Comment

Ronald W. Butler

I welcome this papêr by Professor Bjørnstad because it calls attention to the important practical problem of prediction which, of course, has been of scientific interest long before the subject of statistical science itself. This paper is also timely because I believe there will be much future interest in this area driven by the current concern for quality control. Indeed, most of the stated goals and objectives in the quality control area are predictive aims to which the predictive methodology herein might be more appropriately and profitably applied.

The first portion of Bjørnstad's article summarizes efforts to produce a likelihood-based approach to predictive inference. Lauritzen (1974), Hinkley (1979) and Butler (1986) all conditioned on sufficient statistics in order to judge the compatibility of future observable values with the data. Such conditional inference methodology is consistent with Fisher's (1973) use of conditioning for parametric inference in two-by-two tables. Section 1 below elaborates on this comparison and also motivates conditional predictive likelihood (denoted L_c by Bjørnstad) for discrete data. In addition, various profile-based predictive likelihoods can in turn be motivated as saddlepoint approximations to these conditional predictive likelihoods (see Butler, 1989).

Section 4 of Bjørnstad's article introduces new material on predictive likelihood assessment. These assessment procedures are based on the accuracy of certain unconditional coverage probabilities which I do not believe are either relevant or useful for assessing and choosing among the various predictive likelihood recipes. Section 2 below discusses an assessment procedure based on the accuracy of conditional coverages given the appropriate ancillary statistics. Since the original motivation for predictive likelihood is founded on ideas of conditional inference, it seems fitting and indeed more meaningful (to me at least) that assessment should be conditional as in Barnard (1986), Butler (1989) and in Section 2 below.

1. CONDITIONAL PREDICTIVE LIKELIHOOD

The conditional predictive likelihood recipe

$$(1)\qquad L_c(z\,|\,y) = \frac{f(y, z; \theta)}{f(r(y, z); \theta)}$$

first appeared in Hinkley (1979) in a rather disguised form that precluded its general usage. Subsequently it was written in the form (1) in Butler (1986, page 4) and suggested therein for use with discrete data only. Motivation for (1) was also provided in the same article (page 3) which I now expand upon.

The Bayesian analyst uses the marginal distribution of the data for model criticism (Box, 1980), i.e.,

$$(2)\qquad f(y) = \int f(y\,|\,\theta)f(\theta)\,d\theta,$$

where $f(\theta)$ denotes a prior distribution. In prioritizing a generic value z of future observable Z, the Bayesian analyst also includes the value z with the data y and criticizes the model with

$$(3)\qquad f(y, z) = \int f(y, z\,|\,\theta)f(\theta)\,d\theta,$$

which is proportional in z to $f(z\,|\,y)$, the Bayesian predictive density. Criticism in (3) is not concerned with whether the theta associated with the distribution of Y is the same as that of the distribution of Z given $Y = y$; these have been assumed to be the same. What is being criticized is the level of agreement between the generically assumed value z and the observed data y.

From a likelihood perspective model criticism generally proceeds by conditioning the data on a minimal sufficient statistic $r(y)$ for the parameter (see Cox and Hinkley, 1974, pages 37–38), i.e.,

$$(4)\qquad f(y\,|\,r(y)) = \frac{f(y; \theta)}{f(r(y); \theta)}.$$

In prioritizing a generic value z of a future observable we incorporate it into the data as does the Bayesian analyst and use

$$(5)\qquad f(y, z\,|\,r(y, z)) = L_c(z\,|\,y),$$

or conditional predictive likelihood, to assess the compatibility of the value z with data y.

We can illustrate these principles using a simple example in which $y = (x_1, \cdots, x_n)$ is assumed to be an iid sample of Bernoulli (θ) trials. Conditioning y on $r(y) = \sum x_i = r$ leads to a uniform distribution over all $\binom{n}{r}$ subsets or configurations of the r successes (ones) and $n - r$ failures (zeros). Suppose we are concerned that $\theta = \text{pr}\{\text{success}\}$ might be increasing with trial number and wish to measure such concern with data $y = (0, 1, 0, 0, 1, 1)$. Then among $\binom{6}{3} = 20$ configurations, we count those which are at least as

Ronald W. Butler is Associate Professor, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523.

extreme as what is observed where extremity is measured by counting the minimal number of 0-1 interchanges necessary to convert a configuration into the most extreme case (000111). Configuration (001011) takes 1 interchange, (001101) and (010011) both take 2, and all others require 3 or more. We can therefore measure our concern as 4/20 using probability conditional on $r = 3$. The distribution of the number of interchanges associated with the 20 configurations is

no. interchanges	0	1	2	3	4	5	6	7	8	9
no. configurations	1	1	2	3	3	3	3	2	1	1

.

