MODEL DETERMINATION USING PREDICTIVE DISTRIBUTIONS WITH IMPLEMENTATION VIA SAMPLING–BASED METHODS

by

Alan E. Gelfand

Dipak K. Dey

Hong Chang

TECHNICAL REPORT No. 462

DECEMBER 4, 1992

Prepared Under Contract

N00014-92-J-1264 ((NR-042-267)

FOR THE OFFICE OF NAVAL RESEARCH

Professor Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA 94305-4065

MODEL DETERMINATION USING PREDICTIVE DISTRIBUTIONS WITH IMPLEMENTATION VIA SAMPLING–BASED METHODS

by

Alan E. Gelfand

Dipak K. Dey

Hong Chang

TECHNICAL REPORT No. 462

DECEMBER 4, 1992

Prepared Under Contract

N00014-92-J-1264 ((NR-042-267)

FOR THE OFFICE OF NAVAL RESEARCH

Professor Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA 94305-4065

Model Determination Using Predictive Distributions

With Implementation Via Sampling–Based Methods

Alan E. Gelfand, Dipak K. Dey and Hong Chang
University of Connecticut

Summary

Model determination is divided into the issues of model adequacy and model selection. Predictive distributions are used to address both issues. This seems natural since, typically, prediction is a primary purpose for the chosen model. A cross–validation viewpoint is argued for. In particular, for a given model, it is proposed to validate conditional predictive distributions arising from single point deletion against observed responses. Sampling based methods are used to carry out required calculations. An example investigates the adequacy of and rather subtle choice between two sigmoidal growth models of the same dimension.

Key Words: Model adequacy, model choice, predictive distributions, cross–validation, sampling based methods, sigmoidal growth model, logistic growth curve model, Gompertz model.

1. Introduction

Responsible data analysis must address the issue of model determination which divides into two components: model assessment or checking and model choice or selection. That is, apart from rare situations, the model specification is never "correct". Rather the questions are (i) is a given model adequate? (ii) within a collection of models under consideration, which model is the best choice? The questions are distinct. We need not envision a collection of models to address (i). Conversely, amongst a collection of models several may be adequate (or perhaps all may be inadequate) but we still seek the "best" one. Nonetheless in practice the questions are typically investigated concurrently. This paper adopts a predictive viewpoint for answering them with resultant overlapping methodology.

The literature on model assessment and model choice is, by now, enormous. Of course, our modeling framework here is Bayesian whence our approach to these problems is as well. We restrict ourselves to parametric models where the amount of data is fixed and which are expressible in the form Likelihood × Prior where the Likelihood is an explicit readily evaluable function of both the data and the parameters with the prior a readily evaluable function of the parameters.

Model adequacy is considered with regard to the form of the product. Box (1980) is persuasive on behalf of the predictive stance in arguing that the adequacy of a model can not be assessed from the posterior distribution of the model parameters. Berger (1985) observes that "Bayesians have long used [predictive distributions] to check assumptions". A few additional references are Jeffreys (1961), Box and Tiao (1973), Dempster (1975), Rubin (1984) and Geisser (1985).

Turning to model selection we need to clarify our objective. For a fixed likelihood the prior may be varied. This case is typically not viewed as one of model selection but rather one of model robustness (see Berger (1984) for a review). More recent work is summarized in Gelfand and Dey (1991). In this work the estimative side using the resultant posterior is usually investigated with regard to sensitivity to such variation.

Rather, the usual business of model selection is the specification of the likelihood or joint distribution of the data. For instance, in the case of response models one needs to specify the parametric form of the error distribution (equivalently, a transformation of the response variable), a parametric form for the mean and a parametric form for the variance. (See in this regard the recent work of Carlin and Polson (1991) and George and McCulloch (1991). There is an implicit trade–off here (Smith, 1986) e.g. complex form for the mean with pure Gaussian error versus simple specification of the mean with say a mixture of Gaussians as error. Hence judicious choice of the collection of models entertained is essential.

Our discussion is concerned with model choice in this broad sense emphasizing that predictive distributions enable arbitrary model comparisons. Moreover the predictive viewpoint is typically consonant with the intended use of the model. In the special case where models under consideration are nested so that a subset of components of the parameter vector may be viewed as a discrepancy parameter measuring departure from a baseline model (Box 1980) model choice reduces to posterior inference for this subset of parameters. In the absence of nesting, comparison of posteriors between models typically conveys little information. The models need not share the same dimensionality. Moreover, even if they do, from one to another, the parameters will have different interpretations.

Model determination is closely related to other data analytic issues such as residual analysis, influence measures and outlier detection. We offer no elaboration here but note the recent papers of Pettit and Smith (1985), Geisser (1987), Chaloner and Brant (1988), Verdinelli and Wasserman (1991) and Guttman (1991).

The entire proposed enterprise rests upon our ability to obtain desired predictive distributions and to calculate expectations under these distributions. Analytic evaluation of the required integrals is generally hopeless with effective approximations only available for simpler cases. To accomplish needed integrations we propose the use of sampling–based methods as discussed in Rubin (1988) and in Gelfand and Smith (1990). Such calculation is very computer–intensive, particularly when undertaking more complex and realistic

modeling. However, the routine availability of enormous computing power makes this no obstacle. In particular this implies that many contending models may eventually be considered. As Geisser (1988) notes, "this could create havoc.... for those who adhere to stringent versions of.... Bayesian approaches." However we feel that the *art* of data analysis should not be bound by overly formal inferential frameworks.

This paper is essentially a synthesis and extension of earlier work in Bayesian model determination. Our main contributions are the fleshing out of the cross–validation approach and the presentation of straightforward computing procedures to implement such cross–validation. In Section 2 we detail the proposed predictive approaches for model determination. In Section 3 we describe the Monte Carlo techniques for performing the required calculations. An illustrative example is discussed in Section 4 and we offer brief conclusions in Section 5.

2. Predictive Approaches for Model Determination

2.1 Predictive densities

Box (1980) discusses the complementary roles of the predictive and posterior distributions in Bayesian data analysis noting that the posterior distribution provides a basis for "*estimation* of parameters conditional on the adequacy of the entertained model" while the predictive distribution enables "*criticism* of the entertained model in the light of current data." We concur noting further that in comparing models predictive distributions are directly comparable while posteriors are not. The predictive distribution (or marginal likelihood) is the joint marginal distribution of the data. This distribution may be used in various ways to examine model determination questions. Our approach, which is argued for below, is a cross–validation one and leads to the examination of a collection of conditional distributions arising from this joint distribution.

We use customary notation in letting upper case letters denote random variables and lower case letters denote the observed realizations of these variables in our sample. In particular, the observed value of the r^{th} response, Y_r, in our sample is y_r. Let $\mathbf{Y}$ denote the

$n\times 1$ data vector, and let $Y_{(r)}$ denote the $n-1 \times 1$ data vector with Y_r deleted. Let $X_{n\times p}$ denote the matrix of explanatory variables whose r^{th} row, X_r is associated with y_r. Let $X_{(r)}$ denote the matrix which is X with the r^{th} row deleted. All densities for the data will be denoted by f, all densities for the parameters will be denoted by π with arguments providing clarification. More precisely for a given model let θ denote the vector of model parameters. Then $f(Y|\theta,X)$ is the joint density of the data given θ and $\pi(\theta)$ is the prior density for θ whence $f(Y|\theta,X) \cdot \pi(\theta)$ is the model specification. Both Y and θ are presumed continuous.

Assuming that the integral exists the predictive density is $f(Y) = \int f(Y|\theta,X)\,\pi(\theta)d\theta$. In model selection, interest focuses on $f(Y|\theta,X)$ and π is often chosen to be vague. But if π is improper then f(Y) necessarily is, making it awkward to use in model checking. Note however, that

$$f(Y_r|Y_{(r)}) = \frac{f(Y)}{f(Y_{(r)})} = \int f(Y_r|\theta,Y_{(r)},X)\,\pi(\theta|Y_{(r)})d\theta \qquad (1)$$

is proper since $\pi(\theta|Y_{(r)})$ is. The density $f(Y_r|\theta,Y_{(r)},X)$ is immediate if the Y_r are conditionally independent given θ as well as, for example, if $f(Y|\theta,X)$ is multivariate normal.

Suppose that f(Y) is proper and strictly positive over its domain. Then f(Y) is equivalent to the set $\{f(Y_r|Y_{(r)}) : r=1,\cdots,n\}$ in the sense that each uniquely determines the other (Besag, 1974). Hence, in terms of model assessment, examining the observed y with respect to f(Y) is the same as with respect to the set of $f(Y_r|Y_{(r)})$. It may be easier to work with the latter distributions since each is univariate.

We briefly argue that, in checking against the observed y_r, $f(Y_r|Y_{(r)})$ is the preferred univariate predictive distribution for Y_r. If f(Y) is proper we could consider the univariate marginal $f(Y_r)$. Of course the $f(Y_r)$ do not determine f(Y) but more importantly $f(Y_r)$ ignores the remaining observations, $y_{(r)}$. In practice, were we to attempt prediction of a new, not necessarily independent, Y_r at X_r we would use a posterior distribution for θ

in creating the desired predictive distribution. In assessing the model this seems appropriate as well; we should check how well the model predicts in the manner in which we would use it to predict. But then should we use $f(Y_r|y)$ or $f(Y_r|y_{(r)})$? The former is used for prediction at a new vector say X_0 but for validation at an X_r for which we have already observed y_r the latter seems preferable. That is, if we propose to check the predictive distribution for Y_r against an observed y_r we should not use that y_r to determine this distribution. Note that $f(Y_r|y_{(r)})$ may be quite different from $f(Y_r|y)$ even in the case that the Y_r are conditionally independent given θ.

Such cross validation is well established in the Bayesian literature dating at least to Stone (1974) and Geisser (1975). Frequentist model diagnostic approaches adopt a similar point of view (see e.g. Belsley, Kuh and Welsch, 1980 or Cook and Weisberg, 1982). Cross validation schemes other than single point deletion may be helpful and will share the same advantages described above. However in the sequel we use $f(Y_r|y_{(r)})$ exclusively.

