J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 45–97

Assessment and Propagation of Model Uncertainty

By DAVID DRAPER†

University of Bath, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, March 16th, 1994,* Professor V. S. Isham *in the Chair*]

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 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 assess and propagate structural uncertainty fully and may lead to miscalibrated uncertainty assessments about y given x. When miscalibration occurs it will often result in 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 US space shuttle.

Keywords: BAYES FACTORS; CALIBRATION; FORECASTING; HIERARCHICAL MODELS; INFERENCE; MODEL SPECIFICATION; OVERFITTING; 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 the 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 we typically appeal 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) and Hodges (1987)) in two parts as $M=(S, \theta)$, where S represents one or more sets of *structural* assumptions—such

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

 0035-9246/95/57045

as a particular link function in a generalized linear model, or a particular form of heteroscedasticity or time dependence with data that are not independent and identically distributed (IID)—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 by using this notation, but that does not affect the discussion that follows.) Once S has been chosen, θ typically follows 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, Hodges, Mallows and Pregibon (1993)). 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 non-random 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 has been chosen, but it is less routine to acknowlege structural uncertainty about S itself. A widely used approach 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 30 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 $(\hat{y} - y_{\text{actual}})/\widehat{\text{SD}}(\hat{y})$ across one or more such assessments is unacceptably far from (say) $N(0, 1)$). When such miscalibration occurs it often results in anticonservatism: in retrospect one notices that one's uncertainty bands were not sufficiently wide.

1.2. *Overfitting*

This problem, which is often referred to as *overfitting* the available data, is well known but has yet to receive a fully satisfying treatment in statistical research and pedagogy. Most of the leading text-books on applied statistics (e.g. Cox and Snell (1981)) and regression (e.g. Weisberg (1985)) include warnings against overfitting, 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 (1994)) is in time series modelling, 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 modelling and validation data sets—as a partial solution to the overfitting 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 modelling and the validation activities in a stable way. *Bootstrapping the modelling process* (e.g. Efron and Gong (1983))—creating bootstrap copies of the available data, conducting independent modelling activities on each copy and combining the results in a way that is sensitive to the modelling uncertainty thus uncovered—may help, but as yet little is known about the performance of this approach.

2. CONSEQUENCES OF UNACKNOWLEDGED STRUCTURAL UNCERTAINTY

There are many recent references on the degree of overconfidence generated by basing inferences and predictions on the same data set on which the search for structure occurred; see, for example, 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 selection of variables, 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 0.75, 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) selection of variables by 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 data sets—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 selection of variables yielded average nominal values below 0.01. The degree of similarity between some of the most egregious strategies in Adams's experiment and standard text-book prescriptions for empirical regression model building is disquieting.

2.2. *Forecasting 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 (Energy Modeling Forum, 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 concluded, 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 the scenario) about how to translate the inputs into forecasts and *predictive* uncertainty, conditional on the 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 downplay scenario uncertainty informally as well, but model uncertainty conditional on the reference scenario was evident in the report's tables and figures. Fig. 1, 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.

Averaging across models—giving them equal weight, since the EMF summary report treats them even-handedly—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 the 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 \$500000 million were invested worldwide by governments and private companies in the early 1980s on the strength of forecasts

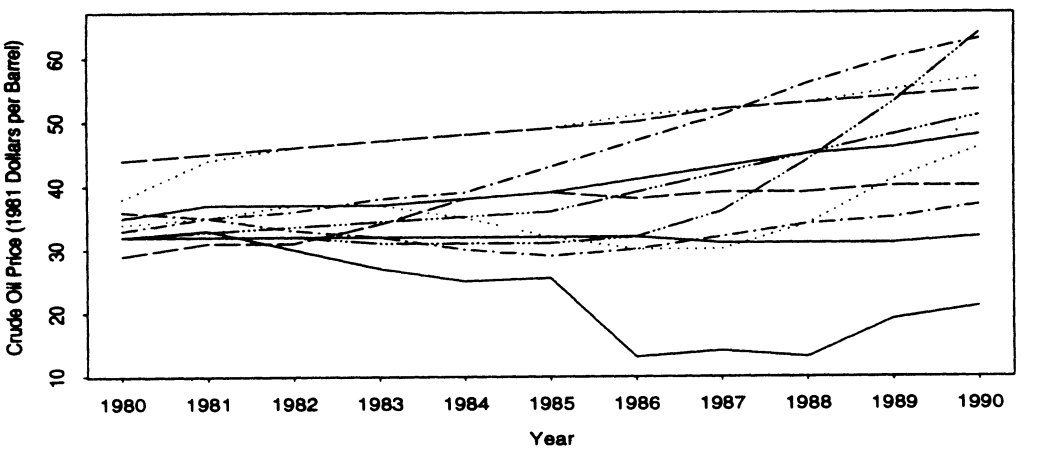

Fig. 1. Forecasts of the price of oil by each of the 10 EMF models under the reference scenario, 1980–90: the lower full curve is the actual price

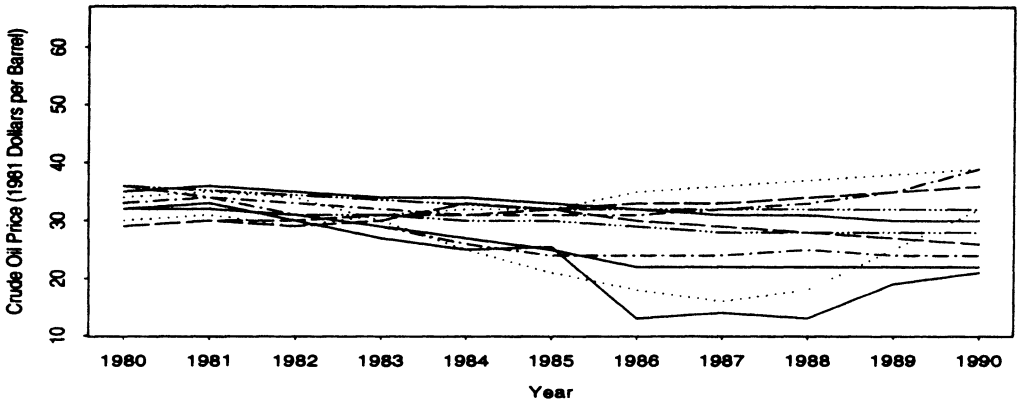

Fig. 2. Forecasts of the price of oil by each of the 10 EMF models under one of the 11 non-reference scenarios, 1980–90: the lower full curve is the actual price

and informal uncertainty assessments like those in Fig. 1. The actual 1986 world average spot price of oil (see the lower full curve 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 happened bore no relation to any of the 12 scenarios that the EMF's working group examined, and people 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 I examine the extent to which assessing and propagating between-scenario and predictive uncertainty improves predictive calibration in this example.

3. STANDARD BAYESIAN SOLUTION

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 to integrate over uncertainty about both S and θ, as in the expression

$$p(y|x, \mathscr{M}') = \int_{\mathscr{M}'} p(y|x, M)p(M|x)\,dM = \int\int p(y|x, \theta, S)p(\theta, S|x)\,d\theta\,dS. \quad (1)$$

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) and Madigan and Raftery (1994)), econometrics (e.g. Geisel (1974) and Leamer (1978))

and artificial intelligence (e.g. Self and Cheeseman (1987) and Mackay (1992)). In the past the implementation of equation (1) in practice has presented major computational challenges, but advances in the last 10 years have greatly reduced this burden. I discuss computational issues in Sections 4 and 5. But first, what should one take for the range of integration $\mathscr{M}'$ 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^*:

$$p(y|x, \mathscr{M}') = p(y|x, S^*) = \int p(y|x, \theta^*, S^*)\,p(\theta^*|x, S^*)\,\mathrm{d}\theta^*. \tag{2}$$

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

$$p(M|x) = c\,p(S)\,p(\theta|S)\,p(x|\theta, S),$$

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 on the S^*-approach to modelling 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, we might consider specifying $p(S)$ much more diffusely, hoping that the updating process from $p(S)$ to $p(S|x)$ would automatically identify plausible modelling 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) and Diaconis (1977) is just a joint probability distribution for the observables, i.e. a single point in the $(2^n - 1)$-dimensional simplex $\{(p_{0\ldots0}, \ldots, p_{1\ldots1}): 0 \leqslant p_{i_1 i_2 \ldots i_n} \leqslant 1, p_{0\ldots0} + \ldots + p_{1\ldots1} = 1\}$. 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 $\mathscr{M}$ 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 non-linear 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. We 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 modelling 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 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) and 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 our 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 modelling* (e.g. Lindley and Smith (1972) and DuMouchel and Harris (1983)), by adding to the top of the hierarchy a level that corresponds to the structural uncertainty: the usual Bayesian formulation

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

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-hand side of formulation (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. *Hierarchical Model for Location Inference*

An early example of continuous model expansion was given by Box and Tiao (1962), who reanalysed 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 modelling the

pairwise differences $x=(x_1, \ldots, x_n)$ to condition on the structural assumptions $\mathscr{S}_0=\{x_i=\mu+\sigma e_i,\ e_i \text{ IID symmetric about } 0\}$, but there is no *a priori* reason to insist on a specific distributional choice for the e_i. Fisher (1935) had previously analysed these data by conditioning on the Gaussian distribution; Box and Tiao (1962) expanded Fisher's model continuously, by embedding the Gaussian distribution in the symmetric power-exponential family $p(e|\alpha)=c\exp(-\frac{1}{2}|e|^{2/(1+\alpha)})$, 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 $\mathscr{S}_1$ (say) as an expansion of $\mathscr{S}_0$, note that the 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 $\mathscr{S}_1$, 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

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

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 our 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).

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 distribution 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, we 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' owing 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 our structural judgments:

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

where the V_i are regarded as known for convenience (typically each x_i is based on a sufficiently large sample of patients that this provides an adequate approximation). Fixed effects models are a special case of equation (5) in which all the θ_i are assumed equal and correspond to random effects models in which the between-study variance parameter τ^2 is set to 0. 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 we plot a time series of the currently accepted value of c with uncertainty bands obtained from the standard fixed effects measurement error model, we notice 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 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, for example, Draper, Gaver, Goel, Greenhouse, Hedges, Morris, Tucker and Waternaux (1993) for other uses of model (5) in physics and chemistry.

4.3. *Computation and Calibration Issues*

Gaussian fixed effects models are easy to fit by 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, for example, Efron and Morris (1973)). Thus practitioners tend to favour fixed effects models when appropriate, so much so that a common modelling 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* that are sometimes used in econometrics (e.g. Waikar *et al.* (1984)). Methods of this type have been shown to be 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, for example, Sclove *et al.* (1972) and Greenland (1993)).

There is a direct analogy between preliminary test methods and the S^*-approach to modelling described in Section 1: in the S^*-approach we search for a single best structure, test its adequacy and adopt 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 modelling 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 10 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 increase greatly the routine feasibility of continuous model expansion in applied work. The hierarchical structure in the right-hand side of formulation (3) is particularly well suited to MCMC methods; see, for example, Seltzer (1994) 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 the following:

(a) *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;
(b) *discrete propagation of scenario uncertainty*, as in the EMF oil example of Section 2.2, in which 12 distinct scenarios meriting non-zero 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 (1962) 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 shall concentrate on the discrete case.

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

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

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

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

(a) 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 (1982) calls Cromwell's rule, which reminds us that any possibility receiving prior probability 0 must also have posterior probability 0. The main way to avoid noticing after the fact that a set of modelling assumptions, different from those originally assumed, turned out to be correct is for one's model prospectively to have been sufficiently large to encompass the retrospective truth. This argues for the routine use of 'big' models. In deciding how big is sufficiently big, we may undertake a kind of preposterior 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.

(b) $\Sigma_{i=1}^{m} p(S_i|x)p(y|x, S_i)$ is intended to be a discrete approximation to

$$p(y|x, \mathscr{M}') = \int_{\mathscr{M}'} p(M|x)p(y|x, M)\,\mathrm{d}M.$$

To improve on the less satisfactory approximation $p(y|x, S^*)$, we can try to include structures S_i' alternative to S^* satisfying two criteria:

(i) S_i' would have high posterior probability $p(S_i'|x)$ (if not given zero prior probability) and
(ii) 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'. This idea may be employed 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 modelling activities on each sample. Also see George and McCulloch (1993), who used Gibbs sampling to produce posterior probabilities for subsets of predictor variables in regression, and Madigan and Raftery (1994), who used 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 of the set $\mathscr{S}$ has been made, the numerical specification of the prior probabilities $p(S_i)$ will also typically be context specific. In situations that are not strongly guided by contextual considerations, we may again proceed by preposterior analysis, e.g. starting with constant $p(S_i)$ and computing forwards with various possible data sets x to see whether the composite result $p(y|x, \mathscr{S})$ appears realistically to 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 we may

(a) choose a variety of points in the past and pretend temporarily that they are the present,
(b) make predictions into the known 'future', building up a history of forecast errors, and
(c) 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 or 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 we would have had to compute these distributions anyway as part of our 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 (especially in prediction problems, where parametric uncertainty on the variance scale is $O(n^{-1})$ but overall prediction uncertainty is $O(1)$). For an example of a more accurate approximation of $p(y|x, S_i)$ see equation (15) later.

5.3. *Computing 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 (1994) for an excellent review. I focus here on two Laplace approximations (e.g. Lindley (1961), Cox (1961), Leonard (1982) and Raftery (1993)), of which the first is

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

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

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

The second and third terms on the right-hand side of equation (11) are recognizable as the basis of the Bayesian information criterion for model selection (Schwarz (1978); also see Rissanen (1986)). The first term on the right-hand 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 that I have examined involving the comparison of structural choices S_i whose θ_i have unequal k_i (see Kashyap (1982)). The main way in general to be sure when n is sufficiently large to use equation (11) instead of equation (10) is to compute them both and to compare them, 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 undefined constants in approximation (10); see, for example, 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 preposterior analysis.

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

To summarize this section, a simple large sample approximation to

$$p(y|x, \mathscr{S}) = \sum_{i=1}^{m} p(S_i|x)\, p(y|x, S_i)$$

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 we may then take

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

and

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

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 $p(S_i|x) = \pi_i$,

$$E(y|x, \mathscr{S}) = E_S[E(y|x, S)] = \sum_{i=1}^{m} \pi_i \mu_i \equiv \mu,$$

$$V(y|x, \mathscr{S}) = E_S[V(y|x, S)] + V_S[E(y|x, S)]$$

$$= \sum_{i=1}^{m} \pi_i \sigma_i^2 + \sum_{i=1}^{m} \pi_i (\mu_i - \mu)^2$$

$$= \begin{pmatrix} \text{within-} \\ \text{structure} \\ \text{variance} \end{pmatrix} + \begin{pmatrix} \text{between-} \\ \text{structure} \\ \text{variance} \end{pmatrix}. \tag{12}$$

The 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.

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 $(\lambda_1, \ldots, \lambda_m) = (0.1, \ldots, 0.1)$. 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) 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 $\bar{y}_i = \Sigma_{j=1}^m \lambda_j \hat{y}_{ij}$ and standard deviations $\hat{o}_i = \{\Sigma_{j=1}^m \lambda_j (\hat{y}_{ij} - \bar{y}_i)^2\}^{1/2}$ for 1986, together with scenario descriptors and a probability assessment $(\pi_1, \ldots, \pi_s)$ based on how many non-standard conditions (relative to the 'reference' scenario) must occur simultaneously to produce each scenario. Other probability specifications that I examined, ranging as far away from that in Table 1 as $\pi = (0.2, 0.1, 0.05, 0.05, 0.1, 0.1, 0.05, 0.1, 0.05, 0.1, 0.05, 0.05)$ and (0.49, 0.06, 0.06, 0.03, 0.06, 0.06, 0.03, 0.06, 0.03, 0.03, 0.03, 0.06), yielded conclusions that are qualitatively similar to those presented here.

Attempting to go beyond the implied uncertainty assessment in Figs 1 and 2 requires acknowledging three levels of uncertainty:

(a) between scenarios,

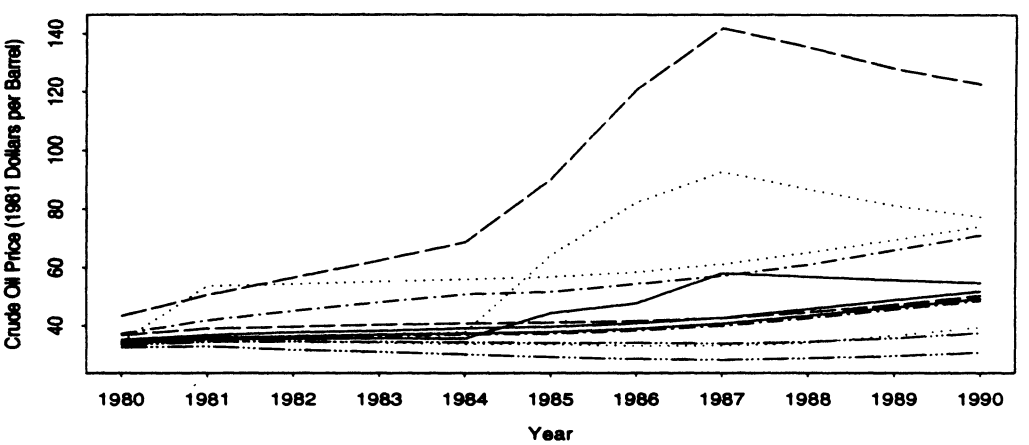

Fig. 3. Scenario-specific forecasts obtained by averaging across models, giving them equal weight

TABLE 1
Scenario-specific summaries of the oil price data†

Scenario i	*Mean* $\bar{y}_i$ *($)*	*Standard deviation* $\hat{\sigma}_i$ *($)*	*Probability* π_i
1, reference	39	8	0.32
2, oil demand reduction	33	8	0.08
3, low demand elasticity	54	22	0.08
4, combination of 2 and 3	42	16	0.04
5, low economic growth	34	7	0.08
6, restricted backstop	41	9	0.08
7, drop in OPEC production	82	44	0.04
8, technological breakthrough	38	7	0.08
9, combination of 3 and 7	121	67	0.04
10, optimistic	29	5	0.04
11, combination of 2 and 7	48	11	0.04
12, high oil price	59	12	0.08

†'Restricted backstop' means slow growth of alternative energy sources; 'optimistic' combines scenarios 2 and 8, plus the assumption of expanded OPEC capacity.

(b) between models within scenarios and
(c) between predictions within models and scenarios.

