

To make a posterior predictive assessment of the fitness, it is necessary to define a reference set of replications; i.e., a set of "fixed features" in the notion of Rubin (1984). For this dataset, we defined replications in which the (observed) population size and (unobserved) mortality rates at each age stayed the same, with only the number of deaths varying, according to their assumed Poisson distributions. For each draw from the posterior distribution of θ, we simulated a replication; a random sample of nine replicated datasets corresponding to the previous nine draws of θ is plotted as dotted lines in Figure 5, with the observed frequencies from Figure 2 displayed as a solid line for comparison.

The fitness of the model can be checked directly by examining a graph such as Figure 5—is the solid line an outlier in the forest of dotted lines?—or quantitatively, by defining a statistic such as y_{64}, the number of deaths at age 64, and comparing it to the distribution of simulated values of y_{64}^{rep}. We emphasize that whenever feasible, graphical assessment, including those corresponding to p-value calculations, should be made, because they not only provide direct inspections of various magnitudes, but also help to detect problems that may not be easily "visible" otherwise (e.g., Figure 5 shows that the lack of fit is much more prominent for later ages; we will return to this point shortly).

Checking residuals, especially in the form of the χ^2 discrepancy, is another standard practice for detecting lack of fit, and here we illustrate the minimum and realized χ^2 discrepancies discussed in Section 2.4. Besides illustrating the posterior predictive approach graphically, we also estimate the associated p-value as it provides a useful probability statement (when interpreted correctly) supplementing the graphical assessment. For each simulated replication, y^{rep}, the optimization routine was run to find the minimum χ^2 discrepancy, $X^2_{\min}(y^{\mathrm{rep}})$. A histogram of these minimum χ^2 values—the reference distribution for $X^2_{\min}(y)$—is displayed in Figure 6. With a mean of 21.1 and a variance of 39.6, this posterior predictive reference distribution has lower values than the approximate distribution based on the MLE and displayed in Figure 3. The posterior predictive p-value of the minimum χ^2 is 10%, which is more extreme than the maximum likelihood approximation, as expected.

For the realized discrepancy, Figure 7 shows a scatterplot of the realized discrepancy, $X^2(y;\theta)$ and the predictive discrepancy, $X^2(y^{\mathrm{rep}};\theta)$, in which each point represents a different value of $(\theta, y^{\mathrm{rep}})$ drawn from the posterior distribution, as described in Section 2.3. The tail-area probability of the realized discrepancy is just the probability that the predictive discrepancy exceeds the realized discrepancy, which in this case equals 6.3%, the proportion of points above the 45° line in the figure. The realized discrepancy p-value is more extreme than the minimum discrepancy p-value, which perhaps suggests that, given the constraint of increasing convexity, it is the uniform prior distribution, not necessarily the likelihood, that does not fit the data. (The analysis of linear models in Section 2.4 suggests that if the likelihood were the source of the poor fit, the minimum discrepancy test would give the more extreme tail-area probability. Assessing sources of lack of fit is an important topic that requires further investigation; related discussions are provided in Section 4.1.)

Even if we do not overhaul the model immediately, it is useful to note that the posterior predictive datasets were mostly higher than the observed data for the later ages (see Figure 5), and to consider this information when reformulating the model or setting a prior distribution for a similar new dataset. Even if the assumption of convex mortality rate is true in the natural population, it is very likely that the insurance company has screened out some high-risk older people, and thus destroys the convexity for later ages. In addition, there are general problems with using the uniform prior distribution for parameters constrained to be increasing, as discussed by Gelman (1995) in the context of this example.

3.3 Assessing fitness of a mixture model in psychology

Stern, Arcus, Kagan, Rubin, and Snidman (1995) fit a latent class model to the data from an infant temperament study. Ninety-three infants were scored on the degree of motor activity and crying to stimuli at 4 months and the degree of fear to unfamiliar stimuli at 14 months. Table 2 gives the data, y, in the form of a $4 \times 3 \times 3$ contingency table. The latent class model specifies that the population of infants is a mixture of relatively homogeneous subpopulations, within which the observed variables are independent of each other. The parameter vector, θ, includes the proportion of the population belonging to each mixture class and the multinomial probabilities that specify the distribution of the observed variables within a class. Psychological and physiological arguments suggest two to four classes for the mixture, with specific predictions about the nature of the infants in each class.

