NONPARAMETRIC GOODNESS-OF-FIT

Tim Swartz

Department of Mathematics and Statistics
Simon Fraser University
Burnaby, BC Canada V5A1S6

Keywords: Monte Carlo; hypothesis testing; Dirichlet process; prior elicitation.

ABSTRACT

This paper develops an approach to testing the adequacy of both classical and Bayesian models given sample data. An important feature of the approach is that we are able to test the practical scientific hypothesis of whether the true underlying model is close to some hypothesized model. The notion of closeness is based on measurement precision and requires the introduction of a metric for which we consider the Kolmogorov distance. The approach is nonparametric in the sense that the model under the alternative hypothesis is a Dirichlet process.

1. INTRODUCTION

Although Bayesian applications have seen unprecedented growth in the last 10 years, there is no consensus on the correct approach to Bayesian model checking. A selection of diverse approaches that address Bayesian model checking includes Guttman (1967), Guttman, Dutter and Freeman (1978), Chaloner and Brant (1988), Weiss (1994), Gelman, Meng and Stern (1996), Verdinelli and Wasserman (1996), Albert and Chib (1997), Evans (1997), Hodges (1998) and Dey, Gelfand, Swartz and Vlachos (1998).

In classical model checking, informal methods are based on the inspection of graphical displays such as residual plots. Formal methods, which often go by the name "goodness-of-fit" rely on a p-value and involve the test of a null hypothesis without the specification of an alternative hypothesis. To many, goodness-of-fit methods are appealing since the space of alternative hypotheses is rarely known. In principle, Bayesian testing cannot mimic the classical goodness-of-fit approach since Bayesian methods require the specification of an alternative hypothesis and the associated prior. D'Agostino and Stephens (1986) is a comprehensive source for classical goodness-of-fit techniques.

The classical goodness-of-fit approach considers null hypotheses such as "H_0 : the underlying model is normal". It is widely accepted that such hypotheses are rarely true, and that given a large enough sample, one will obtain a sufficiently small p-value to reject the null hypothesis. In this case, what most experimenters really want to assess is the actual scientific hypothesis of whether the underlying model is close to normal. Thus there is a major gap between the classical goodness-of-fit approach and what experimenters really want to test.

In this paper we develop a systematic approach to model checking that is in the spirit of classical goodness-of-fit (i.e. avoids formulating a parametric alternative hypothesis) yet addresses the actual scientific hypothesis of interest (i.e. closeness). Our approach is fully Bayesian but is applicable to both classical and Bayesian models. The main tool in our methodology is the Dirichlet process which puts us in the nonparametric Bayesian framework and implicitly assigns an alternative hypothesis. The notion of closeness requires the introduction of a metric for which we consider the Kolmogorov distance. Our approach then is straightforward: Based on sample data, the theory of the Dirichlet process provides the posterior of the true underlying model. Using the Kolmogorov metric, we calculate the posterior distance of the true underlying model from the hypothesized model. Inference is then based on the posterior distribution of the Kolmogorov distance. The methods are highly computational.

The idea of assessing closeness to a null hypothesis has previously been explored by Evans, Gilula and Guttman (1993) in the analysis of Goodman's RC model. It has also been investigated by Evans, Gilula, Guttman and Swartz

(1997) in tests of stochastic order for contingency tables.

Our nonparametric goodness-of-fit approach requires the user-specification of a single parameter m. Although "most Bayesians rely on the subjectivist foundations articulated by De Finetti and Savage" (Kass and Wasserman, 1996), few are willing and able to go through the pains of prior elicitation. In this paper, 2 reasonable questions are asked of the experimenter to elicit the required prior opinion. For a review of the current state of prior elicitation, see Kadane and Wolfson (1998).

Sections 2 through 4 deal with classical models based on univariate sample data. In Section 2, we develop a test for the Bernoulli model which does not require the use of the Dirichlet process. This test is instructional for the more general tests that follow. We also include a discussion of prior selection. Section 3 develops a test for precise hypotheses with arbitrary support and provides further discussion on prior selection. Of particular importance in Section 3 is the reduction of all continuous precise tests to a test of uniformity via the probability integral transformation. Tests of composite hypotheses are considered in Section 4. The natural extension to Bayesian models is presented in Section 5 along with a generalization to the case of multivariate sample data. Some concluding remarks are then given in Section 6.

2. A TEST FOR THE BERNOULLI MODEL

To fix ideas we illustrate the approach in the simplest context before moving on to more general problems. Here we test the adequacy of a hypothesized Bernoulli model. More formally, we test $H_0 : P = F$ where P is the true underlying distribution and F is the hypothesized Bernoulli(θ_0) distribution. The observed sample is $x_1, \ldots, x_n$ where $P(X_i = 1) = \theta$ and $P(X_i = 0) = 1-\theta$ for $i = 1, \ldots, n$. The data are summarized by $y = \sum_{i=1}^{n} x_i \sim$ Binomial(n, θ).

The parameter space is one-dimensional and we assign a Beta$(\alpha(1), \alpha(0))$ prior on θ where $\alpha(0) > 0$ and $\alpha(1) > 0$ are specified. The Beta family is a special case of the Dirichlet family used in Sections 3, 4 and 5. Routine calculations give the posterior distribution

$$\theta | x_1, \ldots, x_n \sim \text{Beta}(y + \alpha(1), n - y + \alpha(0)).$$

The metric $d = |\theta - \theta_0|$ is used to assess the distance between the true underlying distribution and the hypothesized distribution. It follows that the posterior distribution function of the metric d is given by

$$P(d < \epsilon | x_1, \ldots, x_n) = \int_{\max(0,\theta_0-\epsilon)}^{\min(1,\theta_0+\epsilon)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, u^{a-1}(1-u)^{b-1} du \qquad (1)$$

where $a = y + \alpha(1)$, $b = n - y + \alpha(0)$ and $0 < \epsilon < \max(\theta_0, 1-\theta_0)$. The integral in (1) is known as a truncated beta and is readily evaluated.

