Assessment and Propagation of Model Uncertainty

By DAVID DRAPER[†]

University of Bath, UK

SUMMARY

In most examples of inference and prediction, the expression of uncertainty about unknown quantities y on the basis of known quantities x is based on a *model* M that formalizes assumptions about how x and y are related. M will typically have two parts: *structural* assumptions S, such as the form of the link function and the choice of error distribution in a generalized linear model, and *parameters* θ whose meaning is specific to a given choice of S. It is common in statistical theory and practice to acknowledge parametric uncertainty about θ given a particular assumed structure S; it is less common to acknowledge structural uncertainty about S itself. A widely used approach, in fact, involves enlisting the aid of x to specify a plausible single "best" choice S^* for S, and then proceeding as if S^* were known to be correct. In general this approach fails to fully assess and propagate structural uncertainty, and may lead to miscalibrated uncertainty assessments about y given x. When miscalibration occurs it will often be in the direction of understatement of inferential or predictive uncertainty about y, leading to inaccurate scientific summaries and overconfident decisions that do not incorporate sufficient hedging against uncertainty. In this paper I discuss a Bayesian approach to solving this problem that has long been available in principle but is only now becoming routinely feasible, by virtue of recent computational advances, and examine its implementation in examples that involve forecasting the price of oil and estimating the chance of catastrophic failure of the U.S. Space Shuttle.

Keywords: BAYES FACTORS; CALIBRATION; FORECASTING; HIERARCHICAL MODELS; INFERENCE; MODEL SPECIFICATION; OVER-FITTING; PREDICTION; ROBUSTNESS; SENSITIVITY ANALYSIS; UNCERTAINTY ASSESSMENT

1. INTRODUCTION

The general framework of problems in inference and prediction involves two sets of ingredients: unknown(s) y—such as the causal effect of a treatment in inference, or the price of something next year in prediction—and known(s) x, which will typically include both data and context. The desire is usually to express uncertainty about y in light of x, for instance through a probability specification of the form $p(y|x)$. Specifications of this type that involve conditioning only on things that are known are rare, even in comparatively simple settings (e.g., Lindley, 1982); instead one typically appeals to a *model* M that formalizes judgments about how x and y are related.

1.1. *Structural Uncertainty*

The model may be expressed (e.g., Draper et al., 1987; Hodges, 1987) in two parts as $M = (S, \theta)$, where S represents one or more sets of *structural* assumptions—such as a particular link function in a generalized linear model, or a particular form of heteroscedasticity or time dependence with non-IID data—and θ represents parameters whose meaning is specific to the chosen structure(s). (It will often be possible to express a given model M in more than one way using this notation, but that does not affect the discussion that follows.) Once S is chosen, θ typically follows

[†]*Address for correspondence*: Statistics Group, School of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (d.draper@maths.bath.ac.uk).

fairly unambiguously, apart from technical concerns about reparameterization; but how is S arrived at in practice?

Often the design by which the data in x were gathered renders some structural assumptions compelling. For instance, the randomization employed in designed experiments and sample surveys may be regarded as serving the dual purpose of promoting comparability of treated (sampled) and untreated (unsampled) units and of supporting the assumption of a particular form of conditional exchangeability of the relevant outcome values (e.g., Draper et al., 1993a). But even in controlled experiments and randomized sample surveys, key aspects of S—such as distributional choices for residuals and functional forms for dose-response relationships—will usually be uncertain, and this is even more true with observational studies and data gathered with nonrandom sampling plans.

Thus in practice the model often contains aspects that are not known with certainty: M is not necessarily a part of x. It is a routine feature of most statistical methods to acknowledge *parametric* uncertainty about θ once a particular form for S is chosen, but it is less routine to acknowledge *structural* uncertainty about S itself. A widely used approach, in fact, involves examining the data in x to identify a single "best" choice S^* for S, and then proceeding as if S^* were known to be correct in making inferences and predictions. The field of data analysis, for instance, which has grown considerably in the last thirty years (e.g., Hoaglin et al., 1985), is devoted to the development of graphical and numerical methods, often based on the examination of residuals from the fit of a single standard model, that facilitate a data-driven search for S^*. The very fact of this search, however, implies structural uncertainty that in general is not fully assessed and propagated with the S^* approach, and the result can be uncertainty assessments about y given x whose *calibration* is poor (e.g., in the sense that the empirical distribution of $(\hat{y} - y_{\text{actual}})/\widehat{SD}(\hat{y})$ across one or more such assessments is unacceptably far from (say) $N(0,1)$). When such miscalibration occurs it is often in the direction of anti-conservatism: in retrospect one notices that one's uncertainty bands were not wide enough.

1.2. *Over-Fitting*

This problem, which is often referred to as *over-fitting* the available data, is well known, but has yet to receive a fully satisfying treatment in statistical research and pedagogy. Most of the leading textbooks on applied statistics (e.g., Cox and Snell, 1981) and regression (e.g., Weisberg, 1985) include warnings against over-fitting, but also contain examples of empirical model-building of the S^* form. Another applied area in which the problem has potential to arise (e.g., Chatfield, 1993) is in time series modeling, where model identification, fitting, and forecasting are all routinely based on the same data.

Good regression texts (e.g., Mosteller and Tukey, 1977) offer advice on the value of *cross-validation*—splitting the data into independent modeling and validation data sets—as a partial solution to the over-fitting problem (e.g., Picard and Cook, 1984), but model uncertainty will typically remain even after cross-validation. Moreover, with small samples of data—precisely when structural uncertainty is greatest—cross-validation may not be feasible, because there are too few data values with which to carry out both the modeling and validation activities in a stable way. *Bootstrapping the modeling process* (e.g., Efron and Gong, 1983)—creating bootstrap copies of the available data, conducting independent modeling activities on each copy, and combining the results in a way that is sensitive to the modeling uncertainty

thus uncovered—may help, but as yet little is known about the performance of this approach.

2. CONSEQUENCES OF UNACKNOWLEDGED STRUCTURAL UNCERTAINTY

There is a considerable recent literature on the degree of overconfidence generated by basing inferences and predictions on the same dataset on which the search for structure occurred; see, e.g., Freedman et al. (1986), Hjorth (1989), Miller (1990), Pötscher (1991), and Faraway (1992). Instances may also be found in decision-making in which structural uncertainty is documented by analysts but ignored by consumers of the analysis. Examples of each of these phenomena follow.

2.1. *Model Selection in Regression*

Adams (1991) has conducted perhaps the most comprehensive investigation to date of the effects of the search for S^* on inference in regression. He used simulation to estimate the combined effects of variable selection, transformation of outcome and predictor variables, and deletion of outliers on the nominal observed significance level of R^2. He varied the sample size from 10 to 70, the number of predictors x from 5 to 30, and the degree of correlation among the predictors from 0 to .75, and simulated random error and predictor values from t-distributions with degrees of freedom from 1 to ∞. He examined 114 regression strategies, each based on a different pattern of presence or absence of (a) a simple Bonferroni-based outlier rejection rule, (b) variable selection using a stepwise algorithm or C_p, (c) transformation of the x values with the Box-Tidwell method, and (d) transformation of the outcome y with the Box-Cox approach. Averaging over characteristics of the datasets—all in null situations in which y was unrelated to x, so that the average p-value for judging the significance of the observed R^2 should have been 0.5—he found that the most opportunistic of the 114 strategies produced average nominal p-values well below 0.001, and that every strategy involving either stepwise- or C_p-based variable selection yielded average nominal values below 0.01. The degree of similarity between some of the most egregious strategies in Adams's experiment and standard textbook prescriptions for empirical regression model-building is disquieting.

2.2. *Forecasting the Price of Oil*

In 1980 the Energy Modeling Forum (EMF) at Stanford University assembled a 43-person working group of economists and energy experts, whose goal was to forecast world oil prices from 1981 to 2020 to aid in policy planning. The group generated predictions based on each of 10 leading econometric models, under each of 12 scenarios embodying a variety of assumptions about inputs to the models, such as supply, demand, and growth rates of relevant quantities. One scenario, the so-called "reference," was identified as a "plausible median case" and as "representative of the general trends that might be expected," although readers of the group's summary report (EMF, 1982) were cautioned not to interpret point predictions based on the reference scenario as "[the working group's] 'forecast' of the oil future, as there are too many unknowns to accept any projection as a forecast." The summary report did conclude, however, that most of the uncertainty about future oil prices "concerns not whether these prices will rise ... but how rapidly they will rise."

One may identify three sources of uncertainty in this situation (Draper et al., 1987): *scenario* uncertainty about the inputs to the models; *model* uncertainty (conditional on scenario) about how to translate the inputs into forecasts; and *predictive* uncertainty, conditional on scenario and model. The working group did not attempt

to assess predictive uncertainty, and their final report concentrated on the reference scenario, which—despite their warning above—tended to informally downplay scenario uncertainty as well, but model uncertainty conditional on the reference scenario was evident in the report's tables and figures. Fig. 1 below, for example, is a plot of the yearly point predictions from each of the 10 econometric models under the reference scenario from 1980 to 1990.

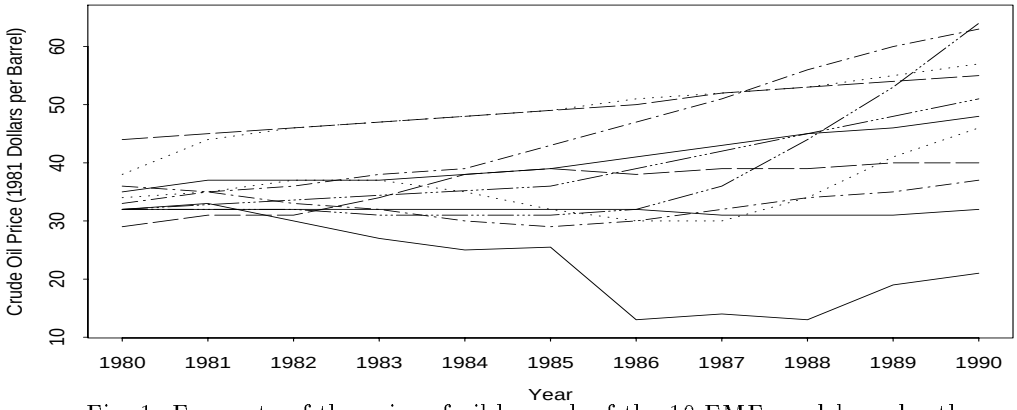

Fig. 1. Forecasts of the price of oil by each of the 10 EMF models under the reference scenario, 1980–1990; lower solid line is actual price.