Table 3 summarizes the results of fitting one through four-class models using the EM algorithm. A discrepancy variable usually associated with contingency tables is the log likelihood ratio (with respect to the saturated model), $D(y;\theta) = 2\sum_i y_i \log[y_i/E(y_i|\theta)]$, where the sum is over the cells of the contingency table. The final column in Table 3 gives $D_{\min}(y)$ for each model. The two-class mixture model provides an adequate fit that does not appear to improve with additional classes. The maximum likelihood estimates of the parameters of the two-class model (not shown) indicate that the two classes correspond to two groups: the uninhibited children (low scores on all variables) and the inhibited children (high scores on all variables). It is well known that the usual asymptotic reference distribution for the likelihood ratio test (the χ^2 distribution) is not appropriate for mixture models (e.g., Titterington, Smith, and Makov, 1985).

At one level, this is a model selection problem (i.e., choosing the number of classes) for which a complete Bayesian analysis, incorporating the uncertainty in the number of classes, could be carried out. However, such an analysis is complicated by the fact that the parameters of the various probability models (e.g., the two- and four-class mixture models) are related, but not in a straightforward manner. To be more explicit, theory suggests at least two fundamentally different groups of children, the inhibited and the uninhibited. Additional classes, if they are present, represent a bifurcation of one or both of these classes, and the parameters of such classes are related to the parameters of the two-class model in a way that is difficult to model explicitly. Given the small amount of data, we restrict our attention to only assessing the fit of the two-class model.

The prior distribution of the parameters of the latent class model is taken to be a product of independent Dirichlet distributions: one for the class proportions, and one for each set of multinomial parameters within a mixture class. The Dirichlet parameters were chosen so that the multinomial probabilities for a variable (e.g., motor activity) are centered around the values expected by the psychological theory but with large variance. The use of a weak but not uniform prior distribution helps identify the mixture classes (e.g., the first class of the two-class mixture specifies the uninhibited infants). With this prior distribution and the latent class model, draws from the posterior distribution are obtained using the data augmentation algorithm of Tanner and Wong (1987). Ten widely dispersed starting values were selected and the convergence of the simulations was monitored using the method of Gelman and Rubin (1992). The draws from the posterior distribution of the parameters for the two-class model were centered about the MLE. Rubin and Stern (1994) describe the prior distribution and resulting Bayesian analysis more fully. Their analysis of the data includes a posterior predictive evaluation of the likelihood ratio statistic for testing a one-class model versus a two-class model but does not directly address the fit of the model.

3.4 Posterior predictive assessment for the psychology example

To assess the quality of fit of the two-class model, we define replications of the data in which the parameters of the latent class model are the same as those responsible for the available data. These replications may be considered as data sets that would be expected if new samples of infants were to be selected from the same population. For each draw from the posterior distribution, a replicated data set y^{rep} was drawn according to the latent class sampling distribution. The reference distribution of the minimum discrepancy, $D_{\text{min}}(y^{\text{rep}})$, based on 500 replications, is shown in Figure 8 with a dashed line indicating the observed value $D_{\text{min}}(y)$. The mean of this distribution, 23.4, and the variance, 45.3, differ nontrivially from the χ^2_{20} distribution that would be expected if the usual asymptotic results applied. The posterior predictive p-value is 93% based on these replications.

Using D_{min} with mixture models is often problematic because of multi-modality of the likelihood. For the data at hand, two modes of the two-class mixture likelihood were found, and for larger models the situation can be worse. Model assessments based on only a single mode, such as with D_{min}, may ignore important information.