Inference is based on the posterior distribution of d which is the synthesis of prior information and the observed data concerning d. Note that the posterior distribution displays results over the range of distances $0 < \epsilon < \max(\theta_0, 1-\theta_0)$ and is therefore more informative than goodness-of-fit procedures that rely on a single number summary. Although distribution functions are intrinsic to probability measures, statisticians are generally more experienced and comfortable when viewing probability density functions. For this reason, we plot the posterior density when studying the posterior distribution of d. When appropriate, we also calculate posterior and prior probabilities that $d \leq \epsilon$ for various ϵ.

When specifying the prior parameters $\alpha(1)$ and $\alpha(0)$, we take the position that testing of $H_0 : P = F$ is only done when we have some prior view that P is in the vicinity of F. We therefore take $\alpha(1) = m\theta_0$ and $\alpha(0) = m(1-\theta_0)$ such that $E(P) = F$. This leaves us with only the specification of the prior mass m. From a subjective Bayesian point of view we specify ϵ^* and $0 < q < 1$ such that the subjective prior probability $P(d \leq \epsilon^*) = q$. Letting ϵ^* represent the value of the metric $d = |\theta - \theta_0|$ describing practical equivalence of P to F, it then follows that the equation $P(d \leq \epsilon^*) = .5$ represents "ignorance" concerning the hypothesis of practical equivalence. For example, a pharmaceutical company may only report success rates in round percentages (e.g. values such as 76%, 83%, etc.). In this case, prior indifference concerning practical equivalence of the underlying model to the hypothesized Bernoulli(θ_0) model involves setting $\epsilon^* = .005$. From a robust perspective, the experimenter may wish to elicit a range of probabilities q for a given ϵ^*. Ideas such as these have been used by Swartz (1993) to obtain subjective priors for the Dirichlet process.

Given θ_0, ϵ^* and q, our problem of specifying the prior therefore reduces to

solving for m in the equation

$$\int_{\max(0,\theta_0-\epsilon^*)}^{\min(1,\theta_0+\epsilon^*)} \frac{\Gamma(m)}{\Gamma(m\theta_0)\Gamma(m(1-\theta_0))}\, u^{m\theta_0-1}(1-u)^{m(1-\theta_0)-1}du = q \qquad (2)$$

As the left hand side of (2) is an increasing function of m, the equation is easily solved via bisection. Note that a solution does not exist for sufficiently large values of ϵ^*.

Example 1. Consider the test of whether a coin is fair ($\theta_0 = 1/2$) and suppose that we observe $y = 28$ heads in $n = 40$ flips of the coin. In this example, we are apriori indifferent to the closeness of θ to $\theta_0 = 1/2$ as defined by accuracy in the first digit of θ. We therefore take $q = .5$, $\epsilon^* = .05$ and obtain the prior mass $m = 45.76$. Figure 1 gives the posterior density of d. Whereas the standard two-tailed test based on the normal approximation to the Binomial gives a p-value of .018 and rejects the null hypothesis, Figure 1 is less conclusive. For example, the posterior probability that $d \leq \epsilon^*$ is .202 where $d \leq \epsilon^*$ corresponds to the null hypothesis of practical equivalence (i.e. $.45 \leq \theta \leq .55$). To check the sensitivity of the prior specification m, the prior probability, posterior probability and Bayes factor corresponding to $d \leq \epsilon^*$ are calculated for various values of m and reported in Table I. As expected, the prior and posterior probabilities increase as m increases. However, the Bayes factor defined as the ratio of the prior odds to posterior odds is relatively constant as it falls in the range from 2.0 to 6.0. According to Jeffreys (1961), such values do not provide strong evidence against the null hypothesis.

Table I

Prior, posterior and Bayes factor corresponding to $d \leq \epsilon^*$ in Example 1.

m	Prior	Posterior	Bayes factor
1	.064	.026	2.56
10	.243	.053	5.75
20	.342	.090	5.24
50	.519	.221	3.81
100	.683	.425	2.92
200	.843	.692	2.40
500	.974	.948	2.13

3. TESTS OF PRECISE HYPOTHESES

We now consider the precise hypothesis $H_0 : P = F$ where P is the true underlying distribution and F is some specified continuous distribution. D'Agostino and Stephens (1986) refer to this as the Case 0 situation. We assume that the support for both P and F is $\mathcal{R}$ and that the data consist of a sample $x_1, \ldots, x_n$ from P. Our approach is the same as in Section 2: We introduce a distance measure d between the true underlying distribution P (which is unknown and random) and the hypothesized distribution F. Based on sample data, the posterior distribution of d then provides the basis for determining fit.

Ferguson (1973, 1974) introduced the Dirichlet process as a tool for carrying out nonparametric Bayesian inference. A review of the Dirichlet process is given by Ferguson, Phadia and Tiwari (1992). For testing the precise hypothesis $H_0 : P = F$ let P be a Dirichlet process on $(\mathcal{R}, \mathcal{B})$ with parameter α where $\mathcal{B}$ is the Borel-σ-algebra on $\mathcal{R}$. Thus the Dirichlet process defines the prior distribution on P. Although it is well known that P is discrete with probability 1, this technicality can be overcome. By imposing the weak topology, the closure of the support of the Dirichlet process is extended to the space of all probability measures absolutely continuous with respect to α. This larger space is more in keeping with the spirit of classical goodness-of-fit which does not specify an alternative hypothesis (i.e. any alternative is possible). Therefore we need a measure of discrepancy which metrizes the weak topology. Amongst the possible measures, we choose the Kolmogorov distance as it is computationally simple and is readily interpretable as the maximum difference in cumulative probability between 2 distributions. More precisely, letting F_1 and F_2 be distribution functions, the Kolmogorov distance between F_1 and F_2 is given by

$$d(F_1, F_2) = \sup_{x \in \mathcal{R}} |F_1(x) - F_2(x)|.$$

In testing the Bernoulli model, the Kolmogorov distance between the true underlying $P \sim \text{Bernoulli}(\theta)$ and the hypothesized $F \sim \text{Bernoulli}(\theta_0)$ reduces to $d = |\theta - \theta_0|$ as in Section 2. We note that we have experimented with other metrics such as the Lévy distance and have obtained similar results.