Averaging across models—giving them equal weight, since the EMF summary report treats them evenhandedly—to obtain a predicted value for 1986, for instance, would yield a figure of about \$39, with implied 90% uncertainty limits (across models, conditional on the reference scenario, and ignoring predictive uncertainty) of about (\$27,\$51). This uncertainty band is consistent with those produced by other efforts parallel to EMF's at the time (e.g., Energy Information Administration, 1982); indeed, as Syme (1987) puts it, "[many] reputable institutions and individuals made forecasts of 1986 oil prices in the 1970s and early 1980s, predicting prices over \$40." She goes on to report that an estimated \$500 billion was invested worldwide by governments and private companies in the early 1980s on the strength of forecasts and informal uncertainty assessments like those in Fig. 1. The actual 1986 world average spot price of oil (see the lower solid line in the plot) was about \$13.

What went wrong? It is not fair to criticize forecasters after the fact for making a sharply inaccurate prediction—no one can see into the future—but it is fair to note that both scenario uncertainty, which might be expected to dominate, and predictive uncertainty were missing in uncertainty assessments like that implicit in Fig. 1. In particular, anyone relying only on Fig. 1 to produce predictive intervals would in effect be assigning zero weight to the 11 non-reference scenarios. This observation may seem nothing more than hindsight—after all, perhaps what actually happened bore no relation to any of the 12 scenarios EMF's working group examined, and one can hardly be faulted for not anticipating something totally new—but in fact one of the non-reference scenarios was rather like what actually occurred (Fig. 2). In Section 6.1 below I examine the extent to which assessing and propagating between-scenario and predictive uncertainty improves predictive calibration in this example.

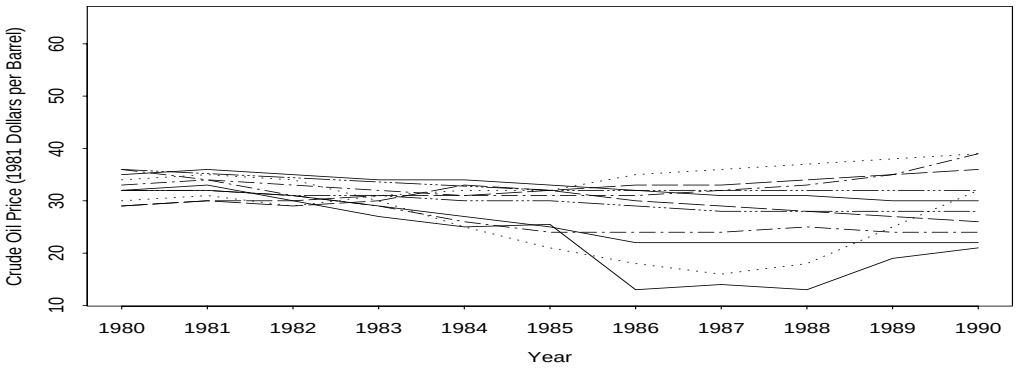

Fig. 2. Forecasts of the price of oil by each of the 10 EMF models under one of the 11 non-reference scenarios, 1980–1990; lower solid line is actual price.

3. A STANDARD BAYESIAN SOLUTION, REVISITED

In theory there is a straightforward Bayesian approach to solving the problem of failure to assess and propagate structural uncertainty, namely to treat the entire model $M = (S, \theta)$ as a nuisance parameter and integrate over uncertainty about both S and θ, as in the expression

$$p(y|x, \mathcal{M}') = \int_{\mathcal{M}'} p(y|x, M)\, p(M|x)\, dM = \iint p(y|x, \theta, S)\, p(\theta, S|x)\, d\theta\, dS\,. \qquad (1)$$

One forms a weighted average of the conditional inferential or predictive distributions $p(y|x, M)$, using as weights the posterior model probabilities $p(M|x)$. This idea is present, implicitly or explicitly, in the writings of workers in at least three fields: statistics (e.g., Box and Tiao, 1962; de Finetti, 1972; Davis, 1979; Geisser and Eddy, 1979; Smith and Spiegelhalter, 1981; Stewart and Davis, 1986; Brown and Lindley, 1986; Draper et al., 1987; Hodges, 1987; Lavine, 1988, 1992; Raftery, 1988; Madigan and Raftery, 1992); econometrics (e.g., Geisel, 1974; Leamer, 1978); and artificial intelligence (e.g., Self and Cheeseman, 1987; Mackay, 1992). In the past the implementation of equation (1) in practice has presented major computational challenges, but advances in the last ten years have greatly reduced this burden. I discuss computational issues in Sections 4 and 5 below. But first, what should one take for the range of integration $\mathcal{M}'$ in this equation?

Writing the posterior model probabilities $p(M|x)$ as $p(\theta, S|x) = p(S|x)p(\theta|x, S)$, it may be seen that the S^* approach described in Section 1 is a special case of equation (1), in which acting as if the structural assumptions in S^*, chosen after a data-driven search, are "correct" corresponds to conditioning on S^*:

$$p(y|x, \mathcal{M}') = \; p(y|x, S^*) = \int p(y|x, \theta^*, S^*)\, p(\theta^*|x, S^*)\, d\theta^*. \qquad (2)$$

This approach correctly assesses parametric uncertainty given S^*—through the integration over θ^* with respect to the posterior distribution $p(\theta^*|x, S^*)$—and *inferential* or *predictive* uncertainty about y conditional on $M^* = (S^*, \theta^*)$, through the distribution $p(y|x, \theta^*, S^*)$. But the search for S^* implies structural uncertainty that has not been fully assessed and included in the uncertainty about y contained in $p(y|x, S^*)$.

Working backwards from $p(M|x) = p(S|x)\, p(\theta|x,S)$ to the prior distributions on which the posterior model probabilities are based gives $p(M|x) = c\, p(S)\, p(\theta|S) \cdot p(x|\theta,S)$, where c is a constant of proportionality. This expression includes two familiar ingredients, a prior distribution $p(\theta|S)$ on the parameters and the likelihood $p(x|\theta,S)$—both specific to a given structural choice S—but it also includes the unfamiliar $p(S)$, a prior distribution on the set of all possible structural assumptions. The key issue in improving upon the S^* approach to modeling is how to specify $p(S)$.

In effect the S^* approach solves this specification problem by equating $p(S)$ to point mass on S^*, a choice that may be too concentrated on a single set of structural assumptions to lead to well-calibrated inferences and predictions. At the other extreme, one might consider specifying $p(S)$ much more diffusely, hoping that the updating process from $p(S)$ to $p(S|x)$ would automatically identify plausible modeling choices. However (e.g., Diaconis and Freedman, 1986), in even the least complicated applied problems with any hint of realism, the space of all possible models is too large to guarantee the success of this updating.

For example, consider perhaps the simplest case of all, a finite sequence $x = (x_1, \ldots, x_n)$ of binary outcomes with no predictors. A model for these data (e.g., Fienberg and Gilbert, 1970; Diaconis, 1977) is just a joint probability distribution for the observables, i.e., a single point in the $(2^n - 1)$–dimensional simplex $\{(p_{0\cdots0}, \ldots, p_{1\cdots1})\colon 0 \le p_{i_1 i_2 \cdots i_n} \le 1, p_{0\cdots0} + \ldots + p_{1\cdots1} = 1\}$. Making standard structural choices—such as taking the x_i to have an IID, exchangeable, or Markovian character—corresponds to conditioning on subspaces of this simplex of very low dimension. With only 10 observations, for instance, an amount of data insufficient to support any but the crudest comparisons of model plausibility, the set $\mathcal{M}$ of all possible models has dimensionality more than 1000, whereas making a standard structural assumption such as "IID Bernoulli with success probability p" corresponds to conditioning on a nonlinear subspace of dimension only 1. The problem is that the dimensionality of $\mathcal{M}$ increases exponentially with n, a rate much faster than that at which information about the relative plausibility of alternative structural choices accumulates. One cannot count on "the data to swamp the prior" when what is at issue is the structural specification of how known and unknown quantities are related.

Thus the space of all models is "too big" to support a diffuse $p(S)$: the promise of inference unconditional on a specific set of modeling assumptions—which appears to be offered by making the range of integration in equation (1) all of $\mathcal{M}$—is unrealizable. However, although it will always be necessary to set $p(S)$ to 0 over most of model space, a single structural choice S^* chosen by a data-driven search amounts to a specification of $p(S)$ that may well be "too small" to be well-calibrated. Is there a compromise between S^* and all of $\mathcal{M}$?

A reasonable intermediate position might be based in practice on *model expansion* (e.g., Box, 1980; Smith, 1984), i.e., starting with a single structural choice such as S^* and expanding it in directions suggested by context, by the data-analytic search that led to S^*, or by other considerations. Good applied work already features *sensitivity analyses* (e.g., Skene et al., 1986), in which the assumptions in S^* are challenged by qualitatively exploring how much one's conclusions would change if an alternative set of plausible assumptions were made. Equation (1) takes this process a step further, by integrating over structural uncertainty rather than simply examining it qualitatively.

4. CONTINUOUS MODEL EXPANSION

Model expansion fits naturally into the framework of *hierarchical modeling* (e.g., Lindley and Smith, 1972; DuMouchel and Harris, 1983), by adding to the top of the hierarchy a level that corresponds to the structural uncertainty: the usual Bayesian formulation on the left of (3)

$$\left\{ \begin{array}{lcl} \theta & \sim & p(\theta) \\ (x|\theta) & \sim & p(x|\theta) \\ (y|x,\theta) & \sim & p(y|x,\theta) \end{array} \right\} \quad \begin{array}{c} \text{is} \\ \text{replaced} \\ \text{by} \end{array} \quad \left\{ \begin{array}{lcl} S & \sim & p(S) \\ (\theta|S) & \sim & p(\theta|S) \\ (x|\theta,S) & \sim & p(x|\theta,S) \\ (y|x,\theta,S) & \sim & p(y|x,\theta,S) \end{array} \right\}. \tag{3}$$

Two cases arise, *discrete* and *continuous*, according to whether the embedding of S^* in a larger subset of model space—by including the top level in the right side of (3)—is indexed discretely or continuously. In the continuous case let α be the expansion index and M_α be the expanded model, of which $S^* = M_0$ (say) is a special case.