In contrast, the realized discrepancy, $D(y;\theta)$, uses the entire posterior distribution rather than a single mode. In addition, the realized discrepancy requires much less computation, as discussed in Section 2.3. Figure 9 is a scatterplot of the discrepancies for the observed data and for the replications under the two-class model. The p-value for the realized discrepancy is 74% based on 500 Monte Carlo draws. We thus have not obtained any evidence suggesting lack of fit of the two-class model. Again, this says nothing about the correctness of the model, only that the values of the discrepancy measures we have examined are reasonable under the posited model.

To indicate the type of result that we might obtain for a clearly inadequate model, we performed the same calculations for the one-class mixture model (equivalent to the independence model for the contingency table), yielding p-values for the minimum and realized discrepancies of 2.4% and 5.8%, respectively. Here the realized discrepancy gives the less extreme p-value, which seems to confirm that the lack of fit is due to the likelihood, not the prior distribution (as we mentioned in Section 3.2, this assertion requires validation from further studies). At this point, a person who is using posterior predictive p-values in the Neyman-Pearson fashion might be troubled by the fact that the two p-values are on the opposite sides of the conventional cut-off point 5%. We emphasize again that we are merely using them as probability measures of lack of fit of various aspects of the posited model, which we know is at best a useful inferential approximation to the underlying true model. Nevertheless, empirical and theoretical studies so far do suggest that posterior predictive p-values generally have reasonable long-run frequentist properties. For example, the theoretical results on Type I error presented in Meng (1994) can be easily extended to replications involving ancillary statistics, as presented in an early version of this paper, Gelman, Meng, and Stern (1992).

4 Making Choices

4.1 The differences among replications and comparison to the prior predictive distribution of Box (1980)

Choosing replications amounts to specifying the joint distribution, $P[y, \theta, y^{\mathrm{rep}} \mid A(y^{\mathrm{rep}}) = A(y)]$, from which all reference distributions can be derived by conditioning on y, with the ancillary statistics being fixed at their observed values. Three different replications are illustrated in Figure 1. Figure 1a is the posterior predictive replication, the focus of this paper, corresponding to repeating the experiment tomorrow with *the same (unknown) value of* θ that produced today's data, y. In contrast, Figure 1b shows the prior predictive replication, advocated by Box (1980), in which *new values of both* θ *and* y are assumed to occur tomorrow. Figure 1c shows a mixed predictive replication that can be useful with hierarchical models, where we may choose to assume that the same value of the hyperparameters α defines the replications but that new values of the parameters θ are assumed to occur. In practice, the choice of a model for y^{rep}, as well as the choice of ancillary statistics, should depend on which hypothetical replications are of interest.

Unlike the posterior predictive replication, the prior predictive replication is undefined under improper prior distributions, which limits its use in practice. More generally, if the parameters θ are well-estimated from the data, posterior predictive tests give results similar to classical procedures for reasonable prior distributions. In contrast, results of prior predictive tests are sensitive to the prior distribution, even in analyses with a large amount of data in which the prior distribution is essentially irrelevant to posterior inferences.

Both the examples in Section 3 consider replications in which the values of the model parameters are fixed (although unknown) and therefore use the posterior predictive replication. To see the consequence of choosing different replications, we compare the prior and posterior predictive distributions for a simple theoretical example. Consider 100 observations, $y_1, \ldots, y_n$, modeled as independent samples from a $N(\theta, 1)$ distribution with a diffuse prior distribution, say, $p(\theta) = \frac{1}{2A}$ for $\theta \in [-A, A]$ with some extremely large value of A, such as 10^5. We wish to check the model using, as a test statistic, $T(y) = \max_i |y_i|$: is the maximum absolute observed value consistent with the normal model? We choose this test statistic as an illustration of the kind of measure that might be used to identify outliers or extreme points. Consider a data set in which $\overline{y} = 5.1$ and $T(y) = 8.1$. To perform the posterior predictive check, we first determine the posterior distribution of θ, $N(5.1, 1/100)$, and then compute the posterior predictive distribution of $T(y^{\text{rep}})$ by simulation: for each posterior simulation of θ, we draw 100 values $y_i^{\text{rep}} \sim N(\theta, 1)$ and compute the maximum of their absolute values. Figure 10 displays a histogram of 1000 values of $T(y^{\text{rep}})$; 149 of the 1000 values are greater than 8.1, giving an estimated p-value of 15%. Thus the observed $T(y)$ is larger than usual under the model, but not surprisingly so.