2.2 Model adequacy

The predictive distributions, $f(Y_r|y_{(r)})$, are to be checked against y_r for $r=1,2,\cdots n$ in the sense that, if the model holds, y_r may be viewed as a random observation from $f(Y_r|y_{(r)})$. To do this we consider $g(Y_r; y_r)$, called a checking function by Box (1980), whose expectation under $f(Y_r|y_{(r)})$ we will calculate and denote by d_r. The set of d_r will be used for model assessment. Computation of the d_r is discussed in Section 3. Once obtained the approach is exploratory. In fact, since each d_r is a function of the entire data vector Y they will be strongly dependent making formal inference very difficult. Our strategy is a Bayesian analogue to well accepted frequentist strategy of examining studentized residuals, DFFITS, DFBETAS etc., (again see Belsley, Kuh and Welsch, 1979 or Cook and Weisberg, 1982).

We will look at several choices of g. For example

(i) $g_1(Y_r; y_r) = y_r - Y_r$ yielding $d_{1r} \equiv y_r - \mu_r$ where $\mu_r = E(Y_r|y_{(r)})$. The d_{1r} are natural deviations or residuals mentioned in Geisser (1987, p. 138). With

$\sigma_r^2 = \text{var}\ (Y_r|\mathbf{y}_{(r)})$, standardizing yields $d'_{1r} = d_{1r}/\sigma_r$. The quantity $\Sigma(d'_{1r})^2$ could be used as an index of model fit. Many large $|d'_{1r}|$ cast doubt upon the model but retaining the sign of d'_{1r} allows patterns of under or over fitting to be revealed. If the $f(Y_r|\mathbf{y}_{(r)})$ are assumed approximately normal then a normal plot of the d'_{1r} may be informative as well.

(ii) $g_2(Y_r; y_r) = 1_{A_r}(Y_r)$ where $A_r = (-\infty, y_r]$ yielding $d_{2r} = P(Y_r \le y_r|\mathbf{y}_{(r)})$. Viewing y_r as a random observation from $f(Y_r|\mathbf{y}_{(r)})$ implies $d_{2r} \sim U(0,1)$. Because of the dependence amongst the d_{2r} it would be wrong to expect them to exhibit the spread associated with independent uniform samples. Nonetheless an adequate model should manifest d_{2r} which are roughly centered about .5 without many extreme values. Evidence to the contrary calls the model to question.

(iii) $g_3(Y_r; y_r) = 1_{B_r}(Y_r)$ where $B_r = \{Y_r : f(Y_r|\mathbf{y}_{(r)}) \le f(y_r|\mathbf{y}_{(r)})\}$ yielding $d_{3r} = P(B_r|\mathbf{y}_{(r)})$. Again viewing y_r as a random observation from $f(Y_r|\mathbf{y}_{(r)})$ implies $f(y_r|\mathbf{y}_{(r)})$ is, itself, a random realization of $f(Y_r|\mathbf{y}_{(r)})$ whence $d_{3r} \sim U(0,1)$. Again the d_{3r} should be roughly centered about .5 without many extreme values. The d_{3r}, adapted from Box (1980) also appear in Geisser (1987). Note that Y_r need not be univariate in this definition; g_3 may be used for other cross–validation schemes. In fact Box proposed use of g_3 to assess the entire joint predictive distribution. Unfortunately, calculation of this multivariate probability will generally be difficult. This same measure is referred to as the surprise index in Aitchison and Dunsmore (1975).

Assuming that Y_r is univariate and that $f(Y_r|\mathbf{y}_{(r)})$ is unimodal, the d_{3r} calculate a set of tail areas. If we further assume that $f(Y_r|\mathbf{y}_{(r)})$ is approximately a normal density, then the event B_r is approximately the event $\{(Y_r-\mu_r)^2/\sigma_r^2 \ge (d'_{1r})^2\}$. Thus $d_{3r} \approx P(\chi_1^2 \ge (d'_{1r})^2)$ relating d_{1r} and d_{3r}. In retaining the sign of the deviation, d_{1r} is preferable to the induced d_{3r}.

(iv) $g_\epsilon(Y_r; y_r) = (2\epsilon)^{-1}\ 1_{C_r(\epsilon)}(Y_r)$ where $C_r(\epsilon) = \{Y_r : y_r - \epsilon \le Y_r \le y_r + \epsilon\}$ yielding $d_{4r}(\epsilon) = (2\epsilon)^{-1}\ P(C_r(\epsilon)|\mathbf{y}_{(r)})$. To avoid specification of ϵ we take the limit as

$\epsilon \to 0$ obtaining $d_{4r} = f(y_r|y_{(r)})$. This quantity dates at least to Geisser and Eddy (1979) and is computed in the definition of B_r. In the case of conditionally independent Y_r given θ, they suggest the use of $\prod_{r=1}^{n} d_{4r}$ as a modification of $f(y)$ for assessing comparative validity of models. We might call $\prod_{r=1}^{n} d_{4r}$ a pseudo–predictive distribution or pseudo–marginal likelihood.

Many small d_{4r} criticize the model but it may not be obvious what a small d_{4r} is. Following an idea of Berger (1985, p 201) we might instead consider a relative likelihood leading to a modified d_{4r} such as $d'_{4r} = d_{4r}/\sup_y f(y|y_{(r)})$ or $d''_{4r} = d_{4r}/E(f(Y_r|y_{(r)})|y_{(r)})$.

2.3 Model Choice

The standard Bayesian approach for model selection goes as follows. Suppose there are J proposed models with model M_j denoted as $f(Y|\theta_j; X, M_j) \cdot \pi(\theta_j)$. If w_j denotes the prior probability of M_j then, by Bayes theorem, the posterior probability of M_j is

$$p(M_j|Y) = f(Y|M_j) \cdot w_j / \sum_{j=1}^{J} f(Y|M_j) \cdot w_j \qquad (2)$$

where $f(Y|M_j)$ is the predictive or joint marginal distribution of the data under model M_j. For observed y the model yielding the largest $p(M_j|y)$ is selected. Calculation of (2) is discussed in Section 3. Geisser and Eddy (1979) suggest a cross validation version replacing $f(Y|M_j)$ in (2) by the pseudo–predictive distribution.

There is a fundamental complication engendered in this formalism which was recognized as early as Bartlett (1957) and elegantly clarified by Pericchi (1984). Some models are implicitly disadvantaged relative to others using this approach even under a state of presumed indifference towards the models i.e. $w_j = 1/J$, $j=1,\cdots,J$. Section 2.3.1

further investigates this complication including an extension of Pericchi's remedy. An alternative remedy is considered in Section 2.3.2 also using cross–validation ideas.

Another criticism is that, in most practical situations, we doubt that anyone including Bayesians would select models in this fashion. One doesn't really believe that any of the proposed models are correct whence attaching a prior probability to an individual model's correctness seems silly. The selection process is typically evolutionary with comparisons often made in pairs until a satisfactory choice (in terms of both parsimony and performance) is made. Such pairwise decisions would be made using the Bayes factor, $f(Y|M_1)/f(Y|M_2)$. But if at least one of the $\pi(\theta_j)$ is vague interpretation of this factor is problematic. A possible remedy is suggested in Spiegelhalter and Smith (1982) using a reserved or imaginary training data set. In Section 2.3.3 we suggest simpler validation criteria based on the ideas in Section 2.2 and in the spirit of Box (1980, p. 427).

2.3.1 Neutralizing differential expected increase in information

A simple example due to Bartlett (1957) reveals a difficulty with the Bayes factor and the standard procedure. Suppose under Model 1, that $Y_1,\cdots,Y_n$ are i.i.d. N(0,1) while, under Model 2, $Y_1,\cdots,Y_n$ are i.i.d. $N(\theta,1)$ with $\theta \sim N(0,\tau^2)$. Then regardless of the data Y, of n, and of w_1, as $\tau^2 \to \infty$, $f(Y|M_1)/f(Y|M_2) \to \infty$ and $p(M_1|Y) \to 1$. This example was extended to more general nested normal models in Smith and Spiegelhalter (1980). Pericchi (1984) identified the source of the complication: for a given experiment the expected increase in information about the model parameters varies with the specification of the model. His remedy is to weight the Bayes factor or to revise the prior probabilities w_j to achieve neutral discrimination with regard to what is expected to be learned about θ_j under model M_j.

In particular, using the usual information entropy measure, the information in the prior is $-\int \pi(\theta) \log\pi(\theta)\, d\theta$ (making the rather strong assumption that this integral exists), the information in the posterior is $-\int \pi(\theta|Y) \log \pi(\theta|Y)\, d\theta$ whence the expected increase in information about θ from the experiment (Lindley, 1956) is

$$I(f,\pi) = \int \left(\int \pi(\theta|Y) \log \pi(\theta|Y)\, d\theta \right) f(Y)\, dY - \int \pi(\theta) \log \pi(\theta)\, d\theta.$$

For two models with $I(f_1,\pi_1)$, $I(f_2,\pi_2)$ Pericchi (1984) proposes revision of w_1 to $w_1' = \exp(I(f_1,\pi_1))/\{\exp(I(f_1,\pi_1)) + \exp(I(f_2,\pi_2))\}$. Equivalently the Bayes factor would be multiplied by $\rho = w_1'/1\text{-}w_1')$.

