Prequential Analysis, Stochastic Complexity and Bayesian Inference

A. P. DAWID
University College London, UK

SUMMARY

Prequential Analysis addresses the empirical assessment of statistical models, and of their associated forecasting techniques, using techniques borrowed from the methodology of Probability Forecasting. In the theory of Stochastic Complexity, the empirical assessment of a model is based on the minimal length of a coded message needed to transmit the data. It turns out that this is essentially the same as prequential assessment based on the logarithmic scoring rule. These approaches are particularly well suited to model selection, where they provide further justification for the use of Bayes factors, or the asymptotically equivalent Jeffreys-Schwarz-BIC penalized log-likelihood criterion.

This paper surveys the current state of understanding of these new assessment methodologies, emphasizing connexions with Bayesian inference. In particular a general argument for the consistency of a Bayesian model selection procedure is given.

Keywords: PROBABILITY FORECASTING; OPTIMAL CODING; EMPIRICAL ASSESSMENT; CODE-LENGTH; LOGARITHMIC SCORE; LIKELIHOOD; CONSISTENT MODEL SELECTION.

1. INTRODUCTION

Recent years have seen the development of two new general approaches to problems of statistical inference: prequential analysis (Dawid (1984), Dawid (1991a), Dawid (1991b)) and stochastic complexity (Solomonoff (1978), Rissanen (1987), Rissanen (1989)). Although both approaches can be developed and applied in entirely non-Bayesian ways, they are in close sympathy with Bayesian thinking—prequential analysis in fact arose directly out of such thinking, whilst strong Bayesian connexions have become apparent during the development of stochastic complexity. But these two approaches also share a “Popperian” philosophical attitude more commonly associated with non-Bayesian theories of hypothesis testing: that any attempt at describing reality must be measured against empirical evidence, and be discarded if it proves inadequate. Such critical examination may be applied to subjective beliefs about the world no less than to statistical models of it.

I consider prequential analysis and stochastic complexity as conduits linking Bayesian and non-Bayesian approaches to inference, along which insights can flow in both directions. This paper presents a Bayesian overview of these theories and of the relationship between them.

2. PHILOSOPHICAL STANDPOINT

Prequential Analysis and Stochastic Complexity share the view that any theory or model (we use the terms interchangeably) is merely a human attempt to describe or explain reality; and that such models are to be assessed in terms of their success at this task. It is misguided, according to this view, to believe in Nature as “obeying” some theory or model (even an
unknown one). Even if we can find a “completely successful” theory, this does not mean we have identified Nature’s “true model”—some other, distinct theory might be just as successful.

Since, in this view, theories can only be distinguished by means of their predictions about observables, theories of seemingly different form which necessarily make identical predictions must be treated as identical. But it is also possible that two non-identical theories happen to make identical predictions in the particular circumstances of the empirical world, in which case we may call them contingently identical. We shall never, in this world, be able to distinguish between two contingently identical theories—but then again, we shall never need to (for any predictive purpose). Conversely, if two theories do ever make differing predictions, we should be able to distinguish them on the basis of empirical evidence, and thus prefer one over the other as explanations of the observed data. This property—that theories can be empirically distinguished when and only when it makes a difference that they should be—has been termed Jeffreys’s Law by Dawid (1984). It is simultaneously trivial and profound.

My acceptance of the above philosophy in Probability and Statistics has been greatly influenced by de Finetti (de Finetti (1975)). In his view, it is meaningless to regard a coin as having an “intrinsic probability” of landing heads. Probability is a subjective (human) construction (theory), which can meaningfully be attached only to observable events, such as a specified sequence of outcomes. Even though, under exchangeability, the de Finetti representation theorem might be invoked to conjure up a “chance of landing heads” for the coin, this remains a mathematical artifact of the theory, and need not have any counterpart in the real world: while we may believe that the relative frequency of heads will tend to a limit, Nature is not obliged to humour us, and even if such a limit does exist, any particular coin toss could be embedded in different exchangeable sequences, with varying “chances” of heads (Dawid (1985b)). More generally, the (uniquely defined) subjective predictive model can be decomposable in several distinct ways into a “statistical model” and a prior distribution for its parameter (Dawid (1986b)), so that the concept of the “true model” must be a fiction.

De Finetti has also considered carefully the relationship between subjective opinions and the empirical world (see Dawid (1986a)). In particular, proper scoring rules provide the Bayesian with a means of comparing the success of different models and/or prior distributions in forecasting observables, while ideas of calibration and refinement may be used to measure the absolute success of a model.

3. THE PREQUENTIAL APPROACH TO PROBABILITY

In many contexts it is reasonable to regard Nature as producing, sequentially, an infinite data-string \( x = (x_1, x_2, \ldots) \). Of course, this is an idealization, since we shall never observe all of \( x \)—some of the difficulties resulting from this have been discussed by Schervish (1985). But this will be our sole idealization—in particular, since we do not regard Nature as obeying any laws, probabilistic or otherwise, any consideration of “alternative data-strings which might have been produced” must be strictly theory-dependent. The relevant empirical evidence (past and future) is entirely contained in \( x \). In this view Nature, like History, is “just one darn thing after another”.