In contrast, the prior predictive distribution of $T(y)$, given the diffuse uniform prior distribution on θ, is spread roughly uniformly from 0 to A (with some nonuniformity at the boundaries). For a large value of A, the observed $T(y) = 8.1$ is in the extreme left tail of the distribution, with a p-value close to 1 (or, equivalently, a p-value close to 0 for the test statistic $-T(y)$). The difference between the posterior and prior predictive replications is that the former treats the prior distribution as an outmoded first guess, whereas the latter treats the prior distribution as a true "population distribution." It is commonly useful in Bayesian analysis to use a diffuse prior distribution for parameters in a well-understood model with the understanding that posterior inferences will be reasonable. (For this particular example, the prior and posterior checks could be made more similar by choosing a test statistic that is approximately pivotal, such as $T(y) = \max_i |y_i| - |\overline{y}|$, but our general desire is to be able to check the fit of the model for *any* function of data and parameters, not just pivotal quantities.)

4.2 The impact of prior distributions

As we stated in the previous section, in the posterior predictive framework, the prior distribution for the parameters of the model need not be especially accurate (and it is often not accurate in routine practice), as long as the posterior predictions are "near" the data. This relates to the observation that Bayesian methods based on convenient prior models (normality, uniformity, etc.) can often yield useful inference in practice.

To check the impact of prior distributions, we calculated posterior predictive p-values for the mixture example using one- and two-class models under a variety of prior distributions. Two properties of the prior distribution were varied. The center of each class of the prior distribution was chosen either to match the values suggested by the psychological theory, or to represent a uniform distribution over the levels of each multinomial variable. The strength of the prior information was also varied (by changing the scale of the Dirichlet distributions as measured by the sum of the Dirichlet parameters). As long as the prior distributions are not particularly strong, the size of the p-values and the conclusions reached remained essentially unchanged. This was true for p-values based on both minimum and realized discrepancies.

If the prior distribution is strongly informative, however, it affects the tail-area probabilities of different discrepancies in different ways. Realized discrepancies are naturally quite sensitive to such prior distributions. The posterior predictions obtained under strong incorrect prior specifications may be quite far from the observed data. For example, in Section 3.3, a strong prior distribution specifying two mixture classes, but not corresponding to inhibited and uninhibited children, leads to a tail-area probability of essentially zero. By comparison, minimum discrepancies are much less sensitive to the prior distribution, because the original dataset is judged relative to the best-fitting parameter value rather than to the entire posterior distribution. Conversely, a strong prior distribution, if trusted, can be used to more effectively assess the fit of the likelihood part of the model.

4.3 Choosing discrepancy variables

In the traditional framework of testing the correctness of a null hypothesis, the choices of test statistics are strongly influenced by the consideration of power, which is necessarily governed by the specifications of alternative hypotheses. The reason for preferring a more powerful test is to increase the chance of rejecting a null hypothesis when it is wrong. In the context of checking a whole model (in contrast to testing a few parameters within a model), we know that virtually all models are wrong, and thus a more relevant focus is how the model fits in aspects that are important for our problems at hand. Therefore, the choice of discrepancy variables should reflect our inferential interests, as well as some standard checks on overall fitness (such as those measured by the χ^2 discrepancy). For example, Gelman, Carlin, Stern, and Rubin (1995, §18.3) check the fit of a model used to estimate a population total from a simple random sample; they use the sample total as a test statistic and find poor fits under some reasonable-looking models (such as the log-normal and power-transformed normal). They find that models that poorly fit the sample total yield poor inferences for the population total, although the same models can perform excellently for estimating the population median. We also note that there is no need of adjusting for multiple comparisons when we use more than one discrepancy variable because here we are merely assessing the fitness of a model from various perspectives.