With y as the actual 1986 oil price, x as the means and standard deviations (SDs) in Table 1 and σ_{ij}^2 as the predictive variance conditional on the scenario and model, the analogue of equation (12) in this case (with M standing for econometric model and S for scenario) is

$$E(y|x, \mathscr{S}) = E_S[E_M\{E(y|x, M, S)\}] = \sum_{i=1}^{s} \pi_i \bar{y}_i \equiv \bar{y},$$

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

The 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 (0.05, 0.3). To obtain a composite predictive distribution for y I simulated $n_{ij} = 100000\,\pi_i\lambda_i$ Gaussian random variates with mean $\hat{y}_{ij}$ and SD σ_{ij} and merged the resulting sample of 100000 values together. The full curve in Fig. 4 is a density trace for a typical result with $c = 0.25$; this may be compared with the density (broken curve) 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 full curve in Fig. 4 is about \$46, with an SD of about \$30, and the (0.01, 0.05, 0.5, 0.95, 0.99) 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

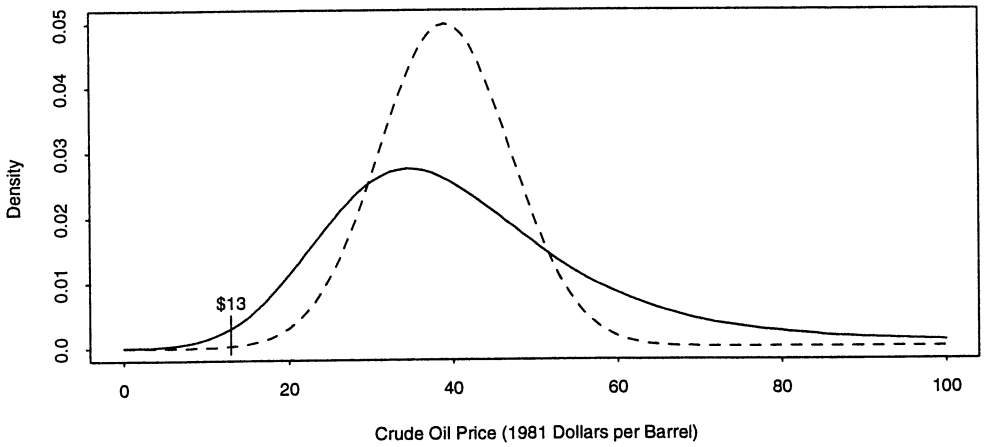

Fig. 4. Density of the simulated predictive distribution for the 1986 price of oil, including scenario, model and prediction uncertainty (———) and conditioning on the reference scenario and ignoring predictive uncertainty (· · · · ·)

given the scenario and 20% to predictive uncertainty given the model and scenario. Only the second of these terms is present in Figs 1 and 2.

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, much more hedging against uncertainty would have been built into their actions, and there was nothing to prevent this retrospectively happier outcome: all the information needed to carry out this analysis was available in 1980.

6.2. *Challenger Space Shuttle Disaster*

On January 28th, 1986, the US space shuttle Challenger exploded shortly after take-off, 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 *blow-by* 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 predicted 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 the probability of failure may rise monotonically as the 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 *et al.* (1989) performed an extensive risk analysis of Challenger's field joint system, restricting themselves to data available on the night before the launch. A key

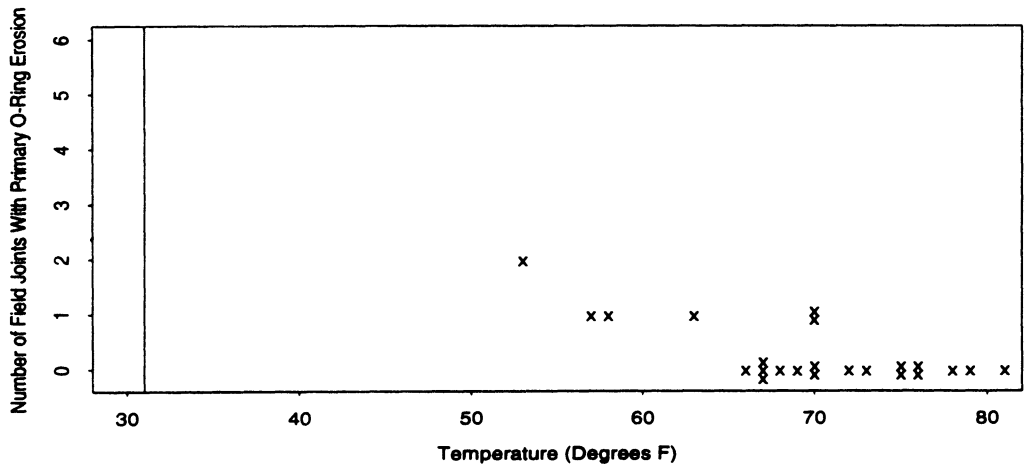

Fig. 5. Scatterplot of number of field joints with primary O-ring erosion *versus* launch temperature for the 23 shuttle flights before the Challenger launch

step in that analysis was the assessment of the probability p_t^a of primary O-ring erosion at $t = 31$ °F. 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 before Challenger's. It may be seen that the shuttle had never flown at a temperature lower than 53 °F, so that the assessment of the unknown $y = p_{31}^a$ requires considerable extrapolation from the body of existing data. Dalal *et al.* (1989) presented a lucid analysis of the data relevant to p_t^a employing the S^* modelling 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 °F. Here I offer a reanalysis of these data that focuses on model uncertainty, without (for brevity) bringing in the important ingredient of utility. For related alternative analyses see Lavine (1991), who does touch on utility, and Martz and Zimmer (1992).

In the model of Dalal *et al.* (1989) field joint failures were independent, both between and within shuttle flights, so that we may regard the data x as consisting of $n = 6 \times 23 = 138$ binary failure observations, together with the associated values of temperature t and leak check pressure s (see Table 1 in Dalal *et al.* (1989) for the raw data values). Dalal *et al.* noted

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

This distribution has a median of 0.95, a mean of 0.88 and a variance of 0.028, and is equivalent in information content to $\alpha + \beta = 2.52 + 0.36 \doteq 3$ binary field joint failure observations at 31 °F, 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 the logistic formulation of Dalal *et al.* (1989).

An examination of the sensitivity analysis of Dalal *et al.* reveals that the following structural variations S_i are good candidates for inclusion in a discrete model expansion:

(a) three link functions—logit, probit and complementary log-log;
(b) three functional forms for the temperature variable t—linear, quadratic and no temperature effect at all, which was a conclusion favoured by some involved in the Challenger decision-making process;
(c) two functional forms for leak check pressure s—linear or no effect.

The $m=6$ structures $\mathscr{S}=\{$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 shall use this set of S_i in what follows. Continuous model expansion from Dalal *et al.*'s (1989) 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 $\mathscr{S}$ all have the same generalized linear model structure,

$$(x_j|\theta_i, S_i) \overset{\text{indep}}{\sim} B(p_j), \qquad F_i^{-1}(p_j)=t_{ij}'\theta_i, \qquad j=1, \ldots, n, \tag{14}$$

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

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

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 $\{p(p_{31}^a|x, S_i), S_i \in \mathscr{S}\}$, which differ substantially in both centre 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 approximation (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 $\mathscr{S}$ differing from constant $p(S_i)$ multiplicatively by a factor of 2 in any component, and all these choices yield conclusions that are 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 $\mathscr{S}$ (see 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

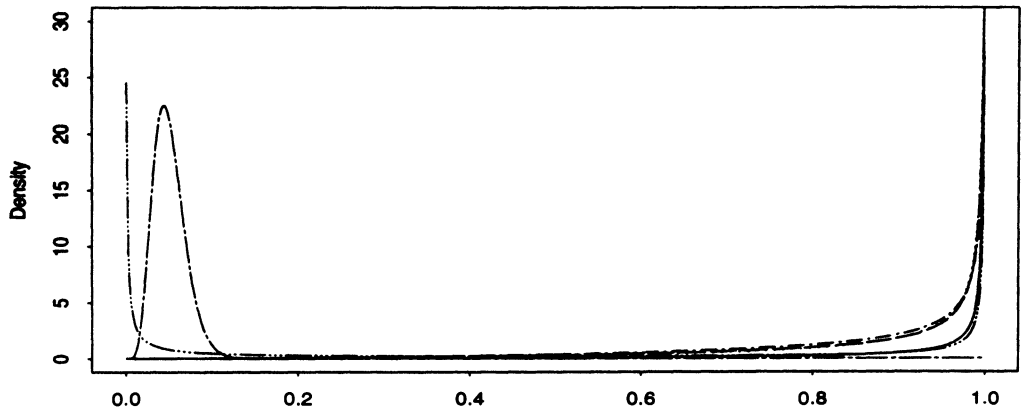

Fig. 6. Conditional posterior distributions $p(p_{31}^a \mid x, S_i)$ for the six structural choices in $\mathscr{S}$

TABLE 2
Discrete model expansion results for the Challenger data

S_i	$p(p_{31}^a \mid x, S_i)$					$p(S_i \mid x)$
	α	β	*Mean*	*Median*	*Variance*	
Cloglog-t	2.0	0.06	0.971	1.0	0.009	0.282
Logit-t	2.66	0.294	0.900	0.96	0.0227	0.286
Probit-t	2.40	0.410	0.854	0.93	0.0327	0.300
Logit-(t, s)	2.17	0.302	0.878	0.95	0.0307	0.064
Logit-(t, t^2)	0.116	0.1	0.537	0.69	0.204	0.063
No effect	7.0	131.	0.051	0.05	0.0003	0.005
Composite	1.11	0.155	0.88	0.98	0.0473	—

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 $\mathscr{S}$ lead to quite different extrapolations at 31 °F.

The posterior structural distribution (the last column in Table 2) differs considerably from {point mass on logit-t}, the implicit result of Dalal *et al.*'s (1989) 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 the light of the data to deserve inclusion in the overall uncertainty assessment for p_{31}^a. The composite posterior distribution $p(p_{31}^a \mid x, \mathscr{S})$ (see Fig. 8) is well approximated by a beta distribution with parameters 1.11 and 0.155; this distribution has median 0.98, mean 0.88 and variance

$$V_{\text{within structure}} + V_{\text{between structure}} = 0.0338 + 0.0135 = 0.0473,$$

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 Dalal *et al.*'s (1989) result but includes considerably more uncertainty: $p(p_{31}^a \mid x, \mathscr{S})$ is equivalent to only about one binary observation at 31 °F, an implied information

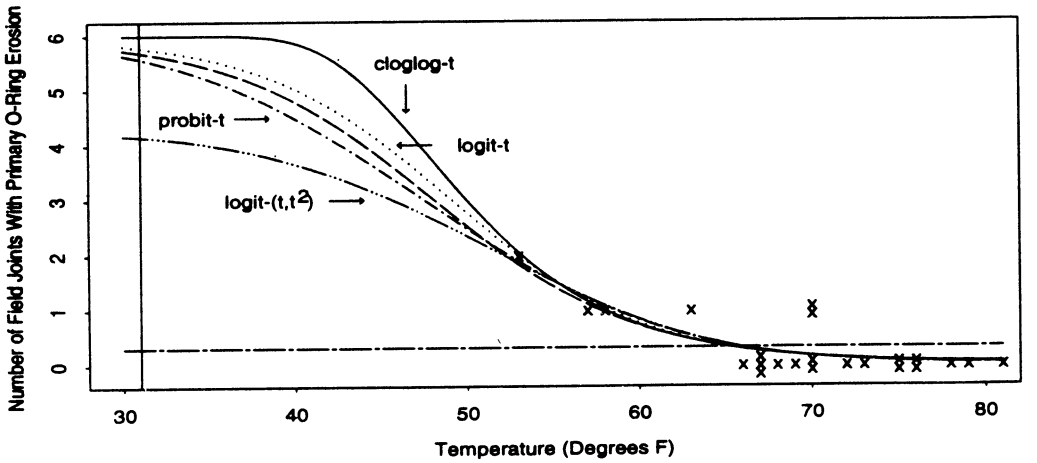

Fig. 7. Expected number of field joints with primary O-ring erosion, conditional on each of the structural choices in $\mathscr{S}$

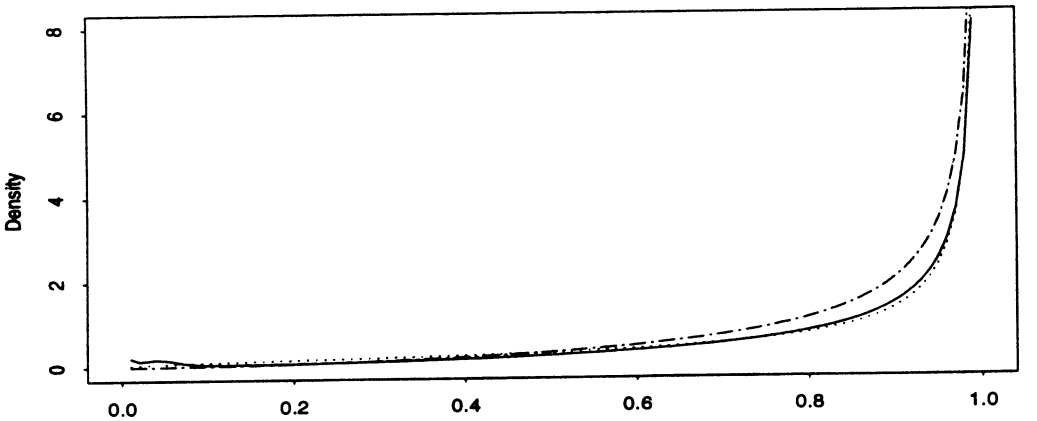

Fig. 8. Posterior distributions $p(p_{31}^a|x, \mathscr{S})$: - · - ·, result of Dalal *et al.* (1989); ———, exact result from the discrete model expansion (equation (6)); · · · · ·, beta approximation to equation (6)

content 56% smaller than Dalal *et al.*'s value, and the 90% central interval for p_{31}^a based on the discrete model expansion runs from 0.33 to 1, compared with their interval (0.5, 1).

The model uncertainty audit presented here is not the only possible analysis of these data; for instance, $V(p_{31}^a|x, \mathscr{S})$ could easily increase somewhat more if more structures S_i were to receive non-zero prior probability. This possibility raises the following question: in the limit as increasingly more model uncertainty is acknowledged, will the composite posterior distribution degenerate to *beta*(0, 0), i.e. no information at all at 31 °F? 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 that presented here if other variations on the monotone theme were included in the model expansion (see Lavine (1991), whose analysis conditioning only on independence and monotonicity resulted in a nonparametric MLE for p_{31}^a of (0.33, 1)).

Note that in this problem the results of the discrete model expansion only reinforce the overall conclusion of Dalal *et al.* (1989): 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, we may arrive at different substantive conclusions about what constitutes a sensible decision after model expansion than before. It is also noteworthy 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 that other models, with different inferential or predictive consequences, fit equally well or better.

7. DISCUSSION

7.1. *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 our actual uncertainty. Thus *calibration* is also a goal in successful inference and prediction. These two goals compete: by making sufficiently strong modelling assumptions we may easily produce narrow intervals that look good on accuracy grounds, but what use are they if they consistently miss the truth?

Most statistical theory has focused on a kind of *conditional calibration*, in which we make a set of modelling assumptions M and then figure 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 that we can obtain. The problem is that, if the particular set of modelling assumptions chosen to produce our intervals turns out in retrospect not to have been correct, it does not necessarily help much to have verified that our inferences assuming that M is true were conditionally accurate and well calibrated. This makes choosing a single M on which to condition seem like a bad idea.

As the discussion in Section 3 indicates, however, the space $\mathscr{M}$ of all possible models relating knowns x to unknowns y is too big to avoid conditioning on a subset $\mathscr{M}'$ of it. The inability of the data—when the prior distribution on $\mathscr{M}$ is specified too diffusely—to identify reliably which modelling 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 $\mathscr{M}'$. This can lead to decisions based on better-calibrated uncertainty assessments.

7.2. *Alternative Approaches*

There are various techniques for dealing with model uncertainty that differ in spirit or implementation from the approach presented here, e.g. *robustness* methods based on solving a minimax problem over a neighbourhood of S^* in model space rather than integrating over such a neighbourhood (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) and 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 naïve 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, for example, 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.

7.3. *Value of Calibration Assessment*

The proportion of inferential and predictive applications in which an attempt is made to assess calibration, by direct comparison of our 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, for example, Dawid (1986)). In some applications the actual value is difficult or impossible to observe, making such comparisons problematical, but in many cases it is both possible and desirable to check our 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, for example, Shlyakhter and Kammen (1992) for a catalogue of appallingly bad uncertainty assessments in physics, energy policy and demography). Such an increase would be non-trivial, requiring the explicit setting aside of study resources that would have been used in some other way, but the long-term benefits of investment in calibration monitoring would often outweigh the costs. Examples in which this cost–benefit trade-off is formalized would be useful.

7.4. *Presentation of Structural Uncertainty*

At a minimum consumers of analyses like those in Section 6 need to be able to examine the conditional inferential or predictive distributions (e.g. Fig. 6) and the posterior structural probabilities (e.g. Table 2), so that they may decide for themselves whether the composite result is sensible. The already pressing need for a software system that encourages the realtime exploration of the mapping from assumptions to conclusions (e.g. Dickey (1973) and Smith *et al.* (1987)) is only heightened by the acknowledgement of structural uncertainty in addition to parametric and predictive uncertainty. A 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.

7.5. *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 of future observables this amounts to combining forecasts, an activity with many references (e.g. Clemen (1989) and 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 a *smaller* uncertainty than any input forecast. Such an outcome would contrast with the findings of Section 6, where the 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 (see Table 1). The situation is identical with 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).

7.6. *Category 'Other'*

Model expansion is not a panacea; in particular it cannot protect us from 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 little 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, Hodges, Mallows and Pregibon (1993)). 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.'

7.7. *Unpleasant (Short Run) Outcome*

A greater acknowledgement of model uncertainty often has the consequence of widening our 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?

ACKNOWLEDGEMENTS

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 contributions 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 assistance with the data. 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. Comput. Statist. Sect. Am. Statist. Ass.*, 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 modelling and robustness (with discussion). *J. R. 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 Decis.*, **20**, 133–154.

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

Chatfield, C. (1994) Model uncertainty, data mining and statistical inference. *Statistics Research Report 94:01*. University of Bath, Bath.

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

Cox, D. R. (1961) Tests of separate families of hypotheses. In *Proc. 4th Berkeley Symp. Mathematical Statistics*, vol. 1, pp. 105–123. Berkeley: University of California Press.

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. Am. Statist. Ass.*, **84**, 945–957.

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

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

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

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

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. R. Statist. Soc.* B, **35**, 285–305.

Draper, D. (1993) A note on the relationship between model uncertainty and inferential/predictive uncertainty. *Statistics Research Report 94:03*. University of Bath, Bath.

Draper, D., Gaver, D. P., Goel, P. K., Greenhouse, J. B., Hedges, L. V., Morris, C. N., Tucker, J. and Waternaux, C. (1993) *Combining Information: Statistical Issues and Opportunities for Research*. Alexandria: American Statistical Association.

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. *Report N-2683-RC*. Rand Corporation, Santa Monica.

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

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

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. Am. Statist. Ass.*, **78**, 293–315.