Let now \( P \) be a joint distribution for a sequence of random variables \( X = (X_1, X_2, \ldots) \). We can enquire how successful \( P \) is as an explanation of the realized outcomes \( x_1, x_2, \ldots \). It is natural to tackle this question sequentially. Let \( P_t \) be the conditional distribution, under \( P \), of \( X_t \), given \( X^{i-1} = x^{i-1} \equiv (x_1, \ldots, x_{i-1}) \). Thus, after observing the first \( i-1 \) elements of \( x \), the “theory” \( P \) outputs the probability forecast \( P_t \) for the next observable \( X_t \). Considered as a rule for generating a distribution \( P_t \) of \( X_t \) for any values of \( i \) and \( x^{i-1} \),
P is a Probability Forecasting System (PFS); and conversely any PFS determines a joint probability distribution P. We can now compare the realized forecasts (P_i) with the realized values (x_i), using probability assessment techniques such as scoring rules, calibration plots etc., to obtain a global assessment of the success of P at explaining x. This predictive, sequential methodology is termed prequential assessment. Note that since such assessments depend on P only through the actual sequence P = (P_1, P_2, . . .) of probability forecasts it makes in the light of the empirical data x, we do not need to know the full structure of P as a PFS—we make no use of conditional probabilities given histories which have not materialized.

3.1. The Prequential Principle

The last property, regarded as a desideratum for a method of assessing P against x, is termed the Prequential Principle. We can consider various assessment methods, and investigate to what extent they respect it.

For example, suppose that, under P, X_i has mean \( \mu_i \) and variance \( \sigma_i^2 \) (these values generally depending on \( X^{i-1} \)). A simple and intuitively sensible statistic for testing the fit of P, based on \( X^n \), is

\[
Y_n = \frac{\sum_{i=1}^{n} (X_i - \mu_i)}{(\sum_{i=1}^{n} \sigma_i^2)^{1/2}}.
\]

Since the value \( y_n \) of \( Y_n \), depends only on P and x, calculation of \( y_n \) respects the Prequential Principle. However, any attempt to use this value to test the validity of P, by comparing \( y_n \) with the "null" distribution of \( Y_n \) under P, would appear to involve aspects of P over and above P. Indeed, if we were only given the actual forecast sequence P we could not know, for instance, whether or not, under P, the \( (X_i) \) were being modelled as independent. If they were, then \( \mu_i \) and \( \sigma_i^2 \) would not depend on \( X^{i-1} \), and we could (under mild conditions) take \( Y_n \) as asymptotically standard normal, thus yielding an approximate test of fit in which \( y_n \) was referred to the \( N(0,1) \) distribution. But what if P did not incorporate such independence? Seillier-Moiseiwitsch and Dawid (1992) have shown, using martingale arguments, that in this case also the limiting distribution of \( Y_n \) under P will (again under mild conditions) be \( N(0,1) \). Consequently, both the value and the asymptotic null reference distribution for \( Y_n \) may be calculated without any knowledge of P beyond its associated forecast sequence P, and the test does therefore respect the Prequential Principle.

3.2. Prequential Probability

Results such as that above, which are numerous, suggest that it might be possible to define "prequential probabilities" for events defined only in terms of two sequences, one being a string of data x, and the other an associated string of probability forecasts P, no overall probability distribution being given. Such a programme, a genuine extension of the standard Kolmogorov framework, is currently being undertaken by Vovk (1990a, 1990b, 1991), who has given rigorous definitions of prequential probability and expectation, and has shown that the traditional limit theorems (laws of large numbers, central limit theorem, law of the iterated logarithm) can all be given valid interpretations within this new framework.
3.3. Successful Explanations

Kolmogorov and others (see Kolmogorov and Uspenskii (1987)), using ideas from algorithmic information theory, have addressed the idea that a specific infinite data-sequence \( \mathbf{x} \) might be considered “random” with respect to a probability model \( P \). In the light of Section 2, we might reinterpret this as stating that a posited model \( P \) is a “completely successful” explanation of the empirical data-sequence \( \mathbf{x} \).

There are various ways of explicating this. Dawid (1985a) formulates a criterion based on the concept of computable calibration. An alternative (stronger) requirement is that, for every alternative model \( Q \), the observed likelihood ratio \( (dQ^n/dP^n)(\mathbf{x}^n) \), based on the first \( n \) terms of \( \mathbf{x} \), be bounded above as \( n \to \infty \) (here both \( P \) and \( Q \) must be restricted to be suitably computable). This is reasonable if we interpret the likelihood ratio as measuring preference for \( Q \) over \( P \), such preference becoming definitive when its value becomes infinite. Since \( (dQ^n/dP^n)(\mathbf{x}^n) \) is an arbitrary (computable) non-negative martingale under \( P \), this requirement may be termed the martingale criterion (Ville (1939)). Both these definitions of the success of \( P \) respect the Prequential Principle.

Both the above properties will be satisfied with probability 1 for sequences generated by \( P \), thus respecting a compelling desideratum for any success criterion. In particular, if \( P \) and \( Q \) are mutually absolutely continuous over the space of infinite sequences, so that they agree as to which events have probability 1, then each will expect the other, as well as itself, to provide a fully successful explanation of the data to be observed. In this case we may regard \( P \) and \( Q \) as (essentially) equivalent.