4.1. *A Hierarchical Model for Location Inference*

An early example of continuous model expansion was given by Box and Tiao (1962), who reanalyzed Darwin's data on the heights of self- and cross-fertilized plants. These data are in the form of a paired comparison, so that it is reasonable in modeling the pairwise differences $x = (x_1, \ldots, x_n)$ to condition on the structural assumptions $\mathcal{S}_0 = \{x_i = \mu + \sigma\, e_i,\ e_i \text{ IID symmetric about } 0\}$, but there is no *a priori* reason to insist on a specific distributional choice for the e_i. Fisher (1935) had previously analyzed these data by conditioning on the Gaussian; Box and Tiao expanded Fisher's model continuously, by embedding the Gaussian in the symmetric power-exponential family $p(e|\alpha) = c \exp\left\{-\frac{1}{2}|e|^{2/(1+\alpha)}\right\}$, which includes the double exponential ($\alpha = 1$), Gaussian ($\alpha = 0$), and uniform ($\alpha \to -1$) distributions as special cases. Regarding Box and Tiao's structural assumptions $\mathcal{S}_1$ (say) as an expansion of $\mathcal{S}_0$, note that the three quantities μ, σ, and α may be viewed as playing three different roles in this formulation: α may be thought of as indexing one aspect of the structural assumptions in $\mathcal{S}_1$, and μ, the location parameter of interest (the quantity y in equation (1)), and σ, a nuisance (scale) parameter, are components of $\theta = (\mu, \sigma)$. Equation (1) in this context becomes

$$p(\mu|x, \mathcal{S}_1) = \iint p(\mu|x, \sigma, \alpha)\, p(\sigma, \alpha|x)\, d\sigma\, d\alpha, \tag{4}$$

in which the integration over α may be regarded as acknowledging a form of structural uncertainty unaddressed in Fisher's formulation. Interestingly, even though Fisher's model corresponds to placing all one's prior mass on $\alpha = 0$ in the Box and Tiao model, so that Box and Tiao expressed greater model uncertainty than did Fisher, it is possible to have *less* posterior uncertainty about μ in Box and Tiao's formulation than in Fisher's; see Draper (1993).

Note that in model expansion applications involving parametric inference it is important for the quantity of interest, in this case μ, to have the same meaning for each value of α in the expanded model M_α, so that for instance it would have been problematic in Box and Tiao's analysis to embed the Gaussian in a family including asymmetric distributions. In predictive applications this sort of restriction does not arise, because the quantity of interest, a future observable y, is automatically common to all models M_α.

4.2. *Fixed- and Random-Effects Models for Combining Information From Related Experiments*

A more recent example of continuous model expansion, which arises in the combining of information from related experiments, is the case of so-called *fixed-effects* and *random-effects* models in meta-analysis (e.g., Wachter and Straf, 1990). Given data from k experiments or studies designed to measure essentially the same outcome, such as the change in mortality rate caused by a treatment in medical research, one may wish to pool the information from these k sources, to create a better summary of what is known about the effects of the treatment in question than that available from any single source. Letting θ_i be the underlying treatment effect in study i, which may differ from that in study i' due to unmeasured differences in patient cohorts or treatment protocols, and letting x_i be the corresponding data summary in study i, a hierarchical Gaussian random-effects model like the following may approximate one's structural judgments:

$$M_\alpha\colon \begin{cases} (\mu, \alpha \equiv \tau^2) & \sim & p(\mu)\, p(\tau^2) \\ (\theta_i | \mu, \tau^2) & \overset{\text{IID}}{\sim} & N(\mu, \tau^2), \\ (x_i | \theta_i) & \overset{\text{indep}}{\sim} & N(\theta_i, V_i), \end{cases} \tag{5}$$

where the V_i are regarded as known for convenience (typically each x_i is based on a large enough sample of patients that this provides an adequate approximation). Fixed-effects models are a special case of equation (5) in which all of the θ_i are assumed equal, and correspond to random-effects models in which the between-study variance parameter τ^2 is set to zero. Expanding the model from a fixed-effects formulation to one in which $\tau^2 > 0$ implies a net increase in uncertainty about the underlying effect of interest, arising from the between-studies component of variance; failing to adopt a random-effects formulation when necessary may therefore lead to miscalibration.

Model (5) has an interesting application in the physical sciences, in the determination of fundamental constants such as the speed of light c. As Henrion and Fischhoff (1986) and others have noted, if one plots a time series of the currently accepted value of c with uncertainty bands obtained from the standard fixed-effects measurement error model, one notices that every 20 years or so a new value for c is accepted that is inconsistent with the previous uncertainty assessments, demonstrating the presence of bias in the measurement process in addition to the "random" error present in the fixed-effects formulation. With i indexing experiment and j indexing replication within experiment, hierarchically expanding the usual measurement model $x_{ij} = \mu + e_{ij}$ to account for the bias, as in the two-stage model $x_{ij} = \mu + b_i + e_{ij}$, $b_i = \theta + \epsilon_i$, leads to better-calibrated uncertainty assessments than those obtained from the fixed-effects model. See, e.g., Draper et al. (1993b) for other uses of model (5) in physics and chemistry.

4.3. *Computation and Calibration Issues*

Gaussian fixed-effects models are easy to fit using weighted least squares, and when appropriate lead to particularly simple pooling rules by which information from the available sources may be effectively combined. In contrast, even a relatively straightforward empirical-Bayes approach to the random-effects model (5) involves an iterative estimate of τ^2 (see, e.g., Efron and Morris, 1973). Thus practitioners tend to favor fixed-effects models when appropriate, so much so that a common modeling approach involves performing a test of heterogeneity of the θ_i and only adopting the random-effects formulation if the test rejects the null hypothesis H: $\tau^2 = 0$ of homogeneity (see DuMouchel, 1990, for criticisms of this strategy). This is a so-called *preliminary-test* method, similar in spirit to *testimators* sometimes used in econometrics (e.g., Waikar et al., 1984). Methods of this type have been shown inferior in both accuracy and calibration to random-effects methods, such as the empirical-Bayes approach mentioned above, that deal more smoothly with the uncertainty about τ^2 (see, e.g., Sclove et al., 1972; Greenland, 1993).

There is a direct analogy between preliminary-test methods and the S^* approach to modeling described in Section 1: in the S^* approach one searches for a single "best" structure, tests its adequacy, and adopts it unless it fails the test. Using model expansion to embed S^* in a larger class of models, motivated by the structural assumptions in S^* that are most in doubt, treats the modeling uncertainty more smoothly, and—as in the case of empirical Bayes improvements to testimators—may be expected in general to yield better calibration.

Computation in hierarchical models has been difficult until recently, in most settings other than that treated by Lindley and Smith (1972): Gaussian linear models with a conjugate prior structure, in which closed-form expressions for many of the quantities of interest are available. The application of a variety of approximation methods in the last ten years to hierarchical models—including the EM algorithm (e.g., Wong and Mason, 1985), Monte Carlo integration (e.g., Stewart, 1987), and Gibbs sampling and related Markov-chain Monte Carlo (MCMC) methods (e.g., Smith and Roberts, 1993)—promises to greatly increase the routine feasibility of continuous model expansion in applied work. The hierarchical structure in the right side of (3) is particularly well suited to MCMC; see, e.g., Seltzer (1993) for educational applications.

5. DISCRETE MODEL EXPANSION

Although it is often preferable to perform model expansion continuously, so that all the structural uncertainty in the expanded model formulation is accounted for, it is not always possible to index departures from a single structural choice S^* smoothly. Examples include

- *Dynamic linear models* with discrete state spaces (e.g., West and Harrison, 1989). In many applications of dynamic linear models it is natural to regard the state space as continuous, but in other problems (e.g., Smith and West, 1983) it is more fruitful to view the underlying process of interest as moving over time among a finite set of states that have direct substantive meaning; and

- *Discrete propagation of scenario uncertainty*, as in the EMF oil example of Section 2.2, in which 12 distinct scenarios meriting nonzero prior probability but not readily indexed continuously were available.

Discrete model expansion may also be used to approximate a continuous expansion, as in Spiegelhalter's (1981) approximation of the power-exponential model in Box and Tiao's approach in Section 4.1 by the three-point distributional family {Gaussian, uniform, double exponential} to produce a robust location estimator. Recent applied examples of discrete model expansion include Racine et al. (1986), Taylor (1989), and Moulton (1991). For the remainder of the paper I will concentrate on the discrete case.

With a finite set $\mathcal{S} = \{S_1, \ldots, S_m\}$ of structural alternatives in the expanded model, equation (1) becomes

$$p(y|x, \mathcal{S}) = \sum_{i=1}^{m} \int p(y|x, S_i, \theta_i)\, p(S_i, \theta_i|x)\, d\theta_i = \sum_{i=1}^{m} p(S_i|x)\, p(y|x, S_i). \qquad (6)$$

There are thus three ingredients in the computation of $p(y|x, \mathcal{S})$:

- The choice, and prior plausibility, of the S_i over which model uncertainty is assessed and propagated;
- The conditional inferential or predictive distributions $p(y|x, S_i)$ given structural choices S_i; and
- The posterior structural probabilities $p(S_i|x)$.

Each of these components is addressed in the subsections that follow. The second and third components are essentially technical; the first is substantive, and includes the greatest possibility for a retrospective judgment of error.

5.1. *Alternative Structural Choices: Specifying* $p(S)$

As the examples in Section 6 below indicate, the choice of the alternative structures S_i in equation (6) is highly context-specific, but several general comments may be made in any case.