We now state the main result from Ferguson (1973) which is used in developing our methodology: For every $k = 1, 2, \ldots$ and any measurable partition $(A_1, \ldots, A_k)$ of $\mathcal{R}$, the posterior distribution of $P(A_1), \ldots, P(A_k)$ is Dirichlet$(\alpha(A_1) + \sum_1^n I_{A_1}(x_i), \ldots, \alpha(A_k) + \sum_1^n I_{A_k}(x_i))$ where I_Q is the indicator function on the set Q. As in Bernoulli testing, we choose the parameter $\alpha = mF$ such that $E(P) = F$ (Ferguson, 1973).

Unlike expression (1), the posterior distribution function of d can no longer be expressed as a simple one-dimensional integral. Our approach then is Monte Carlo. We generate posterior distributions P_i, from which we calculate $d(P_i, F)$ and build up the posterior distribution of the Kolmogorov distance d.

The algorithm begins with the recognition that we can generate right continuous step functions $\hat{P}$ which approximate the random posterior distribution function P to any required accuracy (in Kolmogorov distance). Perhaps the simplest way of doing this is given by Muliere and Tardella (1998) whose method involves a truncation of the Sethuraman (1994) construction of the Ferguson-Dirichlet distribution. That is, we generate $\alpha_j \sim \text{Beta}(1, m + n)$ and $y_j \sim (mF + I_{\underline{x}})/(m + n)$ independently for $j = 1, \ldots, k$ until $(1 - \alpha_1) \cdots (1 - \alpha_{k-1})$ is less than some prescribed tolerance. The random step function $\hat{P}$ is then given by the finite mixture $\sum_{j=1}^k w_j I_{y_j}$ where $w_1 = \alpha_1$, $w_k = (1-\alpha_1) \cdots (1-\alpha_{k-1})$ and $w_j = (1-\alpha_1) \cdots (1-\alpha_{j-1})\alpha_j$ for $j = 1, \ldots, k-1$. Thus $\hat{P}$ is a finite discrete distribution on the set $\{y_1, \ldots, y_k\}$.

Having generated $\hat{P}$, we calculate the Kolmogorov distance $d = d(\hat{P}, F)$. Letting $y_0 = -\infty$, we calculate $d_i^{(1)} = |\hat{P}(y_{i-1}) - F(y_i)|$ and $d_i^{(2)} = |\hat{P}(y_i) - F(y_i)|$ for $i = 1, \ldots, k$. Then

$$d(\hat{P}, F) = \max\, (d_1^{(1)}, d_1^{(2)}, \ldots, d_k^{(1)}, d_k^{(2)}). \tag{3}$$

Of special interest is the test for uniformity (i.e. $F \sim \text{Uniform}(0, 1)$). Here the support is compact, we define $y_0 = 0$ and note that (3) simplifies via $F(y) = y$. The importance of the test for uniformity stems from the observation that for a given precise continuous hypothesis $H_0 : P_X = F$ with sample data $x_1, \ldots, x_n$ we can make a change of variables $U = F(X)$ via the probability integral transformation. This leads to an equivalent test of $H_0 : P_U = U$ with sample data $u_1, \ldots, u_n$ where $u_i = F(x_i)$, $i = 1, \ldots, n$ and $U \sim \text{Uniform}(0, 1)$. Therefore only a single program is needed for the general

testing of continuous precise hypotheses.

How do we elicit the prior mass m in these general tests of precise hypothesesΓ We suggest that the experimenter consider the initial measurement precision p_0 of the original data $x_1, \ldots, x_n$. For example, if the data are measured in feet and $p_0 = .5$, then we are stating that the x's are rounded to the nearest foot. Alternatively, an experimenter may measure the data in feet to several decimal points but then reason that for practical purposes a measurement of 283.648 feet (for example) is essentially the same as a measurement of 284 feet. In this case we would also set $p_0 = .5$. Having specified p_0, we then investigate the maximum effect of the precision p_0 on the uniform scale. Mathematically, we calculate

$$p^* = \max_{x \in \mathcal{R}} \{F(x + p_0) - F(x)\}. \tag{4}$$

It is clear that p^* satisfies a desirable location-scale invariance based on the initial measurement scale. For example, we would obtain the same value of p^* having measured the x's in feet with $p_0 = .5$ or having measured the x's in yards with with $p_0 = 3(.5) = 1.5$. Now the transformed precision p^* has the maximum effect of shifting the line $y = x$ (corresponding to the Uniform$(0, 1)$ distribution function) a "practically equivalent" horizontal distance p^*. This, in turn, defines the Kolmogorov distance $\epsilon^* = p^*$ which we view as practical equivalence. Therefore, the suggested procedure results in posterior inferences regarding the Kolmogorov distance d that are invariant to location-scale transformations of the data $x_1, \ldots, x_n$.

In summary, prior elicitation is straightforward as the experimenter is required to answer the following 2 questions:

> (A) What is the measurement precision p_0 of the data $x_1, \ldots, x_n$ that I care aboutΓ In other words, what is the maximum value p_0 such that a measurement $x \pm p_0$ could be considered practically equivalent to xΓ
>
> (B) What is my prior belief q that the true underlying distribution P is practically equivalent (in the sense of (A)) to the specified distribution FΓ

Using (4) to obtain p^* and letting $\epsilon^* = p^*$, the prior mass m is then obtained iteratively. To carry out the iteration, begin with an initial value $m = m_0$ and generate N random d's from the prior distribution. Estimate $P(d \leq \epsilon^*)$ by the proportion of the random d's that are less than or equal to ϵ^*. If this estimate is smaller (larger) than q, increase (decrease) m and repeat the procedure. Terminate the algorithm when the estimate is within a certain number of standard errors of q. Naturally, accuracy will increase as N is increased.