Under such a success criterion, two not necessarily equivalent probability models \( P \) and \( Q \) may both turn out to be completely successful explanations of the empirical data \( \mathbf{x} \). For consistency with Jeffrey's Law, we should expect that, in this case, \( P \) and \( Q \) yield essentially identical forecast sequences \( \mathbf{P} \) and \( \mathbf{Q} \) for the data-sequence \( \mathbf{x} \). In fact, this holds in an asymptotic sense: under the complete calibration criterion we can show \( \rho(P_i, Q_i) \to 0 \) (Dawid (1985a)), while under the martingale criterion we obtain the stronger result \( \sum_{i=1}^{\infty} \rho(P_i, Q_i) < \infty \) (where \( \rho \) denotes Hellinger distance) (Vovk (1987)). In particular, when \( P \) and \( Q \) are equivalent, each assigns probability 1 to such essential identity of their forecasts (Blackwell and Dubins (1962), Shiryaev (1981)).

4. THE CODING APPROACH TO PROBABILITY

A seemingly quite different approach to assessing the success of a distribution \( P \) is based on the connections between probability distributions and coding systems. We imagine a sender, who observes a (finite) data-string \( \mathbf{x} \), and wishes to transmit this, by means of a coded message, to a receiver. We suppose that the communication channel supports error-free serial transmission of sequences of binary digits, and that, in the case of real \( (x_i) \), the values may be rounded to some chosen precision, so that the set of possible messages is discrete: the actual process of discretization will prove to be asymptotically unimportant. We generically use the corresponding lower-case symbol (e.g. \( p \)) to denote the density or probability mass function of a distribution symbolized by an upper-case symbol (e.g. \( P \)).

Since data-transmission is expensive, the message should be as short as possible. More important, a short code has little redundancy, and so must capture any “pattern” which the data might possess. We therefore search for a coding system \( C \), i.e. an injective map \( C \) from the space \( \mathcal{X} \) of possible message strings into the space \( B^* \) of all binary strings, which will allow short encoding of \( \mathbf{x} \). This coding system must be agreed upon between sender and receiver before transmission can begin. After observation of \( \mathbf{x} \), the success of \( C \) is measured by the shortness of its code for \( \mathbf{x} \).
Clearly, the choice of \( C \) must depend on the nature and extent of the receiver’s prior knowledge about \( \omega \): for example, if he already knows \( \omega \), a message of length 0 will suffice. More generally, the receiver will have certain expectations about \( \omega \), which could be expressed as a subjective probability distribution \( P \) for a random variable \( X \) over \( \mathcal{X} \). Then we might choose \( C \) to minimize the expected code-length under \( P \).

4.1. Prefix Coding Systems

Often it is desirable to transmit a string \((x_1, x_2, \ldots)\) by encoding each symbol separately and concatenating their codewords. If this is to admit unique decoding, then the coding system used for each symbol must be a prefix coding system, with the property that no code-word is the same as an initial segment of any other. This is also desirable when transmitting a single symbol, if the receiver is to know when transmission is over and he can proceed to decoding.

For a finite binary string \( x \), let \( l(x) \) denote the length of \( x \). If \( Q \) denotes the model of Bernoulli trials with probability parameter \( \frac{1}{2} \), then \( Q(Z_i = z_i \text{ for } i = 1, \ldots, l(x)) = 2^{-l(x)} \).

Given a prefix coding system \( C \) over \( \mathcal{X} \), the associated length function \( L \) is defined by \( L(y) \overset{\text{def}}{=} l(C(y)) \) \((y \in \mathcal{X})\). Since the events “an initial sequence of \((Z_1, Z_2, \ldots)\) forms a code-word for \( y \)” are disjoint as \( y \) ranges over \( \mathcal{X} \), the sum of their \( Q \)-probabilities cannot exceed 1, so that we must have

\[
\sum_{y \in \mathcal{X}} 2^{-L(y)} \leq 1. \tag{1}
\]

If \( C \) is complete in the sense that every infinite binary string has an initial segment that is a code-word, then we shall have equality in (1). It is not difficult to show that, whenever (1) holds for an integer-valued function \( L \), there exists a prefix coding system \( C \) (complete if we have equality in (1)) having \( L \) as its length function.

4.2. Codes and Distributions

Suppose that \( L \) is a non-negative but not necessarily integral function on \( \mathcal{X} \), satisfying (1) with equality. Since the integer-valued function \( L' \overset{\text{def}}{=} \lfloor L \rfloor \) satisfies (1), we can construct a prefix coding system whose length-function exceeds \( L \) by at most 1. Applying this argument to coding systems defined over the space \( \mathcal{X}^n \) of sequences of \( n \) symbols from \( \mathcal{X} \), we see that it is possible to encode long sequences with a per-symbol length function differing negligibly from that obtained using \( L \). We shall therefore ignore the fact that \( L \) may be non-integral, and treat it as an achievable length function.

Let \( P \) be a distribution over \( \mathcal{X} \). The function

\[
L_P(y) \overset{\text{def}}{=} -\log p(y) \quad (y \in \mathcal{X}), \tag{2}
\]

(where \( \log \) denotes logarithm to base 2) satisfies (1) with equality, and hence may be considered as defining a complete prefix coding system. Conversely, given any such system \( C \), with length function \( L \), we can treat it as defining a distribution \( P_C \) over \( \mathcal{X} \), where \( p_C(y) = 2^{-L(y)} \).