To configuration (010101) we would therefore attach the probability 7/20.

Predicting the seventh Bernoulli (θ) trial based on data (010011) using conditional predictive likelihood (1) amounts to computing the probability of (0100111) given $r = 4$, the predictive likelihood that $Z \sim \text{Bern}(\theta)$ is 1, and comparing it with the probability of (0100110) given $r = 3$, the predictive likelihood Z is 0. The former is $\binom{7}{4}^{-1}$ while the latter is $\binom{7}{3}^{-1}$ so the odds in favor of $Z = 1$ are $L(z = 1 \mid y)/L(z = 0 \mid y) = 1$, a quite reasonable answer.

This Bernoulli discussion is very similar to that of Fisher (1973, Chapter 4, Section 4) concerning the two-by-two table

$$(6) \qquad \boxed{\begin{matrix} 3 & 0 \\ 0 & 3 \end{matrix}}$$

In criticizing the independence model he conditions on row and column totals, i.e., statistics sufficient for the parameters under the current model, as in (4). The conditional probability of data as extreme as in (6) is $1/20 = \binom{6}{3,0,3,0}/\binom{6}{3}^2$.

Similar probabilities are involved when conditional predictive likelihood is used in this context. Suppose that our data consisted of six iid multivariate Bernoulli trials and we wish to predict the seventh in the sequence. Under the assumption of an independence model, the occurrences of the seventh trial in the various cells have conditional predictive likelihoods $\binom{7}{4}^{-2} = \binom{7}{4}^{-1}\binom{7}{3}^{-1} = \binom{7}{3}^{-1}\binom{7}{4}^{-1} = \binom{7}{3}^{-2}$, so we get the intuitively reasonable result that each cell is equally likely to contain the future observable.

The version of conditional predictive likelihood for continuous data (denoted as L_I) is (1) with an additional Jacobian factor. This factor was included because otherwise with (1) alone one gets different predictive likelihoods depending on which minimal sufficient statistic is conditioned upon. With the inclusion of this factor one gets the same result no matter which minimal sufficient statistic is used; thus conditional predictive likelihood becomes a well-defined quantity. The usage of this predictive likelihood is pretty much restricted to regular exponential families along with a few other examples. Approximate conditional (AC) predictive likelihood (L_{a2}) was suggested in Butler (1986, Rejoinder) as an attempt to generalize this usage to most any parametric setting that might arise. The agreement between AC predictive likelihood and conditional predictive likelihood is quite close in the regular exponential setting; the former is a saddlepoint approximation for the latter (Butler, 1989, Section 2.2).

2. PREDICTIVE LIKELIHOOD ASSESSMENT

Cox (1986) has noted the need for calibrating or assessing various predictive likelihoods in terms of hypothetical long-run properties. By doing so, we shall be able to assess the worth of the various recipes in specific applications. The calibration to be discussed here proceeds as recommended in Barnard (1986) and Butler (1989) which I have chosen to illustrate with a sequence of simple examples.

EXAMPLE 1. Suppose Y and Z are iid with a logistic density having mean θ and variance 1 as in Section 4 of Bjørnstad. Then a classical solution to predicting Z from Y can be based on the predictive pivotal statistic $A_p = 1/2(Z - Y) = 1/2[(Z - \theta) - (Y - \theta)]$ whose density does not depend on θ. Since $\hat{\theta}_z = 1/2(Y + Z)$ then transformation $(Y, Z) \to (\hat{\theta}_z - \theta, A_p)$ leads to a marginal density for A_p as

$$(7) \qquad f(a_p) = \int_{-\infty}^{\infty} f(\hat{\theta}_z - \theta, a_p)\, d(\hat{\theta}_z - \theta).$$

A 90% high-density region for A_p from (1) can be computed numerically as (± 1.155) which leads to ($Y \pm 2.311$) as a 90% predictive interval for Z. With sufficient assurance in the logistic model I would use predictive pivot (7) instead of the various predictive likelihood recipes because predictive intervals based on (7) always have their exact preset coverage; i.e.,

$$\text{pr}\{Y - 2.311 < Z \leq Y + 2.311; \theta\}$$
$$= \text{pr}\{-1.155 < A_p \leq 1.155\} = .9$$

for all $\theta \in (-\infty, \infty)$.