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

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

Energy Information Administration (1982) *Outlook for World Oil Prices*. Washington DC: US Department of Energy.
Energy Modeling Forum (1982) World oil: summary report. *EMF Report 6*. Energy Modeling Forum, Stanford University, Stanford.
Faraway, J. J. (1992) On the cost of data analysis. *J. Comput. Graph. Statist.*, **1**, 215–231.
Fienberg, S. E. and Gilbert, J. P. (1970) The geometry of a two by two contingency table. *J. Am. Statist. Ass.*, **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. *Lect. Notes Econ. Math. Syst.*, **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* (eds S. E. Fienberg and A. Zellner), pp. 227–256. New York: North-Holland.
Geisser, S. and Eddy, W. F. (1979) A predictive approach to model selection. *J. Am. Statist. Ass.*, **74**, 153–160; corrigendum, **75** (1980), 765.
George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **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. Med.*, **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. Planng 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. Pattn Anal. Mach. Intell.*, **4**, 99–104.
Kass, R. E. and Raftery, A. E. (1994) Bayes factors. *J. Am. Statist. Ass.*, **89**, in the press.
Lavine, M. (1988) Prior influence in Bayesian statistics. *J. Am. Statist. Ass.*, to be published.
——(1991) Problems in extrapolation illustrated with space shuttle O-ring data. *J. Am. Statist. Ass.*, **86**, 919–922.
——(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. Am. Statist. Ass.*, **77**, 657–658.
Lindley, D. V. (1961) The use of prior probability distributions in statistical inference. In *Proc. 4th Berkeley Symp. Mathematical Statistics*, vol. 1, pp. 453–468. Berkeley: University of California Press.
——(1982) The Bayesian approach to statistics. In *Some Recent Advances in Statistics* (eds J. T. de Oliveira and B. Epstein), pp. 65–87. London: Academic Press.
Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**, 1–41.
Mackay, D. J. C. (1992) Bayesian interpolation. *Neural Computn*, **4**, 415–447.
Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Statist. Ass.*, **89**, in the press.
Martz, H. F. and Zimmer, W. J. (1992) The risk of catastrophic failure of the solid rocket boosters on the space shuttle. *Am. Statistn*, **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: Addison-Wesley.
Moulton, B. R. (1991) A Bayesian approach to regression selection and estimation, with application to a price index for radio services. *J. Econometr.*, **49**, 169–193.

Palm, F. C. and Zellner, A. (1992) To combine or not to combine?: issues of combining forecasts. *J. Forecast.*, **11**, 687–701.
Picard, R. R. and Cook, R. D. (1984) Cross-validation of regression models. *J. Am. Statist. Ass.*, **79**, 575–583.
Pötscher, B. M. (1991) Effects of model selection on inference. *Econometr. Theory*, **7**, 163–185.
Racine, A., Grieve, A. P., Flühler, 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 121*. Department of Statistics, University of Washington, Seattle.
——(1993) GLIB: Bayesian generalized linear modeling. *S-PLUS Function*. Statlib, Carnegie Mellon University, Pittsburgh.
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. In *Proc. 3rd Wkshp Uncertainty in Artificial Intelligence, Seattle*, pp. 61–69.
Seltzer, M. H. (1994) Sensitivity analysis for fixed effects in the hierarchical model: a Gibbs sampling approach. *J. Educ. Statist.*, **19**, in the press.
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. *Statistician*, **35**, 281–288.
Smith, A. F. M. (1984) Bayesian statistics. *J. R. 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. R. Statist. Soc.* B, **55**, 3–23.
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. *Statistician*, **36**, 75–82.
Smith, A. F. M. and Spiegelhalter, D. J. (1981) Bayesian approaches to multivariate structure. In *Interpreting Multivariate Data* (ed. V. Barnett), pp. 335–348. New York: Wiley.
Smith, A. F. M. and West, M. (1983) Monitoring renal transplants: an application of the multiprocess Kalman filter. *Biometrics*, **39**, 867–878.
Spiegelhalter, D. J. (1981) Adaptive inference using a finite mixture model. *PhD Thesis*. University College London, London.
Spiegelhalter, D. J. and Smith, A. F. M. (1982) Bayes factors for linear and log-linear models with vague prior information. *J. R. 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. *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. *Statistician*, **35**, 175–182.
Syme, J. (1987) Forecast models and policy analysis: the case of oil prices. *Report N-2524-RC*. Rand Corporation, Santa Monica.
Taylor, J. M. G. (1988) The cost of generalizing logistic regression. *J. Am. Statist. Ass.*, **83**, 1078–1083.
——(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: Sage.
Waikar, V. B., Schuurmann, F. J. and Raghunathan, T. E. (1984) On a two-stage shrinkage testimator of the mean of a normal distribution. *Communs Statist. Theory Meth.*, **13**, 1901–1913.
Weisberg, S. (1985) *Applied Linear Regression*, 2nd edn. New York: Wiley.
West, M. and Harrison, J. (1989) *Bayesian Forecasting and Dynamic Linear Models*. New York: Springer.
Williams, W. H. and Goodman, M. L. (1971) A simple method for the construction of empirical confidence limits for economic forecasts. *J. Am. Statist. Ass.*, **66**, 752–754.
Wong, G. and Mason, W. (1985) A hierarchical logistic regression model for multilevel analysis. *J. Am. Statist. Ass.*, **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. Econometr.*, **25**, 365–393.

DISCUSSION OF THE PAPER BY DRAPER

David J. Spiegelhalter (Medical Research Council Biostatistics Unit, Cambridge): It was recently brought home to me that one should pay attention to between-model as well as within-model uncertainty. I was looking through the annual report of the Director of Public Health of a certain regional health authority and found projections made on the basis of a linear regression, with a predictive estimate of around -12 and uncertainty covering a range from around -28 to 3. It is quite possible that, given the model assumptions, these calculations were immaculately carried out. The fact, however, that these were projections for the 1995 rate of gonorrhoea per 100000 population suggests that some attention might have been paid to extramodel uncertainty.

Dr Draper's paper is an extremely timely contribution to the burgeoning references on serious applications of Bayesian methods. Since, at last, computation is no longer a hindrance to Bayesian methods, it is natural that discussion of Bayesian model criticism and choice has come to greater prominence. One could argue that, had more of such work already been carried out, then we would be better placed to answer the questions of the new generations of statisticians who want to use Bayesian methods for their practical flexibility and care little for theoretical optimality properties.

Section 5 is, to me, the guts of the paper. However, if these techniques are to become routine then certain guidelines may be appropriate in deciding the family $\mathscr{S}$ of structural alternatives. I would like to ask two main questions which, in the usual style of these discussions, I shall then immediately attempt to answer. First, when is it reasonable to select a single model from the family $\mathscr{S}$ being considered, and, second, should $\mathscr{S}$ vary with sample size?

For selected $\mathscr{S}$, Draper's equation (6) shows that inference on a quantity of interest y should be based on

$$p(y \mid x, \mathscr{S}) = \sum_{i=1}^{m} p(y \mid x, S_i)\, p(S_i \mid x).$$

Bernardo (1979) showed that it is reasonable to measure the expected utility of the distribution $p(y \mid x, \mathscr{S})$ by

$$U_{\mathscr{S}}\{p(y \mid x, \mathscr{S})\} = E_{\mathscr{S}}\{\log p(y \mid x, \mathscr{S})\}$$

when the expectation is taken with respect to $p(y \mid x, \mathscr{S})$. Suppose that we wanted to select a single model S_i and hence base our inferences on $p(y \mid x, S_i)$. Then the *fall* in expected utility is

$$U_{\mathscr{S}}\{p(y|x, \mathscr{S})\} - U_{\mathscr{S}}\{p(y|x, S_i)\} = E_{\mathscr{S}}\left\{\frac{\log p(y|x, \mathscr{S})}{\log p(y|x, S_i)}\right\}$$

which can be recognized as the Kullback–Leibler directed divergence between $p(y \mid x, \mathscr{S})$ and $p(y \mid x, S_i)$. This non-negative quantity may be calculated numerically and could be thought of as representing the inadequacy of a single choice S_i.

Some insight into this measure comes from assuming a normal approximation $y \mid x, S_i \sim N(\mu_i, \sigma_i^2)$. Then it is straightforward to show that

$$E_{\mathscr{S}}\{-\log p(y \mid x, S_i)\} = \frac{1}{2}\log(2\pi\sigma_i^2) + \frac{1}{2}\sum_{j=1}^{m} p(S_j \mid x)\left\{\frac{\sigma_j^2}{\sigma_i^2} + \frac{(\mu_j - \mu_i)^2}{\sigma_i^2}\right\}.$$

The larger this quantity the more inadequate is S_i, and hence no single model will be adequate if there is another model S_j that both has reasonable posterior plausibility *and* either has a relatively large associated variance σ_j^2/σ_i^2 or has a very different estimate μ_j from that under S_i. This just formalizes the comments in Section 5.1 that models should be included if they are plausible in the light of the data *and* lead to substantially different conclusions from those following from a 'null' model.

We shall illustrate this idea with location estimation for Darwin's data (a data set that perhaps should be honourably retired after many years' service to statistical science). Fig. 9 shows the posterior distributions that arise from consideration of a family of t-distributions with degrees of freedom 1 (Cauchy), 2, 4, 8, 16 and ∞ (normal). Table 3 shows the posterior weights (calculated by using the approximation in Draper's expression (10)) assuming equal priors of 1/6 for each model, together with the above measure of inadequacy of each single assumption. We note that a t_4-distribution has the highest posterior support, but a narrower tailed t_8-distribution is the most adequate single choice of distribution, since its conditional conclusions are closest to those obtained by using the whole family.

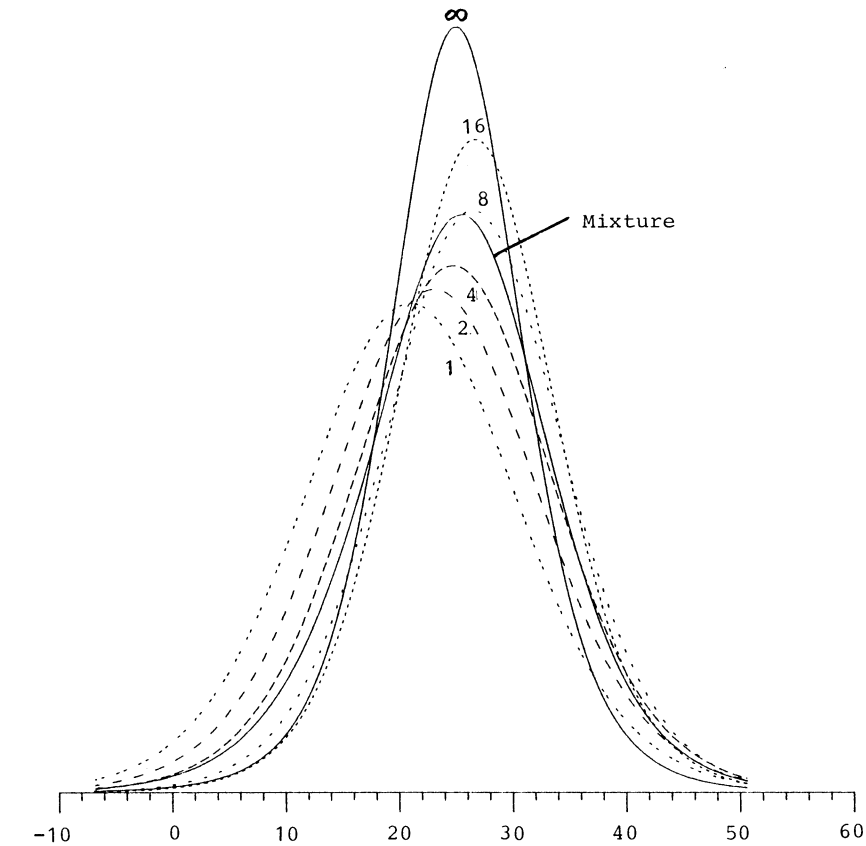

Fig. 9. Posterior distributions for the location parameter under different distributional assumptions for Darwin's data: the annotations indicate the degrees of freedom of each t-distribution, together with the 'mixture' posterior distribution formed by using the posterior weights shown in Table 3

TABLE 3
Alternative assumptions for inference on location parameter for Darwin's data

Distributional shape	*Posterior weight*	*Inadequacy (× 10000)*
Cauchy—t_1	0.11	543
t_2	0.24	385
t_4	0.29	159
t_8	0.18	48
t_{16}	0.11	313
Normal—t_∞	0.06	1009

This warns us not to be overconcerned with probabilities of models, but to pay more attention to their consequences.

My second question concerns the large sample properties of this technique. Berk (1966) showed that under broad conditions, assuming a true model S_0 which does not necessarily lie in $\mathscr{S}$, the model $\hat{S} \in \mathscr{S}$ will receive asymptotic posterior weight 1, and the parameter estimates will converge to $\hat{\theta}$, where $\hat{S}$, $\hat{\theta}$ maximize $E_0\{\log p(x \mid \hat{\theta}, \hat{S})\}$. This leads the ratio of posterior weights $p(S_j \mid x)/p(\hat{S} \mid x)$ to be $O_p(c^n)$ for $S_j \neq \hat{S}$, $c < 1$, and, as those who use Bayes factors are well aware, this entails a sometimes disquieting

tendency to obtain huge Bayes factors and hence all posterior weight plumping onto a single model. All this suggests that the granularity of the model space should crucially depend on sample size, with a few quite distinct models being sufficient for small samples and much finer gradations of model being appropriate for large samples. Thus perhaps Draper's comments on 'staking out the corners of model space' should include the observation that the size of this search area should decrease with n.

Draper has been admirably careful in citing previous work and communicating with many people on this excellent paper. Preparing this discussion led me back to an old draft of a paper, buried in the bowels of my filing cabinet, with a 1980 letter from the *Journal of the Royal Statistical Society*, Series B, suggesting resubmission. I never got round to resubmitting the paper, so I am grateful to Dr Draper for giving me the opportunity to exhume some old work and at last, by the back door, to sneak it into Series B.

It gives me great pleasure to propose the vote of thanks.

A. P. Grieve (ZENECA Pharmaceuticals, Macclesfield): I welcome this paper because it addresses an issue which has for too long been neglected by both the academic and the applied statistical communities. It provides an opportunity for brethren from all inferential persuasions to reassess their own particular strategies for intra-experiment assessment of assumptions. I shall comment on three aspects: pretesting; continuous *versus* discrete model expansion; sensitivity to prior model beliefs.

Many statisticians reading the description of Adams's experiment in Section 2.1 will regard it as a parody of the way that they work—discussants of Racine *et al.* (1986) suggested that 'no competent statistician' would use an essentially similar pretesting strategy. Yet such strategies are not only typical of 'standard text-book prescriptions' they are also redolent of what might be termed a regulatory prescription. As illustration consider stability studies which are carried out to estimate the shelf-life of a drug. In such studies random samples of, for example, tablets are taken from either a single production batch or a series of batches, and are stored under known conditions of temperature and humidity. Periodically a sample is taken and assayed for content. The data are modelled by a regression line and the shelf-life is defined to be the time at which the lower 95% confidence limit about the fitted line crosses the lower limit of acceptable drug content, typically 90%. When data are available from multiple batches an analysis of covariance is carried out to test the null hypothesis of equal regression slopes. The prescription proposed by the Food and Drug Administration (1987) is that the null hypothesis should be tested at the 25% level, thereby ensuring a more sensitive test! The paradox in this prescription is well illustrated in two hypothetical examples given by Ruberg and Stegeman (1991). Studies which are well conducted will generally lead to low variability and are therefore likely to reject small interbatch differences as being statistically significant, whereas for poorly conducted studies the converse is true. Thus this prescriptive approach penalizes good experimentation.

In Section 5 the comment is made that 'it is preferable to perform model expansion continuously'. I am not convinced that this is in general true and in particular I believe that it is often of interest to treat null models discretely rather than as part of a continuum. In Racine *et al.* (1986) we considered the two-period, two-treatment crossover in which interest centres on the treatment effect τ with the carry-over effect λ a nuisance parameter. For known variance parameters the posterior distribution of τ given λ has the form

$$p(\tau \mid \lambda, \sigma^2, \rho, X) \sim N\{\hat{\tau} + \lambda/2, m\sigma^2(1-\rho)/8\}$$

where m is a function of the sample sizes in each randomization group and from which the posterior for τ under the null model $\lambda = 0$ is given by

$$p(\tau \mid \lambda = 0, \sigma^2, \rho, X) \sim N\{\hat{\tau}, m\sigma^2(1-\rho)/8\}.$$

However, equation (4) gives

$$\begin{aligned} p(\tau \mid \sigma^2, \rho, X) &= \int p(\tau \mid \lambda, \sigma^2, \rho, X)\, p(\lambda \mid \sigma^2, \rho, X)\, \mathrm{d}\lambda \\ &\sim N(\hat{\tau} + \hat{\lambda}/2, m\sigma^2/4) \end{aligned}$$

and the variance of this posterior is based on between-patient variability rather than on within-patient variability as is the case for the null model. Therefore the posterior distribution under the null model is structurally very different from the marginal posterior distribution. Indeed if the latter is more

appropriate then it casts doubt on the use of the crossover since its primary advantage is usually thought to be that the treatment effect is estimated within patient. Similar arguments exist for the stability study already described and also for the random effects models of Section 4.2.

The analysis of the Challenger disaster data is a particularly good example of the discrete model expansion approach. However, not all people looking at Table 2 will *a priori* believe that the six models are equally likely and therefore it is advisable to present posterior inferences as a function of prior model beliefs. In Racine *et al.* (1986) it was simple to provide a graphical representation of the relationship between a particular posterior inference $P(\tau>0\,|\,X)$, the posterior probability of a positive treatment effect, and the prior belief in the non-null model, $P(M_1)$. When there are more than two competing models it is no longer so simple to look at the relationship as the following example illustrates.