Now, Ferguson (1974) describes $m = m_{p_0,q}$ as the prior sample size. This interpretation is immediate from the Sethuraman (1994) construction of the Ferguson-Dirichlet distribution. Therefore, as a check on prior elicitation, one may consider the ratio n/m. For smaller ratios, posterior inferences are not as sensitive to the data as more of the $\hat{P}$-sampling is from F. In these cases, one may consider decreasing q to increase the ratio $n/m_{p_0,q}$.

From a theoretical perspective, it is clear from the Sethuraman (1994) construction of the Ferguson-Dirichlet distribution that as the sample size increases (i.e. $n \to \infty$), the Ferguson-Dirichlet distribution samples from the ecdf (empirical cumulative distribution function). Therefore, in large samples, the Kolmogorov distance d is insensitive to the prior specification m. Moreover, since the ecdf converges in distribution to the true underlying distribution P, one can establish the consistency of the Kolmogorov metric.

In the case of precise discrete hypotheses $H_0 : P = F$, the theory is exactly the same as in the continuous case except that there is no probability integral transformation to uniformity. Here, the simulations and distance calculations are done on the F scale.

Example 2. We consider a common situation where a statistical procedure (e.g. regression) gives rise to residuals that are checked against the standard normal distribution. With large samples, it is often the case that even though the residuals appear satisfactory, formal statistical tests reject the null hypothesis of normality. We fit the simple autoregressive model $y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$ where $y_1, \ldots, y_{1150}$ are heights measured at 1 micron intervals along the drum of a roller and the ϵ_t are independent Normal$(0, \sigma^2)$ errors. The dataset was studied in Laslett (1994) and is available from the *jasadata* section of Statlib (http://lib.stat.cmu.edu/). Using the standardized residuals from the fitted model the corresponding qq plot is given in Figure 2. Although the residu-

als clearly show departures from normality, for some practitioners, apart from outliers in the left tail, the residual plot would appear adequate. Yet, in this example, standard goodness-of-fit procedures such as the Anderson-Darling test and the Cramer-von Mises test emphatically reject the hypothesis of normality with p-values near zero. Using our methodology, we consider measurement precisions of $p_0 = .05$ and $p_0 = .01$ where the former corresponds to accuracy in the standardized residuals to the first decimal place. Using (4), these values translate to $\epsilon^* = p^* = .02$ and $\epsilon^* = p^* = .04$ respectively as meaningful distances on the uniform scale. To investigate prior sensitivity, Table II gives posterior probabilities that $d \leq \epsilon^*$ for a wide range of values of the prior mass m. We see that the posterior probabilities are fairly robust with respect to the prior specification and that in the case of $\epsilon^* = .04$, there is no reason to reject the hypothesis of approximate normality.

Table II
Posterior probability that $d \leq \epsilon^*$ in Example 2.

m	$\epsilon^* = .02$	$\epsilon^* = .04$
1	0.00	0.76
50	0.00	0.83
100	0.00	0.88

4. TESTS OF COMPOSITE HYPOTHESES

We now consider the composite hypothesis $H_0 : P = F_\theta$ for some $\theta \in \Omega$ where P is the true underlying distribution and F_θ is a member of a family of distributions indexed by the parameter $\theta \in \Omega$. Our setup and approach is the same as in the general case presented in Section 3 with the addition of a prior distribution $\pi(\theta)$ on θ.

For composite hypotheses, we observe that there is no transformation of the data which leads to a test of uniformity since such a transformation would depend on the unknown parameter θ. This means that a special program needs to be written for every hypothesized family of distributions. Fortunately, only modules of a standard program need to be modified. The situation is the same,

if not more daunting, in the classical goodness-of-fit framework (see Chapter 4 of D'Agostino and Stephens (1986)).

Given $\alpha_\theta = mF_\theta$, a hyper-prior $\pi(\theta)$ must be chosen to complete the prior specification. We continue to advocate a subjective Bayesian approach and attempt to elicit priors from the experimenter. Standard default priors can also be considered although many of these are improper. The elicitation of the prior mass m is again guided by the notions of measurement precision and practical equivalence between distributions. If p_0 is the measurement precision of the x's, then we recommend setting

$$\epsilon^* = \max_{x \in \mathcal{R}} \{F_{\mathrm{E}(\theta)}(x + p_0) - F_{\mathrm{E}(\theta)}(x)\} \tag{5}$$

as this represents the Kolmogorov distance between the hypothesized distribution and a practically equivalent distribution evaluated at the expected value of θ.

As before, we generate right continuous step functions $\hat{P}$ which approximate a random posterior distribution function to any required accuracy (in Kolmogorov distance). However, there are now two steps involved in generating $\hat{P}$. We must first generate θ_0 from the distribution of $\theta \mid \underline{x}$ and then generate $\hat{P}$ from the distribution of $P \mid \theta_0, \underline{x}$. Whereas the latter distribution is a Dirichlet process, the density corresponding to the distribution of $\theta \mid \underline{x}$ is

$$[\theta \mid \underline{x}] \;\propto\; (\prod_{i=1}^{n} f_\theta(x_i))\; \pi(\theta) \tag{6}$$

where f_θ is the density corresponding to F_θ (see the Appendix). Sampling from the non-standard distribution given by (6) may require specialized techniques such as rejection sampling (Devroye (1986)), adaptive rejection sampling (Gilks and Wild (1992)) or Metropolis-Hastings (Tierney (1994)).