The calibration of a predictive likelihood recipe proceeds in the same manner. Determine the 90% high-density region for Z from $L(z \mid y)$, the predictive likelihood, and call it $I_{.9}(y)$. Then the coverage $\text{pr}\{\mathbf{Z} \in I_{.9}(\mathbf{Y}); \theta\}$ can be graphed against θ and compared to the target .9. Closeness of this graph to .9 gives assurance in the particular recipe and stands as a basis for its calibration. Recipes for L_{a2} and L_{a3}, which I shall refer to as approximate conditional (AC) and modified profile (MP) predictive likelihoods, respectively, are the same here and result in $I_{.9}(Y) = (Y \pm 2.459)$ with a coverage function of constant value .918 for all θ. By comparison profile (P) predictive likelihood and the suggestion of Barndorff-Nielsen (BN), L_{P} and L_{PC} in this article, are the same (but different from AC and MP) producing $I_{.9}(Y) = (Y \pm 2.045)$ with coverage .859 for all θ. Thus AC = MP is better calibrated and preferable to P = BN for this example. A closer examination of the forms of AC = MP and P = BN suggests why. The former is

$$f(\hat{\theta}_z - \theta, a_p)\Big|_{\theta=\hat{\theta}_z} \left\{ -\frac{\partial^2 \log f(\hat{\theta}_z - \theta, a_p)}{\partial \theta^2} \right\}^{-1/2}_{\theta=\hat{\theta}_z} \tag{8}$$

or Laplace's approximation to the integral in (7) while P = BN is $f(\hat{\theta}_z - \theta, a_p)\,|_{\theta=\hat{\theta}_z}$ and lacks the information correction term of Laplace's method. Further discussion of this is in Butler (1989).

Sensitivity of AC = MP and P = BN to the choice of the logistic density over the normal density as input to the recipes can be viewed by comparison of 90% predictive intervals $I_{.9}(Y)$ from each model. In the normal case all four recipes reproduce the predictive pivot $Z - Y \sim N(0, 2)$ leading to $I_{.9}(Y) = (Y \pm 2.326)$. This compares with $(Y \pm 2.459)$ for AC = MP and $(Y \pm 2.045)$ for P = BN in the logistic case. The former appears less sensitive in this example.

EXAMPLE 2. Suppose X_1, X_2, and Z are iid uniform $(\theta - ½, \theta + ½)$ with $Y = (X_1, X_2)$ as the data. This example differs from the last one because an ancillary statistic $X_2 - X_1 = A_a$ exists based on Y. As we shall see, relevant and meaningful predictive pivots as well as relevant predictive coverage probabilities for predictive likelihood assessment require that we condition on the observed value of A_a (see Barnard, 1985, 1986; and Butler, 1989, for further discussion). Consider the transformation $(X_1, X_2, Z) \rightarrow (\bar{X}, A_a = X_2 - X_1, A_p = Z - \bar{X})$ where $\bar{X}$ is an estimator of θ and (A_p, A_a) is ancillary with A_p a predictive pivotal statistic. The conditional density of A_p given A_a is our predictive pivot. We shall see that the marginal density of A_p does not lead to sensible predictive intervals.

Let us suppose, for example, that $x_1 = .06$ and $x_2 = .98$ so that $A_a = .92$. Such data is highly informative about θ having parametric likelihood $f(x_1, x_2, \theta) \propto \chi\{.48 < \theta < .56\}$ where $\chi\{\cdot\}$ is an indicator. Our precise knowledge that $\theta \in (.48, .56)$ should therefore convert into a tight predictive interval for Z. In fact the support of Z, $(\theta - ½, \theta + ½)$, must be contained in $(-.020, 1.060)$ so this range should have 100% coverage for Z. The marginal density of A_p is

$$f(a_p) = \begin{cases} 1 - 2a_p^2 & \text{if } 0 \le |a_p| \le ½, \\ 2(1 - a_p)^2 & \text{if } ½ \le |a_p| < 1, \\ 0 & \text{otherwise,} \end{cases}$$

which is bell-shaped with a 95th percentile of .578 so $(\bar{X} \pm .578)$ is a 90% predictive interval with unconditional coverage .9. For our data this works out to be $(-.058, 1.098)$. This "predictive interval" is not sensible because it contains the known support of Z as a proper subset so it should have coverage 100% and not 90%. Note that alternative data $x_1^* = .50$ and $x_2^* = .54$ which is quite uninformative about θ, leads to the same nonsensical "predictive interval" and the interval's length is unable to vary with A_a so as to reflect the informativeness of the data about the model.