Under the linear model $Y = X\theta + \epsilon$ with $X_{n\times p}$, $\epsilon \sim N(0,\sigma^2 I)$, σ^2 known and $\theta \sim N(\mu_\theta, \sigma^2 V_\theta)$, μ_θ, V_θ known, it follows from Stone (1959) that

$$I(f,\pi) = \tfrac{1}{2} \log(|I + V_\theta X^T X|). \tag{3}$$

For the example in Section 4 we propose model choice between two nonlinear models with normal errors. But if we replace the mean of Y_r, $X_r\theta$, by $\varphi(X_r;\theta)$, $I(f,\pi)$ no longer has the closed form (3). However a first order approximation may be readily obtained. Assuming $\partial\varphi/\partial\theta_i$ exists for all θ_i, $i=1,2,\cdots,p$ we may write

$$\varphi(X_r;\theta) = \varphi(X_r;\theta_o) + \sum_{i=1}^{p} (\theta_i - \theta_{oi}) \cdot \left.\frac{\partial\varphi}{\partial\theta_i}\right|_{\theta_o}$$

so that $Y_r' \approx \Sigma\, a_{ri}(X)\cdot\theta_i + \epsilon$ where $Y_r' = Y_r - \varphi(X_r;\theta_o) - \Sigma a_{ri}(X)\cdot\theta_{oi}$ and $a_{ri}(X) = \left.\dfrac{\partial\varphi(X_r;\theta)}{\partial\theta_i}\right|_{\theta_o}$. Hence using (3) $I(f,\pi) \approx \tfrac{1}{2}\log(|I + V_\theta A^T A|)$ where A is an $n\times p$ matrix such that $A_{ri} = a_{ri}(X)$. A practical choice for θ_o might be the MLE. For two nonlinear models the resultant weight $\rho = \{|I + V_{\theta_1} A_1^T A_1|/|I + V_{\theta_2} A_2^T A_2|\}^{\frac{1}{2}}$. We may pass to noninformative prior specifications by setting $V_{\theta_1} = V_{\theta_2} = V$ and letting $V^{-1} \to \emptyset$ whence $\rho = \{|A_1^T A_1|/|A_2^T A_2|\}^{\frac{1}{2}}$.

2.3.2 A maximum expected utility approach

An alternative remedy modifies examination of (2) by formulating the problem of

model choice as one of maximizing expected utility. Several authors (Box and Hill, 1967; San Martini and Spezzaferri, 1984; Poskitt 1987) discuss such an approach. The crucial concept is the introduction of a utility functional to capture "the utility of a model given the data". Utility structures incorporating posterior distributions as an argument will have limited applicability for model choice since the parameter vector may be interpreted differently from model to model. Use of predictive distributions avoids this problem. San Martini and Spezzaferri (1984) take the utility of the predictive distribution at a future unobserved Y_o, $U(f(Y_o|Y), y_o)$, where y_o is the true unknown future value. From an argument in Bernardo (1979) they recognize that the unique proper local utility function has the form $b_0 f(y_o|Y) + b_1(y_o)$.

In choosing between two models the expected utility solution is to select the model yielding the larger expected utility. It turns out that, regardless of b_0 and $b_1(y_0)$, we choose $M_1(M_2)$ if $w_1 K(f(Y_o|y,M_1), f(Y_o|y,M_2)) > (<)\ w_2 K(f(Y_o|y,M_2), f(Y_o|y,M_1))$ where $K(f_1,f_2) = \int f_1 \log(f_1/f_2)$ denotes the Kullback–Leibler divergence between the densities f_1 and f_2. This criterion is appealing since $K(f_1,f_2)$ is interpreted as the expected or average information for discriminating in favor of f_1 against f_2. In the cross validation context we replace $K(f(Y_o|Y,M_j), f(Y_o|Y,M_{j'}))$ with $\sum_{r=1}^{n} K(f(Y_r|y_{(r)},M_j),f(Y_r|y_{(r)},M_{j'}))$. This substitution arises by replacing $f(Y_o|y,M_j)$ with the pseudo–predictive distribution, $\prod_{r=1}^{n} f(Y_r|y_{(r)},M_j)$, as discussed in (iv) of Section 2.2. An alternate form is

$$\text{choose } M_1(M_2) \text{ if } E_{f^*}\left[\log \frac{\prod f(Y_r|y_{(r)},M_1)}{\prod f(Y_r|y_{(r)},M_2)}\right] > (<)0 \qquad (4)$$

where $f^* = w_1 \prod f(Y_r|y_{(r)},M_1) + w_2 \prod f(Y_r|y_{(r)},M_2)$.

Calculation of the Kullback–Leibler divergences is discussed in Section 3. Other information measures (see e.g. Csiszár, 1977) could be investigated as well. The expected utility approach is readily extended to $J > 2$ models.

2.3.3 Ad hoc procedures

Each of the criteria developed in Section 2.2 may be converted to an ad hoc model choice procedure. Given two models, for k=1, 2, 3, 4 associate $d_{kr}(M_j)$ with model M_j, j=1, 2.

For $k = 1$ choose M_j with the smaller value of $D_{1j} = \Sigma(d'_{1r}(M_j))^2$

For $k = 2, 3$ choose M_j with the smaller value of $D_{kj} = \Sigma(d_{kr}(M_j) - .5)^2$

For $k = 4$ choose M_j with the larger value of $\Pi d_{4r}(M_j)$. Equivalently

$$\text{choose } M_1(M_2) \text{ according to } D_4 = \log\left[\frac{\Pi d_{4r}(M_1)}{\Pi d_{4r}(M_2)}\right] > (<)0 \qquad (5)$$

Expression (5) may be directly compared with (4). For the latter we calculate *expected* utilities; for the former we calculate *observed* utilities. In fact $\exp(D_4)$ is a pseudo–Bayes factor. Given that the $d_{kr}(M_j)$ will have already been calculated for use in model assessment the additional computation for their use in these ad hoc model choice procedures is negligible.

3. Computational Approaches

We propose the use of sampling–based methodology to calculate the various objects of interest in Section 2. Monte Carlo techniques have significantly advanced our ability to carry out integrations required for Bayesian inference. The literature for noniterative methods is substantial. We mention here the recent papers of Rubin (1988), Geweke (1989) and West (1990). The paper of Geweke provides many further references. Iterative approaches are discussed in Tanner and Wong (1987) and in Gelfand and Smith (1990).

3.1 Monte Carlo estimates of the d_r

For a given model, computational effort focuses on the calculation of the $d_r = E(g(Y_r;y_r)|\mathbf{y}_{(r)}) = \int\int g(Y_r;y_r)\, f(Y_r|\boldsymbol{\theta},\mathbf{y}_{(r)},X)\cdot\pi(\boldsymbol{\theta}|\mathbf{y}_{(r)})\, d\boldsymbol{\theta}\, dY_r$.

If (θ_s, Y_{rs}), $s = 1, \cdots B$ are samples from the joint conditional distribution for θ and Y_r, $f(Y_r|\theta, \mathbf{y}_{(r)}, X) \cdot \pi(\theta|\mathbf{y}_{(r)})$ then a Monte Carlo approximation for d_r is $\hat{d}_r = B^{-1} \sum_{s=1}^{B} g(Y_{rs}; y_r)$. Sampling from $f(Y_r|\theta, \mathbf{y}_{(r)}, X)$ is usually no problem; sampling from $\pi(\theta|\mathbf{y}_{(r)})$ is. We return to this matter shortly.

If $\mathcal{E}_r(\theta; y) = \int g(Y_r; y_r) f(Y_r|\theta, \mathbf{y}_{(r)}, X) \, dY_r$ then $d_r = E(\mathcal{E}_r(\theta; y)|\mathbf{y}_{(r)})$, an expectation with respect to the posterior $\pi(\theta|\mathbf{y}_{(r)})$. In certain cases $\mathcal{E}_r(\theta; y)$ can be calculated explicitly whence $\hat{d}_r = B^{-1} \sum_{s=1}^{B} \mathcal{E}_r(\theta_s; y)$. We need not draw the sample of Y_{rs}. Such savings in random variate generation is referred to as streamlining in Rubin (1988) and in Gelfand and Smith (1990). In fact the estimate of the predictive density itself, $f(Y_r|\mathbf{y}_{(r)})$, requires only the θ_s, i.e., $\hat{f}(Y_r|\mathbf{y}_{(r)}) = B^{-1} \sum_{s=1}^{B} f(Y_r|\theta_s, \mathbf{y}_{(r)}, X)$.

If $h(\theta)$ is an importance sampling density for $\pi(\theta|\mathbf{y}_{(r)})$ and θ_s, $s = 1, \cdots, B$ are drawn from h, the above Monte Carlo estimates are modified to $\hat{d}_r = \sum_{s=1}^{B} g(Y_{rs}, y_r) \cdot v_{rs}$, or $\hat{d}_r = \sum_{s=1}^{B} \mathcal{E}_r(\theta_s; y) \cdot v_{rs}$ and $\hat{f}(Y_r|\mathbf{y}_{(r)}) = \sum_{s=1}^{B} f(Y_r|\theta_s, \mathbf{y}_{(r)}, X) \cdot v_{rs}$ respectively where $v_{rs} = \left[\sum_{s=1}^{B} \pi(\theta_s|\mathbf{y}_{(r)})/h(\theta_s) \right]^{-1} \cdot \pi(\theta_s|\mathbf{y}_{(r)})/h(\theta_s)$. As a related remark, if, for example, $f(Y_r|\theta, \mathbf{y}_{(r)}, X)$ is a normal distribution then $\hat{f}(Y_r|\mathbf{y}_{(r)})$ is a finite mixture of normals. Theory developed in Johnson and Geisser (1983) shows that in such situations $f(Y_r|\mathbf{y}_{(r)})$ is exactly or approximately a t–distribution. But t–distributions arise as scale mixtures of normal distributions which can be arbitrarily well approximated by a finite mixture of normals.

Note that $\pi(\theta|\mathbf{y}_{(r)}) \propto \tau(\theta)/f(y_r|\theta, \mathbf{y}_{(r)}, X)$ where $\tau(\theta) = f(\mathbf{y}|\theta, X) \cdot \pi(\theta)$ so that

$$v_{rs} = \frac{\tau(\theta_s)/(h(\theta_s) \cdot f(y_r|\theta, y_{(r)}, X))}{\left[\sum_{s=1}^{B} \tau(\theta_s)/(h(\theta_s) \cdot f(y_r|\theta, y_{(r)}, X)) \right]^{-1}}.$$

Rather than develop an $h(\theta)$ for each $\pi(\theta|\mathbf{y}_{(r)})$ it would be more efficient to find a simple choice which we could sample and then use for all r. The form of v_{rs} suggests $h(\theta) \propto \tau(\theta)$ i.e. $h(\theta) = \pi(\theta|\mathbf{y})$ would be a natural choice. We recall that the Gibbs sampler, as described in Gelfand and Smith (1990) for application to hierarchical Bayes models, produces observations essentially from the joint posterior $\pi(\theta|\mathbf{y})$. Hence, if the Gibbs sampler is used to carry out Bayesian inference under the given model, the outputted θ_s can be used directly as input to carry out computations needed for studying model adequacy and model choice. Implementation of the Gibbs sampler for challenging models will require tailored versions of the rejection method. See Carlin and Gelfand (1991), Gilks and Wild (1991), Ritter and Tanner (1991). If a noniterative approach has been employed resulting in an importance sampling density $h(\theta)$ for $\tau(\theta)$ then the samples from $h(\theta)$ can as well be used directly in the above formulas. There is considerable literature on the creation of a good importance sampling density. In particular we note the recent work of Geweke (1989) and West (1990).