The entropy of a distribution \( P \) is defined as

\[
H(P) \overset{\text{def}}{=} E_P(L_P(X)) = -\sum_{y \in \mathcal{X}} p(y) \log p(y). \tag{3}
\]
Proposition 1. For any distribution $P$, and prefix coding system $C$ with length function $L$,

i) 
\[ E_P(L(X)) \geq H(P) , \]

with equality if and only if $L \equiv L_P$;

(ii) for all $c > 0$
\[ P(L(X) \leq L_P(X) - c) \leq 2^{-c} . \]

Proof. (i) The “information inequality”,
\[ \sum_{y \in \mathcal{X}} p(y) \log \left( \frac{p(y)}{q(y)} \right) \geq 0 \]  \hspace{1cm} (4)

holds for all probability distributions $P$ and $Q$ over $\mathcal{X}$, with equality if and only if $Q = P$. Take $q(y) \overset{\text{def}}{=} k^{-1}2^{-L(y)}$, with $k \overset{\text{def}}{=} \sum_{y \in \mathcal{X}} 2^{-L(y)} \leq 1$ by (1).

(ii) Whenever $L(y) \leq L_P(y) - c$, $p(y) \leq 2^{-c}2^{-L(y)}$. Sum over all $y$ satisfying this condition and again apply (1). \hfill \triangleleft 

Proposition 1 may be taken as establishing $L_P$ as an optimal length function for a prefix coding system under the distribution $P$ for the unknown message $X$. In particular, if we apply (ii) to the encoding of long sequences of symbols, then the per-symbol message length achieved by any prefix coding system $C$ cannot improve on that given by $L_P$ by more than a negligible amount, with arbitrarily high probability under $P$. These results suggest strongly that, when the receiver’s state of uncertainty about the message to be transmitted is expressed by a distribution $P$, a coding system should be used with length function $L_P$.

A fruitful view of the above argument comes from turning it on its head. Suppose that sender and receiver agree to use a prefix coding system $C$, with length function $L$, to encode the message. We may suppose $C$ to be complete, since otherwise we could shorten every code-word by using the coding system with length-function $L + \log \left\{ \sum_{y \in \mathcal{X}} 2^{-L(y)} \right\}$, which satisfies (1) with equality. The associated distribution $P_C$ is then that distribution under which $C$ is an optimal code. Thus the mere willingness to use $C$ can be regarded as generating a distribution for that object. (Compare this argument with that of Savage (1971), which infers a subjective distribution from choices made in decision problems—viz. that for which these would be the optimal choices.) For example, Rissanen (1983) shows how one can construct a prefix coding system for an arbitrary non-negative integer $n$ whose length-function is essentially $\log c + \log^* (n)$, where $\log^* (n) \overset{\text{def}}{=} \log n + \log \log n + \ldots$ (the sum including all positive iterates) and $c$ is chosen to satisfy (1) with equality. The associated distribution $P^*$ has $p^*(n) = (cn \log n \log \log n \ldots)^{-1}$, and can be shown to be optimal, in a certain sense, amongst all distributions having infinite entropy. Rissanen terms it a “universal prior distribution” for a completely unknown integer.

5. CONNECTIONS

For a particular message $\mathfrak{x}$, we can compare coding systems in terms of the actual length of their codes for $\mathfrak{x}$. If $C$ is a complete prefix coding system, then this length is $- \log p_C(\mathfrak{x})$. This is identical with the negative logarithmic score for assessing the success of the distribution $P_C$ at explaining $\mathfrak{x}$; it can also be interpreted as the observed negative log-likelihood of $P_C$, and as such the difference in code-lengths $C(\mathfrak{x}) - C'(\mathfrak{x})$ is just the observed $\log$
likelihood-ratio in favour of $P_C$ as against $P_C$. We thus see that the criterion of code-length is identical with the use, in Probability Forecasting, of the logarithmic scoring function for assessing a distribution over $\mathcal{X}$ against the empirical data.

The individual terms of a string $x = (x_1, x_2, \ldots)$ can be encoded and transmitted sequentially, using a prefix coding system for the $i$'th term which can be allowed to depend (in a way agreed between the parties before transmission) on the previously transmitted segment $x^{i-1}$. Such a system for encoding strings of arbitrary length may be called a prequential coding system. When prior uncertainty is expressed in a joint distribution $P$ for $\mathcal{X}$, we might try encoding $x_i$ using a prefix code optimal for the conditional distribution, under $P$, of $X_i$ given $X^{i-1} = x^{i-1}$. The length of the code for $x_i$ is then $- \log p(x_i \mid x^{i-1})$, and so the overall code-length for $x^n$ is $- \sum_{i=1}^{n} \log p(x_i \mid x^{i-1}) = - \log p(x^n)$, showing that this prequentially optimal coding is in fact optimal for each $n$. Conversely, any complete prequential coding system determines a probability forecasting system, and hence a joint distribution for $\mathcal{X}$. The length of the code for each term may be interpreted as the contribution of that term to the overall negative logarithmic score, or negative log-likelihood, for the associated distribution. In this way we obtain a fully prequential interpretation of code-length.