- L. J. Savage used to say that one's model should be "as big as a house." One way to express why this is desirable is by appeal to what Lindley (e.g., 1982) calls Cromwell's rule, which reminds us that any possibility receiving prior probability zero must also have posterior probability zero. The main way to avoid noticing after the fact that a set of modeling assumptions, different from those one originally assumed, turned out to be correct is for one's model prospectively to have been large enough to encompass the retrospective truth. This argues for the routine use of "big" models. In deciding how big is big enough, one may undertake a kind of pre-posterior analysis of structural assumptions, with an eye to the avoidance of retrospective regret at not having included all plausible ways in which the unknown and known quantities might be related.
- $\sum_{i=1}^{m} p(S_i|x)\, p(y|x, S_i)$ is intended to be a discrete approximation to $p(y|x, \mathcal{M}') = \int_{\mathcal{M}'} p(M|x)\, p(y|x, M)\, dM$. To improve on the less satisfactory approximation $p(y|x, S^*)$, one can try to include structures S_i' alternative to S^* satisfying two criteria:
  - S_i' would have high posterior probability $p(S_i'|x)$ (if not given zero prior probability), and

– S_i' has inferential or predictive consequences $p(y|x, S_i')$ that differ substantially from those of S^*.

This was referred to in Draper et al. (1987) as "staking out the corners in model space." One may employ this idea to define directions of departure from S^* that are the most relevant for model expansion.

Other possible approaches to the generation of alternative structures S_i were mentioned at the end of Section 1: creating cross-validation or bootstrap samples from the available data and conducting parallel modeling activities on each sample. Also see George and McCulloch (1993), who use Gibbs sampling to produce posterior probabilities for subsets of predictor variables in regression, and Madigan and Raftery (1992), who use ideas from expert systems, together with an implicit $p(S)$ strongly weighted against complicated structural choices, to find parsimonious submodels of high posterior probability in large contingency tables.

Once a choice is made of the set $\mathcal{S}$, the numerical specification of the prior probabilities $p(S_i)$ will also typically be context-specific. In situations not strongly guided by contextual considerations, one may again proceed by pre-posterior analysis, e.g., starting with constant $p(S_i)$ and computing forward with various possible datasets x to see if the composite result $p(y|x, \mathcal{S})$ appears to realistically assess uncertainty about y given x, and then varying $p(S_i)$ as needed. A form of prequential reasoning (Dawid, 1984) referred to in Draper et al. (1987) as *retrospective calibration* may be helpful in specifying the $p(S_i)$ in time series contexts: with enough data one may (1) choose a variety of points in the past and pretend temporarily that they are the present, (2) make predictions into the known "future," building up a history of forecast errors, and (3) adjust the prior weights $p(S_i)$ to bring the predictive distributions into good calibration with the actual values.

5.2. *Computing the Conditional Inferential/Predictive Distributions* $p(y|x, S_i)$

The second ingredient in discrete model expansion is the set of inferential or predictive distributions

$$p(y|x, S_i) = \int p(y|x, S_i, \theta_i)\, p(\theta_i|S_i, x)\, d\theta_i. \tag{7}$$

This aspect of model expansion creates no new computational burden, since one would have had to compute these distributions in any case as part of one's sensitivity analysis. Closed-form expressions for the results of the (possibly high-dimensional) integration in equation (7) exist in important special cases, such as normal linear models (e.g., Zellner, 1971), and approximations—based, for instance, on Monte Carlo integration (e.g., Geweke, 1989)—are also available. For large n the simple approximation

$$p(y|x, S_i) \doteq p(y|x, S_i, \hat{\theta}_i), \tag{8}$$

where $\hat{\theta}_i$ is the maximum likelihood estimate (MLE) of θ_i under structural choice S_i, may be sufficiently precise. For an example of a more accurate approximation of $p(y|x, S_i)$ see equation (15) below.

5.3. *Computing the Posterior Structural Probabilities* $p(S_i|x)$

Evaluating the posterior structural probabilities $p(S_i|x) = c\,p(S_i)\,p(x|S_i)$ comes down to computing Bayes factors $p(x|S_i)/p(x|S_j)$ for structure S_i against structure S_j, by calculating

$$p(x|S_i) = \int p(\theta_i|S_i)\,p(x|\theta_i, S_i)\,d\theta_i. \tag{9}$$

Several methods for approximating Bayes factors are available, including Gaussian quadrature and a variety of simulation methods based on importance sampling, acceptance/rejection techniques, and MCMC; see Kass and Raftery (1993) for an excellent review. I focus here on two Laplace approximations (e.g., Lindley, 1961; Cox, 1961; Leonard, 1982; Raftery, 1993), of which the first is

$$\ln p(x|S_i) = \frac{1}{2}\,k_i \ln(2\pi) - \frac{1}{2}\ln|\hat{I}_i| + \ln p(x|\hat{\theta}_i, S_i) + \ln p(\hat{\theta}_i|S_i) + O(n^{-1}), \tag{10}$$

where k_i is the dimension of θ_i, $\hat{\theta}_i$ is either the mode of the posterior distribution $p(\theta_i|x, S_i)$ or the MLE, and $\hat{I}_i$ is the observed information matrix evaluated at $\hat{\theta}_i$. A simpler approximation that is often somewhat less accurate with small samples is obtained by noting that for large n, $\ln|\hat{I}_i| \doteq k_i \ln(n)$ and the prior contribution $\ln p(\hat{\theta}_i|S_i)$ becomes negligible, leading to

$$\ln p(x|S_i) = \frac{1}{2}\,k_i \ln(2\pi) - \frac{1}{2}\,k_i \ln(n) + \ln p(x|\hat{\theta}_i, S_i) + O(1). \tag{11}$$

The second and third terms on the right side of equation (11) are recognizable as the basis of the Bayesian information criterion (BIC) for model selection (Schwarz, 1978; cf. Rissanen, 1986). The first term on the right side, $\frac{1}{2}k_i \ln(2\pi)$, has been omitted in most other treatments of this approximation, but its inclusion has improved the accuracy of expression (11) in examples I have examined involving the comparison of structural choices S_i whose θ_i have unequal k_i (cf. Kashyap, 1982). The main way in general to be sure when n is large enough to use equation (11) instead of (10) is to compute them both and compare, although routine experience with this approach will yield guidelines that over time will lessen the need for such explicit comparisons.

In small-sample situations with vague prior information about the parameters, care must be taken, if improper priors are used, to avoid the presence of undefined constants in approximations (10); see, e.g., Spiegelhalter and Smith (1982) for an approach to solving this problem. An alternative solution would involve the use of proper but relatively uninformative priors whose specification is guided by pre-posterior analysis.

5.4. *Summary of a Large-Sample Approximation to* $p(y|x, \mathcal{S})$

To summarize this section, a simple large-sample approximation to $p(y|x, \mathcal{S}) = \sum_{\rangle=\infty}^{\Updownarrow} \sqrt{(\mathcal{S}_\rangle|\S)}\ \sqrt{(\dagger|\S, \mathcal{S}_\rangle)}$ may be obtained by computing the MLE $\hat{\theta}_i$ and maximum log likelihood value for each model $M_i = (S_i, \theta_i)$, and setting $k_i = \dim(\theta_i)$. With diffuse structural and parametric prior information and large n one may then take

- $p(y|x, S_i) \doteq p(y|x, S_i, \hat{\theta}_i)$, and

- $\ln p(S_i|x) \doteq \frac{1}{2} k_i \ln(2\pi) - \frac{1}{2} k_i \ln(n) + \text{loglik}_{\max} + c,$

with c chosen to permit accurate normalization of the posterior structural probabilities so that they sum to 1. It is also useful to note that if $p(y|x, S_i)$ has mean μ_i and variance σ_i^2, and $p(S_i|x) = \pi_i$,

$$\begin{aligned} E(y|x,\mathcal{S}) &= E_S[E(y|x,S)] = \sum_{i=1}^{m} \pi_i \mu_i \equiv \mu, \\ V(y|x,\mathcal{S}) &= E_S[V(y|x,S)] + V_S[E(y|x,S)] \\ &= \sum_{i=1}^{m} \pi_i \sigma_i^2 + \sum_{i=1}^{m} \pi_i (\mu_i - \mu)^2 \\ &= \left(\begin{array}{c} \text{within-} \\ \text{structure} \\ \text{variance} \end{array} \right) + \left(\begin{array}{c} \text{between-} \\ \text{structure} \\ \text{variance} \end{array} \right). \end{aligned} \tag{12}$$

This last expression may be used as the basis of a *model uncertainty audit*, in which the overall inferential or predictive uncertainty about y is decomposed into the sum of two terms: the average conditional uncertainty given each structural choice, and the uncertainty about y arising from structural uncertainty itself. With the S^* approach of Section 1 this second term is set to 0, often inappropriately.

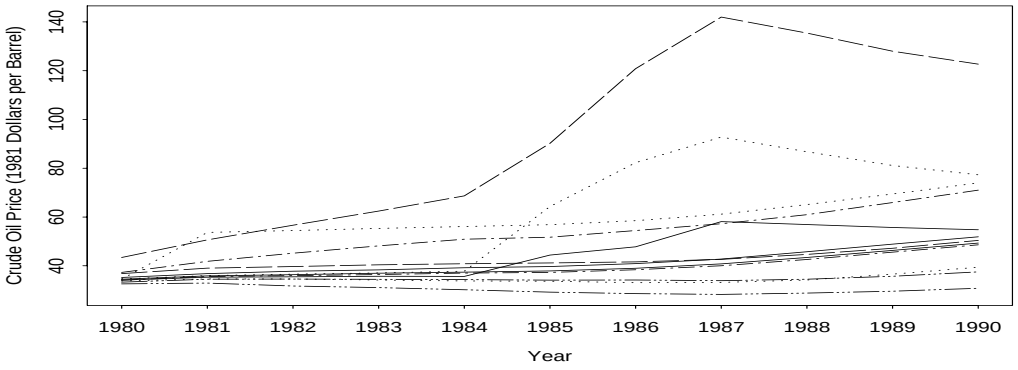

Fig. 3. Scenario-specific forecasts obtained by averaging across models, giving them equal weight.