In Grieve (1994) I have extended the analysis of the two-period two-treatment crossover to include base-line measurements. The use of base-line measurements allows us to disentangle the carry-over effect from period × treatment interaction, which I denote by θ, which are aliased in the previous case. There are now four models of potential interest:

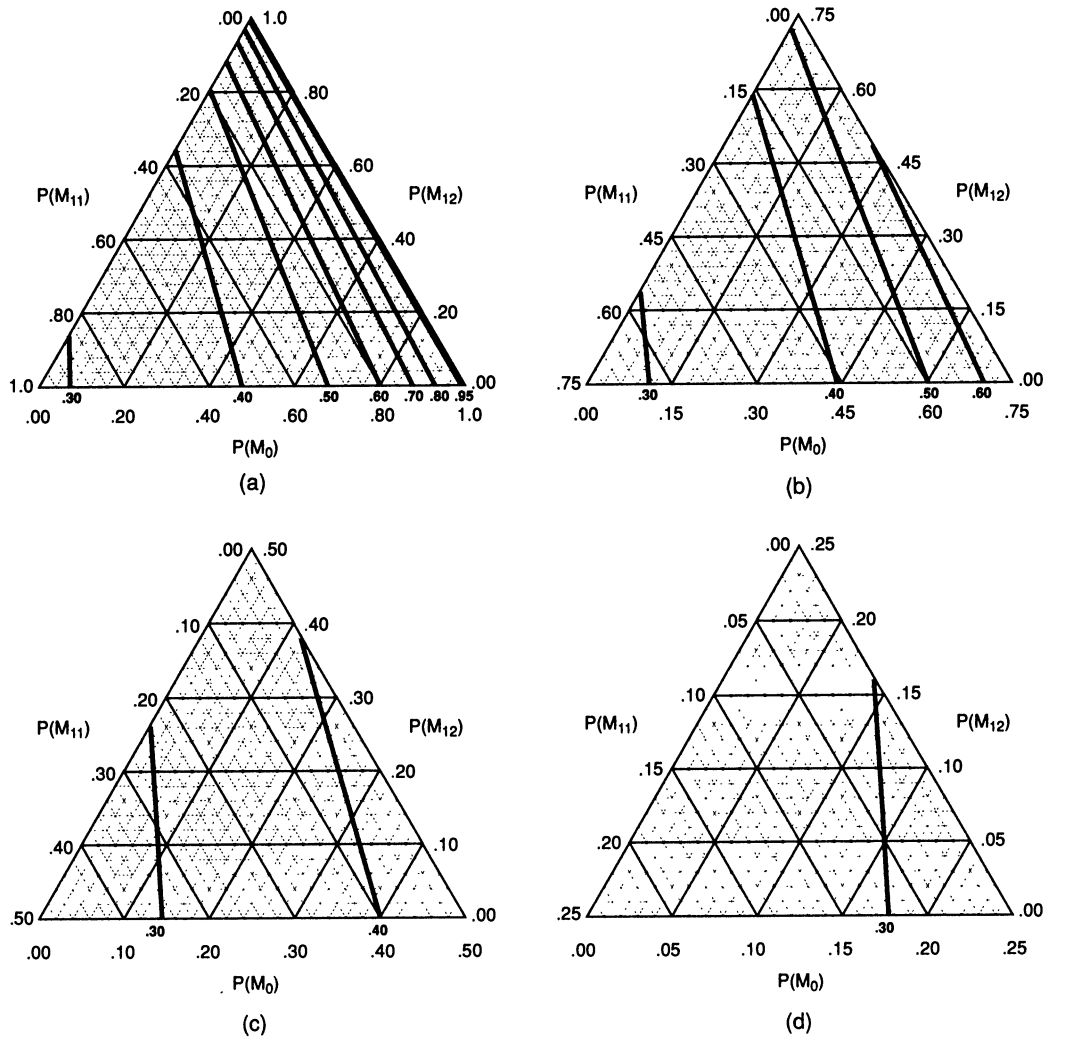

Fig. 10. Posterior probability of a positive treatment effect—$P(\tau>0|X)$—as a function of the prior model beliefs: (a) $P(M_2)=0.00$; (b) $P(M_2)=0.25$; (c) $P(M_2)=0.50$; (d) $P(M_2)=0.75$

M_2: τ, λ, θ (saturated);
M_{11}: τ, $\lambda=0$, θ;
M_{12}: τ, λ, $\theta=0$;
M_0: τ, $\lambda=0$, $\theta=0$.

In Grieve (1994) I developed a graphical method for displaying $P(\tau>0|X)$ on the simplex $P(M_2)+P(M_{11})+P(M_{12})+P(M_0)=1$. An example of this is shown in Fig. 10 which is based on angina attack rate data taken from Nicholls *et al.* (1986) in a study comparing transdermal and oral formulations of nitrate therapy. For a fixed value of $P(M_2)$ contours of equal posterior probability $P(\tau>0|X)$ are displayed on the simplex $P(M_{11})+P(M_{12})+P(M_0)=1-P(M_2)$. The interpretation of the display for these data is that we would need to be *a priori* very certain that there is no period × treatment interaction before we could conclude that there is substantial treatment effect. Given the facilities available in XLISPSTAT we could smoothly vary $P(M_2)$ and update $P(\tau>0|X)$ accordingly.

It gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

D. V. Lindley (Minehead): This admirable paper addresses an important problem and presents a fine method of solution. The examples are integral to the thesis and not merely illustrative.

I shall confine my attention to one aspect, namely overfitting (Section 1.2), illustrating a general argument by reference to multiple regression. There are many published examples where fitting a regression of y on s variables $x_1, \ldots, x_s$ does better than a similar fit involving, in addition, t other variables $x_{s+1}, \ldots, x_{s+t}$, even when the data are plentiful. These practical experiences can be supported by a mathematical argument that calculates the deterioration as the number of variables increases.

However, there is an even simpler argument that demonstrates the opposite. This depends on the easily demonstrated, mathematical theorem that any information is expected to be of value: a result which is supported by intuition. Applying the result, the extra variables $x_{s+1}, \ldots, x_{s+t}$ are expected to be of value in improving the prediction of y. Indeed, this must be so, since, in choosing the best from the class of all regressions using all variables, the one with only s is included by putting the regression coefficients $\beta_i=0$, for $i>s$.

There is direct conflict here. How is it to be resolved? I believe that the resolution lies in the appreciation that the results, both theoretical and practical, that suggest overfitting are all based on likelihood methods and fail to take account of our uncertainties about the roles of the variables. They are based on $p(y|x, \beta)$ and ignore $p(x, \beta)$. They usually use least squares, which can be regarded as either omitting $p(x, \beta)$ or treating it as improper. Outside one and two dimensions, such methods lead to unsatisfactory, technically inadmissible, results, and the dissatisfaction increases with the dimensionality. It is least squares that has produced overfitting, which is not inherent.

I argue that overfitting is not a problem in a world with proper, σ-additive, distributions. These distributions should come from realistic considerations and not from studies of Greek letters, like β above.

J. B. Copas (University of Warwick, Coventry): The paper reminds us of an uncomfortable fact which we know exists, but for a quiet life we would much prefer that it did not, namely that data dependence in an assumed model leads to inferences which are less precise than we pretend them to be. Our Society's new *Code of Professional Conduct* says that we should not knowingly promote misleading inferences from data, and so we should take this whole issue, and this useful paper, very seriously.

George Barnard commented at one of these meetings that we statisticians spend far too much time trying to answer silly questions. To expect any sensible answer when $\mathcal{M}$ is very large is rather silly, as the paper points out. We have to restrict to an assumed subset—an assumption-free inference is an illusion. The very fact that we use probability as the basis for inference, knowing full well that our observational data were not the result of a game of pure chance, is itself an assumption.

Provided that we avoid extrapolations, such as in the alarming space shuttle example, and just want a well-calibrated fit *within the range of our observed data*, then shrinkage methods can sometimes give a practical solution to the problem of overfitting. Perhaps we have binary data y grouped by covariate vector x into counts of s_x successes out of n_x trials. We want to estimate λ_x, the logit of $P(y=1|x)$. The rough estimates are the empirical logits z_x. Simple methods such as point scoring lead to a score T_x which can be calibrated directly on the data to give smooth estimates

$$L_x = \frac{\sum_u w_u z_u I_{T_u = T_x}}{\sum_u w_u I_{T_u = T_x}}$$

where w_x is a set of suitable weights. As in the paper the same data are here being used for both constructing the score and calibrating it.

The plot of L_x against T_x is often remarkably linear: let $C(L, T)$ be the sample covariance of this plot, and let $C(L^*, T)$ be the same quantity but with z_x replaced by logits calculated from an *independent* replication of s_x. The estimates L_x are now shrunk to

$$\tilde{L}_x = \bar{L} + K(L_x - \bar{L})$$

where K is chosen such that $E\{C(\tilde{L}, T) - C(L^*, T)\} = 0$. A simple formula for K is given in Copas (1993).

Andrew J. G. Cairns (Heriot-Watt University, Edinburgh): Firstly I would like to thank the author for writing a most interesting and stimulating paper. While I was reading the paper I identified several points for which I could find direct parallels in my own work on epidemic modelling, particularly of acquired immune deficiency syndrome (AIDS). The problem here is how to fit a complex epidemic model to a relatively simple set of AIDS incidence data. The key I found to solving this problem was to partition the basic parameters ϕ into two sets:

(a) ϕ_P—the (small) set of *primary components* which are functions of the basic parameters which dictate epidemic dynamics;
(b) ϕ_S—the remaining secondary parameters which have little effect on the shape of an epidemic curve if they stay within a realistic range. Such a realistic range may be determined by secondary studies or reflected in a more subjective prior for ϕ_S but still based on such studies. For example we know the *mean* incubation period of the human immunodeficiency virus to be around 10 years rather than 2 or 20.

A number of different problems present themselves of which one is projection of the epidemic curve. In a recent series of papers (Cairns, 1991, 1993, 1994) I argue that in many such contexts if the set of primary components is well chosen then the result of the exercise will not be sufficiently compromised if we make the assumption that the secondary parameters are known and fixed.

Fig. 11 illustrates the situation graphically. In the context of the present paper the vertical axis represents the continuous space of models described in Section 4 and the horizonal axis the space of parameters for a given model. In my own work these represent the secondary parameters ϕ_S and the primary components ϕ_P respectively. The left-hand upturned curve indicates how well different models (different

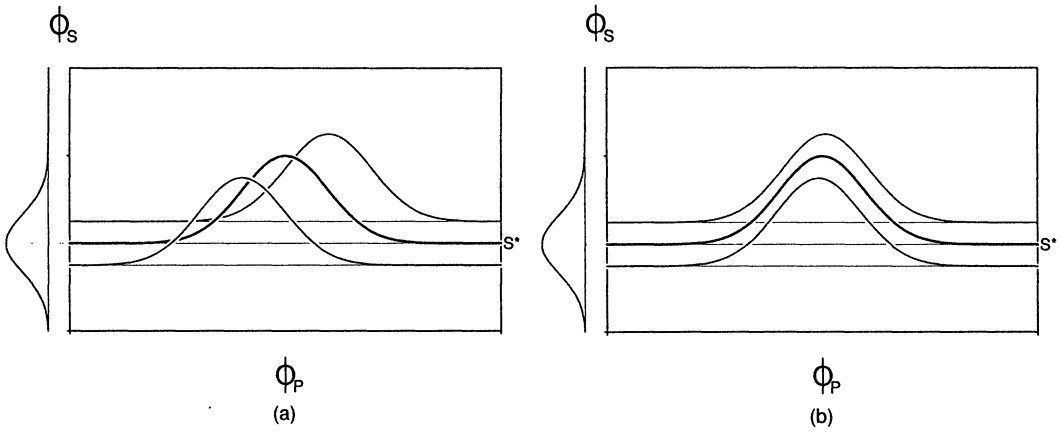

Fig. 11. Effect of uncertainty in fixed secondary parameters, ϕ_S: (a) plausible values of ϕ_S give rise to quite different ranges of predictor variables ϕ_P; (b) with a well-chosen partition of the parameter set ϕ, uncertainty in ϕ_S will not significantly influence the outcome of an uncertainty analysis

ϕ_S) fit the data. In Fig. 11(a) this degree of model uncertainty incorporates significant differences in parameter estimates and in projected quantities. Thus there is significant *between-structure* variance (Section 5.4) which is ignored by fixing attention on the best estimate S^*. In Fig. 11(b) we see that the incorporation of model or secondary parameter uncertainty will not significantly change our view of the future. This is what we should be aiming for. My work has concentrated on how to achieve this with a space of primary components of minimal dimension. The present paper indicates that we can often only achieve this by incorporating model uncertainty.

Chris Chatfield (University of Bath): This paper should be warmly welcomed for tackling a sadly neglected topic. Traditional statistical inference generally assumes a model of *known* form and takes no account of possible uncertainty regarding its structure. Yet few statisticians believe that there really is a 'true' model in any particular situation. Rather a model hopefully provides an approximate, but useful, description of the main qualitative features of a given set of data. Sometimes a model is specified on external subject-matter grounds but computers now allow the analyst to try many different models and to pick the one that fits best. For example time series analysts often search for the 'best' autoregressive integrated moving average model. Although we have admitted model uncertainty by searching for a best fitting model, inferences and forecasts are then usually made as if the model were known in the first place. This is a 'quiet scandal' (Breiman, 1992).

In a frequentist framework it is easy to demonstrate that (large) biases can result from formulating, fitting and checking a model to the *same* set of data, though rather hard to assess and overcome these biases. Limited progress can be made using theoretical and resampling methods (Chatfield, 1994), but the Bayesian model averaging approach reviewed by the author seems more satisfying in many ways. It avoids the necessity of choosing a single best model and gives a more realistic assessment of predictive uncertainty. The author's example of forecasting oil prices by using a weighted average of 12 models is a convincing case in point. In addition Tony O'Hagan has reminded me that the multiprocess models used by Harrison and Stevens (1976) in Bayesian forecasting have much in common with the approach described here.

Perhaps the greatest compliment that I can pay the author is to say that his paper has made me seriously reconsider my attitude to Bayesian methods. For years I have regarded myself as an eclectic or pragmatic statistician. The contortions of Bayesians as they try to cope with prior ignorance or to carry out hideous numerical integrations, of frequentists as they struggle to make sense of P-values and of those who try to pretend that the likelihood function is of fundamental importance when it depends on the model that has been assumed in the first place would be amusing if they were not so serious. I firmly believe that no single philosophical approach can cope with the wide variety of real life situations which confront the applied statistician, but this paper emphasizes that Bayesian methods can be a valuable addition to our toolkit.

George Box (University of Wisconsin, Madison): This is a fine example of how far a first-class statistician can go in the careful reanalysis of dead data sets. But I believe that he should aspire to more.

Statistics has no reason for existence except as a catalyst for scientific enquiry in which only the last stage, when all the creative work has already been done, is concerned with a final fixed model and a rigorous test of conclusions. The main part of such an investigation involves an inductive–deductive iteration with input coming from the subject-matter specialist at every stage. This requires a continuously developing model in which the identity of the measured responses, the factors considered, the structure of the mathematical model, the number and nature of its parameters and even the objective of the study change. With its present access to enormous computer power and provocative and thought-provoking graphical display, modern statistics could make enormous contributions to this—the main body of scientific endeavour. But most of the time it does not.

Statistics, I would like to believe, is finally at the end of the artificial tether imposed by the perceived needs of mathematics. If we confine our thinking to the reanalysis of single 'data sets', the statistician need never leave his office and mathematical statistics as presently prescribed can be king—but of a very small domain—a half-leg of a single iterative step whose optimization may be irrelevant to the final objective.

As I argue in Box (1980) we should instead be prepared to share with experimenters a wider realm of *scientific statistics* with indeterminate and flexible models and sequences of data generation and analysis. To do this would change the present ideas of what statistics is about, how it should be taught and how it should be funded. But, since it would make our subject so much more useful, it might be worth the effort.

David Cox (Nuffield College, Oxford): Fitting different models to the same data is of most interest when the different models correspond to different interpretations of the data, but surely it is often wise even when the objective is the same for all models, the case considered by Dr Draper.

The arguments for incorporating variation between models quantitatively into the formal assessment of uncertainty no doubt depend on the circumstances but seem to me in general uncompelling.

For example, the report written in the summer of 1988 on forecasting the incidence of acquired immune deficiency syndrome in England and Wales (Cox, 1988) showed the results of 11 different models, and many more had been considered. The report, although written for a relatively non-technical audience, distinguished between Poisson errors, parameter estimation errors and errors arising from the model, the last being the most important. Two pages of careful, although qualitative, discussion of the advantages and limitations of the various models were given. Should these have been *replaced* by a Bayesian assessment of uncertainty: certainly not! Should these have been *supplemented* by such an assessment: I am not convinced. Where would the prior probabilities have come from? Would not this have been importing an air of bogus quantification? Other cases might well be different.

John W. Tukey (Princeton University): Draper sends us four messages, the first two of which—

(a) we badly need to take account of model uncertainty and
(b) we can only afford to do this by looking at a small number of separate alternatives—

I believe every statistician and every analyst of data need to take very seriously. The other two—

(c) alternatives should be treated in terms of probability rather than reasonability and
(d) the Bayes mechanism is appropriate for this—

I find it impossible to accept.

Suppose that we have one model that produces quite wide limits and either three or 12 others that produce quite narrow limits, all with consistent point estimates and all with the data appearing reasonable for that model. Draper's approach would provide much narrower limits if there were 12 acceptable models with tight limits rather than three models. This seems to me to be quite wrong, since I must believe in the possibility that the model with loose limits is the nearest to reality so we need to take the most conservative result—for a single model—if we wish to be sure of our conclusions.

It is known, and should be well known, that given data it is easy to make up models that

(a) find the data to be compatible with themselves and
(b) yield very narrow limits.

Although these models are not likely to appeal to either statisticians or analysts of data, their existence is enough to make the question 'Were the models selected before the data were examined, or not?' crucial.

The most acceptable pattern, so far as I am concerned, for the development of a bouquet of models begins with a predata choice of a collection of models likely to be relevant in the field in question, followed by an examination of the reasonability of the data in the light of each model. For those models for which the data seem unreasonable, we have a choice:

(a) drop them from consideration or
(b) move them sufficiently close to a smoothed version of the data to make the data reasonable.

Here reasonability is a yes–no decision, not a probability reduction, and the models are thought as challenges, trying to mark the boundaries of reasonability, not to represent likely outcomes. Taking the worst of what remains is a conservative but, in my judgment, reasonable step.

Adrian E. Raftery (University of Washington, Seattle): It is a pleasure to congratulate David Draper on a fine paper that makes a strong case for accounting for model uncertainty. My comments will focus on software, approximating posterior model probabilities, and previous work.

Software

The GLIB (generalized linear Bayesian modelling) software gives posterior model probabilities and inference that take account of model uncertainty in generalized linear models (Raftery, 1993a); see Raftery and Richardson (1994) for a detailed application to epidemiology. GLIB is an S-PLUS function which

can be obtained free of charge by sending the electronic mail message 'send glib from S' to statlib@stat.cmu.edu.

This can be used to take account of uncertainty about independent variable selection, the link function and the error function. For another case where, as in the Challenger example, conclusions are very sensitive to the link function, see Kass and Raftery (1994).

Approximating posterior model probabilities

In my experience, the term $\frac{1}{2}k_i \ln(2\pi)$ in equation (11) hurts the accuracy of the approximation. As the author points out, by omitting it, we recover the Bayes information criterion (BIC) approximation, which I have found to be more accurate. A reason for this was given by Kass and Wasserman (1992a), who showed, roughly, that if the prior has the same amount of information as one 'average' observation the BIC approximation has error $O(n^{-1/2})$, compared with $O(1)$ for equation (11).

It is possible to improve on the Laplace approximation (10) in its maximum likelihood estimate (MLE) form by taking a single Newton step towards the posterior mode, starting at the MLE (Raftery, 1993a). This is more accurate but still involves only the MLE and the information matrix at the MLE; it is the basis for the GLIB software mentioned above.