These difficulties are eliminated when $A_a = a$ is conditioned upon leading to the conditional density shown in Figure 1. For our data this density has support $(-.54, .54)$ and the horizontal portion spans $(-.46, .46)$. The 95th percentile is .450 so $(\bar{X} \pm .450)$ or $(.070, .970)$ is our 90% predictive interval. Note that this is a proper subset of $(-.02, 1.06)$, the region known to contain Z. Also note that the interval length is .90, the same as the coverage. Thus this interval is only sensible if it is known to be a subset of the support of the density $(\theta - ½, \theta + ½)$. With likelihood $\chi\{.48 < \theta < .56\}$ then $.56 - .50 = .06$ and $.48 + .50 = .98$ are known points of support verifying that $(.07, .97)$ is indeed a subset of $(\theta - ½, \theta + ½)$. The coverage (given $A_a = .92$) is always 90% for any θ since

$$\begin{aligned} &\mathrm{pr}\{\bar{X} - .45 < Z \le \bar{X} + .45 \,|\, A_a = .92; \theta\} \\ &\quad = \mathrm{pr}\{|A_p| \le .45 \,|\, A_a = .92\} = .9. \end{aligned}$$

I contend that predictive coverage conditional on $A_a = .92$ is the relevant probability upon which to base our prediction of Z. Similar conditional coverages are used in constructing confidence intervals for θ as suggested by Fisher (e.g., 1934) with Efron and Hinkley (1978) and Barnard (1985) providing particularly insightful discussion. The ancillary $A_a = .92$ specifies the accuracy the data attains for the estimation of θ; the numerical example above suggests that it has the same role for predicting Z. I believe such probabilities are the most appropriate measure of the worth of various predictive recipes.

This example has been useful for motivating predictive assessment using conditional coverages. Unfortunately none of the predictive likelihood recipes is applicable to this problem. Profile based recipes fail

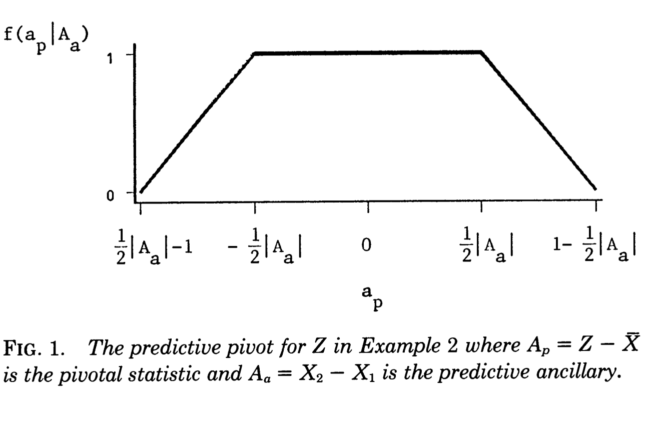

FIG. 1. *The predictive pivot for Z in Example 2 where $A_p = Z - \bar{X}$ is the pivotal statistic and $A_a = X_2 - X_1$ is the predictive ancillary.*

because $\hat{\theta}_z$ is not unique. Recipes which condition on sufficient statistic $R = (R_1, R_2)$, where $R_1 = \min(X_1, X_2, Z)$ and $R_2 = \max(X_1, X_2, Z)$, fail for a more subtle reason; ancillary statistic $R_2 - R_1$ provides predictive information about Z so that conditioning on R conditions out this information preventing us from extracting full predictive information from $f(x_1, x_2, z; \theta)$. In fact, this situation arises in any curved exponential family setting in which $\dim(R) > \dim(\theta)$ so that R consists of $\hat{\theta}$ supplemented by an exact or an affine ancillary (Hinkley, 1980; Barndorff-Nielsen, 1980). As stated in Butler (1986, page 3), conditional predictive likelihoods (given R) should be used only when the density of R bears no extractable information about Z. This is clearly not the case when there are ancillaries or approximate ancillaries within the sufficient statistic.