6. EXAMPLES

6.1. *Predicting Oil Prices*

Continuing the example of Section 2.2, what may be said about the likely price of oil in 1986 (say) from the vantage point of 1980, when scenario and prediction uncertainty are accounted for? Fig. 3 plots the $s = 12$ scenario-specific time series of point predictions from 1980 to 1990 obtained by averaging across the $m = 10$ econometric models described previously, with equal weights $(\lambda_1, \ldots, \lambda_m) = (.1, \ldots, .1)$. With i indexing scenarios and j econometric models, most 1986 forecasts $\hat{y}_{ij}$ ranged from about \$30–60 per barrel, with the exception of those based on two scenarios

(numbered 7 and 9 in Table 1 below) incorporating a large and sudden drop in oil production capacity by the Organization of Petroleum Exporting Countries (OPEC) in the mid-1980s.

Table 1 gives the scenario-specific means $\bar{y}_i = \sum_{j=1}^{m} \lambda_j \hat{y}_{ij}$ and standard deviations ($\hat{\sigma}_i = [\sum_{j=1}^{m} \lambda_j (\hat{y}_{ij} - \bar{y}_i)^2]^{1/2}$) for 1986, together with scenario descriptors and a probability assessment $(\pi_1, \ldots, \pi_s)$ based on how many nonstandard conditions (relative to the "reference" scenario) must occur simultaneously to produce each scenario. Other probability specifications I examined, ranging as far away from that in Table 1 as $\pi = (.2, .1, .05, .05, .1, .1, .05, .1, .05, .1, .05, .05)$ and $(.49, .06, .06, .03, .06, .06, .03, .06, .03, .03, .03, .06)$, yielded conclusions qualitatively similar to those presented here.

TABLE 1

Scenario-specific summaries of the oil price data.

Scenario (i)	Mean ($\bar{y}_i$)	SD ($\hat{\sigma}_i$)	Probability (π_i)
1. Reference	\$39	\$8	.32
2. Oil demand reduction	33	8	.08
3. Low demand elasticity	54	22	.08
4. Combination of 2 and 3	42	16	.04
5. Low economic growth	34	7	.08
6. Restricted backstop	41	9	.08
7. Drop in OPEC production	82	44	.04
8. Technological breakthrough	38	7	.08
9. Combination of 3 and 7	121	67	.04
10. Optimistic	29	5	.04
11. Combination of 2 and 7	48	11	.04
12. High oil price	59	12	.08

Notes: Restricted backstop = slow growth of alternative energy sources;
"Optimistic" combines scenarios 2 and 8, plus the assumption of expanded OPEC capacity.

Attempting to go beyond the implied uncertainty assessment in Figs. 1 and 2 requires acknowledging three levels of uncertainty: (1) between scenarios, (2) between models within scenarios, and (3) between predictions within models and scenarios. With y as the actual 1986 oil price, x as the means and SDs in Table 1, and σ_{ij}^2 as the predictive variance conditional on scenario and model, the analogue of equation (12) in this case (with M standing for econometric model and S for scenario) is

$$\begin{aligned} E(y|x,\mathcal{S}) &= E_S\{E_M[E(y|x,M,S)]\} = \sum_{i=1}^{s} \pi_i \bar{y}_i \equiv \bar{y}, \\ V(y|x,\mathcal{S}) &= (1) + (2) + (3) \\ &= V_S\{E_M[E(y|x,M,S)]\} + E_S\{V_M[E(y|x,M,S)]\} + \\ &\quad E_S\{E_M[V(y|x,M,S)]\} \\ &= \sum_{i=1}^{s} \pi_i (\bar{y}_i - \bar{y})^2 \; + \; \sum_{i=1}^{s} \pi_i \hat{\sigma}_i^2 \; + \; \sum_{i=1}^{s} \pi_i \sum_{j=1}^{m} \lambda_j \sigma_{ij}^2. \end{aligned} \tag{13}$$

EMF made no attempt to assess the predictive SDs σ_{ij}. I have chosen values of the form $\sigma_{ij} = c\,\hat{y}_{ij}$ for small to moderate c, in the range (.05,.3). To obtain a

composite predictive distribution for y I simulated $n_{ij} = 100000\,\pi_i\,\lambda_j$ Gaussian random variates with mean $\hat{y}_{ij}$ and SD σ_{ij} and merged the resulting sample of 100,000 values together. The solid curve in Fig. 4 is a density trace for a typical result with $c = 0.25$; this may be compared with the density (dotted line) implied by an analysis of the type examined in Section 2.2, which conditions on the reference scenario and ignores predictive uncertainty. The mean of the solid curve in Fig. 4 is about \$46, with an SD of about \$30, and the (.01,.05,.5,.95,.99) quantiles are approximately (\$14,\$20,\$39,\$92,\$187). The variance of this distribution (895) decomposes into the three terms (scenario, model, prediction) = (354, 363, 178), so that a model uncertainty audit on the variance scale would attribute about 40% of the overall uncertainty to variation across scenarios, 40% to variation across econometric models given scenario, and 20% to predictive uncertainty given model and scenario. Only the second of these terms is present in Figs. 1 and 2.

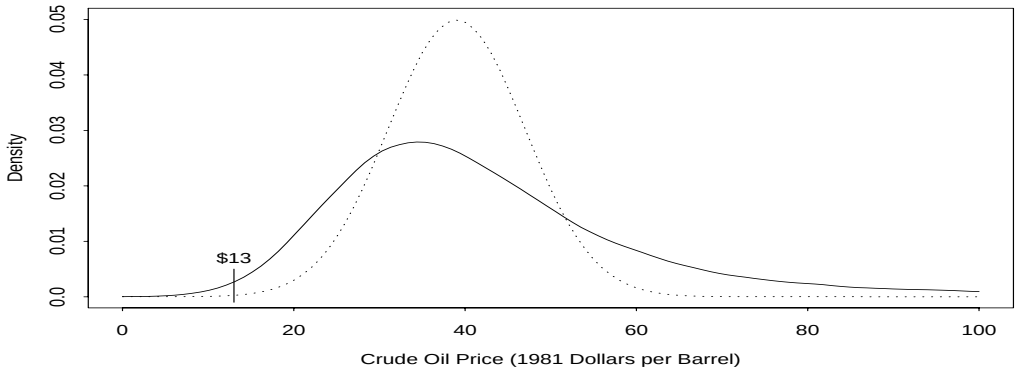

Fig. 4. Density of simulated predictive distribution for 1986 oil price, including scenario, model, and prediction uncertainty (solid curve). Dotted density conditions on the reference scenario and ignores predictive uncertainty.

The actual 1986 oil price of about \$13 is unlikely given the assessment presented here—for example, the ratio of the predictive density at \$13 to its maximum value (at about \$37) is about 1/18. But \$13 is by no means out of the question in the context of this assessment, as it was in the informal assessments of those making decisions on the basis of an implied uncertainty band of (\$27,\$51). If decision-makers had been basing their policies and business choices on something like Fig. 4 instead of Fig. 1, a great deal more hedging against uncertainty would have been built into their actions, and there was nothing to prevent this retrospectively happier outcome: all of the information needed to carry out this analysis was available in 1980.

6.2. *The* Challenger *Space Shuttle Disaster*

On January 28, 1986, the U.S. space shuttle *Challenger* exploded shortly after takeoff, leading to an intensive investigation of the reliability of the shuttle's propulsion system. The explosion was eventually traced to the failure of one of the three *field joints* on one of the two solid booster rockets. Each of these six field joints includes two *O-rings*, designated as primary and secondary, which fail when phenomena called *erosion* and *blowby* both occur.

The night before the launch a decision had to be made regarding launch safety. The discussion among engineers and managers leading to this decision included concern that the probability of failure of the O-rings depended on the temperature t at launch, which was forecast to be 31°F. There are strong engineering reasons based on the composition of O-rings, which are made of rubber, to support the judgment that failure probability may rise monotonically as temperature drops. One other variable, the pressure s at which safety testing for field joint leaks was performed, was available, but its relevance to the failure process was unclear.

Dalal, Fowlkes, and Hoadley (1989, hereafter DFH) performed an extensive risk analysis of the *Challenger*'s field joint system, restricting themselves to data available on the night before the launch. A key step in that analysis was the assessment of the probability p_t^a of primary O-ring erosion at $t = 31°$. Fig. 5 is a plot of the number of field joints experiencing primary O-ring erosion, as a function of launch temperature, on each of the 23 shuttle flights previous to the *Challenger*'s. It may be seen that the shuttle had never flown at a temperature lower than 53°, so that the assessment of the unknown $y = p_{31}^a$ requires considerable extrapolation from the body of existing data. DFH presented a lucid analysis of the data relevant to p_t^a employing the S^* modeling approach of Section 1, and concluded—after relating p_{31}^a to the overall probability of catastrophic failure of the shuttle—that it should have been possible from the available data to foresee the unacceptably high risk created by launching at 31°. Here I offer a reanalysis of these data that focuses on model uncertainty, without (for reasons of space) bringing in the important ingredient of utility. For related alternative analyses see Lavine (1991), who does touch on utility, and Martz and Zimmer (1992).

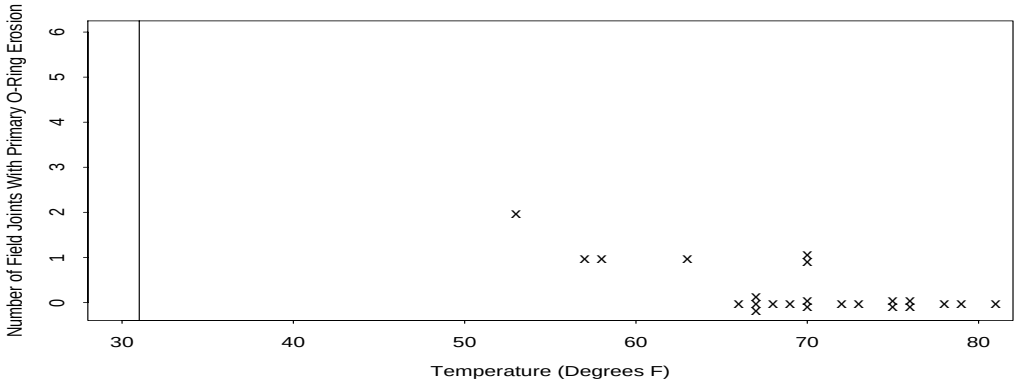

Fig. 5. Scatterplot of number of field joints with primary O-ring erosion versus launch temperature for the 23 shuttle flights prior to the *Challenger*.