Having generated $(\theta_0, \hat{P})$, we are no longer able to calculate the distance metric $d(\hat{P}, F)$ since $F = F_\theta$ depends on the unknown parameter θ. Instead for composite hypotheses, we calculate

$$d^* = d(\hat{P}, F_{\theta_0}) \tag{7}$$

which has intuitive appeal as a diagnostic for fit. It measures the distance between a posterior realization of the model and the hypothesized model evaluated at the same realization of θ.

Substituting the generated θ_0 in (7) is similar to the calculation of the discrepancy variable used in obtaining posterior predictive p-values as discussed in Gelman, Meng and Stern (1996). In their approach, the discrepancy variable $D(x;\theta)$ is also a function of both the data and the parameter. Given the observed data x_{obs}, the parameter θ_j is first generated from the posterior distribution, and given its value, data x^{rep} is drawn from its conditional distribution. The procedure is repeated to build up the reference distribution of the pairs $(D(x_{\text{obs}};\theta_j), D(x^{\text{rep}};\theta_j))$.

Example 3. We consider a composite test of exponentiality. Using our notation, we test $H_0 : P = F_\theta$ where F_θ is the exponential distribution with mean $\theta > 0$. The data consist of a sample of size $n = 40$ generated from the Chi-squared(5) distribution and are presented in Table III. We stipulate a precision of $p_0 = .25$ which corresponds to meaningful measurements in the upper or lower half of the first decimal point. In order to generate from the prior distribution we require that $\pi(\theta)$ be proper, and for this we choose $\theta \sim \text{Normal}(5,1)$ truncated on the left at zero. Letting $q = .2$ represent our prior belief that the true underlying distribution is exponential, we obtain $\epsilon^* = .049$ via (5) and the prior mass $m = 159$ by solving $P(d^* \leq \epsilon^*) = q$ iteratively. Sampling from the distribution of $\theta \mid \underline{x}$ is carried out via the Metropolis-Hastings algorithm using an independence chain with $\pi(\theta)$ as the proposal density. In more detail, our implementation for generating d^* from its posterior distribution involves first generating $\theta^{(0)} \sim \pi(\theta)$. We then generate $u_i \sim \text{Uniform}(0,1)$, $\theta^{(i)} \sim \pi(\theta)$ and set $\theta^{(i)} = \theta^{(i-1)}$ if

$$u_i \; > \; \frac{[\theta^{(i)} \mid \underline{x}]\pi(\theta^{(i-1)})}{[\theta^{(i-1)} \mid \underline{x}]\pi(\theta^{(i)})} \; = \; (\theta^{(i-1)}/\theta^{(i)})^n \exp\{-\sum_{i=1}^{n} x_i(1/\theta^{(i)} - 1/\theta^{(i-1)})\}$$

for $i = 1, \ldots, 1000$. The final variate $\theta_0 = \theta^{(1000)}$ is taken as a realization from the distribution of $\theta \mid \underline{x}$ from which we generate $\hat{P}$ from the distribution of $P \mid \theta_0, \underline{x}$ and calculate $d^* = d(\hat{P}, F_{\theta_0})$. By using a different Metropolis-Hastings chain for each d^* as described here, we have ensured independence of the variates θ_0. Convergence of the individual chains is suggested by standard procedures such as the use of the Gelman and Rubin diagnostic as described in Gelman (1996). A kernel density estimate of the posterior of d^* based on 2000 Monte Carlo simulations is plotted in Figure 3. The kernel density estimate

was obtained using the Splus function "density" with the width parameter set equal to .029. Despite the strong prior, the graph rightly provides some evidence against the composite null hypothesis of exponentiality. Here the posterior probability that $d \leq \epsilon^*$ is .123 which yields the Bayes factor 1.78. For comparison purposes, consider a less informative prior based on $q = .075$ (ie. $m = 110$). Here the posterior probability that $d \leq \epsilon^*$ is .039 which gives the Bayes factor 2.00. The relative stability of the Bayes factor indicates a lack of sensitivity to the prior in this example.

Table III

The data from Example 3 presented in increasing order across rows.

0.277	1.054	1.138	1.946	1.953
2.227	2.293	2.598	2.937	3.000
3.296	3.385	3.501	3.535	3.615
3.616	3.827	4.386	4.399	4.405
4.585	4.779	4.984	5.317	5.331
5.637	6.570	6.808	7.283	7.306
7.413	7.508	8.288	8.638	9.691
10.951	12.017	13.467	17.271	17.477

We remark that we have experimented with diagnostics other than (7) in the context of testing composite hypotheses. For example, we have implemented the diagnostic

$$d_{\text{inf}} = \inf_{\{\theta \in \Omega\}} d(P, F_\theta) \tag{8}$$

as a measure of fit for exponentiality. Note that $d_{\text{inf}} \leq d^*$. The difficulty with the general use of d_{inf} involves the optimization in (8). Typically, the difficulty of the optimization increases as the dimensionality of θ increases.

5. TESTS OF BAYESIAN MODELS

Up until this point we have investigated the adequacy of classical models using Bayesian methods. These methods may serve as a useful screening device as often an experimenter may want to check the underlying distribution of

data (e.g. normality) before proceeding to more specialized procedures (e.g. ANOVA) that depend on the underlying distribution. With classical models we specify a prior mass m, and we also specify a prior distribution $\pi(\theta)$ if the hypothesized distribution is composite. In this context we may think of m and $\pi(\theta)$ as model expansion parameters which allow us to judge departures from the hypothesized model.

The situation is more natural in the case of Bayesian models. Suppose that we have a sample $x_1, \ldots, x_n$ from a hypothesized model F_θ with a proper prior distribution $\pi(\theta)$. Then the methodology follows exactly as before where we need only specify the prior mass m. Here we avoid placing a prior probability mass on a null model which is widely considered one of the more distasteful aspects of Bayesian hypothesis testing. Note also, that in the case of hierarchical models, there is no additional difficulty. For example, in a two-stage hierarchical model, we simply write $\pi(\theta) = \pi(\theta_1 \mid \theta_2)\pi(\theta_2)$ to complete the prior specification.