EXAMPLE 3. In normal linear models (Section 2, Example 5) this assessment procedure leads to the classical predictive intervals as the best. Using the author's notation, the transformation

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \to \begin{bmatrix} \hat{\beta} - \beta \\ n\hat{\sigma}^2/\sigma^2 \\ A_a = \hat{\sigma}^{-1}(Y - C\hat{\beta}) \\ A_p = [(n-p)^{-1}n\hat{\sigma}^2]^{-1/2}V^{-1/2}(Z - C_0\hat{\beta}) \end{bmatrix},$$

where $V = I + C_0(C^TC)^{-1}C_0^T$, leads to a predictive pivot given by the conditional density of A_p given $A_a = a$. When errors $(\varepsilon, \varepsilon_0)$ are iid $N(0, \sigma^2)$, then A_p and A_a are independent so the marginal density of $A_p \sim t_{n-p}(I)$ is the pivot which extends the classical predictive interval given originally by Fisher (1935). Nonnormal errors generally lead to dependence between A_p and A_a so that conditioning on $A_a = a$ leads to the use of $f(a_p \mid A_a = a)$ and not $f(a_p)$ as a pivot when dealing with nonnormal linear models (Barnard, 1986).

There is a distinct preferential ordering of predictive likelihoods for this example using normal errors. Conditional and AC likelihoods agree here, have the most accurate coverage probabilities and treat A_p as $t_{n-p-1}^m(I)$. The MP likelihood is a close second best and treats $A_p \sim t_{n-p-2}^m(I)$ while BN and P likelihoods treat A_p as $t_{n-1}^m(I)$ and $t_n^m(I)$, respectively. The latter two are quite inadequate because they fail to adjust to the degrees of freedom in the model.

EXAMPLE 4. Suppose X_1, X_2, and Z are iid uniform $(0, \theta)$ as discussed in Burridge (1986) and Example 4 of this paper. Transformation

$$\begin{pmatrix} X_1 \\ X_2 \\ Z \end{pmatrix} \to \begin{pmatrix} \hat{\theta}/\theta \\ A_a = \min(X_1, X_2)/\hat{\theta} \\ A_p = Z/\hat{\theta} \end{pmatrix}$$

with $\hat{\theta} = \max(X_1, X_2)$ reveals that A_p and A_a are independent so the pivot is based on A_p with density

$$f(a_p) = \begin{cases} 2/3 & \text{if } a_p \le 1, \\ 2/3\, a_p^{-3} & a_p > 1. \end{cases}$$

A 90% predictive interval is $(0, 1.826\hat{\theta})$ with 90% coverage for all θ.

Profile predictive likelihood exactly reproduces this pivot and is preferred here. Conditional predictive likelihood after normalization is

$$L(z \mid y) = \begin{cases} 1/2\,\hat{\theta}^{-1} & \text{if } z \le \hat{\theta}, \\ 1/2\,\hat{\theta} z^{-2} & z > \hat{\theta} \end{cases}$$

and has 90% predictive interval $(0, 5\hat{\theta})$ with coverage .986. Clearly it is too long. The BN-likelihood agrees with this, while the AC and MP likelihoods are not applicable. The conditional likelihood above is derived as the conditional density of U_1, U_2 given $T = t$ where (U_1, U_2, T) are the order statistics of (X_1, X_2, Z), T is sufficient for θ, and U_1, U_2 are orthogonal to T in (x_1, x_2, z)-space. This does not agree with the result of Bjørnstad who shows that one gets a different answer by taking conditional predictive likelihood as $f(u \mid t)$ where $U = \max(X_1, X_2)$ and $T = \max(X_1, X_2, Z)$.

EXAMPLE 5. This final example is meant to generalize and summarize predictive likelihood assessments. We judge the value of a recipe according to the closeness of its coverage probability function to preset values; e.g., 90% high density predictive interval $I_{.9}(Y)$ has coverage function

$$\text{(9)} \qquad \text{pr}\{\mathbf{Z} \in I_{.9}(\mathbf{Y}) \mid A_a = a; \theta\} \text{ vs. } \theta,$$

which would ideally be .9 for all θ. Meaningful and relevant coverage is computed conditionally on A_a, the value attained by a maximal ancillary statistic based on data Y.

The examples above resulted in preferential orderings of recipes that were uniform in θ and this occurred because the coverage functions could be based entirely on the distributions of ancillary statistics. Examples for which this is not the case can be found in Butler (1989) along with further discussion as to when such coverage functions are flat. Deciding amongst the various recipes when coverage functions are not flat was fairly straightforward for the examples in Butler (1989), although it need not always be so. There is a temptation to compare coverages at and near $\hat{\theta}$, the MLE based on Y. Whether or not this is sensible needs to be considered, however, there is reason to believe that such a procedure would be sensible. This is because coverage functions are determined only from the ancillary statistic A_a (as well as the model) while $\hat{\theta}$, a complimentary portion of the data, is being used to assess the accuracy of these coverage functions. In parametric inference the roles of these statistics are reversed in that the ancillary statistic A_a assesses the accuracy of $\hat{\theta}$ in determining the true model. Practical examples are needed to bear out the sensibility of basing recipe choice on coverages at and near $\hat{\theta}$.