For prequential coding systems, Proposition 1 (ii) can be strengthened as follows:

**Proposition 2.** Let $C$ be a prequential coding system with length function $L_C$, and $P$ a distribution for $\mathcal{X}$. Then, with $P$-probability 1, $L_P(X^n) - L_C(X^n)$ is bounded above as $n \to \infty$.

**Proof:** We may suppose $C$ to be complete, since otherwise we could shorten the code-length term by term. Then $(U_n)$, where $U_n = \exp\{L_P(X^n) - L_C(X^n)\} = p_C(X^n)/p(X^n)$, is a non-negative martingale under $P$, and hence is bounded above with $P$-probability 1. $\triangle$

6. STATISTICAL MODELLING

Suppose Nature produces a data-string $x$. In the prequential approach, we might search for a probability forecasting system for $\mathcal{X}$ which will be a good explanation of $x$; from the coding viewpoint, we would like a coding system which yields a short code-length for $x$.

In a sense, this problem is trivial: use the distribution which gives probability 1 to the actual data sequence—equivalent to a code of length 0. Of course this is only possible when the receiver knows the message before it is sent. However, the very existence of such a distribution or coding system does point up problems in the idea of minimizing code-length, or other probabilistic penalty function, when the possibilities are unrestricted.

We might therefore restrict the search to some parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of joint distributions, or $\mathcal{C} = \{C_\theta : \theta \in \Theta\}$ of coding systems. However, there is still a problem. When the sender has observed $x$, she will be able to discover the value of $\theta$ which yields the shortest code, or minimal penalty, for $x$; but the receiver will not be in a position to apply this optimizing coding system without further information.

6.1. Two-part Codes

The first approaches to this problem from a coding viewpoint (see Wallace and Freeman (1987), Rissanen (1985)) involved constructing a two-part code, having a preamble, encoding a value for $\theta$ using a prefix coding system $C$ defined over some discrete subset $H$ of $\Theta$, and a body, which then encodes the data using the optimal code under $P_\theta$. The sender should clearly choose that value of $\theta \in H$ yielding the minimum total code length $l(C(\theta)) - \log p_\theta(x)$ for
the data \( x \) at hand—the "minimum message length" estimator. The minimized total code-length can then be regarded as a measure of the success of the model (in conjunction with the subset \( H \) and coding system \( C \)) at explaining \( x \), yielding the "minimum description length" criterion.

The choice of \( H \) and \( C \) must be agreed in advance. They should be chosen in some \textit{a priori} optimal way—say, so as to minimize the expected total code-length under some subjective distribution \( P^* \) for \( X \). The density of the discrete set \( H \) in \( \Theta \) is particularly crucial: it is suboptimal to specify \( \theta \) too accurately, since this will necessitate a long preamble. (In contrast, increasing the precision with which \( x \) is specified will add a quantity independent of \( \theta \) to the length of the body, and so will not affect the choice of \( H \) and \( C \).)

Suppose we have a regular statistical model for a sequence \( X = (X_1, X_2, \ldots) \), with \( k \)-dimensional parameter space, such that

(i) \( \hat{\theta}_n \xrightarrow{L} \theta \), a distribution with positive density \( \pi(\cdot) \), under \( P^* \); and
(ii) with \( P^* \)-probability 1, \(- \sup_\theta \log p_{\theta}(X^n) \) is of order \( n \),

where \( P^* \) is the subjective distribution for \( X \), and \( \hat{\theta}_n \) is the maximum likelihood estimator of \( \theta \) based on the length \( n \) initial segment \( X^n \) of \( X \). (These conditions will typically hold for \( P^* = P_\Pi \stackrel{\text{def}}{=} \int P_\theta d\Pi(\theta) \), as would be appropriate for a Bayesian who fully believed in the model \( P \).) The optimal density of \( H \) in \( \Theta \) then turns out to be of order \( n^{k/2} \), the length of \( C(\theta) \) is (constant) \(- \log \pi(\theta) + \frac{1}{2} k \log n \) and the optimized total code-length for a data-string \( x^n \) is then asymptotically

\[- \sup_\theta \log p_{\theta}(x^n) + (k/2) \log n, \tag{5}\]

which can thus be considered as expressing, approximately, the success of the model at explaining the data.

### 6.2. Stochastic Complexity

If the "theory" \( P \) is truly believed, and beliefs about \( \theta \) are expressed by a prior distribution \( \Pi \), then the predictive distribution of \( X \) is \( P_\Pi \), and the optimal code-length for \( x^n \) is \( L_\Pi(x^n) \stackrel{\text{def}}{=} - \log p_{\Pi}(x^n) \). If we are given any prefix coding system \( C \) for \( X^n \), with length-function \( L \), then by Proposition 1 (ii), for any \( \epsilon > 0 \), \( L_\Pi(L(X^n)) - L_\Pi(x^n) \leq 2 \log \epsilon \leq \epsilon^2 \), whence, since \( P_\Pi = \int \theta P_\theta d\Pi(\theta) \),

\[P_\Pi \{ \theta : L(X^n) - L_\Pi(x^n) \leq 2 \log \epsilon \leq \epsilon \} \geq 1 - \epsilon.\]

That is, for "most" \( \theta \), the probability is high under \( P_\theta \) that the code-length obtained from \( C \) will not be significantly shorter than \( L_\Pi \). This holds irrespective of the length \( n \) of the data-string.