In DFH's model field joint failures were independent, both between and within shuttle flights, so that one may regard the data x as consisting of $n = 6 \cdot 23 = 138$ binary failure observations, together with the associated values of temperature t and leak-check pressure s (see Table 1 in DFH for the raw data values). DFH noted (a) that failure probability did not seem to be strongly related to s and (b) that a logistic regression of primary O-ring erosion against temperature t, entered linearly in the model, fits the observed data of Fig. 5 well. After a thorough sensitivity analysis examining alternative models, DFH conditioned on the logistic structural

choice (with linear t and no s) to estimate p_t^a, and assessed uncertainty at 31° with a parametric bootstrap. They obtained a posterior distribution for p_{31}^a given x (see Fig. 8 below) that was well approximated by a beta distribution with parameters $\alpha = 2.52$ and $\beta = 0.36$.

This distribution has a median of .95, a mean of .88, and a variance of .028, and is equivalent in information content to $\alpha + \beta = 2.52 + 0.36 \doteq 3$ binary field-joint failure observations at 31°, an assessment that seems to understate extrapolation uncertainty. Lavine (1991) arrived at a similar judgment; by examining the extrapolated estimates of p_{31}^a based on link functions other than the logit, and by using a nonparametric method that assumes little more than independence of the binary failure outcomes and monotonicity of the relationship between temperature and failure probability, he obtained much wider implied uncertainty bands for p_{31}^a than those produced by DFH's logistic formulation.

An examination of DFH's sensitivity analysis reveals that the following structural variations S_i are good candidates for inclusion in a discrete model expansion:

- Three link functions—logit, probit, and complementary log-log;
- Three functional forms for the temperature variable t—linear, quadratic, and no temperature effect at all, which was a conclusion favored by some involved in the *Challenger* decision-making process; and
- Two functional forms for leak-check pressure s—linear or no effect.

The $m = 6$ structures $\mathcal{S} = \{$cloglog-t, logit-t, probit-t, logit-(t, s), logit-(t, t^2), no effect$\}$ span most of the model uncertainty implied by this list of structural variations. I will use this set of S_i in what follows. Continuous model expansion from DFH's S^* logit-t choice—by embedding the logit in a parametric family of link functions (e.g., Taylor, 1988)—yields results similar to those presented here.

The models in $\mathcal{S}$ all have the same generalized-linear-model structure,

$$(x_j|\theta_i, S_i) \overset{\text{indep}}{\sim} B(p_j), \quad F_i^{-1}(p_j) = t'_{ij}\theta_i, \quad j = 1, \ldots, n, \tag{14}$$

where t_{ij} is the vector of predictor values for observation j assuming structure S_i. With diffuse prior information about the θ_i, Zellner and Rossi (1984) have shown that the required conditional posterior distributions $p(y|x, S_i)$ in this case are given approximately by

$$p(p_t^a|x, S_i) \doteq (2\pi\hat{\phi}_i^2)^{-1/2} e^{-\frac{1}{2\hat{\phi}_i^2}[F_i^{-1}(p_t^a) - t_i'\hat{\theta}_i]^2} \left| \frac{d}{dp_t^a} F_i^{-1}(p_t^a) \right|, \tag{15}$$

where $\hat{\theta}_i$ and $\hat{I}_i$ are the MLE and observed information matrix for structure S_i, $\hat{\phi}_i^2 = t_i'\hat{I}_i^{-1}t_i$, and t_i is the vector of predictors corresponding, under structural choice S_i, to a new temperature t. These conditional densities are well approximated by beta distributions obtained by equating moments. Fig. 6 plots the six densities $\{p(p_{31}^a|x, S_i), S_i \in \mathcal{S}\}$, which differ substantially in both center and spread.

Table 2 presents the results of a discrete model expansion, using equal prior probabilities on the S_i and employing approximation (11) to compute the posterior structural probabilities $p(S_i|x)$. (Changing from approximation (10) to (11), with

and without the $\frac{1}{2} k_i \ln(2\pi)$ term, produces differences in the composite posterior distribution of the same order of magnitude as variations in the prior on $\mathcal{S}$ differing from constant $p(S_i)$ multiplicatively by a factor of 2 in any component, and all of these choices yield conclusions qualitatively similar to those given below.) Fig. 7 plots the expected number of field joints with primary O-ring erosion, conditional on each of the structural choices in $\mathcal{S}$ (cf. Fig. 1 in Lavine, 1991, which motivated the model uncertainty analysis presented here). It may be seen that, with the exception of the no-effect horizontal line, the expected-value traces in Fig. 7 all fit the data well in the observed range—in fact they are virtually coincidental throughout that range—but the various structural assumptions in $\mathcal{S}$ lead to quite different extrapolations at 31°.

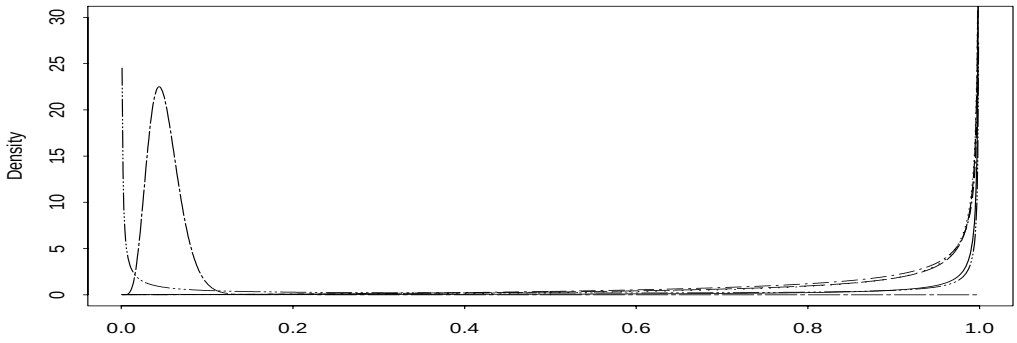

Fig. 6. Conditional posterior distributions $p(p^a_{31}|x, S_i)$ for the six structural choices in $\mathcal{S}$.

TABLE 2
Discrete model expansion results for the Challenger *data.*

	$p(p^a_{31}\|x, S_i)$					
S_i	α	β	Mean	Median	Variance	$p(S_i\|x)$
cloglog-t	2.0	.06	.971	1.0	.009	.282
logit-t	2.66	.294	.900	.96	.0227	.286
probit-t	2.40	.410	.854	.93	.0327	.300
logit-(t, s)	2.17	.302	.878	.95	.0307	.064
logit-(t, t^2)	.116	.1	.537	.69	.204	.063
no effect	7.0	131.	.051	.05	.0003	.005
composite	1.11	.155	.88	.98	.0473	—

The posterior structural distribution (the last column in Table 2) differs considerably from {point mass on logit-t}, the implicit result of DFH's S^*-style analysis: the assumption of no temperature effect is sharply discredited by the evidence, but all five of the other structural choices are sufficiently plausible in light of the data to deserve inclusion in the overall uncertainty assessment for p^a_{31}. The composite posterior distribution $p(p^a_{31}|x, \mathcal{S})$ (see Fig. 8) is well approximated by a beta distribution with parameters 1.11 and 0.155; this distribution has median .98, mean .88, and variance $V_{\text{within-structure}} + V_{\text{between-structure}} = .0338 + .0135 = .0473$, more than

twice the value conditional on the logit-t model (here $V_{\text{between-structure}}$ is about 30% of the total). The resulting assessment of p_{31}^a has about the same mean as DFH's result but includes considerably more uncertainty: $p(p_{31}^a|x,\mathcal{S})$ is equivalent to only about 1 binary observation at 31°, an implied information content 56% smaller than DFH's value, and the 90% central interval for p_{31}^a based on the discrete model expansion runs from .33 to 1, as compared with DFH's interval (.5,1).

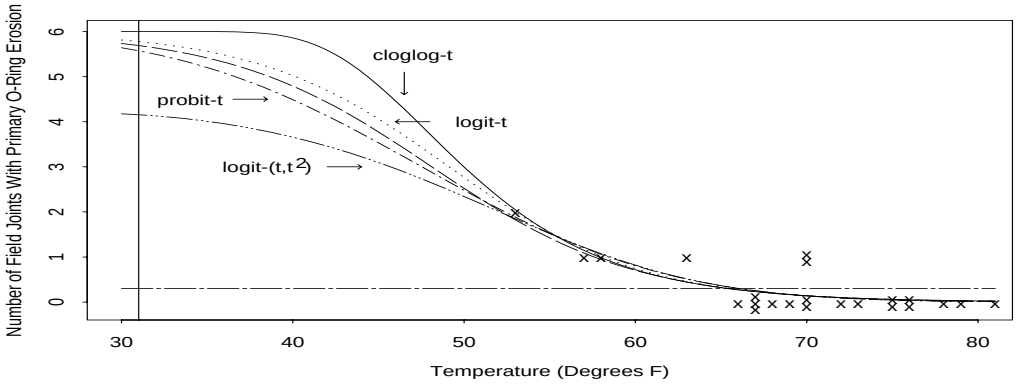

Fig. 7. Expected number of field joints with primary O-ring erosion, conditional on each of the structural choices in $\mathcal{S}$.

The model uncertainty audit presented here is not the only possible analysis of these data; for instance, $V(p_{31}^a|x,\mathcal{S})$ could easily increase somewhat more if more structures S_i were to receive nonzero prior probability. This possibility raises the following question: In the limit as more and more model uncertainty is acknowledged, won't the composite posterior distribution degenerate to $beta(0,0)$, i.e., no information at all at 31°? The answer is no; the available engineering judgment on the monotonicity of p_t^a in t, and the data in Fig. 5 that support this judgment, would together imply an informative distribution like the one presented here if other variations on the monotone theme were included in the model expansion (cf. Lavine, 1991, whose analysis conditioning only on independence and monotonicity resulted in a nonparametric MLE for p_{31}^a of (.33,1)).