Previous work

The large sample approximation summarized in Section 5.4 was proposed in the context of time series by Taplin (1990) and Taplin and Raftery (1991, 1994). The idea of using discrete model expansion to account for model uncertainty can be found in these references and in Taplin (1993).

The following contributions were received in writing after the meeting.

Murray Aitkin (University of Western Australia, Nedlands, and Tel Aviv University): I support the need to allow for model uncertainty by averaging over the models which might reasonably have generated the data rather than conditioning on the best-supported model. However, the standard Bayes factor approach of Section 5.3 suffers the difficulties described by Aitkin (1991) and O'Hagan (1995). The posterior Bayes factor of Aitkin (1991) provides a better framework. I shall illustrate with the O-ring example.

The parameter of interest is the probability of failure at $t=31$: $\psi = p_{31}^{q}$. Each model can be parameterized with $g(\psi)$ as the parameter of interest and nuisance parameter(s) η_j where g is the link function. For example, the linear logit model has

$$g(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 t.$$

Our interest is in $g(\psi) = \beta_0 + 31\beta_1$ so that the logit model can be expressed as

$$\log\left(\frac{p}{1-p}\right) = g(\psi) + \beta_1(t-31)$$

with $\eta = \beta_1$ the nuisance parameter.

I consider as candidate models only the three linear models S_j. For simplicity I omit the model with leak pressure s. Badly fitting models can be omitted, and I exclude the quadratic models as for certain values of ψ they will be non-monotone.

It is straightforward to construct the *profile* likelihood for ψ for each model by defining a grid for ψ, transforming to the corresponding grid for $g(\psi)$ and maximizing over η. Fig. 12 shows the profile likelihoods for the three models. They are surprisingly different from the model conditional posterior densities for ψ shown in Fig. 6. This may be because of poor agreement between the profile likelihoods and the normal approximation of equation (15).

Formally, the average (posterior mean) likelihood in ψ, given data y and prior density $\pi(\eta_j, S_j)$ for model S_j with nuisance parameter η_j of dimension p_j, is

$$A(\psi) = \sum_{j=1}^{m} \int L_j(\psi, \eta_j)\, \pi_j(\eta_j, S_j \mid \psi, y)\, \mathrm{d}\eta_j = \frac{\sum_{1}^{m} \int L_j^2(\psi, \eta_j)\, \pi_j(\eta_j, S_j)\, \mathrm{d}\eta_j}{\sum_{1}^{m} \int L_j(\psi, \eta_j)\, \pi_j(\eta_j, S_j)\, \mathrm{d}\eta_j},$$

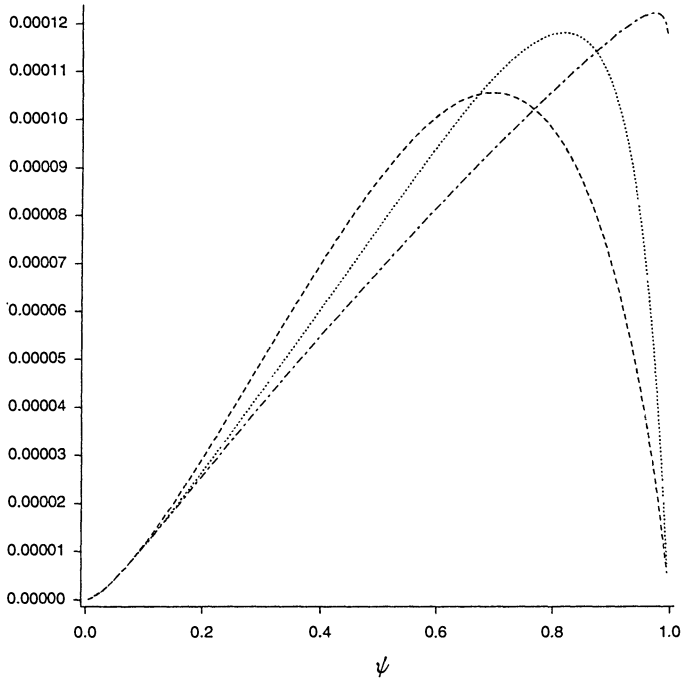

Fig. 12. Profile likelihoods for linear models: ·····, logit; ------, probit; ·–·–, complementary log–log

where $\pi(\eta_j, S_j|\psi, y)$ is the conditional posterior density of η_j and model S_j given the data y and ψ. Taking for convenience diffuse priors on η_j and equal prior probabilities on the three models, and applying the Laplace approximation (10):

$$A(\psi) \doteq \frac{\sum_1^m \pi^{p_j/2} L_j^2\{\psi, \hat{\eta}_j(\psi)\}|I_j\{\hat{\eta}_j(\psi)\}|^{-1/2}}{\sum_1^m (2\pi)^{p_j/2} L_j\{\psi, \hat{\eta}_j(\psi)\}|I_j\{\hat{\eta}_j(\psi)\}|^{-1/2}}$$

where $L_j\{\psi, \hat{\eta}_j(\psi)\}$ is the profile likelihood for model S_j and $I_j\{\hat{\eta}_j(\psi)\}$ is the information matrix at ψ from the profile likelihood. The average likelihood is easily computed from standard generalized linear model output.

Fig. 13 shows the (approximate) average likelihood together with the profile likelihoods for the linear models. The average likelihood is much more diffuse than the linear logit profile likelihood and has a bump at the upper end where the log–log link model has a much higher likelihood than the other linear models.

These results differ substantially from those of Draper, even allowing for his inclusion of the quadratic models, and it is puzzling to see the extremely strong support for $\psi = 1$ in his conclusions.

G. A. Barnard (Colchester): I have space only for a brief comment on the final paragraph of Section 4.2 of this most interesting and important paper. For more in this line see Barnard (1994a, b).

The standard form of what I have called a 'Fechner density' is

$$\phi(u) = K \exp\{-\tfrac{1}{2} M^\alpha(u)\}, \qquad -\infty < u < \infty,$$

where, for $1 \leqslant \alpha < \infty$ and $M \geqslant 0$, we define

$$M^\alpha(u) = \begin{cases} u^\alpha & \text{for } u \geqslant 0, \\ (-Mu)^\alpha & \text{for } u \leqslant 0, \end{cases}$$

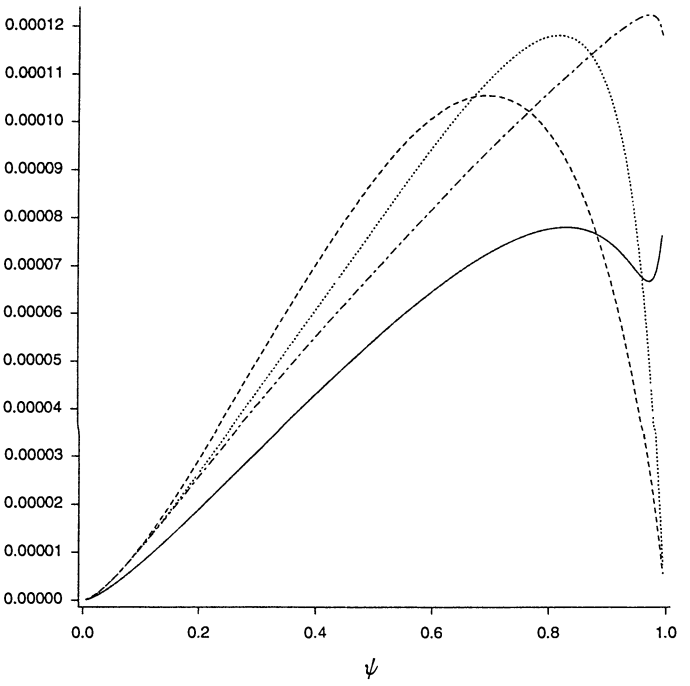

Fig. 13. Average likelihood over three models: ·····, linear logit; ------, linear probit; -·-·-, linear complementary log–log; ———, average

and K, here and below, denotes a norming constant. For $M=1$ variation of α covers the Box and Tiao (1962) expanded model, whereas variation of M covers an arbitrarily high degree of left or right skewness. Since ϕ has mode 0, the natural location parameter is the mode μ of the distribution of independent and identically distributed observables x_i, $i=1, 2, \ldots, n$; moreover it is the parameter of interest if, for example, the x_i denote differences of matched pairs of patients in a clinical trial. The fact that for this family $\log \phi$ is homogeneous of degree α enables us to derive the exact conditional distribution of Student's pivotal $t=(\bar{x}-\mu)\sqrt{n}/s_x$ in the form

$$\psi(t|\underline{c}) = K \Big/ \sum_i \{M^\alpha (t+c_i)\}^{n/\alpha}$$

where $\underline{c}$ with $c_i=(x_i-\bar{x})\sqrt{n}/s_x$ is the maximal ancillary. Denoting by t_0 the result of substituting observed values for the x_i in t and c_i we obtain the confidence distribution for μ, the posterior density relative to a uniform prior, based on the available sample.

The fact that we can write a program with M and α as parameters enables us here to apply serendipity—one thing perhaps not stressed by the author as much as it could be. It will often turn out that varying M and α over the plausible range will make unimportant differences to the confidence distribution, saving much time and effort that would otherwise be devoted to unimportant questions.

For more details on Fechner densities, see Thomas (1993).

Peter Donnelly (Queen Mary and Westfield College, London): The space shuttle example provides a good illustration of a setting in which the primary quantity of interest, the probability, p_F say, that at least one field joint will fail (with the consequent destruction of the shuttle) is a function of parameters (such as the probability of primary O-ring erosion) on which direct data are available. Any quantification of beliefs about this primary quantity thus *necessitates* a quantification of uncertainty about the parameter value. Further, under the assumption of independence of primary and secondary erosion and blow-by (and also under certain dependence assumptions) the mean of p_F, for example, will be *increasing* in the uncertainty about p_{31}^a for fixed $E(p_{31}^a)$ (with the usual variability ordering; see for example Ross (1983)), highlighting the author's concerns about the dangers of under-representing the uncertainty in such estimates.

In deoxyribonucleic acid (DNA) profiling, the natural parameter is the 'match probability', the probability that an individual unrelated to the defendant, chosen randomly from an appropriate reference population, would match the defendant's DNA profile. What the author describes as the S^*-approach is widespread in this context. Hypothesis tests, typically of low power, often fail to reject various assumptions which underpin the match probability calculation. These assumptions are then treated as true and the resulting point estimate is equated to the match probability. This is another setting in which the primary purpose, an assessment of whether the defendant has been proved guilty beyond reasonable doubt, is a function of the match probability. Under various assumptions the probability of innocence is increasing in the uncertainty about the match probability (Balding and Donnelly, 1995). Failure to acknowledge, or quantify, such uncertainty thus prejudices the trial process against the defendant.

More generally, the (expected) costs of various decisions taken subsequently to data analysis will usually depend on the uncertainty involved. Some form of quantification of this uncertainty may thus be essential in advising decision makers. In any case the suggestion of simply presenting the conclusions of different possible models, without commenting on each model's *a priori* plausibility or their consistency with available data, seems undesirable.

A. S. C. Ehrenberg (South Bank Business School, London): I welcome almost any attempt to grapple with issues of model choice, but I have three questions for David Draper at what I think he would call the *scenario* level.

(a) Why focus exclusively on forecasting *catastrophes*? Why not dissect model selection issues for some of the thousands of *success* stories in the Challenger launch, say, or in science more generally?
(b) Why focus so much on 'uncertainty' (lack of data, small samples, weak priors, too many models to choose from, etc.)? Why not help us to deal with problems of model choice in the many cases where there is plenty of data (or even well-digested 'information', submodels and grounded theory)? There are still many problems of model choice, in my experience.
(c) Why accept George Barnard's characterization of the statistical *status quo*: 'We can do no other than suppose that we have spanned the set of possibilities'? Why not follow through on his own (Draper's) advice to 'stake out the corners in the model space', i.e. *explicitly* to establish the range of different conditions under which the chosen model is already known to hold? Is it because that process is fuzzy and does not lend itself to neat logical or mathematical formalizations? But science shows that it is what leads to routine predictability.

Alan E. Gelfand and Bani K. Mallick (University of Connecticut, Storrs): Draper offers some interesting foundational food for thought in observing that a component of the modelling job is the specification of structure, S, i.e. 'exact structure' is not known (and in most cases this is not a meaningful notion since the objective is not a 'correct' model but rather a parsimonious model which approximates reality well). Hence within the Bayesian paradigm we must place a distribution $p(S)$ over S. However, in Draper's examples the set of specifications is treated parametrically implying that a complete specification of model M is still parametric, consisting of θ augmented by structural parameters. Indeed in our own work we have handled structure this way—link functions in generalized linear models (GLMs) (Carlin and Gelfand (1991), p. 126, and Mallick and Gelfand (1994a)), integrated hazards in survival models (Gelfand and Mallick, 1994) and calibration functions in errors-in-variables models (Mallick and Gelfand, 1994b).

Does this render the distinction between parameter and structure semantic? The answer is no. θ is only interpretable given S; the prior specification must take the form $p(\theta|S)\,p(S)$. In this sense, clarity would be added to the presentation if θ were replaced with θ_s. There is no sensible notion of a marginal distribution for 'θ', prior or posterior. A striking example is the case of mixture models where S is the number of components but even in, say, GLMs with unknown link the marginal distribution of a regression coefficient is not sensible.

But then questions arise regarding inference. Posterior inference is sensible only with respect to S so, for instance, in a regression setting we cannot directly investigate the contribution of explanatory variables. ($p(\theta_s|S, x)$ is possible but conditioning over S returns us to the sensitivity analysis, Draper eschews.) Predictive inference, employed by Draper, is also not helpful in this regard. In fact, in the Bayesian framework, predictive distributions are customarily utilized for examining model adequacy and model choice but not for inference activity (Box, 1980). Hence, as food for thought, Draper presents an expanded menu but in the end, a not quite satisfying meal.

We offer a final thought. In many contexts S^* does not arise on the basis of data x, but rather as a 'natural' assumption. Examples include conditional independence, exchangeability, canonical links, identity (or linear) calibration functions. Then, x would be used to check the adequacy of S^* by using predictive distributions.

Andrew Gelman (University of California, Berkeley) **and Xiao-Li Meng** (University of Chicago): We enjoyed this paper and largely agree with the author, even though we would emphasize some of the points slightly differently.

We agree that averaging across competing models is better than choosing just one model, but it is often even better to consider the models as a continuous class (as the author briefly notes at the beginning of Section 5), which to do seriously often requires additional work to ensure that the individual model parameters make sense in the supermodel. We prefer the term 'model improvement' (Gelman and Meng, 1994) instead of mere 'averaging' to indicate the additional information and consideration that goes into creating a sensible larger model.

The author briefly mentions sensitivity analysis as a qualitative method that is improved on by model averaging; we believe that learning about sensitivity of inferences to model assumptions is often an important goal that is not achieved by merely looking at the combined inference. The author seems to recognize this implicitly in Fig. 7. For another example, in the context of dealing with incomplete data, it is typically important to display sensitivity to assumptions about the missing data mechanism; an illustration in an applied context is in Heitjan and Rubin (1990).

The penultimate paragraph of this paper offers a qualitative view of not considering all possibilities in a model. In many cases, we can detect aspects of poor fit by comparing the observed data with their predictions under the model; in Bayesian terms, posterior predictive checks (Rubin, 1981, 1984; Gelman *et al.*, 1995; Gelman and Meng, 1994). Even in the Bayesian context, model checking falls outside the 'model uncertainty' framework. In addition, the larger model used in the uncertainty analysis can, and should, be checked against the data. Even if the large model fits, we must keep an open mind about other possibilities.

Finally, the final paragraph of the paper might leave the impression that there is a trade-off between greater accuracy in a smaller model and better calibration in an expanded model. In both the short and the long term, the expanded model should be superior in both accuracy and calibration. The use of 'accuracy' in Section 7 seems also to include the concept of 'precision', which has quite different implications when considering the desirability of prediction procedures. The paper is somewhat unclear on such a distinction (e.g. Section 7.1). It is not that a statistician wants to 'widen the bands' to cover his neck; it is that real uncertainty governs the widths of the bands. This is especially important when explaining a statistical analysis to decision makers, such as those who authorized the launching of Challenger.

C. A. Glasbey and G. J. Gibson (Scottish Agricultural Statistics Service, Edinburgh): This paper states the obvious, but it is apparent that these are issues which we all need to be reminded of. It is interesting to see a Bayesian approach being used to increase, rather than to decrease, the uncertainty in prior models. In our modelling work in the Scottish Agricultural Statistics Service (SASS) we must often consider uncertainty.

(a) In most models of agricultural systems, scant attention is paid to uncertainty in any guise. The SASS undertook a study of model uncertainty for the Agricultural and Food Research Council (Gibson *et al.*, 1993). We found a range of examples in which failure to take account of uncertainty led to models whose predictions are as misleading as those considered in this paper. We identified a fourth source of uncertainty in addition to those explicitly mentioned by Draper—the intrinsic stochasticity of many models. For example, the number of failed rings on the shuttle has a binomial distribution.

(b) In an on-going study on the feasibility of developing systems models to aid decision-making in agricultural policy, Gibson (1994) considered predicting the detrimental effects associated with the release of genetically modified micro-organisms into the environment. In this case it was apparent that the various sources of uncertainty in the system, not least of which lay in the form of models for the survivability and transmission of organisms, meant that the uncertainty in model predictions would be so great that its utility was negligible.

(c) Glasbey (1987) investigated the effects on estimates of ED50 (the dose at which 50% of subjects respond) from quantal dose–response data of relaxing assumptions about the tolerance distribution.

The distribution was assumed either simply to exist (implying a monotonic response function), or to be unimodal, bell shaped, symmetric or Gaussian. A hierarchy of confidence widths was produced dependent on the strength of the assumptions.

(d) One of the methodological problems raised by the paper is that of specifying prior probabilities for a range of models. In many scenarios these may have to be based on qualitative knowledge elicited from experts. As an aid to this process, the tools developed within the LIKELY project funded by the European Community (Talbot, 1993) could be of value. In particular, the use of cognitive mapping techniques to 'capture' the conceptual models of experts might offer a systematic method for specifying priors.

Richard Glendinning (Defence Research Agency, Great Malvern): The author points out several weaknesses in the conventional approach to model uncertainty which is dominated by the use of a single model. The limitations of this approach and alternative strategies have been noted many times in the time series context; see Poskitt and Tremayne (1987) or Akaike (1985). For a recent survey of model selection for time series, see Glendinning (1993).

The author emphasizes the problem of choosing an appropriate subset of all possible models for further analysis in a Bayesian framework. This problem has two facets. The first issue is purely a question of computational resources and can be mollified in the frequentist approach by the use of an efficient search strategy; see Fernholz and Fernholz (1986).