Now suppose \( C \) is a prequential coding system. By Proposition 2, \( P_\Pi(A) = 1 \), where

\[A \stackrel{\text{def}}{=} (L_\Pi^n(X^n) - L^n(X^n) \text{ is bounded above as } n \rightarrow \infty).\]

Hence \( P_\theta(A) = 1 \). In particular, if \( \Pi \) has an everywhere positive density, then

\[P_\theta(L_\Pi^n(X^n) - L^n(X^n) \text{ is bounded above as } n \rightarrow \infty) = 1 \tag{6}\]

for almost all \( \theta \in \Theta \)—this property is equivalent to the \textit{prequential efficiency} of \( P_\Pi \) as defined by Dawid (1984). V'ieugin and Vovk (1990) have shown a pointwise version of this result:
under weak computability requirements, for every \( \theta \) random with respect to \( \Pi \), the event \( A \) will hold for every data-string \( x \) random with respect to \( P_\theta \) (here "random" means essentially the same as "successfully explained by", under the martingale criterion of Section 3.3). In this strong sense, \( P_\Pi \) thus defines an optimal prequential coding system when the model is supposed true. The corresponding optimal code-length \( L_\Pi(x^n) = -\log p_\Pi(x^n) \) is called the stochastic complexity of the data \( x^n \) relative to the model \( \mathcal{P} \) (this depends on \( \Pi \), but the effect of changing \( \Pi \) is bounded as \( n \to \infty \)). It thus seems compelling to regard \( p_\Pi(x^n) \) as the "likelihood" of the model \( \mathcal{P} \) on data \( x^n \), and to compare different models in terms of their likelihoods, or equivalently the stochastic complexities they yield for the data.

It may be shown (Jeffreys (1976), Schwarz (1976), Akaike (1978)) that, for a regular \( k \)-dimensional statistical model, under conditions (i) and (ii), with \( P^* \)-probability 1 the stochastic complexity is asymptotically equivalent to (5). This may be regarded as establishing the overall optimality of the two-part coding system (although it should be noted that this is not prequential). The negative of (5) can be regarded as a "penalized log-likelihood", the second term correcting for the over-optimism in the first engendered by using that value of \( \theta \) which best fits the data observed. It is interesting to observe this correction arising automatically out of the coding approach.

Rissanen (1986a) showed that, in regular cases, for any prequential coding system \( C \), for all \( \epsilon > 0 \), for almost all \( \theta \in \Theta \)

\[
E_\theta \left( L_n(X^n) - L_\theta^n(X^n) \right) - \left( \frac{1}{2} - \epsilon \right) k \log n > 0
\]  

(7)

for all but finitely many \( n \). This can be regarded as an extension of (i) of Proposition 1. However, the use of the expectation in (7) violates the Prequential Principle. Our simple argument above, which respects that Principle, yields a stronger conclusion, extending Proposition 1 (ii): for almost all \( \theta \), the \( P_\theta \)-probability is 1 that

\[
L_n(X^n) - L_\theta^n(X^n) - \left( \frac{1}{2} - \epsilon \right) k \log n > 0
\]

for all but finitely many \( n \). Moreover, (6) is equally applicable to non-regular cases, where the negative stochastic complexity need not have asymptotic form (5).

6.3. Prequential Coding

For a model \( \mathcal{P} = \{ P_\theta \} \) and data-sequence \( x \), let \( P_{\hat{\theta},i+1} \) be the conditional distribution of \( X_{i+1} \) given \( x^i \) under \( P_\theta \), and \( \hat{\theta}_i \) an efficient estimate (e.g. the maximum likelihood estimate) of \( \theta \) based on data \( x^i \). Dawid (1984) considered the PFS \( P \) defined by \( P_{i+1} = P_{\hat{\theta}_i,i+1} \). The associated prequential coding system encodes \( x^n \) with total length

\[
-\log p(x^n) = \sum_{i=0}^{n-1} -\log p_{\hat{\theta}_i}(x_{i+1} | x^i).
\]  

(8)

Under regularity conditions, (8) will (with \( P^* \)-probability 1) only differ from the stochastic complexity by a bounded quantity as \( n \to \infty \)—this is equivalent to the prequential efficiency of \( P \) for \( \mathcal{P} \). Thus (8) can equally well be considered as the negative log-likelihood of the model \( \mathcal{P} \). In general, for any prequentially efficient \( P \), we can define the prequential likelihood of \( \mathcal{P} \) on data \( x \) as \( p(x) \), all such definitions being in asymptotic agreement up to a (data-dependent) finite positive scale-factor.
6.4. Consistent Model Selection

Suppose we wish to choose between a finite or countable collection \( \{ \mathcal{P}_j : j = 1, 2, \ldots \} \), of regular parametric statistical models for \( X \), with \( \mathcal{P}_j = \{ P_{j, \theta_j} : \theta_j \in \Theta_j \} \), \( \Theta_j \) having finite dimension \( k_j \). For each \( j \), introduce a prior distribution \( \Pi_j \) with positive density \( \pi_j(\cdot) \) over \( \Theta_j \). Let

\[
P_j = \int_{\Theta_j} P_{j, \theta_j} \pi_j(\theta_j) \, d\theta_j
\]

be the marginal distribution of \( X \) derived from \( (\mathcal{P}_j, \pi_j) \), so that the stochastic complexity of \( x^n \) relative to \( \mathcal{P}_j \) may be taken as \(- \log p_j(x^n)\).