For model choice additional calculations we may wish to make are the Kullback–Leibler divergences $K(f(Y_r|\mathbf{y}_{(r)},M_j),\ f(Y_r|\mathbf{y}_{(r)},M_{j'}))$. Since these are expectations with respect to $f(Y_r|\mathbf{y}_{(r)},M_j)$, in principle they can be handled in the same way as calculation of the $d_r(M_j)$. However, in practice, the calculation requires enormous storage, even if n is small, since each f is itself a Monte Carlo estimate and since these estimated f's are created under different models but must be merged to calculate K.

The standard approach to model choice requires updating of w_j to $p(M_j|\mathbf{y})$ as in (2). Except in simple cases the marginalization to obtain $f(\mathbf{Y}|M_j)$ is not available in closed form. Noniterative Monte Carlo integration may be employed as follows. If $\pi(\theta_j)$ is proper and θ_{js}, $s = 1,\cdots,B$ are a sample, $\hat{f}(\mathbf{Y}|M_j) = B^{-1} \sum_{s=1}^{B} f(\mathbf{Y}|\theta_{js};X)$. If $\pi(\theta_j)$ is improper but $h(\theta_j)$ is an importance sampling density for $\tau(\theta_j)$, τ defined above, then $\hat{f}(\mathbf{Y}|M_j) = B^{-1} \sum_{s=1}^{B} \tau(\theta_{js})/h(\theta_{js})$.

Interestingly, the Gibbs sampler is less attractive here. By itself, it does not readily produce an estimator of $f(Y|M_j)$. For a collection of J models it need not uniquely provide the posterior probabilities $p(M_j|Y)$ since Markovian updating using the θ_j, $j = 1, \cdots, J$ and a variable M to label the model may violate conditions for convergence (see e.g., George and McCulloch, 1991).

3.2 Simplified sampling for nonlinear normal models

Suppose the model $Y_r = \varphi(X_r; \beta) + \epsilon_r$, $r = 1, \cdots, n$ where the vector of errors, $\epsilon \sim N(0, \sigma^2 \cdot W)$, W known positive definite. With $\theta = (\beta, \sigma^2)$ suppose $\pi(\theta) = \pi_1(\beta) \cdot \pi_2(\sigma^2)$ where $\pi_2(\sigma^2)$ is inverse gamma IG(a,b) i.e. $\pi_2(\sigma^2) \propto \exp(-b/\sigma^2)/(\sigma^2)^{a+1}$. We allow the improper limiting cases as $a \to 0$ and as $b \to 0$. Then $\pi(\theta|Y) = \pi_1(\beta|Y) \cdot \pi_2(\sigma^2|\beta,Y)$ where $\pi_2(\sigma^2|\beta,Y)$ is $IG(a',b')$ with $a' = a + n/2$, $b' = b + \frac{1}{2}(Y - \varphi)^T W^{-1}(Y - \varphi)$, φ the vector of $\varphi(X_r; \beta)$ and $\pi_1(\beta|Y) \propto \pi_1(\beta)/(b')^{a'}$.

Suppose, after transformation of β to domain R^p, that a noninformative prior is taken for β i.e. $\pi_1(\beta) = 1$ (as in e.g. Johnson and Geisser, 1983, p. 138). If $g(X_r; \beta) = X_r\beta$ then, as is well known (see, e.g., Box and Tiao 1973, p. 117), $\pi_1(\beta|Y)$ is exactly a multivariate student t–distribution and sampling–based approaches are not needed. For the nonlinear case let $\phi(\beta) = (Y - \varphi)^T W^{-1}(Y-\varphi)$ and let $\hat{\beta}$ be the MLE for β whence $\phi(\hat{\beta})$ is the error or residual sum of squares for the model. Assuming derivatives exist, to a second order approximation, $\phi(\beta) \approx \phi(\hat{\beta}) + \frac{1}{2}(\beta - \hat{\beta})^T H(\beta-\hat{\beta})$ where H has entries $H_{ut} = \frac{\partial^2 \phi}{\partial\beta_u \partial\beta_t}\Big|_{\hat{\beta}} \cdot H$ is, of course, proportional to the inverse of the sample Fisher information matrix. At the least standard nonlinear regression software handles the independence case ($W = I$) and routinely provides $\hat{\beta}$, $\phi(\hat{\beta})$, $\hat{\sigma}^2 = \frac{\phi(\hat{\beta})}{n-p}$ and $(H^*)^{-1} = 2\hat{\sigma}^2 H^{-1}$, the estimated asymptotic covariance matrix of β. In $\pi_1(\beta|Y)$, replacing $\phi(\beta)$ by this approximation again yields a multivariate student t–distribution, say $t(\beta)$.

For noniterative Monte Carlo we immediately have a promising importance sampling density for $\pi(\theta|Y)$ namely $t(\beta) \cdot \pi_2(\sigma^2|\beta,Y)$. The work of Van Dijk and Kloek (1985), Geweke (1989) and West (1990) suggests refinements to $t(\beta)$. Simplification occurs for the

Gibbs sampler as well since it may be applied directly to $\pi_1(\beta|\mathbf{Y})$ using $t(\beta)$ as described in Carlin and Gelfand (1991). The resulting Gibbs replicates say β_s would then be used to sample σ_s^2 from $\pi_2(\sigma^2|\beta,\mathbf{Y})$ to obtain θ_s. The illustrative example in Section 4 is handled using noniterative Monte Carlo. Other models may admit similar conjugacies which ameliorate the computing burden.

4. An illustrative example

Our example compares two sigmoidal growth curve models of the same dimension. Consider the following data (Ratkowsky, 1983, p. 88) recording as Y the dry weight of onion bulbs plus tops versus growing time X.

X	1	2	3	4	5	6	7	8	9
Y	16.08	33.83	65.80	97.20	191.55	326.20	386.87	520.53	590.03

X	10	11	12	13	14	15
Y	651.92	724.93	699.56	689.86	637.56	717.41

The data is plotted in Figure 1 suggesting sigmoidal behavior. We propose to investigate model adequacy for and choice between a logistic model, $Y_r = \beta_0(1 + \beta_1\beta_2^{X_r})^{-1} + \epsilon_r$ and a Gompertz model, $Y_r = \beta_0\, e^{-\beta_1\beta_2^{X_r}} + \epsilon_r$ where in either case we assume the ϵ_r iid $N(0,\sigma^2)$, $r = 1,2,\cdots 15$. Under either model β_0 is interpreted as an asymptote while $\beta_2 \in (0,1)$. We take $\beta_1 > 0$ to yield an increasing function of X. In both cases we reparametrize β to R^3 by setting $\beta_1' = \log \beta_1$, $\beta_2' = \log(\beta_2/(1-\beta_2))$ and then taking the prior $\pi(\beta_0, \beta_1', \beta_2', \sigma^2) = (\sigma^2)^{-1}$. In the notation of the previous section $\pi_1(\beta_0, \beta_1', \beta_2') = 1$ and $\pi_2(\sigma^2) = IG(0,0)$.

The results of a standard nonlinear regression fitting package (SAS PROC NLIN) for each model are given in Table 1. These estimates were used to obtain, for each model, a multivariate–t distribution which was then used as an importance sampling density for the

noniterative Monte Carlo approach described in Section 3.2 with B = 2000. Table 2 provides the predictive means, $E(Y_r|\mathbf{y}_{(r)})$, and the d_{ir} for each model.

Table 2 reveals that $X_r = 14$ and, to a lesser extent, $X_r = 11$ are troublesome points under both models. Plots of d_{2r} vs X_r and d_{4r} vs X_r for both models reveal no systematic patterns. For model 1 (logistic) $\bar{d}_2 = .4978$, $\bar{d}_3 = .6076$; for model 2 (Gompertz) $\bar{d}_2 = .5742$, $\bar{d}_3 = .5997$. For illustration, Figure 2 presents boxplots of d_{2r} -.5 for each model. Turning to the criteria of Section 2.3.3 we have $D_{11} = 18.27$, $D_{12} = 16.82$; $D_{21} = .9219$, $D_{22} = 1.3160$; $D_{31} = 1.5657$, $D_{32} = 1.9487$ and $D_4 = 1.6863$. All told, both models seem to provide adequate fit with model 1 being preferable.

5. Conclusions

The predictive techniques proposed here for model checking and model choice are self–contained with respect to the experiment, accommodate both proper and improper priors, employ only univariate distributions and, using sampling–based methods are readily computed. The Monte Carlo technology described here can be straightforwardly modified for use in other predictivist enterprise such as prediction of future observations, diagnostics for outlier/influential point detection (Johnson and Geisser, 1982, 1983) and optimal combination of models for prediction (Min and Zellner, 1990).

Methodology for effective model assessment and selection is available and implementable. As the art of Bayesian data analysis evolves and more challenging problems are tackled, judicious use of this methodology should become a standard component of the data analysis process.

Acknowledgment

The first author's research was supported in part by NSF grant DMS 8918563.

Table 1: Maximum Likelihood Estimation for the Two Sigmoidal Growth Curve Models of Section 4.