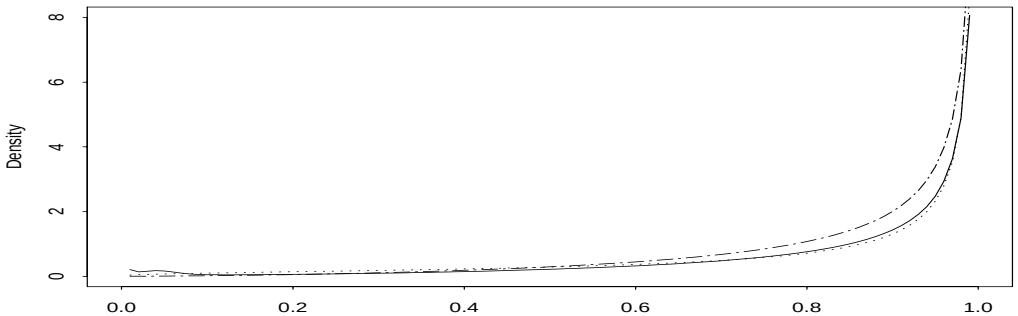

Fig. 8. Posterior distributions $p(p_{31}^a|x,\mathcal{S})$: dashed line is DFH result, solid line is exact result from discrete model expansion (eqn. (6)), dotted line is beta approximation to (6).

Note that in this problem the results of the discrete model expansion only serve to reinforce DFH's overall conclusion: it turns out that for any acceptably small risk r, the posterior distribution for p_{31}, the probability of overall catastrophic failure (not just primary O-ring erosion), concentrates even more of its mass on the interval $(r, 1)$ when the extra structural uncertainty is taken into account. This need not have been so: as the oil price example shows, one may arrive at different substantive conclusions about what constitutes a sensible decision after model expansion than before. Note also that *the good fit of the logit-t model did not imply that model expansion was not needed*—the identification of a single model that fits well does not preclude the possibility of other models, with different inferential or predictive consequences, fitting equally well or better.

7. DISCUSSION

Accuracy and Calibration. Much of statistical theory and practice emphasizes the value of *accurate* inferences and predictions, where accurate means "likely to be close to the truth" in some sense. However, as Dawid (1984, 1985), Hodges (1987), and others have noted, to be fully useful an inference or prediction must also have an uncertainty assessment attached to it, and it is also important for this "give-or-take" to be accurate, because otherwise choices are made that incorporate too little or too much hedging against one's actual uncertainty. Thus *calibration* is also a goal in successful inference and prediction. These two goals compete: by making sufficiently strong modeling assumptions one may easily produce narrow intervals that look good on accuracy grounds, but of what use are they if they consistently miss the truth?

The majority of statistical theory has focused on a kind of *conditional calibration*, in which one makes a set of modeling assumptions M and then figures out how to maximize accuracy subject to calibration constraints given M. This approach is purely deductive: if M is true then the interval (A, B) (say) is the best answer one can obtain. The problem is that if the particular set of modeling assumptions chosen to produce one's intervals turns out in retrospect not to have been correct, it does not necessarily help much to have verified that one's inferences assuming M is true were conditionally accurate and well calibrated. This makes choosing a single M upon which to condition seem like a bad idea.

As the discussion in Section 3 indicates, however, the space $\mathcal{M}$ of all possible models relating knowns x to unknowns y is too big to avoid conditioning on a subset $\mathcal{M}'$ of it. The inability of the data—when the prior distribution on $\mathcal{M}$ is specified too diffusely—to reliably identify which modeling assumptions will retrospectively be seen to be correct argues for making this subset small, but too small runs the risk of poor calibration (e.g., Lindley, 1982). In the oil price example of Sections 2.2 and 6.1, for instance, what decision-makers wanted was the likely price of oil taking all relevant forms of uncertainty into account, not the likely price of oil given that the reference scenario would come to pass. Model expansion permits additional forms of structural uncertainty, whose qualitative treatment in the past has not always led to good decision-making, to enter the probabilistic calculations quantitatively, in effect by permitting more realistic choices of $\mathcal{M}'$. This can lead to decisions based on better-calibrated uncertainty assessments.

Alternative Approaches. There are a variety of techniques for dealing with model uncertainty that differ in spirit or implementation from the approach presented here, for instance *robustness* methods based on solving a minimax problem

over a neighborhood of S^* in model space rather than integrating over such a neighborhood (e.g., Huber, 1981), or Bayesian sensitivity analyses examining the mapping from prior to posterior across a class of prior distributions or likelihoods (e.g., Berger and Berliner, 1986); *nonparametric* methods (e.g., Lehmann, 1975; Friedman, 1991); *data-analytic* methods based on transformations and diagnostics (e.g., Carroll and Ruppert, 1988); and other approaches, including empirical forecast error distributions (Williams and Goodman, 1971). I have argued here that the S^* approach, which may be thought of as a naive data-analytic method, is often inferior to model expansion, but beyond remarks of this type—and theoretical criticism of most of the other methods on, e.g., coherence grounds—little is known about the comparative merits of these various strategies empirically. Theory and case studies closing this gap would have important practical implications.

The Value of Calibration Assessment. The proportion of inferential and predictive applications in which an attempt is actually made to assess calibration, by direct comparison of one's uncertainty assessment for the unknown y with the actual value of y, appears to be fairly low (a notable exception is in weather forecasting; see, e.g., Dawid, 1986). In some applications the actual value is difficult or impossible to observe, making such comparisons problematic, but in many cases it is both possible and desirable to check one's calibration in this way. The ease with which instances of understated uncertainty like those in Section 6 may be found, particularly in situations where substantial extrapolation from the body of available data is necessary for decision-making, makes plausible the speculation that empirical work of a statistical nature would be improved by an increase in calibration activity (see, e.g., Shlyakhter and Kammen, 1992, for a catalogue of appallingly bad uncertainty assessments in physics, energy policy, and demography). Such an increase would be nontrivial, requiring the explicit setting aside of study resources that would have been used in some other way, but it would seem that the long-term benefits of investment in calibration-monitoring would often outweigh the costs. Examples in which this cost-benefit tradeoff is formalized would be useful.

Presentation of Structural Uncertainty. At a minimum consumers of analyses like those in Section 6 need to be able to examine the conditional inferential/predictive distributions (e.g., Fig. 6) and the posterior structural probabilities (e.g., Table 2), so that they may decide for themselves if the composite result is sensible. The already pressing need for a software system that encourages the real-time exploration of the mapping from assumptions to conclusions (e.g., Dickey, 1973; Smith et al., 1987) is only heightened by the acknowledgment of structural uncertainty in addition to parametric and predictive uncertainty. One possible solution is provided by XLISPSTAT (Tierney, 1990), which supports graphical displays in which the prior structural probabilities and prior distributions on the parameters may be smoothly varied and the composite result is updated smoothly.

Combining Forecasts. Model expansion may be thought of as a kind of combining of information from the structures over which model uncertainty is propagated. When the goal is prediction this amounts to combining forecasts, an activity with a large literature (e.g., Clemen, 1989; Palm and Zellner, 1992). Much of this work is devoted to constructing a weighted average composite forecast in the hope that the result will have *smaller* uncertainty than any input forecast. Such an outcome would contrast with the findings of Section 6, where overall uncertainty was *greater* than that implied by any single structural choice. It is worth noting that the uncertainty of the composite forecast will be smaller than that of the inputs

only when all the input forecasts are assumed to be unbiased, a situation that clearly does not hold when substantial structural uncertainty is present (cf. Table 1). The situation is identical to that in choosing between fixed-effects models (which assume no bias) and random-effects models (which allow for bias) in meta-analysis (Section 4.2).

The Category "Other." Model expansion is not a panacea; in particular it cannot protect one from the occurrence of something totally unexpected. In the oil price example of Sections 2.2 and 6.1, for instance, how much prior probability should have been placed on a scenario like the OPEC oil embargo of 1973, several years before it occurred? One is tempted by such events to set aside a bit of probability in model space for "other," but how much probability, and where should it be put? This problem has no solution; inference and prediction always involve an assumption of conditional exchangeability of known and unknown quantities at some level of conditioning (e.g., Draper et al., 1993a). Barnard (1988, personal communication) has put the dilemma well:

> "When the time for decision has arrived, we can do no other than suppose we have spanned the set of possibilities; while at the same time we must allow that we may after all be mistaken—by not closing our minds to that possibility, and so dismissing evidence that may present itself later that our assumptions did not encompass the truth. To come to a decision, while retaining receptiveness to evidence that our decision was wrong, is the only rational course."

An Unpleasant (Short-Run) Outcome. A greater acknowledgment of model uncertainty often has the consequence of widening one's uncertainty bands in pursuit of better calibration. Since hedging against uncertainty is hard work, this is an unpopular turn of events, at least in the short run. But, in view of the oil price example, which is worse—widening the bands now, or missing the truth later?

ACKNOWLEDGMENTS

I am grateful to M. A. Aitkin, G. A. Barnard, C. Chatfield, R. D. Cook, A. Davison, A. P. Dawid, W. H. DuMouchel, S. Greenland, J. Hartigan, M. Lavine, E. E. Leamer, D. V. Lindley, D. Madigan, P. McCullagh, F. Mosteller, J. Nelder, J. W. Pratt, A. E. Raftery, S. Sclove, J. Sedransk, A. F. M. Smith, T. Speed, B. D. Spencer, D. J. Spiegelhalter, P. Stark, J. W. Tukey, A. Zellner, and two referees for comments and references; to J. S. Hodges for his contribution to Draper et al. (1987), which motivated some of this work; to S. Greenland and D. V. Lindley for helpful discussions; to the EMF for providing the oil price data, and to A. Rahman for data assistance; and to Working Group (1988), for permission to reprint Figure 17. Membership on this list does not imply agreement with the ideas expressed here, nor are any of these people responsible for any errors that may be present.

REFERENCES

Adams, J. L. (1991). A computer experiment to evaluate regression strategies. *Proc. Amer. Statist. Assoc. Section on Statist. Comp.* Washington, DC: American Statistical Association, 55–62.

Berger, J. and Berliner, M. (1986). Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. *Ann. Statist.*, **14**, 461–486.

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

Box, G. E. P. and Tiao, G. C. (1962). A further look at robustness via Bayes's theorem. *Biometrika*, **49**, 419–432.

Brown, R. V. and Lindley, D. V. (1986). Plural analysis: Multiple approaches to quantitative research. *Theory and Decision*, **20**, 133–154.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* London: Chapman and Hall.