We shall suppose the \( (P_j) \) to be mutually singular, since otherwise we could have no hope of distinguishing the various models even on the basis of infinite data. This property will generally be easy to verify when the models involve quite different distributions. Alternatively, suppose say \( \mathcal{P}_1 \subset \mathcal{P}_2 \), with \( \Theta_1 \) a lower-dimensional submanifold of \( \Theta_2 \). Suppose further that \( \mathcal{P}_2 \) admits consistent estimation, so that there exists a sequence of statistics \( (T_n) = (T_n(X_n)) \) with \( T_n \overset{a.s.}{\to} \theta_2 \) under \( \mathcal{P}_2, \theta_2 \), all \( \theta_2 \in \Theta_2 \). Then the event \( \lim \frac{T_n}{\Theta_1} = \lim \frac{T_n}{\mathcal{P}_2} \) has \( P_1\)-probability 1 but \( P_2\)-probability 0, so that \( P_1 \) and \( P_2 \) are mutually singular in this case.

Now take a sequence \((\alpha_j)\) with \( \alpha_j > 0 \), \( \sum_i \alpha_i = 1 \), and define \( P_0 = \sum_j \alpha_j P_j \), the overall marginal distribution for \( X \) of a Bayesian who believes that, with probability \( \alpha_j \), the data are generated from a distribution in \( \mathcal{P}_j \), and who, conditional on this, has prior distribution \( \Pi_j \) for its parameter. We can write \( P_0 = \alpha_1 P_1 + (1 - \alpha_1)Q \), where \( Q = \sum_{j \geq 1} \{ \alpha_j / (1 - \alpha_1) \} P_j \), and thus \( P_1 \) and \( Q \) are mutually singular. Hence with \( P_1\)-probability 1 the likelihood-ratio \( q(X^n) / p_1(X^n) \) based on the first \( n \) observation tends to 0 as \( n \to \infty \). From this we deduce that \( (\alpha_j p_j(X^n)) / (\alpha_1 p_1(X^n)) \to 0 \), uniformly in \( i \), with \( P_1\)-probability 1, and hence with \( P_1, \theta_1 \)-probability 1 for almost all \( \theta_1 \in \Theta_1 \). It follows that, with \( P_1, \theta_1 \)-probability 1 for almost all \( \theta_1 \in \Theta_1 \), the “adjusted prequential likelihood” \( \alpha_j p_j(x^n) \) will, for large enough \( n \), be maximized at \( j = 1 \). Since any member of \( \{ \mathcal{P}_j \} \) may be taken instead of \( \mathcal{P}_1 \), we deduce that the model-selection method which proceeds by maximizing the adjusted prequential likelihood, or equivalently minimizing the “adjusted stochastic complexity” of \( x^n \), \(- \log p_j(x^n) - \log \alpha_j \), will be (almost everywhere) consistent.

Note that the above result will continue to hold if we replace \( P_j \) by any other prequentially efficient PFS for \( \mathcal{P}_j \), for example that described at 6.3 above. In the case of a finite number of normal regression or autoregression models (where the correction term \(- \log \alpha_j \) may be ignored) this yields the predictive least squares model selection method (Rissanen (1986b), Hermerly and Davis (1989)). The argument may similarly be used to show the consistency of the method of choosing \( j \) to maximize the penalized log-likelihood \( \sup_{\theta_j \in \Theta_j} \log p_j(\theta_j(x^n)) - (k_j/2) \log n \), which approximates the negative stochastic complexity (Speed and Yu (1989)). When choosing from a countably infinite collection of models, the above argument continues to apply, so long as the correction term is not neglected. This provides a more straightforward consistent model-selection method than, say, that of Hermerly and Davis (1991).

7. FURTHER ISSUES

7.1. Model Failure

Arguments such as in Section 6.4 (amongst many others) are only relevant when the data-string \( x \) is well-explained by a member of one of the model-classes considered. In this case the use of the code-length criterion, with its connections to the logarithmic score, maximum
likelihood estimation and efficiency, can be regarded as optimal. But it is also important to investigate the behaviour of such strategies when the model fails: for example, under the assumption that \( x \) is well-explained by some other distribution \( P^* \), the "true model" (or the subjective distribution of a Bayesian, who will then believe, with probability 1, that this will hold) In such a case alternative strategies, say those based on other prequential assessment criteria, also deserve attention. Dawid (1991a) has investigated the "prequential consistency" of parameter estimates, based on a general loss-function, assuming a false model.

In model selection, we might want to apply predictive least squares, essentially a prequential quadratic score, to regression with non-normal errors: many of the results survive this extension. Alternatively, the "true model" might involve an infinite number of regressors (Breiman and Freedman (1983), Shibata (1980)). Further study of the general behaviour of model-based strategies when the model is false is desirable.