Logistic Model:

$\hat{\beta}_0 = 702.876$ $\hat{\beta}_1' = 4.454$ $\hat{\beta}_2' = -0.008$

$\phi(\hat{\beta}) = 8913.991$ $\hat{\sigma}^2 = 742.833$

$$(H^*)^{-1} = \begin{bmatrix} 193.741 & -1.107 & -0.872 \\ -1.017 & 0.058 & 0.026 \\ -0.872 & 0.026 & 0.0133 \end{bmatrix}$$

Gompertz Model:

$\hat{\beta}_0 = 723.059$ $\hat{\beta}_1' = 2.502$ $\hat{\beta}_2' = 0.564$

$\phi(\hat{\beta}) = 13616.000$ $\hat{\sigma}^2 = 1134.667$

$$(H^*)^{-1} = \begin{bmatrix} 486.053 & -3.842 & 2.311 \\ -3.842 & 0.0813 & 0.039 \\ 2.311 & 0.039 & 0.020 \end{bmatrix}$$

Table 2: Predictive means and d_{ir} for Models 1 and 2

Model 1

X_r	y_r	$E(Y_r \mid \mathbf{y}_{(r)})$	d'_{1r}	d_{2r}	d_{3r}	d_{4r}
1	16.08	15.88	0.0066	0.3846	0.9999	0.0364
2	33.83	31.00	0.0965	0.6125	0.7947	0.0361
3	65.80	59.19	0.2239	0.5091	0.8711	0.0349
4	97.20	110.26	-0.4347	0.3447	0.5418	0.0314
5	191.55	188.42	0.0967	0.5976	0.8754	0.0331
6	326.20	286.56	1.1858	0.8257	0.3467	0.0143
7	386.87	429.51	-1.3458	0.0963	0.2663	0.0110
8	520.53	524.68	-0.1244	0.4985	0.8740	0.0314
9	590.03	602.30	-0.3913	0.3181	0.6780	0.0304
10	651.92	647.43	0.1433	0.5822	0.8910	0.0336
11	724.93	665.16	2.2774	0.9832	0.0210	0.0028
12	699.56	687.16	0.3931	0.6559	0.6893	0.0304
13	689.96	697.74	-0.2425	0.3779	0.6734	0.0323
14	637.56	708.65	-2.9761	0.0012	0.0067	0.0006
15	717.41	698.73	0.5913	0.6784	0.5843	0.0266

Model 2

X_r	y_r	$E(Y_r \mid \mathbf{y}_{(r)})$	d'_{1r}	d_{2r}	d_{3r}	d_{4r}
1	16.08	0.38	0.4745	0.8105	0.4650	0.0280
2	33.83	5.40	0.8779	0.4938	0.6854	0.0205
3	65.80	30.25	1.1041	0.9247	0.1677	0.0158
4	97.20	97.22	-0.0007	0.6426	0.9840	0.0299
5	191.55	202.20	-0.2702	0.4809	0.9438	0.0275
6	326.20	314.97	0.2897	0.7406	0.8339	0.0258
7	386.87	436.20	-1.3458	0.0698	0.1539	0.0107
8	520.53	516.31	0.1157	0.4172	0.9202	0.0289
9	590.03	583.19	0.1916	0.7452	0.8931	0.0292
10	651.92	628.85	0.6543	0.8718	0.6351	0.0234
11	724.93	655.60	2.1501	0.9908	0.0207	0.0026
12	699.56	682.69	0.4533	0.5509	0.7938	0.0265
13	689.96	701.02	-0.2998	0.2456	0.5169	0.0276
14	637.56	717.28	-2.6886	0.0034	0.0038	0.0008
15	717.41	713.76	0.0947	0.6243	0.9778	0.0287

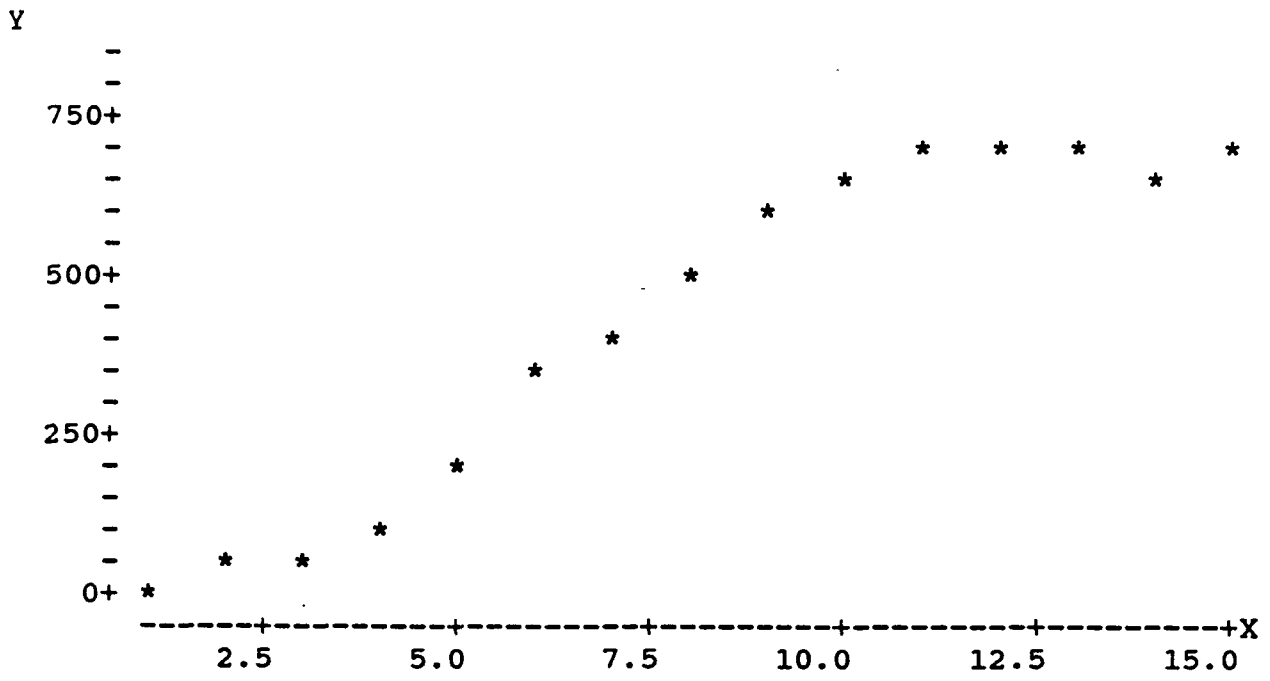
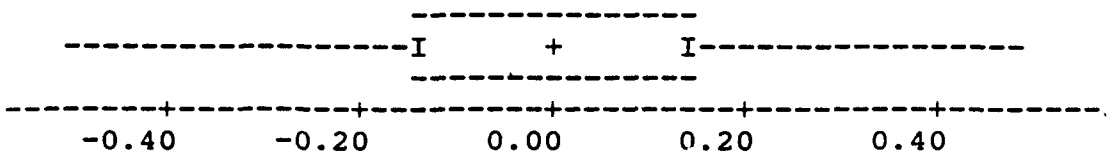


Figure 1: Plot of onion bulb data



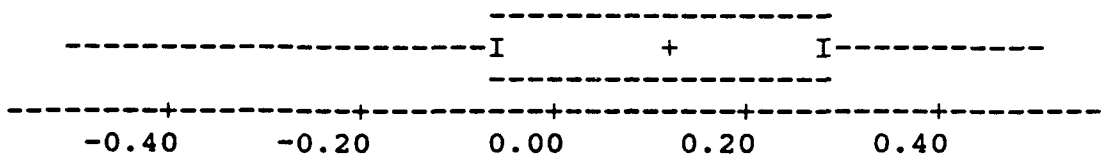


Figure 2: Boxplots of d_{2r}-.5 for Models 1 and 2

References

Aitchison, J. and Dunsmore, I. (1975). *Statistical Prediction Analysis*, University Press, Cambridge, England.

Bartlett, M. (1957). A Comment on D.V. Lindley's Statistical Paradox, Biometrika 44, 533–534.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*, J. Wiley and Sons, New York.

Berger, J. (1984). The robust Bayesian viewpoint. In: *Robustness of Bayesian Analysis*, J. Kadane, ed., p. 63–124, North Holland, Amsterdam.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer–Verlag, New York.

Bernardo, J. (1979). Expected information as expected utility. Ann. Statist. 7, 686–690.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). J.R. Statist. Soc. B, 36, 192–326.

Box, G. (1980). Sampling and Bayes' inference in scientific modeling and robustness (with discussion). J.R. Statist. Soc. A, 143, 382–430.

Box, G. and Hill, W. (1967). Discrimination among mechanistic models. Technometrics, 9, 57–71.

Box, G. and Tiao, G. 91973). *Bayesian Influence in Statistical Analysis*. Addison Wesley, Reading, MA.

Carlin, B. and Gelfand, A. (1991). An iterative Monte Carlo method for nonconjugate Bayesian analysis. Department of Statistics, Statistics and Computing (to appear).

Carlin, B. and Polson, N. (1991). Inference for nonconjugate Bayesian modeling using the Gibbs sampler. Canadian Journal of Statistics (to appear).

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence In Regression*. Chapman and Hall, New York.

Csiszár, (1977). Information Measures: A Critical Survey, Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes – 1974 Reidel Publications, Boston, MA.

Dempster, A. (1975). A subjective look at robustness. ISI Bull. 46, 349–374.

Geisser, S. (1975). The predictive sample reuse method with application. J. Amer. Statist. Assoc. 70, 320–328, 350.

Geisser, S. (1985). On the prediction of observables: a selective update. In: Bayesian Statistics, 2, (J. Bernardo, et. al., eds.), North–Holland, Amsterdam, 203–230.

Geisser, S. (1987). Influential observations, diagnostics and discordancy test. J. Appl. Statist. 14, 133–142.

Geisser, S. (1988). The future of statistics in retrospect. In: Bayesian Statistics, 3, (J. Bernardo, et. al., eds.) Oxford University Press, Oxford, 147–158.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. J. Amer. Statist. Assoc., 74, 153–160.

Gelfand, A.E. and Dey, D.K. (1991). On measuring Bayesian robustness of contaminated classes of priors. *Statistics and Decisions*, 9, 63–80.

George, E. and McCulloch, R. (1991). Variable selection via Gibbs sampling. Graduate School of Business, University of Chicago, Tech. Rpt.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. Econometrica, 57, 1317–1339.

Gilks, W. and Wild P. (1991). Adaptive rejection sampling for Gibbs sampling. Applied Statistics (to appear).

Guttman, I. (1991). A Bayesian look at the question of diagnostics. Tech. Rpt. #9104, Dept. of Statistics, University of Toronto.