The second issue is concerned with the development of procedures which have good statistical properties. How can various techniques for selecting a subset of all possible models be evaluated theoretically or computationally? This issue also affects the frequentist approach to model selection where the performance of penalized 'goodness-of-fit' criteria can be improved by restricting the number of candidate models relative to the data available. Several approaches have been suggested in the time series context. These include the use of the asymptotic properties of $\hat{k}$, where $\hat{k}$ is the index of the selected model. For predictive least squares we can obtain an upper bound on the growth of the number of candidate models which ensure strong consistency; see Hemerly and Davis (1991). Similar results are obtained for penalized goodness-of-fit criteria using Bayesian ideas in Veres (1991). An alternative approach based on complexity is suggested in Smyth (1992). Here the aim is to restrict attention to models with lower complexity than the data themselves. Other *ad hoc* approaches which limit the number of candidate models are based on a variety of techniques including the use of preliminary hypothesis tests; see Wang and Chen (1985).

Urban Hjorth (Linköping University): That modelling uncertainty can be a magnitude worse than parameter uncertainty is now increasingly recognized by statisticians. Also the tools to analyse this uncertainty are now becoming more sophisticated and varied. Classical analysis based on model decision and ignoring competing models simply does not work in many regression, time series and other modelling situations. Draper includes an interesting example of oil price predictions based on different scenarios.

We can choose between an analysis where the statistical effects of a model selection are analysed by computer-intensive methods like cross- or forward validation or the bootstrap (Breiman, 1992; Hjorth, 1994), or an analysis like Draper's where all the models are maintained in a Bayesian formalism. An interesting question is whether recursive model selection or the suggested Bayesian method gives better predictions of time series. When applicable, a Bayesian description from the defined priors to the end result is very elegant. At the same time it has two major difficulties: the computational burden when many models are at hand and the difficulty of setting priors over both model structure and parameters. Draper shows in his examples how a reasonable model subset can be defined. However, the difficulties are also shown and the data-driven cross-validation and bootstrap methods suggested as a solution in Section 5.1 seem to me to be against the general philosophy of his work and may perhaps generate somewhat arbitrary error distributions. More needs to be known about these methods for defining a Bayesian model set and with the present state of knowledge I prefer these methods for the evaluation of more classical modelling approaches. However, the important thing is to find methods which explain and display the modelling uncertainty and Draper has given an important contribution towards this.

Robert E. Kass and Larry Wasserman (Carnegie Mellon University, Pittsburgh): We have one small technical comment on Dr Draper's interesting paper. There are many order $O(1)$ approximations of the form (11). It is not at all clear that Dr Draper's choice, which includes the additional constant $c_1 = \frac{1}{2} k_i \ln(2\pi)$, will be more accurate than any other including Schwarz's. As noted in Kass and

Wasserman (1992b)—see also Kass and Wasserman (1995)—for nested models with an intuitive choice of a normal prior, Schwarz's criterion is accurate to order $O(n^{-1/2})$. This corresponds to replacing c_1 with $c_0 = 0$. Using a Cauchy prior, as in Jeffreys (1961), leads to the constant

$$c_2 = \frac{k}{2}\ln(2n) \ln \Gamma\left(\frac{k+1}{2}\right) - \frac{1}{2}\ln \pi.$$

For $k > 16$, $c_0 \leqslant c_1 \leqslant c_2$ but for smaller values of k Dr Draper's c_1 can be somewhat greater than c_2. In comparing two models for which $3 \leqslant k_2 - k_1 \leqslant 10$, using c_1 in place of c_2 increases the posterior odds by a factor of 10.

Michael Lavine (Duke University, Durham): Draper's point is that accounting for model uncertainty is both important and difficult. We agree, and we would like to borrow one of his examples to illustrate another way in which parametric models and priors over them understate uncertainty.

In the example in Section 6.2 Draper reanalyses the space shuttle O-ring data. Briefly, the probability of O-ring erosion is modelled as a function of launch temperature and leak test pressure. The goal is to calculate a posterior distribution for p_{31}^a, the probability of erosion at 31 °F. Several generalized linear models are considered including three link functions, three functional forms for the temperature variable and two functional forms for the pressure variable. The data used are those available before the launch of Challenger: 23 flights at temperatures ranging from 53 °F to 81 °F. Draper uses a discrete prior over the set of models and computes a posterior variance for p_{31}^a as the sum of variance within models and between models.

But consider what happens in the limit, as data accumulate at temperatures above 53 °F. The posterior concentrates on the model that best fits the data above 53 °F and on the maximum likelihood estimates β for that model. The within-model variance and between-model variance both go to 0 so the posterior variance of p_{31}^a goes to 0.

The result is misleading; the uncertainty due to extrapolating from 53 °F to 31 °F has been hidden by mixing only over a finite number of parametric models. A similar phenomenon occurs with every parametric model. An accumulation of data in one region implies great certainty about another region. Of course, whether this is a disaster depends on the use to which the analysis will be put. In the space shuttle analysis the sample size is small, variances are large and the posterior for p_{31}^a is reasonable. But in general it pays to think about how much information an observation in one region conveys about other regions.

David Madigan (University of Washington, Seattle): It is a pleasure to congratulate David Draper on an important paper. Model uncertainty is the Achilles heel of statistics; to ignore it is to overstate your certainty and to risk making poor predictions. As Draper shows, data analysts *can* account for model uncertainty, at least approximately, by using the output from standard software.

I am concerned that model expansion may be inadequate; model uncertainty is not always restricted to models that are expanded versions of some S^*. Our own work focuses on methods which average over all 'supported' models: Occam's window involves summing over a smaller set of likely models (Madigan and Raftery, 1994), and Markov chain Monte Carlo model composition (MCMCMC) approximates averaging over all models (Madigan and York, 1993). We have applied these algorithms to linear regression (Raftery *et al.*, 1993), graphical models (Madigan and Raftery, 1994), generalized linear models (Raftery, 1993b) and survival analysis (Raftery *et al.*, 1995). A key finding is that, for most of the data sets we analysed, model averaging using Occam's window or MCMCMC had better out-of-sample predictive ability than any single model that an analyst might reasonably have chosen.

Section 2.1 shows that the S^*-strategy can be disastrous in linear regression. Freedman (1983) also made this point with a simple simulation and we call the phenomenon 'Freedman's paradox'. The Occam's window approach largely resolves Freedman's paradox (Raftery *et al.*, 1993).

We have had good results with priors $p(S)$ that are diffuse over the entire model space, so that the criticisms of such priors in Section 3 do not apply. Recently we have also developed a method for eliciting informative prior distributions on model space (Madigan *et al.*, 1994). Eliciting distributions on unobservables such as models is difficult (Kadane *et al.*, 1980); our method starts with a uniform prior distribution on model space, updates it by using imaginary data provided by the domain expert and then uses the updated prior distribution as the prior distribution for the analysis. Laud *et al.* (1992) proposed a somewhat similar approach. The method is simple to implement and, in a challenging graphical models application, predictively outperformed its diffuse counterpart.

Benedikt M. Pötscher (University of Vienna): I would first like to congratulate the author on his interesting and stimulating paper. My first comment concerns the so-called S^*-approach and its relationship to the Bayesian approach suggested by Draper. Both approaches start from a set $\mathcal{S}=\{S_1, \ldots, S_m\}$ of possible structures (in the case of discrete model expansion). In the S^*-approach a structure S^* is selected from $\mathcal{S}$, usually by optimizing a criterion like Akaike's information criterion (AIC), Schwarz's Bayes information criterion or by applying some kind of testing procedure. Inference is then based on the structure S^*. In contrast, the Bayesian approach averages over all structures in $\mathcal{S}$. Both approaches are logically sound as long as in the first approach the inference based on S^* acknowledges the fact that S^* is not fixed *a priori* but is data dependent. Hence, the problem is not in selecting a 'best' structure S^* from the set $\mathcal{S}$ but in a naïve use of classical inference procedures (which assume an *a priori* fixed structure) thereby treating the data-dependent S^* as if it had been chosen before the analysis. If the selection process leading to S^* is properly taken into account in the inference procedures, such a 'non-naïve' S^*-approach will result in correct and not in anticonservative inference. For derivations of the appropriate distributions and inference procedures in this context see, for example, Sen (1979), Kabaila (1995) and Pötscher (1991, 1995). Also Judge and Bock (1978) is relevant here. An alternative approach is via bootstrapping; see Freedman *et al.* (1986) and Hjorth (1989).

The suggestion in Section 5.1 to use a subset of the data to help in specifying the prior probabilities $p(S_i)$ in a time series context seems to lead to a sample reuse problem that is quite similar to the problem that the Bayesian approach discussed in Section 5 tries to avoid in the first place.

In the context of continuous model expansion an alternative to averaging over the model expansion parameter α would be to estimate it jointly with the parameter θ, and to base inference on the (asymptotic) distribution of the estimators in the expanded model.

In time series analysis, parametric models like autoregressive models are sometimes used as 'approximate' models, i.e. the actual distribution of the data is not described by a distribution derived from any of the parametric models; see, for example, Hannan and Deistler (1988), chapter 7. The order of the autoregressive model is then usually determined by the AIC or a similar procedure. It would be interesting to see how the Bayesian approach proposed in Draper's paper would handle such a situation.

John W. Pratt (Harvard University, Boston): My intention is to complement and reinforce, not to disagree with, the paper. I admire analyses like these, but it would be unfortunate if anyone thought them needed to see what Berkson's interocular traumatic test tells you immediately in Fig. 5: unless the focus on temperature was very fishy, even if the points really have variance only $6pq$, the data reveal almost nothing about erosion at 31 °F except that, assuming monotonicity, it is at least 10%, say.

Is there a reason to prefer Bayesian structural choice over continuous model expansion here? The latter is more flexible and less extremist; it expresses uncertainty more smoothly and realistically. The former, unless repeatedly refined with increasing sample size, concentrates exponentially on the Kullback–Leibler closest structure (Berk, 1966). Incidentally, in economics, 'structure' means something about causality, beyond mere statistical relationship or 'reduced form'. Here 'structural' uncertainty differs from parametric uncertainty only in analytical form, not in kind or principle.

Some simple and hence underemphasized points have always underlain my thinking about model uncertainty. In regression, for each variable, adding quadratic or other terms or choosing a transformation from a power or other family buys much realism for just 1 degree of freedom per parameter. This may be unimportant for interpolation or routine forecasting but, for extrapolation, acknowledging uncertainty is far more important than parsimonious explanation, saving degrees of freedom or reducing standard errors of estimates or even predictions (see the last paragraph of Section 4.1). Choosing from an oversupply of variables, interactions, lags and so on is where it becomes difficult. It should not be as bad as Section 2.1 might suggest in the non-null cases that we usually face. Nevertheless, all-or-nothing approaches are inadequate. Shrinkage based on serious judgment is needed. Mosteller and Wallace (1964) is an extreme example, with discrete data and thousands of variables, beautifully analysed. Well-judged hierarchical models help, as would flexible, easily used software, especially for those of us without the skills and perseverance displayed here.

When statistical uncertainty seems small, we rightly start to look for the real uncertainty. I have some worry about acknowledging some but significantly less than all of it. I am also struck by the contingent nature of many of the choices to be made. Even for a given data set, no supermodel or mixture of models is good for all purposes—interpolation, extrapolation, forecasting, explanation. Causation further complicates choices, often requiring the inclusion of variables however much they proxy for one another

and increase standard errors. No recipe, package or expert system will suffice: unbiased professional statistical judgment is needed. But will it be obtained?

Aart F. de Vos (Free University, Amsterdam): This is a nice paper, and the suggestions are useful, though not new. But there are better ways to tackle the problems in the two examples. In the oil as well as the space shuttle case the main problems are caused by extrapolation, in time and functional form respectively. Underestimating uncertainty in these cases is in my view mainly due to the use of deterministic models. Using non-stationary time series models for the oil price or for the parameters in the models, as is easily done by using the Kalman filter, produces realistically diverging forecast intervals.

For the functional form relating the failure rate of rings to temperature, stochastic functional forms may be used, either by using the Kalman filter (Wecker and Ansley, 1983) or 'random direction' schemes like Berger and Chen (1993). These models give rise to huge forecast intervals for extrapolations far from the sample, even when maximum likelihood estimates of the parameters (instead of Bayesian posteriors, leading to even more uncertainty) are used.

Summarizing, if we have no clear prior ideas about deterministic models, we should best leave the idea of determinism altogether, rather than specifying a prior where different deterministic models are possible.

B. J. Worton (University of Essex, Colchester): I would like to comment further on the Fechner distribution mentioned by Professor Barnard in his contribution. This is an expansion of the normal model that allows for departures of both skewness and kurtosis from the normal distribution by replacing the standard normal density function by the density function

$$\phi(u|M,\ \alpha)=\begin{cases} K\exp(-\tfrac{1}{2}u^{\alpha}) & u\geqslant 0, \\ K\exp\{-\tfrac{1}{2}(-Mu)^{\alpha}\} & u<0, \end{cases}$$

where $M>0$, $\alpha\geqslant 1$ and

$$K^{-1}=2^{1/\alpha}\left(1+\frac{1}{M}\right)\frac{\Gamma(1/\alpha)}{\alpha}.$$

Fig. 14 shows the normal probability density function together with some other members of the Fechner family. This family can be used to represent a wide variety of unimodal distributional forms by varying the shape parameters M and α, and thus can be used to investigate how crucial the assumption of normality is for a particular data set. It is straightforward to calculate confidence distributions for the mode parameter or the mean parameter for a range of values of M and α. The former case can be written in closed form, and the latter case requires some numerical integration which is easy to program.

As a simple example consider the data set given by $x_i=\Phi^{-1}\{i/(n+1)\}$, $i=1, \ldots, n$, with $n=10$, where Φ is the distribution function of the standard normal distribution, i.e. the Fechner distribution with $M=1$ and $\alpha=2$. Confidence distributions computed for the mean parameter for this data set are very similar for a range of values of M and α. Thus, for this particular data set, if we inadvertently used the wrong model, then we would still obtain reasonably accurate results. However, corresponding confidence distributions for the mode parameter calculated for this data set depend on the value of M used but are similar over a range of α-values for a given value of M. Generally, it is of interest to study the extent to which the model, the data and the parameter of interest affect the stability of the inferences to be drawn.

Professor Arnold Zellner (University of Chicago): I congratulate David Draper for his thoughtful and useful paper which prompted the following observations.

(a) The subject of overfitting and model selection in regression, as are many other topics, is treated in Jeffreys (1967), pages 253 and 278, under the title 'selection effects'—see also Zellner and Min (1993). Jeffreys provided an explicit correction factor for posterior odds that allows for the possibility that one or some models may fit a given data set well just by chance.

(b) In the work of Geisel (1975) and Reynolds (1980), posterior odds for alternative models were formulated and used to combine models thereby allowing for model uncertainty.

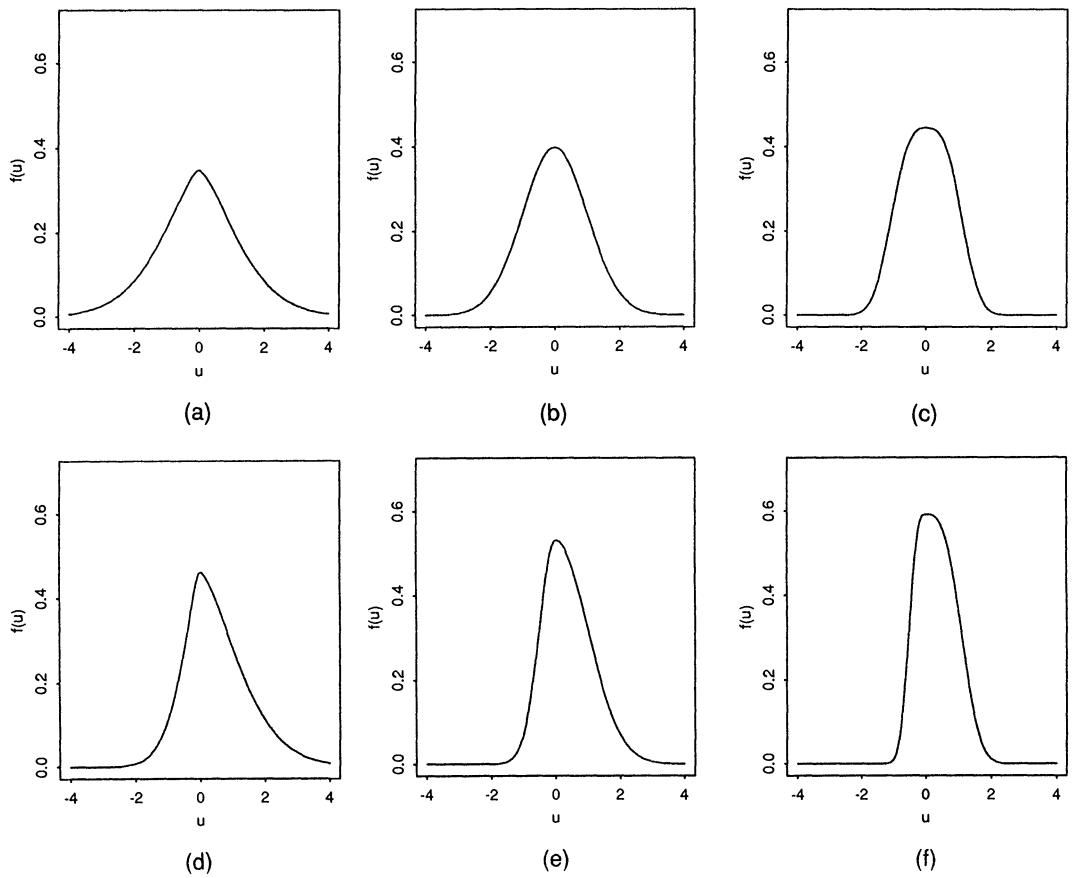

Fig. 14. Probability density functions of distributions from the Fechner family: (a) $M=1$, $\alpha=1.5$; (b) $M=1$, $\alpha=2$, normal model; (c) $M=1$, $\alpha=3$; (d) $M=2$, $\alpha=1.5$; (e) $M=2$, $\alpha=2$; (f) $M=2$, $\alpha=3$

(c) In Palm and Zellner (1992), cited by Draper, a major point of the analysis was to allow forecasts to be biased by the introduction of bias parameters and then to determine under what conditions combining forecasts leads to improved forecasts. Conditions are given under which improvement can be realized even though individual forecasts are biased.

The **author** replied later in writing, as follows.

I am grateful to all the discussants for their interesting remarks, not all of which I am able to address here. The comments seem to fall into 10 broad classes, which are taken up in no particular order below.