Example 4. To highlight the practicality of the methodology for Bayesian models, we address a question that was posed by Seymour Geisser in the model checking session (Session 6) of the 1996 Joint Statistical Meetings held in Chicago. He asked, "Given a sample $x_1, \ldots, x_n$, how can I test the adequacy of the Binomial(N, θ) model with a given prior for θ?" We let $N = 10$, $n = 50$ and consider the prior $\theta \sim \text{Beta}(12, 12)$ such that the prior standard deviation of θ is .10. We simulate data $x_1, \ldots, x_{50}$ from a Poisson(2) distribution. The data appear in Table IV with $T = \sum x_i = 103$. For large N and small θ, the relationship between the Binomial and Poisson distributions is well known. We naturally choose $p_0 = .5$ so as not to alter the value of integer data and we let $q = 1/3$ which assigns prior probability $1/3$ to the binomial model. From (6), we generate θ_0 according to $\theta \mid \underline{x} \sim \text{Beta}(T + 12, 500 - T + 12)$ and we then generate $\hat{P}$ according to the distribution of $P \mid \theta_0, \underline{x}$ in the standard way. We obtain $\epsilon^* = \binom{10}{5}(1/2)^{10} = .246$ via (5) and $m = 25.5$. We note that the posterior probability that $d \leq \epsilon^*$ is .39, a slight increase from the prior probability $q = 1/3$. Here, the affirmation of practical equivalence between the underlying distribution and the binomial distribution is sensible as the Kolmogorov distance between a Binomial$(10, .2)$ distribution and a Poisson(2) distribution is $.032 < \epsilon^*$.

Table IV
The data from Example 4.

Outcome	0	1	2	3	4	5	6
Frequency	7	12	14	9	5	2	1

There is a generalization of the methodology which applies equally well to both classical and Bayesian models. Suppose that the sample $\underline{x} = (x_1, \ldots, x_n)$ is multivariate of dimension r. In principle, there is no need to change the approach. We generate a variate θ_0 from the distribution of $\theta \mid \underline{x}$, we generate $\hat{P}$ from the distribution of $P \mid \theta_0, \underline{x}$ and then calculate $d^* = d(\hat{P}, F_{\theta_0})$. However, whereas the univariate calculation of the Kolmogorov distance in (3) involves the maximization of $2k$ distances, the multivariate calculation ($r > 1$) involves the maximization of up to $2[k + \binom{k}{2}]$ distances where we recall that k is the number of components in the randomly generated step function $\hat{P}$.

6. CONCLUSIONS

In this paper we have developed a theory of goodness-of-fit that allows an experimenter to test the practical scientific hypothesis of whether an underlying distribution is close to some hypothesized distribution. The methodology is useful for testing the adequacy of both classical and Bayesian models and is applicable when we have sample data and proper priors. Unlike many of the recent hybrid techniques that are based upon a synthesis of ideas involving posterior distributions and p-values, our approach is fully Bayesian. We begin with a prior opinion concerning distance between the true and hypothesized model, and via the data, the belief is updated and expressed by the posterior distribution. Moreover, the approach is systematic in the sense that the same steps are followed whether we are testing precise or composite hypotheses and whether the data is univariate or multivariate. This is in sharp contrast to the multitude of goodness-of-fit tests in current statistical practice.

The difficulty with the approach is also its strength. One cannot blindly use the methods as a black box procedure. Rather, the experimenter must be able to sit down and carefully think about meaningful measurement precision.

Clearly, if you want to be able to test closeness, you must be able to define it. Our elicitation procedure is a practical means of achieving this end.

We take the view that testing model adequacy is a difficult problem. There are many ways in which a distribution can depart from a hypothesized model and not every diagnostic will catch every departure. We therefore consider our approach as only one of several that might be part of the toolkit of diagnostics used to check model adequacy. Of particular importance, we have developed a single algorithm for testing the fit of an underlying distribution to any precise continuous hypothesis. Fortran code for this algorithm and for the other examples described in this paper are available from the author upon request.

APPENDIX

We indicate the form of the distribution of $\theta \mid \underline{x}$ as given in (6). The density is given by

$$\begin{aligned} [\theta \mid \underline{x}] &= \int [\theta, P \mid \underline{x}]\, dP \\ &\propto \int [\underline{x} \mid \theta, P]\, [P \mid \theta]\, \pi(\theta)\, dP \\ &= \int [\underline{x} \mid P]\, [P \mid \theta]\, \pi(\theta)\, dP. \end{aligned}$$

We sample P according to the Muliere/Tardella (1998) truncation of the Sethuraman construction. We therefore let $P = (P(A_1), \ldots, P(A_{k+1}))$ where $A_i = (y_{i-1}, y_i]$ with $y_0 = -\infty < y_1 < \cdots < y_{k+1} = \infty$. It follows that $[\underline{x} \mid P]$ is a multinomial density with parameter $P(A_i)$ raised to the power $\sum_{j=1}^n I_{A_i}(x_j)$ and that $[P \mid \theta]$ is a Dirichlet density with $P(A_i)$ raised to the power $\alpha_\theta(A_i) - 1$, $i = 1, \ldots, k+1$. Collecting exponents and integrating, we obtain

$$[\theta \mid \underline{x}] \;\propto\; \frac{\Gamma(\alpha_\theta(A_1) + \sum_{i=1}^n I_{A_1}(x_i)) \cdots \Gamma(\alpha_\theta(A_{k-1}) + \sum_{i=1}^n I_{A_{k-1}}(x_i))}{\Gamma(\alpha_\theta(A_1)) \cdots \Gamma(\alpha_\theta(A_{k-1}))}\, \pi(\theta).$$

Assuming that none of the data values are equal, we let $k \to \infty$ and note that we have either 0 or 1 observations lying in each of the intervals $A_1, \ldots, A_{k+1}$. Since $\Gamma(x+1) = x\Gamma(x)$ and $\alpha_\theta(A_i) = m(F_\theta(y_i) - F_\theta(y_{i-1}))$ we obtain the limiting density

$$[\theta \mid \underline{x}] \;\propto\; (\prod_{i=1}^n f_\theta(x_i))\, \pi(\theta)$$

where f_θ is the density corresponding to F_θ.