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

Johnson, W. and Geisser, S. (1982). Assessing the predictive influence of observations. In: Statistics and Probability Essays in Honor of C.R. Rao. (Kallianpur, Krishnaiah, Ghosh, eds.) North–Holland, Amsterdam, p. 343–358.

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. J. Amer. Statistic. Assoc. 78, 137–144.

Lindley, D. (1956). On a measure of information provided by an experiment. Ann. Math. Statist. 36, 986–1005.

Min C. and Zellner, A. (1990). Bayesian and non–Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. Graduate School of Business, University of Chicago, Tech. Rpt.

Pericchi, L. (1984). An alternative to the standard Bayesian procedure for discrimination between normal linear models. Biometrika, 71, 576–586.

Pettit, L.I. and Smith, A.F.M. (1985). Outliers and influential observations in linear models. In: Bayesian Statistics, 2, (J. Bernardo, et. al., eds.), North Holland, Amsterdam, 473–494.

Poskitt, D. (1987). Precision, complexity and Bayesian model determination. J.R. Statist. Soc. B, 49, p. 199–208.

Ratkowsky, D.A. (1983). *Nonlinear Regression Modeling: A Unified Practical Approach.* M. Dekker, Inc., New York.

Ritter, C. and Tanner, M. (1991). Facilitating the Gibbs Sampler: the Gibbs stopper and the griddy Gibbs sampler. Department of Statistics, University of Wisconsin, Tech. Rpt.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculation for the applied statistician. Ann. Statist. 12, 1151–1172.

Rubin, D. (1988). Using the SIR algorithm to simulate posterior distribution. In: Bayesian Statistics, 3, (J. Bernardo, et. al., eds.), Oxford University Press, Oxford, 395–401.

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion. J. R. Statist. Soc. B, 46, 296–303.

Smith, A.F.M. (1986). Some Bayesian thoughts on modeling and model choice. The Statistician, 35, 97–102.

Smith, A.F.M. and Spiegelhalter, D. (1980). Bayes factors and choice criteria for linear models. J.R. Statist. Soc. B, 42, 213–220.

Spiegelhalter, D. and Smith, A.F.M. (1982). Bayes factors for linear and log–linear models with vague prior information. J. R. Statist. Soc. B, 377–387.

Stone, M. (1959). Application of a measure of information to the design and comparison of regression experiments. Ann. Math. Statist. 39, 55–72.

Stone, M. (1974). Cross–validatory choice and assessment of statistical predictions. J.R. Statist. Soc. B, 36, 111–147.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Statist. Assoc. 82, 528–550.

Van Dijk, H. and Kloek, T. (1985). Experiments with some alternatives for simple importance sampling in Monte Carlo integration. In: Bayesian Statistics, 2, (J. Bernardo, et. al., eds.) North–Holland, Amsterdam, 511–530.

Verdinelli, I. and Wasserman, L. (1991) Bayesian analysis of outlier problems using the Gibbs sampler. Statistics and Computing (to appear).

West, M. (1990). Bayesian computations: Monte Carlo density estimation. ISDS, Duke University, Tech. Rpt.

Discussion

Adrian E. Raftery (University of Washington)

1. Introduction and summary

It is a pleasure to congratulate the authors on an interesting and important paper that points out how sampling–based methods can make Bayesian diagnostics for model checking routinely available. Bayesian diagnostics are often similar to frequentist ones, but they have the great advantage of being systematically available through the predictive distribution, even for complex models. This is in contrast with frequentist diagnostics, which have to be developed from scratch for each new class of models, often requiring considerable ingenuity. The *interpretation* of Bayesian diagnostics is somewhat glossed over by the authors, however.

We part company to some extent on the issue of model choice. I am unconvinced by the arguments against the standard Bayesian procedure, namely that based on posterior model probabilities. New results indicate that posterior model probabilities *can* be readily computed using sampling–based methods. Also, the standard Bayesian procedure *is* based on predictive distributions, in a prequential rather than a cross–validation sense.

2. Bayesian diagnostics for model checking

A real achievement of this paper is to show how sampling–based methods can be used to obtain Bayesian diagnostics systematically and routinely for a very wide class of models. When frequentist diagnostics are available they are often similar to Bayesian diagnostics. The great advantage of Bayesian diagnostics is that they are available quite generally from the predictive distribution, unlike their frequentist counterparts, which can require considerable ingenuity for each new class of models.

The authors have, however, rather glossed over the *interpretation* of their diagnostics. For example, in the nonlinear regression example, they conclude that points 11 and 14 are troublesome but that, all told, both models provide an adequate fit. What is the basis for

this conclusion? Nothing is suggested beyond eyeballing the results, but there are certainly more precise criteria implicitly at work here, and they should be made explicit.

I would suggest that diagnostics not be used to reject the current model, but rather to guide the search for better models by indicating the direction of search, or the way in which the current model is inadequate. If this leads to the specification of an alternative model, then the current model can be compared with alternative one using the posterior odds ratio (or posterior expected utilities of these can be specified); the current model will not be rejected unless the alternative one is decisively preferred. You don't abandon a model unless you have a better one in hand.

Even viewing diagnostics this way, as an exploratory tool rather than as a basis for inference, we still need some yardstick to calibrate our inspection of the results. Here it does seem that frequentist calculations are useful, and I suspect that such calculations implicitly underly the authors' interpretation of the results in their Table 2.

3. Model comparison: In support of the standard Bayesian procedure

The standard Bayesian procedure is given by the authors' equation (3), and amounts to basing inference on the posterior model probabilities. They raise two objections to this procedure, which I will now briefly discuss.

3.1 "Bartlett's paradox"

This is the observation due to Bartlett (1957) that if under M_1 and Y_i are i.i.d. N(0,1), and under M_2 they are i.i.d. $N(\theta,1)$ with $\theta \sim N(0,\tau^2)$, then $p(M_1|Y) \to 1$ as $\tau^2 \to \infty$ regardless of the data; see the authors' section 2.3.1.

This has been presented by the authors and by others that they cite as a major flaw of the standard Bayesian approach, but I do not find it too disquieting. Letting $\tau^2 \to \infty$ implies that $E[|\theta|]$ also becomes arbitrarily large, so it is not too surprising that, for any data set, $E[|\theta|]$ can be set large enough that the data prefer zero. Some prior information is almost always available that will limit the prior variance τ^2, and it is always important to investigate the sensitivity of $p(M_1|Y)$ to changes in τ^2. In practice, $p(M_1|Y)$ tends to

be rather insensitive to changes in τ^2 over a wide range (see, e.g., Raftery, 1988). Thus, Bartlett's paradox seems to me to suggest that the use of highly diffuse priors is not a good idea for model comparison.

It may be objected that it is desirable to have a "reference" procedure for model comparison. However, in my applied experience, reasonable proper priors are often readily accepted, especially when backed up with a serious sensitivity analysis; the likelihood is often the more controversial part of the analysis.

3.2 The more serious criticism

The authors write:

"Another criticism is that, in most practical situations, we doubt that anyone including Bayesians would select models in this fashion. One doesn't really believe that any of the proposed models are correct whence attaching a prior probability to an individual model's correctness seems silly. The selection process is typically evolutionary with comparisons often made in pairs until a satisfactory choice (in terms of both parsimony and performance) is made."

Attaching a prior probability to a model is not any sillier than science as traditionally practiced. Most of science is an attempt to find a model that predicts the observations to date well; it does not claim to have found the "truth" (if such a things exists) or the "correct model". Science typically proceeds by adopting a *paradigm*, which means essentially *conditioning* on a collection of models, often with an explicit parametric form. Prior probabilities conditional on the adopted paradigm, or collection of models, do make sense.

Of course, if one does not so condition, the prior probability, and hence also the posterior probability of most models is zero. Since one does not believe the paradigm to be the "truth", this may make science as a whole seem silly, but its record of success argues in its favor. Note that the marginal likelihood, $f(Y|M_j)$, which is proportional to the

posterior probability of M_j, is just the (predictive) probability of the data given the model M_j, and so is precisely the right quantity for evaluating the scientific theory defined by M_j.

Consider, for example, the question of whether smoking causes lung cancer, and suppose that the currently accepted way of addressing this issue is within the framework of the logistic regression model, logit(Pr[lung cancer]) = γ1[smokes] + β^Tx, where x is a vector of control variables. Conditionally on this framework (or "paradigm"), the issue becomes a comparison of the two models $M_1 : \gamma = 0$ and $M_2 : \gamma > 0$. Then a scientist's natural language statement "I am 90% sure that smoking causes lung cancer" is equivalent, given the framework, to the statement that $p(M_1) = 0.1$ and $p(M_2) = 0.9$. This does seem to make sense even if, unconditionally on the framework, $p(M_1) = p(M_2) = 0$.

Of course, the natural language statement itself can be viewed as not being about "truth", but rather about future data and trends in scientific opinion. It might mean, for example, "I am 90% sure that future data will be better predicted by M_2 than by M_1", or "I am 90% sure that within T years the belief that smoking causes lung cancer will be generally accepted"; note that the latter two statements can be given standard betting interpretations. For an example where scientists might attach substantial prior probability to the smaller ("null") model, consider cold fusion.

The authors describe the standard Bayesian procedure as a model *selection* procedure, but it is considerably richer than that. When comparing two models that genuinely represent rival scientific hypotheses, the posterior odds ratio provides a summary of the evidence for one model against the other; unless the evidence is very strong, one model will not necessarily be selected.

Often, however, model form is not the object of primary scientific interest. The authors did not say what the main scientific question was in their growth curve example, but I suspect that it was not the choice between the two models that they considered. If interest focuses instead on some other quantity, Δ, such as the next observation, Y_{16}, or the asymptote, β_0, then ***model selection is a false problem***, and it is important to take account

of model uncertainty. The Bayesian approach provides an immediate way of doing this using the equation.

$$p(\Delta|Y) = \sum_{j=1}^{J} p(\Delta|Y,M_j)p(M_j|Y). \qquad (1)$$

Hodges (1987) emphasized the importance of taking account of model uncertainty, pointing out that failure to do so leads to the overall uncertainty being underestimated, and hence, for example, to overly risky decisions.