Discrete versus continuous model expansion

I agree with Gelman, Meng and Pratt that continuous model expansion is often preferable to the discrete approach when both are possible, although situations can exist in which it is difficult to embed a structural choice of particular interest continuously in a wider class (e.g. Grieve's example from crossover trials, in which, all the same, it is natural to wonder whether something about the problem formulation has created the discontinuity at $\lambda=0$). Spiegelhalter and Pratt both point to the important result by Berk (1966), which implies that in discrete expansion the set $\mathscr{S}$ of structural possibilities, if viewed as a kind of grey scale, needs to include increasingly finer gradations as the sample size increases to keep the $p(S_i|x)$ from spuriously concentrating on a single structural choice. This also comes up in Lavine's example, where the between-model variance goes to 0 because $\mathscr{S}$ stays fixed while n increases. Evidently

methods, perhaps partially automated, for generating candidate S_i to be considered by the analyst for possible inclusion in $\mathscr{S}$ would be a net gain, and such methods should encourage increased refinement of structural alternatives with increasing sample size.

Tukey makes an important methodological point for anyone attempting a discrete model expansion: if $\mathscr{S}$ includes two (say) models with essentially equivalent predictive consequences and similar plausibilities given the data, the result will be to double-count that particular predictive outcome, unless you either drop one or go back and give each of them (say) half of the prior structural probability which that outcome ought to have had in total. This underscores the desirability in the discrete approach to 'stake out the corners in model space' (Section 5.1) in specifying $\mathscr{S}$.

Hierarchical modelling and shrinkage

I can sympathize with the Food and Drug Administration's (FDA's) concern with model uncertainty as described by Grieve (in this case, one regression line or many?), if not with the regulatory 'solution' to it, which is another version of a preliminary test method (yet another version is Tukey's recommendation to compare models with data in such a way that model 'reasonability is a yes–no decision'). As noted in Section 4.3, we may expect improvement over what the FDA is doing by hierarchically expanding the model in which all lines have the same slope, e.g.

$$Y_{ij} = \beta_{0i} + \beta_{1i} X_{ij} + e_{ij},$$
$$\beta_{ki} = \beta_k + \epsilon_{ki},$$

where i indexes batch, j indexes time point within batch, $k = 0, 1$, and the e_{ij} and ϵ_{ki} are, for example, drawn from mean-zero Gaussian distributions with variances σ^2 and τ_k^2 respectively. This model (see, for example, Draper (1995)) expresses unconditional exchangeability of the regression slopes across batches (which seems to be the situation described by Grieve) and deals with uncertainty about the closeness of τ_1 to 0 more smoothly than the all-or-nothing FDA testing approach. The shrinkage form of the resulting estimates of the βs is related to Copas's shrunken logits.

I am not sure that I fully understand Lindley's comments about overfitting, but I shall use the occasion of his having brought up the issue of variable selection in regression to remark that this, also, is a matter that often ought to be treated smoothly rather than discretely. Many standard methods for variable selection (e.g. Efroymson (1960)) involve a series of all-or-nothing judgments about whether or not a given coefficient β_j exactly equals 0, in spite of the more than 20 years of good empirical Bayes research that we now have showing that methods which treat uncertainty about the β_j more smoothly outperform such discretized approaches. Expanding the usual regression model $Y = X\beta + e$ hierarchically with a second-stage regression $\beta = Z\gamma + \epsilon$, in which the matrix Z of prior covariates can be used to bring in substantive prior information about the signs and monotonic ordering of the β_j (Greenland, 1993; Draper, 1994), yields inferences about β and subsequent predictions of new Ys that can be markedly superior to those provided by preliminary test methods. Then, following Lindley (1968), we may use utility considerations, which weigh such things as parsimony preferences and independent variable data collection costs against predictive accuracy, to cause some of the X_j to drop out if they cost too much in relation to their predictive performance.

Science versus policy; statistics 'versus' science

In response to Copas, yes, the Challenger example involved an alarming extrapolation, but it is an interesting and key difference between science and policy that in the former you can pick problems to work on that you think will have a reasonable chance of a satisfying solution, whereas in the latter the problems, alarming or not, force themselves on you. On this occasion the alarmingness of the policy question was set up by a previous failure to have done some basic science: apparently there were no sufficiently relevant engineering studies available on the night before the shuttle was launched to quantify the decreasing resilience of the rubber O-rings with falling temperatures in the gap between 31 and 53 °F.

With regard to Box's comments, the point of the paper was not to encourage people to 'reanalyse dead data sets' retrospectively, looking for model uncertainty, but to adopt prospectively a viewpoint, in their serious collaborative applied work, in which uncertainty in the modelling process was routinely better acknowledged. I agree with him whole-heartedly that the discipline of statistics would be in trouble if we all sat around in our offices with the door closed deriving optimality results in artificial situations of our own devising, and indeed there have been periods in the recent past in which that kind of work was valued far too highly. We may hope that the pendulum is swinging back towards serious collaborative

efforts leading to new and useful methodologies, in the manner so well exemplified by Fisher in the 1920s and 1930s.

When I worked at the public policy institute Rand in California in the 1980s, my colleagues and I used to comment on the remarkable fact that the great majority of what statisticians do in the large scale collaborative projects at that institute and other places like it—e.g. problem formulation, design and analysis of observational studies, causal inference, sampling, reliability and validity of survey instruments, fieldwork procedures to maximize data quality, and large database management—receives little coverage in typical statistics graduate training, and the part that we are all so carefully trained for (e.g. optimal inference given a particular model) often made up considerably less than 10% of the overall budget in time and money. At its best statistics, construed broadly, is the *study of uncertainty*—how to measure it and what to do about it. If we tried to put that definition into practice in its full scope in our applied work, I believe that Box is right: we would end up playing a broader role in day-to-day science and decision-making.

Accuracy, precision and calibration

I like the terminology that Gelman and Meng suggest on accuracy, precision and calibration; perhaps it would be best to reserve the term 'accuracy' for a retrospective evaluation of the performance of a predictive modelling strategy and to use 'precision' to refer to its *apparent* performance prospectively. For example, consider the first part of the paper's equation (1):

$$p(y|x, \mathscr{M}') = \int_{\mathscr{M}'} p(y|x, M)\, p(M|x)\, \mathrm{d}M.$$

With this new terminology the variance of the first term in the integrand, $p(y|x, M)$, might be said to quantify the *conditional inherent precision* for y arising from model choice M, whereas the second term, $p(M|x)$, monitors the plausibility of that particular model choice given the data. The second term acts calibratively to force us away from models that yield precise conclusions when such models do not fit well, but only if $\mathscr{M}'$ is sufficiently big to include models rather like what will retrospectively be seen to be 'correct'. By choosing $\mathscr{M}'$ sufficiently narrowly—for example by insisting that the data follow a single distribution that is highly informative for location, even if they do not—we can produce quite accurate *looking* inferences and predictions that are wholly out of calibration. This becomes more difficult to get away with as $\mathscr{M}'$ increases in size; thus there can be a trade-off between greater *apparent* accuracy (precision) in a smaller model and better calibration in an expanded model.

Technical alternatives

The Fechner density family mentioned by Barnard and Worton expands the generalized power-exponential distribution of Box and Tiao in an interesting way. One question for Barnard, though: why should we concentrate on Student's pivotal t in this model, when this is the asymptotically optimal summary only for the particular choice ($M = 1$, $\alpha = 2$)? Why not use, for example, Markov chain Monte Carlo (MCMC) methods (e.g. Gilks *et al.* (1994)) to obtain the marginal posterior for μ, given the data and substantively meaningful prior distributions on the parameters?

Aitkin's model uncertainty reanalysis of the Challenger data is welcome. However, his posterior Bayes factor method has been rightfully criticized elsewhere (see the discussion to Aitkin (1991)) for using the data twice (although model expansion itself is not immune to this criticism; see below), and he appears to have employed raw rather than adjusted (or modified) profile likelihoods, which in both the Bayesian and non-Bayesian worlds have been called into question in non-Gaussian problems (e.g. Cox and Reid (1987)). In spite of this it is pleasant to note the qualitative agreement between his analysis and mine that substantial model uncertainty is present and unaccounted for by conditioning on a single structural choice in the Challenger example.

Inference and prediction

With respect to the comments by Gelfand and Mallick, I agree that it would be good to replace θ by θ_S (consider for example, logit *versus* probit on the same data with the same predictors—each X will have a 'β' with both link functions, but they mean quite different things). I do not understand much of the rest of what Gelfand and Mallick say, and I do not agree with most of what I do understand. It seems for them that parametric inference is the grail, and prediction only comes up in model checking. My own view, following de Finetti (1974, 1975), is that prediction of observables is the fundamental activity, and that parameters arise as convenient devices for promoting simplifications, such as conditional independence, in specifying models M (which are just joint probability distributions for observables).

One point of this paper is that it is often helpful in coherently specifying a model to think of it in two pieces, $M=(S, \theta_S)$ (or perhaps $(S, \theta|S)$), so that uncertainty about structure S is assessed and propagated instead of forgotten. When the quantity of interest is a future observable it is *always* possible to view M in this way; when instead the focus is parametric, it is important, in thinking of M as (S, θ_S), that the component(s) of θ_S of real interest have the same meaning across the entire set of structures over which uncertainty is admitted (a kind of identifiability condition; see the last paragraph of Section 4.1). Perhaps we are misunderstanding each other in some way.

The identifiability constraint just mentioned often gives rise in practice to another kind of qualitative difference between inference and prediction as far as model uncertainty is concerned: forcing the meaning of θ_S to be constant in S for parametric inference tends to restrict sufficiently the class of models over which uncertainty may be expressed that the between-model component of variance is often smaller in inference (where this component is typically $O(n^{-1})$) than in prediction (where the effect of this component can be $O(1)$). An example of a predictive $O(1)$ model uncertainty effect arises in linear regression using a greedy algorithm for variable selection (such as all-subsets regression) and then making predictions based on the same data set used to identify the 'best' predictor subset: with diffuse priors on the regression parameters the posterior predictive variance for a future observable y is of the form $(1-R^2)s_y^2\{1+O(n^{-1})\}$, where s_y^2 is the sample variance on the outcome scale and R^2 is the usual coefficient of determination—i.e. sharply overestimating R^2 (as is likely with the S^*-approach in this case) is an $O(1)$ mistake. Thus in prediction problems we will rarely be able to enjoy what Barnard calls 'serendipity', the situation indicated by Cairns's Fig. 11(b). Of course for small and moderate n omitting an $O(n^{-1})$ variance term can also get you into serious calibrative trouble, so in this case model uncertainty can loom large in inference as well.

Posterior model probabilities

Kass, Wasserman and Raftery revisit the topic of approximating posterior model probabilities. It is clear that we shall have to gain additional empirical experience with the success of $O(1)$ approximations to equation (9), bearing in mind that the eventual goal is often to create composite posterior distributions of the form (1) or (6) that exhibit better predictive performance than those obtained with, for example, the S^*-approach. A recent applied example in which equation (11) sharply outperformed S^* is given by Hook and Regal (1994) in the context of capture–recapture estimates of population size in epidemiology (also see Western (1993) for an application of the methods in Section 5 to macrosociology). I do not wish to encourage two wrongs to make a right, but I have seen at least one example in which approximation (11)'s preference for more complex models over that inherent in BIC led to better predictive calibration, because

(a) more complex models tend to have smaller conditional inherent precision when extrapolation is at issue (the situation where propagation of model uncertainty is most urgent) and
(b) the model class $\mathscr{M}'$ employed in that particular example was (as is often the case) retrospectively seen not to be sufficiently rich.

Time series

I have not done full justice here to the subject of model uncertainty in time series, which was raised by Chatfield, Glendinning, Hjorth and Zellner. Dynamic linear models (DLMs, Section 5), which include the Kalman filter as a special case, are well suited to the simultaneous consideration of alternative time series structural specifications and the attendant quantification of between-model uncertainty. For example, consider sequential updating with a discrete state space DLM on the oil price data of Section 6.1. As the yearly actual oil prices in Fig. 1 became known one by one, the posterior structural probability vector would move away from the values in Table 1, which concentrate on the reference scenario, and towards the scenario displayed in Fig. 2 (scenario 10), with an attendant narrowing of the uncertainty bands and a shifting of the 1986 z-score for \$13 from -1.1 towards 0. For more on time series model uncertainty see, for example, Barnett *et al.* (1993a, b), who use MCMC methods to approximate the posterior distribution on (p, d, q) in the ARIMA(p, d, q) modelling class.

Elicitation and graphical presentation of uncertainty

Grieve's comments about graphical displays of the mapping from prior to posterior (Section 7.4) are important—such displays will form a key part of the definitive software test-bed for interactive Bayesian modelling, which still remains to be written (although BUGS (Gilks *et al.* (1994)) is a good start). With regard to elicitation, I strongly support Madigan's attempts to construct non-diffuse structural prior distributions (see the contribution of Glasbey and Gibson). Now that we have a better handle

on computational issues, the next 10 years should see attention shift increasingly, in the implementation of the Bayesian paradigm, to the underemphasized topic of methodology for elicitation of informative priors (e.g. Kadane (1980) and Merkhofer (1987)) and utility functions (e.g. Grayson (1960) and Bernardo and Smith (1994)).

Specifying $p(S)$ diffusely may be a reasonable place to begin (and perhaps even to end) in discrete model expansion, but one point that I was trying to make in Section 3 was that this can lead to problems in continuous expansion when the space $\mathscr{M}$ of all possible models is too big. For instance, I believe that Madigan's good results with diffuse $p(S)$ over all of $\mathscr{M}$ are because he has restricted himself so far to problems in which model uncertainty is finite dimensional. In this regard the example given by Diaconis and Freedman (1986) (mentioned in Section 3) is a cautionary tale which in effect posits a hierarchical model for location inference of the form

$$\alpha \sim \text{Cauchy}, \quad (F|\alpha) \sim \text{symmetrized Dirichlet}(\alpha), \quad (e_i|F) \overset{\text{IID}}{\sim} F, \quad \theta \sim N(0, 1),$$
$$x_i = \theta + e_i, \quad i = 1, \ldots, n,$$

and exhibits a sampling distributionfor the e_i and a true value θ^* that leads to a posterior for θ which does not converge as n increases to point mass on θ^*. Although discussants to Diaconis and Freedman (1986) raised various technical objections to the Dirichlet model, the messages that I would emphasize, in situations like this where $\mathscr{S}$ is infinite dimensional, are that

(a) it is actually quite a strong piece of prior information to claim that $p(S)$ is constant and non-zero over a truly large part of $\mathscr{M}$ and
(b) considerable care is required in arriving at prior probability measures on function spaces of which we are retrospectively proud.

Non-Bayesian approaches to model uncertainty

The example given by Cox provides a nice environment within which to examine the merits of both qualitative and quantitative acknowledgement of modelling uncertainty. Fig. 15 from Cox (1988), to which he refers, dramatically exhibits the effect of various structural assumptions on the likely size of the acquired immune deficiency syndrome (AIDS) epidemic in 1992, when viewed from the perspective of 1987. The most reasonable way to treat structural uncertainty in this situation must depend on the actions to be taken in response to the epidemic's growth. If, for instance, we have only a finite governmental budget to spend on AIDS plus all other pressing problems taken together, so that too much spent on AIDS has adverse consequences that are just as real as those resulting from spending too little, how would Cox recommend hedging appropriately against the structural uncertainty in Fig. 15 except by averaging over it, as maximization of expected utility (MEU) requires? If he has some other behaviour in mind—such as a minimax strategy of some type (compare Tukey's 'taking the worst of what remains')—which he is willing to advocate in this and each of the other governmental problems, I would be happy to put up MEU in a head-to-head simulation, with distance from the retrospectively optimal way to spend the constrained resources as the aggregate criterion. We already know from standard decision theoretic results, frequentist or Bayesian, which method would win: you cannot afford the luxury of minimax—or any other non-MEU choice—simultaneously across many separate subproblems, when the goal is sensible allocation of constrained resources.

With regard to bogus quantification, one strategy for producing what might be termed bogus *non*-quantification is for the statistician to present something like Fig. 15 to the scientist or policy analyst with whom he or she is collaborating and to say, 'This is as far as statistics can go in helping you to figure out what decision to take'. What happens when you do this is that people who must make difficult choices in the face of uncertainty are forced to rely on intuition and heuristics, and the results may easily not be as good as they would have been had we collaborated all the way to the finishing line (see Donnelly's comments).

Incidentally, I am not suggesting, for example, in Cox's AIDS example that we should use the composite (weighted average) predictive distribution for 1992 to decide once and for all, say, the appropriate number of new AIDS hospitals to begin to build in 1987. Adaptive hedging strategies are almost always available that will dominate such a choice, e.g. beginning in 1987 to build facilities that can be converted from AIDS hospitals to some other type of public use structure (and vice versa), and revisiting the situation in (say) 1990, when some of the possibilities in Fig. 15 will have demonstrated themselves to be implausible, to modify the end use of some of the buildings.