Many practical models such as generalized linear models do not admit exact ancillary A_a upon which to condition. In such instances we must find approximate ancillaries as has been done in Hinkley (1980) and Barndorff-Nielsen (1980, 1983).

I do not agree with Bjørnstad's suggestion that $\mathrm{pr}\{\mathbf{Z} \in I_{.9}(\mathbf{Y}); \theta\}$ as an unconditional probability can be used to meaningfully assess the various recipes. Also measuring the worth of an interval (or its associated recipe) by its guarantee of 90% coverage, $\inf_\theta \mathrm{pr}\{C_\theta(\mathbf{Y}) \geq .9\}$ where $C_\theta(y) = \mathrm{pr}\{\mathbf{Z} \in I_{.9}(y) \mid y; \theta\}$, amounts to a worst case scenario assessment. This could be a very unrepresentative assessment measure to use as a basis for recipe choice.

ADDITIONAL REFERENCES

BARNARD, G. A. (1985). A coherent view of statistical inference. Technical Report, Dept. Statistics, Univ. Waterloo, Waterloo, Ontario, Canada.

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

BURRIDGE, J. (1986). Discussion of "Predictive likelihood inference with applications" by R. W. Butler. *J. Roy. Statist. Soc. Ser. B* **48** 29–30.

COX, D. R. (1986). Discussion of "Predictive likelihood inference with applications" by R. W. Butler. *J. Roy. Statist. Soc. Ser. B* **48** 27.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, New York.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with discussion). *Biometrika* **65** 457–487.

FISHER, R. A. (1934). Two properties of mathematical likelihood. *Proc. Roy. Soc. Ser. A* **144** 285–307.

FISHER, R. A. (1935). The fiducial argument in statistical inferences. *Ann. Eugen.* **6** 391–398.

FISHER, R. A. (1973). *Statistical Methods and Scientific Inference,* 3rd ed. Hafner, New York.

HINKLEY, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* **67** 287–292.

Comment

Tom Leonard, Kam-Wah Tsui and John S. J. Hsu

Professor Bjørnstad is to be congratulated on an excellent review of an important area. Previous statistical practice largely referred to point predictions and estimated standard errors when predicting future observations from current data. When analyzing time series, contingency tables or nonlinear regression models, it is often thought necessary to refer to asymptotics, even to obtain an approximate standard error. However, methods are now available permitting precise predictions based upon finite samples. Moreover, the applied statistician can refer to an entire predictive likelihood or density or probability mass function, summarizing the information in the data about any future observation. This broadens the type of nonlinear model, with several parameters, which may yield useful predictions. These predictions can now be expressed in terms of probability statements, thus enhancing their interpretability, e.g., for noisy data sets.

Let $p(\mathbf{y} \mid \boldsymbol{\theta})$ denote our density (or probability mass function) for an $n \times 1$ vector $\mathbf{y}$ of current observations, given a $p \times 1$ vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$ of unknown parameters, and $p(\mathbf{z} \mid \boldsymbol{\theta})$ represent the corresponding density for an independent $m \times 1$ vector $\mathbf{z}$ of future observations. If $\pi(\boldsymbol{\theta})$ is the prior density of $\boldsymbol{\theta}$, for $\boldsymbol{\theta}$ lying in the parameter space Θ, then the predictive distribution

$$p(\mathbf{z} \mid \mathbf{y}) = \int_\Theta p(\mathbf{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y}) \, d\boldsymbol{\theta} \tag{1}$$

of $\mathbf{z}$ given $\mathbf{y}$ is also representable in the form

$$p(\mathbf{z} \mid \mathbf{y}) = \frac{p(\mathbf{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y})}{\pi(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{z})}, \quad \boldsymbol{\theta} \in \Theta. \tag{2}$$

Here we have

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta, \tag{3}$$

denoting the *posterior density* of $\boldsymbol{\theta}$, given $\mathbf{y}$, and

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{z}) \propto p(\mathbf{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y}), \quad \boldsymbol{\theta} \in \Theta, \tag{4}$$

denoting the *postposterior density* of $\boldsymbol{\theta}$, given $\mathbf{y}$ and $\mathbf{z}$.

Tom Leonard is Associate Professor, Kam-Wah Tsui is Professor and John S. J. Hsu is a graduate student, Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin 53706.