If the posterior probability of one of the models is close to unity, or if the posterior distribution of Δ is almost the same for the models that account for most of the posterior probability, then $p(\Delta|Y)$ may be approximated by conditioning on a single model, namely by $p(\Delta|Y,M_i)$ for some i. This seems to be the main situation in which model selection, as such, is a valid exercise. The "evolutionary" process to which the authors refer is in reality an informal search method for finding the main models that contribute to the sum in equation (1), and in this sense may be viewed as an approximation to the full (standard) Bayesian procedure. Clearer recognition of this might lead to more satisfactory model search strategies.

4. The standard Bayesian procedure and sampling–based methods.

The key quantity for the implementation of the standard Bayesian procedure is the marginal likelihood, $f(Y|M_j) = \int f(Y|\theta_j,X,M_j)\pi(\theta_j)d\theta_j$. The authors say that the Gibbs sampler does not readily produce an estimator of $f(Y|M_j)$. However, Newton and Raftery (1991) have recently pointed out the existence of a simple and general such estimator. They show that, given a sample from the posterior, *the marginal likelihood may be (simulation–consistently) estimated by the harmonic mean of the associated likelihood values.* This result applies no matter how the sample was obtained, whether directly using the analytic form of the posterior, by importance sampling, the Gibbs sampler, the SIR algorithm or the weighted likelihood bootstrap. There can be stability problems with this estimator, and slight modifications that avoid these are discussed in the cited reference.

The standard Bayesian procedure *is* a predictive approach since the marginal likelihood can be written

$$f(Y|M_j) = \prod_{r=1}^{n} f(Y_r|Y^{r-1},M_j), \tag{2}$$

where $Y^{r-1} = (Y_1,\cdots,Y_{r-1})$. Note that the conditional densities on the right–hand side of equation (2) are conditional on the first (r-1) observations, and *not* on all the other (n-1) observations. Thus the standard Bayesian procedure is a "prequential" method in the sense of Dawid (1984), and not a cross–validation approach. Each conditional density on the right–hand side of equation (2) may be evaluated in a sampling–based way, using the same methods as the authors propose for their d_{4r}. It follows that this provides an alternative sampling–based way of calculating the marginal likelihood, and hence of implementing the standard Bayesian procedure.

Note also that equation (2) remains valid even if the observations are permuted. Thus, even if the model does not impose a natural ordering on the observations, "prequential diagnostics" may be obtained by sampling from the set of all permutations of the observations and averaging over diagnostics based on the conditional densities on the right–hand of equation (2).

If one replaces the conditional densities on the right–hand side of equation (2) by densities conditional on all the observations except the rth one, one obtains the quantity that the authors denote by $D_4 = \prod_{r=1}^{n} d_{4r}$. This could be called a "pseudo–marginal likelihood", by analogy with the pseudo–likelihood concept introduced by Besag (1975). Using D_4 rather than $f(Y|M_j)$ is similar to using the pseudo–likelihood rather than the likelihood when the latter is available, which does not seem to be a very good choice. As an argument in favor of D_4, however, the authors point out that with improper priors D_4 is defined whereas $f(Y|M_j)$ is not. This strikes me as a disadvantage of improper priors rather than of the standard marginal likelihood.

I will attempt to summarize the various analogies and equivalences discussed in the following table.

Prequential analysis	Cross–validation
Likelihood	Pseudo–likelihood
Marginal likelihood ($f(Y\|M_j)$)	"Pseudo–marginal likelihood" (D_4)
Posterior model probability/ Bayes factor	Fixed–level significance test
BIC (Schwarz, 1978)	AIC, C_p

Entries in the same column are regarded as being related, either by being motivated by the same approach or by being asymptotically equivalent. Entries in the same row are viewed as different approaches to the same task or concept. I prefer the entries in the left–hand column, headed "prequential analysis", while the authors seem to incline to the entries in the right–hand column. Note that the difference can be important, especially with large samples.

References

Besag, J.E. (1975) Statistical analysis of non–lattice data. *Statistician* 24, 179–195.

Dawid, A.P. (1984) Present position and potential developments: some personal views. Statistical theory. The prequential approach (with Discussion). *J.R. Statist. Soc.* A 147, 178–292.

Hodges, J.S. (1987) Uncertainty, policy analysis and statistics (with Discussion). *Statist. Sci.* 2, 259–291.

Newton, M.A. and Raftery, A.E. (1991) Approximate Bayesian inference by the weighted likelihood bootstrap. Technical Report no. 199, Department of Statistics, University of Washington.

Raftery, A.E. (1988) Approximate Bayes factors for generalized linear models. Technical Report no. 121, Department of Statistics, University of Washington.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.* 6, 461–464.

Daniel Peña (Universidad Carlos III de Madrid)

This paper contains three features that I really like: (1) it stresses the importance of model determination and model diagnosis in statistical analysis; (2) it advocates a cross–validatory assessment of the model using the predictive distribution; (3) it points out the difficulties of using a naive Bayesian analysis for Model choice.

Model diagnosis has received an increasing interest in the Bayesian literature and I would like to add the works by Zellner (1975), Johnson and Geisser (1985) and Guttman and Peña (1988) to the references given in the paper.

The information about the model adequacy is contained in the joint predictive distribution f(y) and a key problem is to devise simple procedures to reveal this information. A sensible first step is the one used in this paper, as

$$f(y) = f(y_i|y_{(i)})f(y_{(i)})$$

we can look at the cross–validatory predictive distribution $f(y_i|y_{(i)})$ that is unidimensional. However, this procedure, although very useful, does not show some of the interesting multivariate features of the data, for instance, sets of similar points that are different from the rest of the data and which cannot be identified by univariate analysis because of masking. Also a set of outliers can produce that some other good points appear as outlying, and this situation has been called swamping in the statistical literature. To avoid these problems, we need to consider either the whole predictive density or the distributions $f(y_I|y_{(I)})$ and $f(y_{(I)})$, where y_I is a subset of observations. Of course, looking at all the possible decompositions of the data is an impossible task and we need to develop procedures to search for interesting combinations of points. My joint work with George Tiao in this volume may be a first step on this direction.

As far as the computation of $f(y_I|y_{(I)})$ is concerned it should be pointed out that the easiest way to understand its structure is to use:

$$f(y_I|y_{(I)}) = \int f(y_I|\theta, y_{(I)})f(\theta|y_{(I)})d\theta \qquad (1)$$

instead of $f(y)/f(y_{(I)})$. The reason is that (1) is similar to the standard marginalization to compute the predictive, and therefore, standard techniques can be applied to obtain the distribution in a compact way. For instance, if $y \sim N(\mu,\sigma^2)$ with σ^2 known and $\mu \sim N(\mu_0,\sigma_0^2)$, it is straightforward to show that

$$f(y_I|y_{(I)}) = (\frac{1}{\sqrt{2\pi}})^I \sigma^{-(I-1)}(I\sigma^2_{(I)} + \sigma^2)^{-1/2} \exp\left[-\frac{I}{2}\left[\frac{s_I^2}{\sigma^2} + \frac{(\bar{y}_I - \hat{\theta}_{(I)})^2}{I\sigma^2_{(I)} + \sigma^2}\right]\right]$$

where

$$\hat{\theta}_{(I)} = \frac{(n-I)\bar{y}_{n-I}\sigma^{-2} + \mu_0\sigma_0^{-2}}{(n-I)\sigma^{-2} + \sigma_0^{-2}}$$

and

$$\sigma^2_{(I)} = ((n-I)\sigma^{-2} + \sigma_0^{-2})^{-1}$$

Therefore, the cross–validation predictive density for the subset I depends on the ratio $s_I^2/\sigma^2 = \Sigma_{i\epsilon I}(y_i-\bar{y}_I)^2/(I\sigma^2)$ that is a key factor in the analysis of this subset.

References

Guttman, I. and Peña, D. (1988) Outliers and Influence: Evaluation by posteriors of parameters in the linear model. *Bayesian Statistics III.* Bernardo, J.M. et al (editors). Oxford University Press, 631–640.

Johnson, W. and Geisser, S. (1985) Estimate influence measure of the multivariate general Linear Model. *Journal of Statistical Planning and Inference*, 11, 33, 56.

Zellner, A. (1975) Bayesian analysis of regression error terms. *Journal of American Statistical Society*, 70, 138–44.

L.R. Pericchi (Simon Bolivar Universidad, Caracas)

It would be a promising theoretical exercise to investigate the relationship between your interesting suggestions for selecting a model and the dimension of the model. As an extreme case, one may think of a model that encompasses both models 1 and 2 in your illustrative example, and then compare with models 1 and 2. Which model would your suggestions select?

L.I. Pettit (Goldsmiths' College, London)

I would firstly like to comment on the measure of model adequacy discussed in section 2.2 and fill in some of their history.

The possibility of using d_{1r} and other similar 'Bayesian residuals' was discussed by Pettit (1986) and Geisser (1987). Chaloner and Brant (1988) suggest a different definition of Bayesian residuals using an idea going back to Zellner (1975). Geisser (1987) also discusses the use of d_{3r} which he describes as a discordancy ranking. The quantity d_{4r}, usually called the conditional predictive ordinate (CPO), was proposed by Geisser (1980) and used by Pettit and Smith (1983, 1985) and Pettit (1988) as a tool in outlier modelling. Pettit (1990) presents a number of results about the CPO for the normal distribution. The quantity d'_{4r} is called the ratio ordinate measure by Pettit (1990). I think the idea of comparing a predictive distribution to its mode goes back to Roberts (1965).

As far as model choice goes, I have found the use of Bayes factors, which do not require a prior probability of an individual model's correctness (§2.3), to be very useful. The approach of Spiegelhalter and Smith (1982) to the problem of improper priors is important. Measuring the effect of individual observations on Bayes factors (Pettit and Young, 1990) leads to an expression which is the difference in logarithms of the CPO's for the two models and so ties in with the model adequacy ideas.

The computational methods discussed in this paper will be very useful in calculating all these quantities and it is therefore for me a very welcome paper.

References

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

Geisser, S. (1980). In discussion of Box, G.E.P. (1980). *J. Roy. Stat. Soc., A*, 143, 416–417.