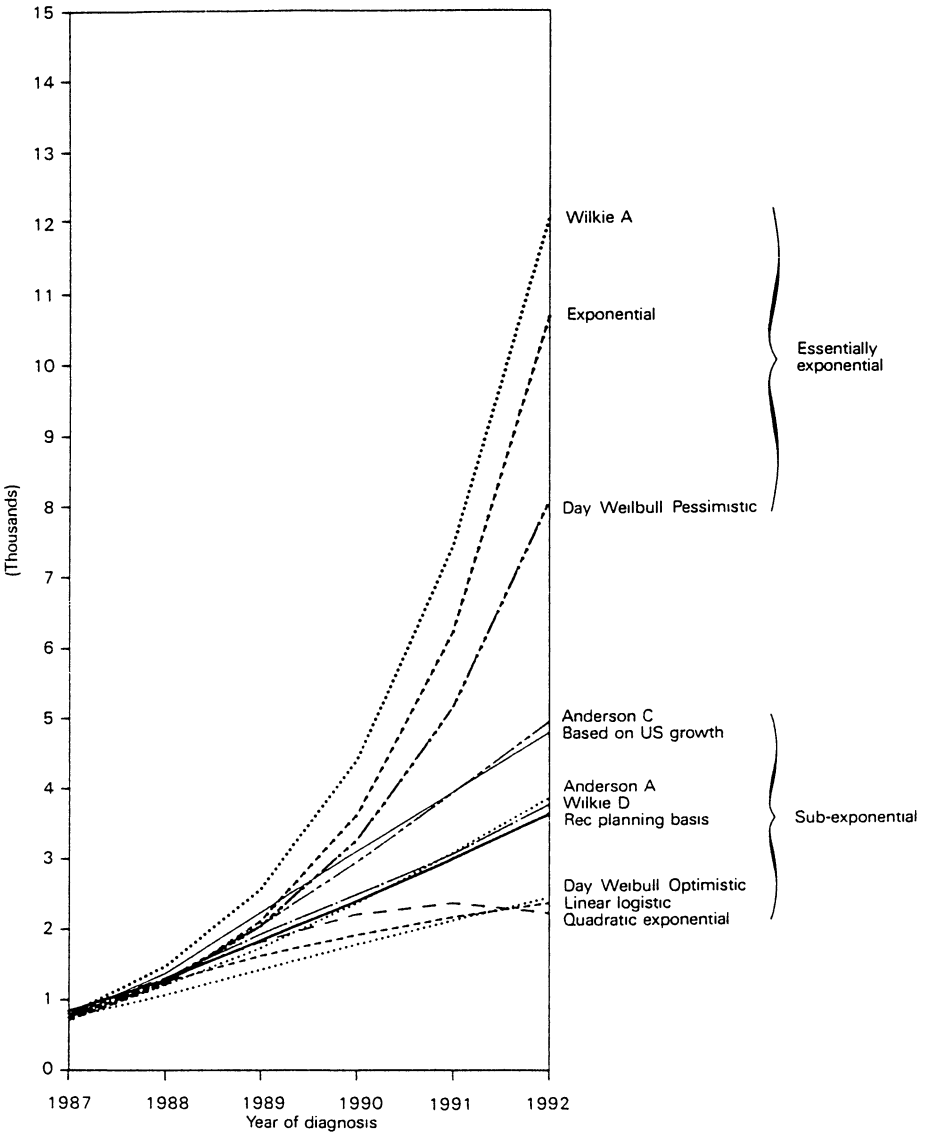

Fig. 15. Summary of predictions of numbers of new cases of AIDS diagnosed in England and Wales (© Crown copyright)

Why, by the way, does Cox regard the specification of prior structural probabilities negatively, when (like all of us) he must face a similar question, particularly when modelling observational data, about his likelihood functions or sampling distributions ('Where do they come from?')? Whatever our statistical faith may be, we are all attempting to use the metaphor of probability, often when no one formally introduced randomness into the data gathering, to help people to obtain useful answers to difficult questions (see Copas's comments). No matter whether the data are experimental or observational in character, it is rarely true in problems of realistic complexity that the data collection process fully specifies the probability model, so we are often faced with Cox's question, and being non-Bayesian does not avoid it. My attitude about prior structural probabilities is that they are an ingredient that arises in employing the metaphor in a way that better accounts for model uncertainty, and—when not much substantive input is available (see the comments by Glasbey and Gibson)—I have the same kind of

engineering pragmatism about this aspect of the modelling process that I would guess Cox has about his sampling distributions with observational data: you scratch something down and see whether it produces sensible looking answers; if not you change it. In this iterative process (e.g. Box (1980) and Rubin (1984)), Bayesian coherence keeps you internally consistent, and predictive calibration—which has an inherent relative frequency character—helps to ensure that you are not too far out of step with the world. I do not see why prior probabilities continue to be such a stumbling-block for people who have not tried to work with them.

Another related pragmatic tool, used to good effect in frequentist applied work (e.g. the nuclear power example in Cox and Snell (1981)) and equally useful here, is sensitivity analysis: in this case, see how much varying the prior structural probabilities across plausible ranges affects the final answer you care about—if their precise specification does not matter much, fine; if it does, add this new uncertainty as a layer in the hierarchy to be averaged over as well. The infinite regress problem hinted at by possibly needing to repeat this programme indefinitely turns out not to be much of a hindrance in practice—a kind of empirical diminishing returns phenomenon typically causes the effect of the new layers in the hierarchy to damp out as their proximity to the final answer decreases.

Pötscher is correct that, from the frequentist point of view, the problem with the S^*-approach is not with the choice of a single 'best' model but with the failure to acknowledge the search on which the choice is based. He is also correct that the suggestion in Section 5.1 to employ the data to help to specify $p(S)$ is circular; indeed model expansion itself suffers from the same flaw if it uses the search leading to S^* as part or all of its inspiration in choosing $\mathscr{S}$. On this issue we face a dilemma: from the Bayesian perspective the ideal is to specify $p(S)$ without looking at the data, but you will then run foul of Cromwell's rule (see Section 5.1) if anything crops up in the data that has already been ruled impossible by your prior. The frequentist approach does not appear entirely satisfying, either, because nobody has yet been able to formalize the full data analytic process, leading to S^*-choices in real world problems of realistic complexity, sufficiently well to pin down its operating characteristics. Model expansion based in part on the search leading to S^* uses the data twice, but to a considerably lesser extent than the S^*-approach itself. The only infallible alternative is omniscience, which outside the world of simulations appears to be in short supply.

Postscript

I would like to close with a historical analogy. It seems to me that we are in about the same position with respect to model uncertainty now as we were in 1910 (say), after Gossett had introduced inference based on the t-distribution but before this relatively new machinery had made its way very far into day-to-day applied work. In the language of this paper, Gossett had noted that the usual practice before 1905—of using the data to estimate the population standard deviation (SD) σ in the denominator of the usual z-ratio, but failing to account for this in the inference—was an instance of the S^*-approach, and a more recent interpretation of his t-distribution remedy, recalling that the t-family is a scale mixture of Gaussian distributions, would be to say that he hierarchically expanded the $N(\mu, \sigma^2)$ model for known σ by adding a layer of uncertainty about σ to the model, and integrated over this new layer rather than using point mass on the sample SD. It took people a while to make use of this new technology because it required an enhancement of their previous computing methods (in Gossett's case, calculation and dissemination of the t-table)—in the same way that it has taken people a while to use Laplace approximations and Gibbs sampling in their routine problem solving—but within 10 or 20 years of Gossett's advance the new approach was well established.

It may be objected in this analogy that structural and parametric uncertainty are of two quite different types, so that it is not fair to draw a correspondence between the choice of (say) a logistic link—or a single AIDS projection in Cox's example—on the one hand and the choice of a particular value of σ on the other, but as Pratt notes these two types of uncertainty, both in effect about nuisance parameters when standard models are expanded hierarchically, do not differ in any basic way. By virtue of recent advances in Bayesian computation analogous to the use of t in place of z, we can now make the same sort of incremental increase in the realism of our modelling that slowly followed Gossett's breakthrough almost a century ago. It is my hope that the next decade will see a parallel advance in more fully accounting for model uncertainty.

REFERENCES IN THE DISCUSSION

Aitkin, M. (1991) Posterior Bayes factors (with discussion). *J. R. Statist. Soc.* B, **53**, 111–142.

Akaike, H. (1985) Prediction and entropy. In *A Celebration of Statistics* (eds A. C. Atkinson and S. E. Fienberg), pp. 1–24. New York: Springer.

Balding, D. J. and Donnelly, P. (1995) Inference in forensic identification (with discussion). *J. R. Statist. Soc.* A, **158**, in the press.

Barnard, G. A. (1994a) Pivotal inference illustrated on the Darwin maize data. In *Aspects of Uncertainty* (eds A. F. M. Smith and P. R. Freeman). Chichester: Wiley.

———(1994b) Pivotal models and the fiducial argument. In *Three Contributions to the History of Statistics: A. W. F. Edwards, Anders Hald, George A. Barnard.* Copenhagen: University of Copenhagen.

Barnett, G., Kohn, R. and Sheather, S. (1993a) Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *Working Paper 93-012.* Australian Graduate School of Management, University of New South Wales, Kensington.

———(1993b) A Bayesian analysis of integrated moving average models. *Working Paper 93-015.* Australian Graduate School of Management, University of New South Wales, Kensington.

Berger, J. O. and Chen, M.-H. (1993) Predicting retirement patterns: prediction for a multinomial distribution with constrained parameter space. *Statistician*, **42**, 427–443.

Berk, R. H. (1966) Limiting behaviour of posterior distributions when the model is incorrect. *Ann. Math. Statist.*, **37**, 51–58.

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

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory.* Chichester: Wiley.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. 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.

Breiman, L. (1992) The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Am. Statist. Ass.*, **87**, 738–754.

Cairns, A. J. G. (1991) Model fitting and projection of the AIDS epidemic. *Math. Biosci.*, **107**, 451–489.

———(1993) Primary components of epidemic models. In *Epidemic Models: Their Structure and Relation to Data* (ed. D. Mollison). Cambridge: Cambridge University Press.

———(1994) Primary component analysis of epidemic models. Submitted to *J. R. Statist. Soc.* B.

Carlin, B. P. and Gelfand, A. E. (1991) An iterative Monte Carlo method for nonconjugate Bayesian analysis. *Statist. Comput.*, **1**, 119–128.

Chatfield, C. (1994) Model uncertainty, data mining and statistical inference. Submitted to *J. R. Statist. Soc.* A.

Copas, J. B. (1993) The shrinkage of point scoring methods. *Appl. Statist.*, **42**, 315–331.

Cox, D. R. (Chairman) (1988) *Short-term Prediction of HIV Infection and AIDS in England and Wales.* London: Her Majesty's Stationery Office.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B, **49**, 1–39.

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

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

Draper, D. (1994) Hierarchical models and variable selection. *Statistics Research Report 94:02.* University of Bath, Bath.

———(1995) Inference and hierarchical modelling in the social sciences (with discussion). *J. Educ. Statist.*, to be published.

Efroymson, M. A. (1960) Multiple regression analysis. In *Mathematical Methods for Digital Computers* (eds A. Ralston and H. S. Wilf), pp. 191–203. New York: Wiley.

Fernholz, L. T. and Fernholz, R. (1986) Autoregressive model identification for multivariate time series. *J. Statist. Computn Simuln*, **24**, 231–243.

de Finetti, B. (1974) *Theory of Probability*, vol. 1. New York: Wiley.

———(1975) *Theory of Probability*, vol. 2. New York: Wiley.

Food and Drug Administration (1987) *Guideline for Submitting Documentation for Stability Studies of Human Drugs and Biologics.* Rockville: Food and Drug Administration.

Freedman, D. A. (1983) A note on screening regression equations. *Am. Statistn*, **37**, 152–155.

Freedman, D. A., Navidi, W. and Peters, S. C. (1986) On the impact of variable selection in fitting regression equations. *Lect. Notes Econ. Math. Syst.*, **307**, 1–16.

Geisel, M. S. (1975) Bayesian comparison of simple macroeconomic models. In *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage* (eds S. E. Fienberg and A. Zellner), pp. 227–256. Amsterdam: North-Holland.

Gelfand, A. E. and Mallick, B. K. (1994) Bayesian analysis of semiparametric proportional hazards models. *Biometrics*, to be published.

Gelman, A. and Meng, X. L. (1994) Model checking and model improvement. In *Practical Markov Chain Monte Carlo* (eds W. Gilks, S. Richardson and D. Spiegelhalter). London: Chapman and Hall. To be published.

Gelman, A., Meng, X. L. and Stern, H. (1995) Bayesian model checking using tail area probabilities. *Statist. Sin.*, to be published.

Gibson, G. J. (1994) System modelling feasibility study. *Interim Report to Scottish Office Agriculture and Fisheries Department.*

Gibson, G. J., Glasbey, C. A., Kempton, R. A. and Elston, D. A. (1993) Modelling in the SARIs—the role of uncertainty. *Report to Agricultural and Food Research Council.*

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994) A language and program for complex Bayesian modelling. *Statistician*, **43**, 169–177.

Glasbey, C. A. (1987) Tolerance-distribution-free analyses of quantal dose–response data. *Appl. Statist.*, **36**, 251–259.

Glendinning, R. H. (1993) Model selection for time series: part 1, a review of criterion based methods. *Memorandum 4730.* Defence Research Agency, Malvern.

Grayson, C. J. (1960) *Decisions under Uncertainty: Drilling Decision by Oil and Gas Operators.* Cambridge: Harvard University Press.

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. Med.*, **12**, 717–736.

Grieve, A. P. (1994) Extending a Bayesian analysis of the two-period crossover to allow for baseline measurements. *Statist. Med.*, **13**, 905–929.

Hannan, E. J. and Deistler, M. (1988) *The Statistical Theory of Linear Systems.* New York: Wiley.

Harrison, P. J. and Stevens, C. F. (1976) Bayesian forecasting (with discussion). *J. R. Statist. Soc.* B, **38**, 205–247.

Heitjan, D. F. and Rubin, D. B. (1990) Inference from coarse data via multiple imputation with application to age heaping. *J. Am. Statist. Ass.*, **85**, 304–314.

Hemerly, E. M. and Davis, M. H. A. (1991) Recursive order estimation of autoregressions without bounding the model set. *J. R. Statist. Soc.* B, **53**, 201–210.

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

——— (1994) *Computer Intensive Statistical Methods, Validation, Model Selection and Bootstrap.* London: Chapman and Hall.

Hook, E. B. and Regal, R. R. (1994) Small-sample adjustments, optimal model selection, and adjustment for model uncertainty in three-source capture–recapture estimates of closed populations. *Technical Report.* School of Public Health, University of California, Berkeley.

Jeffreys, H. (1967) *Theory of Probability*, 3rd edn. London: Oxford University Press.

Judge, G. G. and Bock, M. E. (1978) *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics.* Amsterdam: North-Holland.

Kabaila, P. (1995) The effect of model selection on confidence regions and prediction regions. *Econometr. Theory*, **11**, in the press.

Kadane, J. B. (1980) Predictive and structural methods for eliciting prior distributions. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (ed. A. Zellner), pp. 89–109. Amsterdam: North-Holland.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980) Interactive elicitation of opinion for a normal linear model. *J. Am. Statist. Ass.*, **75**, 845–854.

Kass, R. E. and Raftery, A. E. (1994) Bayes factors. *J. Am. Statist. Ass.*, **89**, in the press.

Kass, R. E. and Wasserman, L. A. (1992a) The surprising accuracy of the Schwarz criterion as an approximation to the log Bayes factor. *Technical Report.* Department of Statistics, Carnegie Mellon University, Pittsburgh.

——— (1992b) A reference Bayesian test for nested hypotheses with large samples. *Technical Report.* Department of Statistics, Carnegie Mellon University, Pittsburgh.

——— (1995) Discussion on Fractional Bayes factors for model comparisons (by A. O'Hagan). *J. R. Statist. Soc.* B, **57**, 131.

Laud, P. W., Ibrahim, J. G., Gopalan, R. and Ramgopal, P. (1992) Predictive variable selection in generalized linear models. *Technical Report.* Division of Statistics, Northern Illinois University, DeKalb.

Lindley, D. V. (1968) The choice of variables in multiple regression (with discussion). *J. R. Statist. Soc.* B, **30**, 31–66.

Madigan, D., Gavrin, J. and Raftery, A. E. (1994) Enhancing the predictive performance of Bayesian graphical models. To be published.

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Statist. Ass.*, **89**, in the press.

Madigan, D. and York, J. (1993) Bayesian graphical models for discrete data. Submitted to *Int. Statist. Rev.*

Mallick, B. K. and Gelfand, A. E. (1994a) Generalized linear models with unknown link functions. *Biometrika*, **81**, 237–245.

——— (1994b) Semiparametric errors-in-variable models: a Bayesian approach.

Merkhofer, M. W. (1987) Quantifying judgemental uncertainty: methodology, experiences, and insights. *IEEE Trans. Syst. Sci. Cyb.*, **17**, 741–752.

Mosteller, F. and Wallace, D. L. (1964) *Inference and Disputed Authorship: the Federalist.* Reading: Addison-Wesley.

Nicholls, D. P., Moles, K., Gleadhill, D. N. S., Booth, K., Rowan, J. and Morton, P. (1986) Comparison of transdermal nitrate and isosorbide dinitrate in chronic stable angina. *Br. J. Clin. Pharm.*, **22**, 15–20.

O'Hagan, A. (1994) Fractional Bayes factors for model comparisons (with discussion). *J. R. Statist. Soc.* B, **57**, 99–138.

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

Poskitt, D. S. and Tremayne, A. R. (1987) Determining a portfolio of linear time series models. *Biometrika*, **74**, 125–137.

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

———(1995) Comment on The effect of model selection on confidence regions and prediction regions, by P. Kabaila. *Econometr. Theory*, **11**, in the press.

Racine, A., Grieve, A. P., Flühler, 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. (1993a) Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Technical Report 255*. Department of Statistics, University of Washington, Seattle.

———(1993b) Bayesian model selection in structural equation models. In *Testing Structural Equation Models* (eds K. A. Bollen and J. S. Long). Beverly Hills: Sage.

Raftery, A. E., Madigan, D. and Hoeting, J. (1993) Accounting for model uncertainty in linear regression. *Technical Report 262*. Department of Statistics, University of Washington, Seattle.

Raftery, A. E., Madigan, D. and Volinsky, C. (1995) Model uncertainty. To be published.

Raftery, A. E. and Richardson, S. (1994) Model selection for generalized linear models via GLIB, with application to epidemiology. In *Bayesian Biostatistics* (eds D. A. Berry and D. K. Stangl). To be published.

Reynolds, R. A. (1980) A study of the relationship between health and income. *Doctoral Dissertation*. Department of Economics, University of Chicago, Chicago.

Ross, S. (1983) *Stochastic Processes*. New York: Wiley.

Ruberg, S. J. and Stegeman, J. W. (1991) Pooling data for stability studies: testing the equality of batch degradation slopes. *Biometrics*, **47**, 1059–1069.

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

Sen, P. K. (1979) Asymptotic properties of maximum likelihood estimators based on conditional specification. *Ann. Statist.*, **7**, 1019–1033.

Smyth, P. (1992) Admissible stochastic complexity models for classification problems. *Statist. Comput.*, **2**, 97–104.

Talbot, M. (1993) LIKELY—linking informal knowledge and expertise to forecasting models. *Report*. Scottish Agricultural Statistics Service, Edinburgh.

Taplin, R. H. (1990) Modeling agricultural field trials in the presence of outliers and fertility jumps. *PhD Dissertation*. Department of Statistics, University of Washington, Seattle.

———(1993) Robust likelihood calculation for time series. *J. R. Statist. Soc.* B, **55**, 829–836.

Taplin, R. H. and Raftery, A. E. (1991) Analysis of agricultural field trials in the presence of outliers and fertility jumps. *Technical Report 218*. Department of Statistics, University of Washington, Seattle.

———(1994) Analysis of agricultural field trials in the presence of outliers and fertility jumps. *Biometrics*, to be published.

Thomas, K. J. (1993) Fechner distributions and the assumption of normality in statistical inference. *MSc Dissertation*. University of Essex, Colchester.

Veres, S. M. (1991) *Structure Selection of Stochastic Dynamic Systems: the Information Criterion Approach*. New York: Gordon and Breach.

Wang, S.-R. and Chen, Z. G. (1985) Estimation of the order of ARMA model by linear procedures. *Chin. Ann. Math.* B, **6**, 53–70.

Wecker, W. E. and Ansley, C. F. (1983) The signal extraction approach to nonlinear regression and spline smoothing. *J. Am. Statist. Ass.*, **78**, 81–89.

Western, B. (1993) Vague theory in macrosociology. *Technical Report*. Department of Sociology, Princeton University, Princeton.

Zellner, A. and Min, C.-K. (1993) Bayesian analysis, model selection and prediction. In *Physics and Probability: Essays in Honor of Edwin T. Jaynes* (eds W. T. Grandy, Jr, and P. W. Milonni), pp. 195–206. Cambridge: Cambridge University Press.