ACKNOWLEDGEMENTS

This work was initiated during a sabbatical visit in 1995 to the Institute of Statistics and Decision Sciences (ISDS) at Duke University. The author thanks the ISDS for its hospitality. The author also thanks Michael Lavine, Xiao-Li Meng, an associate editor and two referees for helpful comments. Partial support was provided by a grant from the Natural Sciences and Engineering Research Council of Canada.

BIBLIOGRAPHY

Albert, J.H. and Chib, S. (1997), "Bayesian tests and model diagnostics in conditionally independent hierarchical models", *Journal of the American Statistical Association*, 92, 916-925.

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

D'Agostino, R.B. and Stephens, M.A. (1986), *Goodness-of-Fit Techniques*, Marcel Dekker.

Devroye, L. (1986), *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

Dey, D.K., Gelfand, A.E., Swartz, T.B. and Vlachos, P.K. (1998), "Simulation based model checking for hierarchical models", *Test*, 7, 325-346.

Evans, M. (1997), "Bayesian hypothesis testing procedures derived via the concept of surprise", *Communications in Statistics - Theory and Methods*, 26, 1125-1143.

Evans, M., Gilula, Z. and Guttman, I. (1993), "Computational issues in the Bayesian analysis of categorical data: log-linear and Goodman's RC model", *Statistica Sinica*, 3, 391-406.

Evans, M., Gilula, Z., Guttman, I. and Swartz, T.B. (1997), "Bayesian analysis of stochastically ordered distributions of categorical variables", *Journal of the American Statistical Association*, 92, 208-214.

Ferguson, T.S. (1973), "A Bayesian analysis of some nonparametric problems", *Annals of Statistics*, 1, 209-230.

Ferguson, T.S. (1974), "Prior distributions on spaces of probability measures", *Annals of Statistics*, 2, 615-629.

Ferguson, T.S., Phadia, E.G. and Tiwari, R.C. (1992), "Bayesian nonparametric inference", in *IMS Lecture Notes - Monograph Series Volume 17*, editors M. Ghosh and P.K. Pathak.

Gelman, A. (1996), "Inference and monitoring convergence", in *Markov Chain Monte Carlo in Practice*, editors W.R. Gilks, S. Richardson and D.J. Spiegelhalter, Chapman and Hall, 131-143.

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

Gilks, W.R. and Wild, P. (1992), "Adaptive rejection sampling for Gibbs sampling", *Applied Statistics*, 41, 337-348.

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.

Guttman, I., Dutter, R. and Freeman, P.R. (1978), "Care and handling of multivariate outliers in the general linear model to detect spuriousity - a Bayesian approach", *Technometrics, 20, 187-193.*

Hodges, J. (1998), "Some algebra and geometry for hierarchical models applied to diagnostics", *Journal of the Royal Statistical Society, Series B*, 60, 497-536.

Jeffreys, H. (1961), *Theory of Probability (3rd ed.)*, Oxford: Oxford University Press.

Kadane, J.B. and Wolfson, L.J. (1998), "Experiences with elicitation", *Journal of the Royal Statistical Society, Series D*, 47, 3-19.

Kass, R.E. and Wasserman, L. (1996), "The selection of prior distributions by formal rules", *Journal of the American Statistical Association*, 91, 1343-1370.

Laslett, G.M. (1994), "Kriging and splines: an empirical comparison of their predictive performance in some applications (with discussion)", *Journal of the American Statistical Association*, 89, 391-409.

Muliere, P. and Tardella, L. (1998), "Approximating distributions of random functionals of Ferguson-Dirichlet priors", *Canadian Journal of Statistics*, 26, 283-297.

Sethuraman, J. (1994), "A constructive definition of the Dirichlet prior", *Statistica Sinica*, 2, 639-650.

Swartz, T.B. (1993), "Subjective priors for the Dirichlet process", *Communications in Statistics - Theory and Methods*, 22, 2999-3011.

Tierney, L. (1994), "Markov chains for exploring posterior distributions (with discussion)", *Annals of Statistics*, 22, 1701-1762.

Verdinelli, I. and Wasserman, L. (1996), "Bayesian goodness of fit using infinite dimensional exponential families", Manuscript.

Weiss, R.E. (1994), "Bayesian model checking with applications to hierarchical models", Manuscript.

FIG.1: The posterior density of the distance d in Example 1.