Pettit, L.I. (1986). Diagnostics in Bayesian model choice. *The Statistician*, 35, 183–190.

Pettit, L.I. (1988). Bayes methods for outliers in exponential samples. *J. Roy. Statist. Soc., B*, 50, 371–380.

Pettit, L.I. (1990). The conditional predictive ordinate for the normal distribution. *J. Roy. Statist. Soc., B*, 52, 175–184.

Pettit, L.I. and Smith, A.F.M. (1983). Bayesian model comparison in the presence of outliers. *Bull. Int. Statist. Inst.*, 50, 292–309.

Pettit, L.I. and Young, K.D.S. (1990). Measuring the effect of observations on Bayes factors. *Biometrika*, 77, 455–466.

Roberts, H.V. (1965). Probabilistic prediction. *J. Amer. Statist. Assoc.*, 60, 50–62.

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

Françoise Seillier–Moiseiwitsch (University of North Carolina, Chapel Hill)

The fundamental, and welcomed, stance of this paper is the shift of emphasis, in model determination, from parameters to observables: goodness–of–fit criteria are abandoned in favour of an assessment based on the model's ability to produce decent predictions. The selected model will indeed often be used on new observables.

Several questions arise regarding the checking functions the authors adopted. Which one did they find most useful? In particular, g_1 focuses on a single characteristic of the predictive distribution whereas g_3 and g_ϵ take into account the whole distribution function. Situations where the former is more informative are likely to be few. For model comparison, have they found the difference in logarithmic scores more revealing than looking at the difference in other scores?

Scoring rules can also be of use in checking the adequacy of a single model. The results of Seillier–Moiseiwitsch & Dawid (1990), developed for discrete outcomes and bounded scoring rules, can be adapted for continuous variables. These results assume a natural ordering in the realizations. The probability range can be partitioned into a fixed number of bins and the probabilities, under the predictive distribution, that the observable falls in each of these bins can be compared to the indicators of the realized Bernoulli process. The score constructed over a number of outcomes, once normalized, behaves asymptotically like a N(0,1) random variable. The normalization is carried out with respect to the predictive distribution at each point.

The authors mention the difficulty in drawing formal inference from $\{d_r\}$ due to the strong dependency among these. Transformations that yield independent random variables could be considered. for instance, by conditioning on sufficient statistics in the probability integral transform, one is provided with a set of i.i.d. residuals (O'Reilly & Quesenberry, 1973). Let $\tilde{F}_n(y) = F(y|T_n)$ where T_n is a minimal sufficient statistic, $\{\tilde{F}_n(Y_i)\}$ are i.i.d. uniforms on [0,1]. Furthermore, $\tilde{F}_n(Y_1)$, $\tilde{F}_n(Y_2|Y_1),\cdots,\tilde{F}_n(Y_\alpha|Y_1,\cdots,Y_{\alpha-1})$ also generate a set of α i.i.d. U[0,1] random variables, where α is the number of components in the vector

of minimally sufficient statistics. This conditional transform fits particularly well in a sequential sampling framework (Seillier–Moiseiwitsch, 1990). Indeed, if T_n is doubly transitive and adequate, then $\tilde{F}_{n-\alpha+1}(Y_{n-\alpha+1}),\cdots,\tilde{F}_n(Y_n)$ have the same distributional properties. If no natural ordering exists, the transform should be applied to the ordering sample (O'Reilly & Stephens, 1982).

References

O'Reilly, F.J. & Quesenberry, C.P. (1973). The conditional probability integral transform and its applications to obtain composite chi–square goodness–of–fit tests. *The Annals of Statistics*, 1, 74–83.

O'Reilly, F.J. & Stephens, M.A. (1982). Characterizations and goodness–of–fit tests. *Journal of the Royal Statistical Society*, B, 44, 353–360.

Seillier–Moiseiwitsch, F. (1990). Sequential probability forecasts and the probability integral transform. Submitted to the *International Statistical Review.*

Seillier–Moiseiwitsch, F. & Dawid, A.P. (1990). On testing the validity of probability forecasts. Tentatively accepted by the *Journal of the American Statistical Association.*

Reply to the discussion:

We thank the discussants for their kind and generally positive remarks. We knew that our reference list for this active research area was very incomplete and appreciate the additional citations provided in the discussion. Pettit's historical perspective is a particularly welcome supplement.

Several discussants comment upon the close relationship between the model determination problem and the issues of diagnosing and modeling outliers. Also see Draper and Guttman (1987). We note that sampling–based approaches expedite calculations associated with these issues. See for instance, the recent paper of Verdinelli and Wasserman (1991). Peña encourages us to investigate cross–validation schemes other than single point deletion in particular with regard to identifying masking and swamping. He suggests that $f(\mathbf{Y}_I | \mathbf{Y}_{(I)})$ be computed. We concur noting that the methodology in section 3 is pertinent to such computation. Our only reservation involves possible combinatoric problems as indicated in Peña and Tiao (1991).

Pericchi raises an interesting question which does not appear to have a simple answer. The difficulty is that, in general, it is not obvious what the model which "encompasses both models 1 and 2" is. In customary linear models it is clear; we merely augment the design matrix to do this. However consider the two nonlinear models discussed in section 4 i.e. model 1: $Y = \beta_0(1 + \beta_1 \beta_2^x)^{-1} + \epsilon$, model 2: $Y = \gamma_0 e^{-\gamma_1 \gamma_2^x} + \epsilon$. The encompassing model which is additive in the mean structure will not be identifiable; the asymptote is $\beta_0 + \gamma_0$. If we remedy this by setting $\beta_0 = \gamma_0$ we can no longer retrieve model 1 or model 2 as a reduced model. Suppose we try a multiplicative form for the encompassing model $Y = \beta_0(1+\beta_1 \beta_2^{\,x})^{-1}\ e^{-\gamma_1 \gamma_2^x} + \epsilon$. Now the reduced models are not identifiable; $\beta_1 = 0$ or $\beta_2 = 0$ produces model 2, $\gamma_1 = 0$ or $\gamma_2 = 0$ produces model 1.

Turning to the remarks of Seillier–Moiseiwitsch we agree that the checking function g_1 may be less informative than the others. Nonetheless examination of residuals is standard and familiar. Moreover, the resulting d_{1r} have an immediate connection with Bayesian residuals as discussed in Chaloner and Brant (1988). They consider the posterior

distribution of the unobserved errors which, in our setting, leads to the distribution of $\epsilon_r | Y_{(r)} = y_r - \varphi(X_r; \beta) | \gamma_{(r)}$. The mean of this distribution is d_{1r}.

Her suggestion to transform the $\{d_{ir}; r=1, \cdots, n\}$ to a set of i.i.d. $U(0,1)$ variates is interesting but we suspect feasible only in certain simple cases. That is, preliminary reading of O'Reilly and Quesenberry (1973) yields several concerns. Their approach requires the joint predictive distribution, f(Y), to be proper, requires an explicit expression for f(Y) and in fact, appears to require that f(Y) be an exponential family to effectively bring (minimal) sufficiency into play. A separate problem is that, even were we able to carry out the calculations, we worry about the inherent order dependence of the results since for general response model data no natural ordering need exist.

Finally, Raftery offers the lengthiest and most penetrating discussion. One of his main points concerns the computation of f(Y). We agree that this can be done and, in fact, at the end of section 3 mention the use of importance sampling densities to do so. Whether the posterior is a good choice is unclear since the resulting harmonic mean estimator may be unstable. Calculation of f(Y) in a sequential fashion seems silly. In most cases the effort to compute f(Y) directly would not be much greater than that required to compute an individual term in the factorization. We also note the aforementioned concern regarding the inherent order dependence which is not mitigated computationally by the suggestion to randomly sample permutations.

More importantly, we criticized the use of f(Y) when it is not integrable and not because we couldn't compute it. We completely agree that the choice of likelihood is the critical problem and in fact say so in the introduction. We are less sanguine about the availability of proper priors. If they are developed through training data (imaginary or otherwise) is this not really similar in spirit to cross–validation?

Turning to our criticism of the standard Bayesian model choice procedure there are no doubt situations where we may knowledgeably assign prior weights to models in which case we would certainly do so and obtain posterior odds. But "garden variety" specification of the likelihood with regard to features discussed in our introduction doesn't

seem to readily lend itself to such weighting. However we thoroughly agree that Bayes factors (when interpretable) or pseudo–Bayes factors are vital objects to compute in comparing models. Still these factors may disadvantage some models relative to others. Hence we value the information obtained through other checking functions. A question requiring further analytic and empirical elaboration is, in the case of proper priors, how different will the Bayes factor and the pseudo–Bayes factor be particularly as n increases?

In conclusion we are invigorated by all of the discussion, critical or otherwise. Model determination is obviously a fundamental data analytic task. Illumination of its aspects in the Bayesian framework, particularly contentious ones, necessarily enhances our understanding of the task.

Additional references:

Draper, N.R. and Guttman, I. (1987). A common model selection criterion. In: Probability and Bayesian Statistics, (R. Viertl, ed.) Plenum Publishing Corp., Innsbruck, Austria, p. 134–150.

Peña, D. and Tiao, G.C. (1991). Bayesian outliers functions for linear models. In: Bayesian Statistics 4, (J. Bernardo, et. al., eds.) (to appear).

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 462	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Model Determination using Predictive Distri--utions with Implementation via Sampling-Based Methods		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Alan E. Gelfand, Dipak K. Dey, Hong Chang		8. CONTRACT OR GRANT NUMBER(s) N0025-92-J-1264
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305-4065		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 111		12. REPORT DATE December 4, 1992 13. NUMBER OF PAGES 40
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Model adequacy, model choice, predictive distributions, cross–validation, sampling based methods, sigmoidal growth model, logistic growth curve model, Gompertz model.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See Reverse Side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Summary

Model determination is divided into the issues of model adequacy and model selection. Predictive distributions are used to address both issues. This seems natural since, typically, prediction is a primary purpose for the chosen model. A cross–validation viewpoint is argued for. In particular, for a given model, it is proposed to validate conditional predictive distributions arising from single point deletion against observed responses. Sampling based methods are used to carry out required calculations. An example investigates the adequacy of and rather subtle choice between two sigmoidal growth models of the same dimension.