We emphasize that the posterior predictive approach is suitable for assessing the fitness of a *single* model (which could be a super-model that incorporates various alternative models, as in Leamer, 1978, Madigan and Raftery, 1994, and Draper, 1995), to the available data. As we illustrate here and elsewhere (e.g., Gelman, Carlin, Stern, and Rubin, 1995), it is entirely possible to construct sensible discrepancy variables to detect the lack of fit of a single model, in the absence of *explicit* alternative models. We disagree with the opinion that one should never reject a model unless there is an available alternative. A consumer surely can refuse to buy a defective product, even if it is the only item available. If the consumer does decide to purchase it despite its deficiencies, he or she surely still would like to know about its defects and their possible consequences. The posterior predictive assessment provides an (imperfect) method for detecting the "defective products" of applied statistics—invalid models. Whenever we choose to work with a model, we should be aware of its deficiencies, always being aware of the difference between practical significance and statistical significance. We should always report any known defects of our chosen models, so as not to mislead others who might be inclined to use the same models, as they may use the model for purposes for which the model defects may be far more damaging. Indeed, Bayesian inference is a powerful tool for learning about model defects, because we have the ability to examine, as a discrepancy measure, any function of data and parameters.

Acknowledgments

We thank Art Dempster, Donald Rubin, George Tiao, and Michael Stein for comments and criticism, Giuseppe Russo for the mortality rate example, Jerome Kagan, Nancy Snidman, and Doreen Arcus for the infant temperament study data, and the National Science Foundation for grants DMS-9204504, DMS-9404305, DMS-9457824, DMS-9404479, and DMS-9505043. We also thank suggestions from many reviewers that led to a more

concise presentation.

References

Belin, T. R., and Rubin, D. B. (1995). The analysis of repeated-measures data on schizophrenic reaction times using mixture models. *Statistics in Medicine*, to appear.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood (with discussion). *Ann. Stat.* **8**, 457–487.

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

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

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

Chaloner, K. (1991). Bayesian residual analysis in the presence of censoring. *Biometrika* **78**, 637–644.

Chaloner, K., and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75**, 651–659.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Stat.* **25**, 573–578.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Ann. Math. Stat.* **23**, 315–345.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B* **57**, 45–97.

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

Gelman, A. (1990). Topics in image reconstruction for emission tomography. Ph.D. thesis, Dept. of Statistics, Harvard University.

Gelman, A. (1992). Statistical analysis of a medical imaging experiment. Tech. Report #349, Dept. of Statistics, University of California, Berkeley.

Gelman, A. (1995). Bayesian model-building by pure thought: some principles and examples. *Stat. Sinica*, to appear.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* New York: Chapman and Hall.

Gelman, A., Meng. X. L., and Stern, H. S. (1992). Bayesian model checking using tail area probabilities. Tech. Report #355, Dept. of Statistics, University of Chicago.

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

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

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley.

Madigan, D., and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Stat. Assoc.* **89**, 1535–1546.

McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models. *Int. Stat. Rev.* **53**, 61–67.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *J. Amer. Stat. Assoc.* **81**, 104–107.

Meng, X. L. (1994). Posterior predictive p-values. *Ann. Stat.* **22**, 1142–1160.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *J. Educ. Stat.* **6**, 377–400.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

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

Stern, H. S., Arcus, D., Kagan, J., Rubin, D. B., and Snidman, N. (1995). Using mixture models in temperament research. To appear in *Int. J. Behav. Devel.*

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Stat. Assoc.* **52**, 528–550.

Tiao, G. C., and Xu, D. (1993). Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika* **80** 623–641.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

Tsay, R. S. (1992). Model checking via parametric bootstraps in time series analysis. *Appl. Stat.* **41**, 1–15.

Tsui, K. W., and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *J. Amer. Stat. Assoc.* **84**, 602–607.