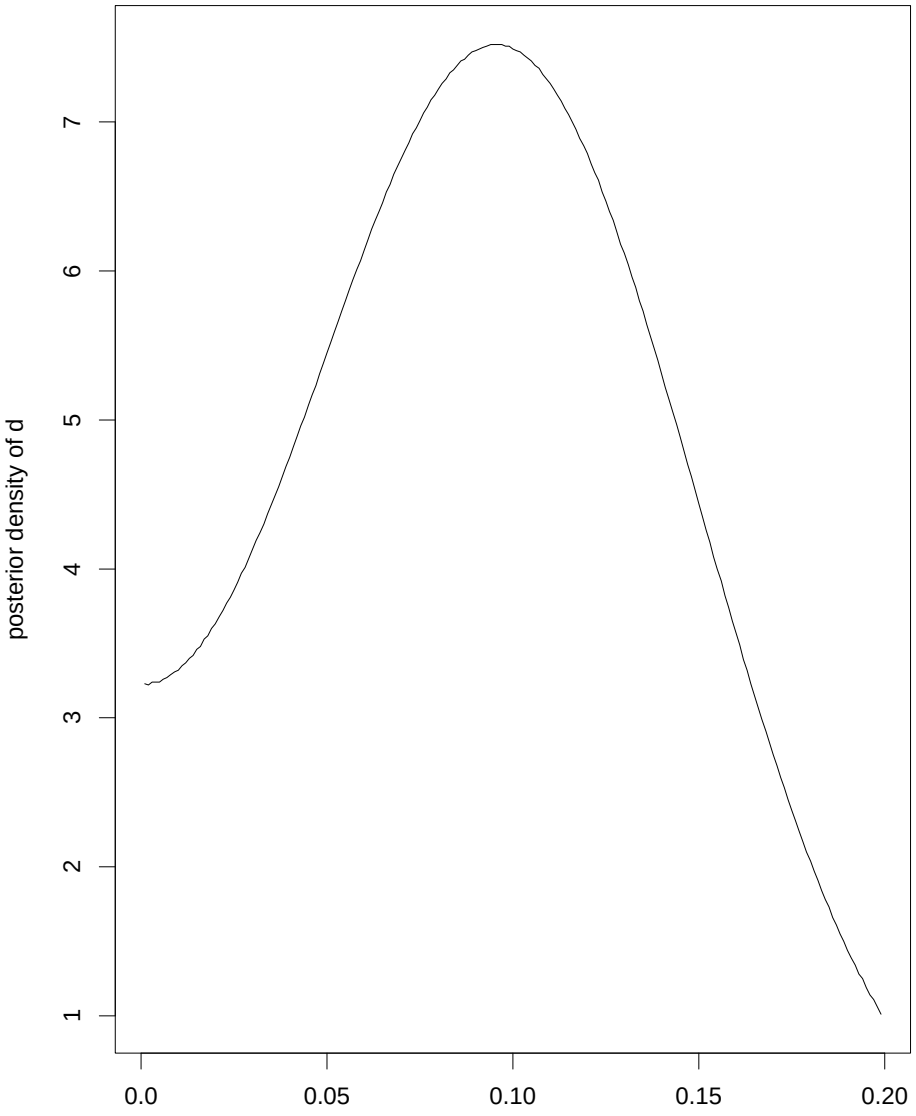

FIG. 2: The qq plot from the model in Example 2.

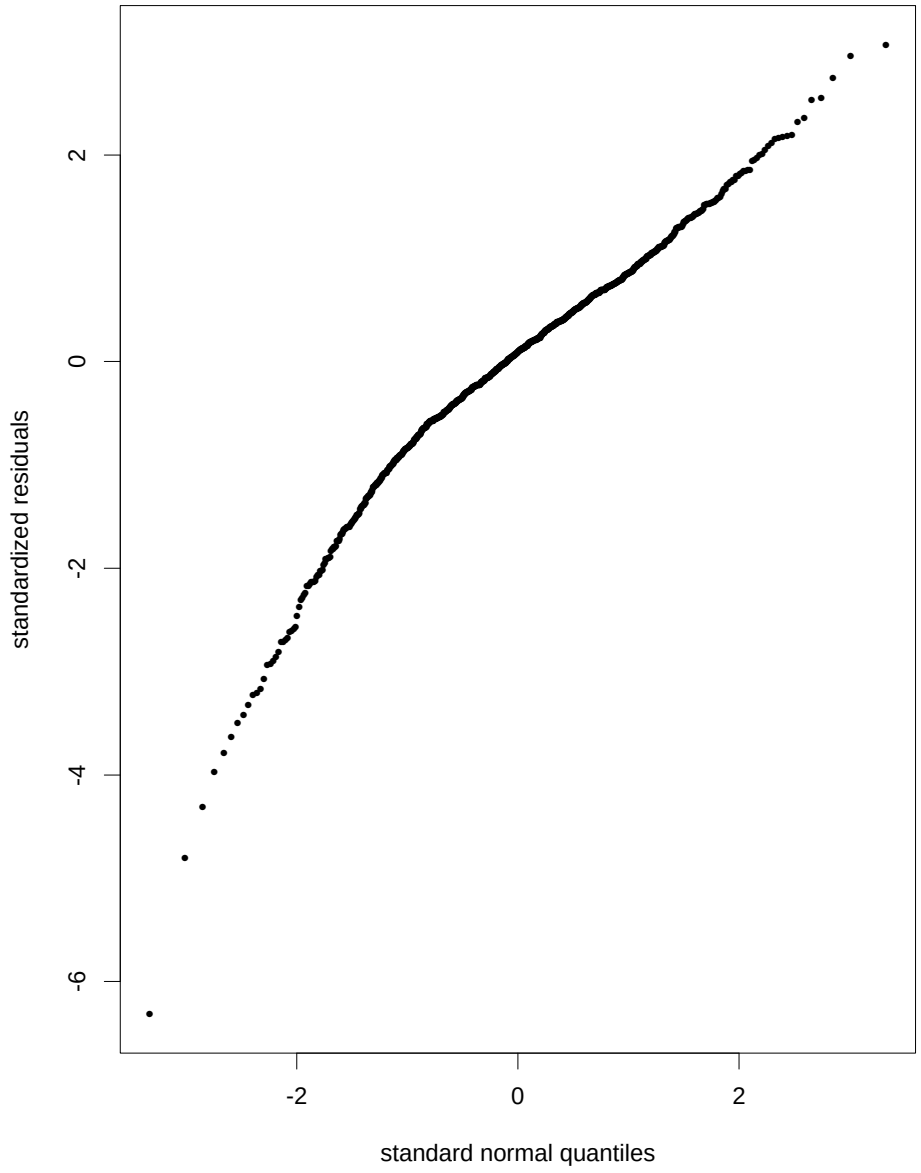

FIG. 3: A kernel density estimate of the posterior distance d^* in Example 3.