Chatfield, C. (1993). Model uncertainty: A review. Manuscript.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography (with discussion). *Int. J. Forecasting*, **5**, 559–608.

Cox, D. R. (1961). Tests of separate families of hypotheses. *Proc. 4th Berkeley Symp.*, **1**, 105–123.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics: Principles and Examples.* London: Chapman and Hall.

Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: pre-*Challenger* prediction of failure. *J. Amer. Statist. Assoc.*, **84**, 945–957.

Davis, W. W. (1979). Approximate Bayesian predictive distributions and model selection. *J. Amer. Statist. Assoc.*, **74**, 312–317.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *J. Roy. Statist. Soc. A*, **147**, 278–292.

Dawid, A. P. (1985). Calibration-based empirical probability. *Ann. Statist.*, **13**, 1251–1285.

Dawid, A. P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences*, Vol. 7, Kotz, S. and Johnson, N. L., eds. New York: Wiley, 210–218.

Diaconis, P. (1977). Finite forms of de Finetti's theorem on exchangeability. *Synthese*, **36**, 271–281.

Diaconis, P. and Freedman, D. A. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.*, **14**, 1–67.

Dickey, J. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *J. Roy. Statist. Soc. B*, **35**, 285–305.

Draper, D. (1993). A note on the relationship between model uncertainty and inferential/predictive uncertainty. Under revision for *Biometrika*.

Draper, D., Gaver, D. P., Goel, P. K., Greenhouse, J. B., Hedges, L. V., Morris, C. N., Tucker, J., and Waternaux, C. (1993b). *Combining Information: Statistical Issues and Opportunities for Research.* Contemporary Statistics Series, No. 1. American Statistical Association, Alexandria VA.

Draper, D., Hodges, J. S., Leamer, E. E., Morris, C. N., and Rubin, D. B. (1987). *A Research Agenda for Assessment and Propagation of Model Uncertainty.* N–2683–RC, Santa Monica, CA: RAND.

Draper, D., Hodges, J. S., Mallows, C. L., and Pregibon, D. (1993a). Exchangeability and data analysis (with discussion). *J. Roy. Statist. Soc. A*, **156**, 9–37.

DuMouchel, W. H. (1990). Bayesian meta-analysis. In *Statistical Methodology in the Pharmaceutical Sciences*, D. Berry, ed. New York: Marcel-Dekker, 509–529.

DuMouchel, W. H. and Harris, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *J. Amer. Statist. Assoc.*, **78**, 293–315.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Statist.*, **37**, 36–48.

Efron, B. and Morris, C. N. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc*, **68**, 117–130.

Energy Information Administration (1982). *Outlook for World Oil Prices.* Washington DC: U. S. Dept. of Energy.

Energy Modeling Forum (1982). *World Oil: Summary Report.* EMF Report 6, Energy Modeling Forum, Stanford University, Stanford, CA.

Faraway, J. J. (1992). On the cost of data analysis. *J. Comp. Graph. Statist.*, **1**, 215–231.

Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of a two by two contingency table. *J. Amer. Statist. Assoc.*, **65**, 694–701.

de Finetti, B. (1972). *Probability, Induction, and Statistics.* New York: Wiley.

Fisher, R. A. (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Freedman, D. A., Navidi, W., and Peters, S. C. (1986). On the impact of variable selection in fitting regression equations. In *On Model Uncertainty and its Statistical Implications*, T. K. Dijkstra, ed. Berlin: Springer-Verlag, Lecture Notes in Economics and Mathematical Systems, Vol. 307, 1–16.

Friedman, J. J. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

Geisel, M. S. (1974). Bayesian comparisons of simple macroeconomic models. In *Studies in Bayesian Econometrics and Statistics*, Fienberg, S. E. and Zellner, A., eds. New York: North-Holland, 227–256.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.*, **74**, 153–160 (corrigendum **75**, 765, 1980).

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, **88**, 881–889.

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

Greenland, S. (1993). Methods for epidemiologic analyses of multiple exposures: A review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Statist. in Medicine*, **12**, 717–736.

Henrion, M. and Fischhoff, B. (1986). Assessing uncertainty in physical constants. *Am. J. Phys.*, **54**, 791–798.

Hjorth, U. (1989). On model selection in the computer age. *J. Statist. Planning Inf.*, **23**, 101–115.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1985). *Exploring Data Tables, Trends, and Shapes*. New York: Wiley.

Hodges, J. S. (1987). Uncertainty, policy analysis, and statistics (with discussion). *Statist. Sci.*, **3**, 259–291.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Trans. PAMI*, **4**, 99–104.

Kass, R. E. and Raftery, A. E. (1993). Bayes factors and model uncertainty. Manuscript.

Lavine, M. (1988). Prior influence in Bayesian statistics. *J. Amer. Statist. Assoc.*, to appear.

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle O-ring data. *J. Amer. Statist. Assoc.*, **86**, 919–922.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Ann. Statist.*, **20**, 1222–1235.

Leamer, E. E. (1978). *Specification Searches*. New York: Wiley.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Leonard, T. (1982). Comment on "A simple predictive density function" (by M. Lejeune and G. D. Faulkenberry). *J. Amer. Statist. Assoc.*, **77**, 657–658.

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference. *Proc. 4th Berkeley Symp.*, **1**, 453–468.

Lindley, D. V. (1982). The Bayesian approach to statistics. In *Some Recent Advances in Statistics*, J. T. de Oliviera and B. Epstein, eds. London: Academic Press, 65–87.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. B*, **34**, 1–41.

Mackay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, **4**, 415–447.

Madigan, D. and Raftery, A. E. (1992). Model selection and accounting for model uncertainty in graphical models using Occam's window. Technical Report No. 213, Department of Statistics, University of Washington, Seattle WA.

Martz, H. F. and Zimmer, W. J. (1992). The risk of catastrophic failure of the solid rocket boosters on the space shuttle. *Am. Statist.*, **46**, 42–47.

Miller, A. J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

Moulton, B. R. (1991). A Bayesian approach to regression selection and estimation, with application to a price index for radio services. *J. Econometrics*, **49**, 169–193.

Palm, F. C. and Zellner, A. (1992). To combine or not to combine? Issues of combining forecasts. *J. Forecasting*, **11**, 687–701.

Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *J. Amer. Statist. Assoc.*, **79**, 575–583.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, **7**, 163–185.

Racine, A., Grieve, A. P., Fluehler, H. and Smith, A. F. M. (1986). Bayesian methods in practice: Experiences in the pharmaceutical industry (with discussion). *Appl. Statist.*, **35**, 93–150.

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

Raftery, A. E. (1993). GLIB: Bayesian Generalized Linear Modeling. *S-PLUS* function, Statlib, Carnegie-Mellon University.

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.*, **14**, 1080–1100.

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

Sclove, S., Morris, C. N., and Radhakrishnan, R. (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.*, **43**, 1481–1490.

Self, M. and Cheeseman, P. (1987). Bayesian prediction for artificial intelligence. *Proceedings of the Third Workshop on Uncertainty in AI*, Seattle WA, 61–69.

Seltzer, M. H. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *J. Educ. Statist.*, forthcoming.

Shlyakhter, A. I. and Kammen, D. M. (1992). Sea-level rise or fall? *Nature*, **357**, 25.

Skene, A. M., Shaw, J. E. H., and Lee, T. D. (1986). Bayesian modeling and sensitivity analysis. *The Statistician*, **35**, 281–288.

Smith, A. F. M. (1984). Bayesian statistics. *J. Roy. Statist. Soc. A*, **147**, 245–259.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov-chain Monte Carlo methods. *J. Roy. Statist. Soc. B*, **55**, 3–23.

Smith, A. F. M. and Spiegelhalter, D. J. (1981). Bayesian approaches to multivariate structure. In *Interpreting Multivariate Data*, V. Barnett, ed. New York: Wiley, 335–348.

Smith, A. F. M. and West, M. (1983). Monitoring renal transplants: an application of the multiprocess Kalman filter. *Biometrics*, **39**, 867–878.

Smith, A. F. M., Skene, A. M., Shaw, E. H., and Naylor, J. C. (1987). Progress with numerical and graphical methods for practical Bayesian statistics. *The Statistician*, **36**, 75–82.

Spiegelhalter, D. J. (1981). Adaptive inference using a finite mixture model. Ph. D. dissertation, University College London.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. B*, **44**, 377–387.

Stewart, L. (1987). Hierarchical Bayesian analysis using Monte Carlo integration: Computing posterior distributions when there are many possible models. *The Statistician*, **36**, 211–219.

Stewart, L. and Davis, W. W. (1986). Bayesian posterior distributions over sets of possible models with inferences computed by Monte Carlo integration. *The Statistician*, **35**, 175–182.

Syme, J. (1987). *Forecast Models and Policy Analysis: The Case of Oil Prices*. N–2524–RC, Santa Monica, CA: RAND.

Taylor, J. M. G. (1988). The cost of generalizing logistic regression. *J. Amer. Statist. Assoc.*, **83**, 1078–1083.

Taylor, J. M. G. (1989). Models for the HIV infection and AIDS epidemic in the United States. *Statist. Med.*, **8**, 45–58.

Tierney, L. (1990). *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. New York: Wiley.

Wachter, K. W. and Straf, M. L. (1990). *The Future of Meta-Analysis.* New York: Russell Sage Foundation.

Waikar, V. B., Schuurmann, F. J., and Raghunathan, T. E. (1984). On a two-stage shrinkage testimator of the mean of a normal distribution. *Comm. Statist. Th. Meth. A*, **13**, 1901–1913.

Weisberg, S. (1985). *Applied Linear Regression* (Second Edition). New York: Wiley.

West, M. and Harrison, J. (1989). *Bayesian Forecasting and Dynamic Linear Models.* New York: Springer-Verlag.

Williams, W. H. and Goodman, M. L. (1971). A simple method for the construction of empirical confidence limits for economic forecasts. *J. Amer. Statist. Assoc.*, **66**, 752–754.

Wong, G. and Mason, W. (1985). A hierarchical logistic regression model for multilevel analysis. *J. Amer. Statist. Assoc.*, **80**, 513–524.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

Zellner, A. and Rossi, P. E. (1984). Bayesian analysis of dichotomous quantal response models. *J. Econometrics*, **25**, 365–393.