Upadhyay, S. K., and Smith, A. F. M. (1993). A Bayesian approach to model comparison in reliability via predictive simulation. Tech. Report, Dept. of Mathematics, Imperial College, London.

Wermuth, N., and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. Roy. Stat. Soc. Ser. B* **52**, 21–50.

West, M. (1986). Bayesian model monitoring. *J. Roy. Stat. Soc. Ser. B* **48**, 70–78.

Zellner, A. (1975). Bayesian analysis of regression error terms. *J. Amer. Stat. Assoc.* **70**, 138–144.

Table 1: Mortality rate data from Broffitt (1988)

age, t	number insured, N_t	number of deaths, y_t
35	1771.5	3
36	2126.5	1
37	2743.5	3
38	2766.0	2
39	2463.0	2
40	2368.0	4
41	2310.0	4
42	2306.5	7
43	2059.5	5
44	1917.0	2
45	1931.0	8
46	1746.5	13
47	1580.0	8
48	1580.0	2
49	1467.5	7
50	1516.0	4
51	1371.5	7
52	1343.0	4
53	1304.0	4
54	1232.5	11
55	1204.5	11
56	1113.5	13
57	1048.0	12
58	1155.0	12
59	1018.5	19
60	945.0	12
61	853.0	16
62	750.0	12
63	693.0	6
64	594.0	10

Table 2: Infant temperament data

motor	cry	fear=1	fear=2	fear=3
1	1	5	4	1
1	2	0	1	2
1	3	2	0	2
2	1	15	4	2
2	2	2	3	1
2	3	4	4	2
3	1	3	3	4
3	2	0	2	3
3	3	1	1	7
4	1	2	1	2
4	2	0	1	3
4	3	0	3	3

Table 3: Comparing latent class models for the data of Table 2

Model Description	Degrees of Freedom	$D_{\text{min}}(y)$
Independence (= 1 class)	28	48.761
2 Latent Classes	20	14.150
3 Latent Classes	12	9.109
4 Latent Classes	4	4.718
Saturated	—	—

Figure 1a: The posterior predictive distribution

Figure 1b: The prior predictive distribution

Figure 1c: A mixed predictive distribution

Figure 2: Observed mortality rates and the maximum likelihood estimate of the mortality rate function, under the constraint that it be increasing and convex.

Figure 3: Histogram of 1000 simulations from the reference distribution for $T(y^{\text{rep}})$, the minimum χ^2 statistic for the mortality rates: classical approximation with θ set to the maximum likelihood estimate. The vertical line represents $T(y)$, the observed value of the test statistic.

Figure 4: Nine draws from the posterior distribution of increasing, convex mortality rates, with the maximum likelihood estimate (solid line) as a comparison.

Figure 5: Nine draws from the posterior predictive distribution of mortality rates, corresponding to the nine draws of Figure 4, with the raw data (solid line) as a comparison.

Figure 6: Histogram of 1000 simulations from the reference distribution for $T(y^{\mathrm{rep}})$, the minimum χ^2 statistic for the mortality rates, using the posterior predictive distribution. The vertical line represents $T(y)$, the observed value of the test statistic.

Figure 7: Scatterplot of predictive vs. realized χ^2 discrepancies for the mortality rates, under the joint posterior distribution; the p-value is estimated by the proportion of points above the 45° line.

Figure 8: Histogram of 1000 simulations from the reference distribution for $T(y^{\text{rep}})$, the log likelihood ratio statistic for the latent class example, using the posterior predictive distribution. The vertical line represents $T(y)$, the observed value of the test statistic.

Figure 9: Scatterplot of predictive vs. realized log likelihood ratio discrepancies for the latent class model, under the joint posterior distribution; the p-value is estimated by the proportion of points above the $45°$ line.

Figure 10: Histogram of 1000 simulations from the posterior predictive distribution for $T(y^{\mathrm{rep}})$, the maximum absolute observed value for the hypothetical normal example. The vertical line represents $T(y)$, the observed value of the test statistic.