7.2. Infinitely Many Parameters

For a finitely parametrized regular model, any choice of prior distribution \( \Pi \) is essentially equivalent to any other—so long as it has full support, i.e. admits a positive density. Any two such priors will be mutually absolutely continuous on the parameter-space, implying the same for their associated predictive distributions over sequence-space, and allowing an essentially unique definition of stochastic complexity. However, when we turn to models having infinitely many parameters, there is no clear analogue of the concept of full support: any prior puts all its mass on some "thin" event. (For example, Dawid (1988) considers a normal regression with infinitely many potential predictor variables, and shows that any natural conjugate prior distribution assigns probability 1 to the event that the parameters have a certain "determinism" property which will, in many contexts, be unacceptable as a description of real beliefs.) Correspondingly, any two distinct priors—and likewise their associated predictive distributions—are generically mutually singular. We can still deduce that, given any prefix coding system \( C \), the coding system based on the predictive density \( p_{\Pi} \) will be at least as good as \( C \) under \( P_\theta \) for any \( \theta \in \Theta \), where \( \Phi \subseteq \Theta \) has \( \Pi \)-probability 1; but since \( \Phi \) is "thin" in \( \Theta \), and dependent on \( \Pi \), this result loses the force it had in the finite-parameter case. One possible approach is to set up a hierarchical structure for the prior, specifying the conditional distributions for \( \theta \) given \( \omega \) say, where \( \omega \) is finite dimensional. Integrating out \( \theta \) then yields a finitely parameterized model for \( \mathbf{X} \), to which the earlier results apply. However, the stochastic complexity will still depend on the assumed "hyper-model" for \( \theta \). In full generality, specification and verification of appropriate optimality criteria for the case of infinitely many parameters present a deep challenge.

7.3. Ad hoc Strategies

When optimality theory fails us, or we have no model, we can still investigate the theoretical or empirical performance of intuitively sensible coding systems or prequential strategies. The coding approach is not restricted to data arriving in sequence, and can be applied to complex problems of data analysis where full probabilistic specification of both model and prior would be difficult or impossible: Patrick and Wallace (1982) describe an application to megalithic geometry. For sequence data, the prequential approach is more general, since its probabilistic assessment methods need not be based on the logarithmic scoring rule (Dawid (1991b)). Vovk (1990c) has investigated a form of "Bayesian mixture" over a countable set of strategies for forecasting, when faced with an arbitrary loss-function, and has shown that it will, in a certain sense, be almost as good as any in the original set, for any data. Work is needed further to define and characterize "good strategies" in the absence of a model.
REFERENCES


DISCUSSION

J. RISSANEN (IBM Almaden Research Center, USA)

This is an elegant survey of the two related ideas, prequential analysis and stochastic complexity, and their role in statistical inference, modeling, and the interpretation of probabilities. Of the several deep issues that deserve to be discussed extensively I shall touch here on just a few. I share Professor Dawid's philosophical view that there exist no unique 'true' models non intrinsic probabilities. Instead, any physical interpretation of the probability is to be done relative to a model. Although any model can legitimately be chosen, the selection should be done by criteria which admit a meaningful data dependent interpretation. I know of only two 'scoring rules' which satisfy this requirement, the code length and its special case the prediction error.

As the main point in my discussion I would like to try to clarify the somewhat subtle relationship between these ideas and model selection within the Bayesian framework. Consider an indexed family of classes \( \{p(x|\theta_\alpha, \alpha)\} \) of parametric distributions as 'models' of the data \( x \). The Bayesian principle amounts to picking the class with the maximum posterior probability \( P(\alpha|x) = P(x|\alpha)\pi(\alpha)/P(x) \), for which we need, in addition to the prior \( \pi(\alpha) \) for the classes, also the priors \( \pi_\alpha(\theta_\alpha) \) with which to eliminate the parameters by integration

\[
P(x|\alpha) = \int P(x|\theta_\alpha, \alpha)d\pi_\alpha(\theta_\alpha).
\]

Although the principle has appeal, the difficulty in finding the two types of distributions to capture prior knowledge poses a severe restriction, which has led to a search for 'empirical' data dependent priors. However, since the same principle cannot be used for the task it is unclear what a 'good' prior should be, and instead the procedures employed are more or less ad hoc.

The principle of the minimum description length (MDL), again, declares in broad terms that the best model/class is the one which permits the shortest encoding of the observed data together with the model/class itself, when advantage is taken of the constraints in the data prescribed by the model/class. To make sure that the encoded data can also be decoded an agreement about the model classes as sort of common knowledge is required, which is shared by the imagined encoder and the decoder alike. This is in contrast with the prior knowledge about both the models and their classes in the Bayesian principle, which can even be of private subjective kind. In a widely applicable form the principle calls for evaluating first the code length \( L(x|\alpha) = -\log P(x|\theta_\alpha(x), \alpha) + L(\theta_\alpha(x)) \), where the second term is the code length for the optimally truncated ML parameter estimates, which at the same time serves as a formalization of the elusive idea of model complexity. To get this length, we may use a prior \( \pi_\alpha(\theta_\alpha) \), which, however, may not provide the shortest encoding of \( \theta_\alpha(x) \),
or we may encode the parameters by universal means, which then defines a data dependent 'prior'. In any event, the optimal 'prior' must reflect the complexity of the encoding task, and no integral of type (1) need be evaluated, which permits fitting even 'nonparametric' models. In the second and the final step the class $\hat{\alpha}(x)$ is sought which minimizes the code length $L(x) = L(x|\hat{\alpha}(x)) + L(\hat{\alpha}(x))$, where the second term denotes the code length needed to encode the optimal class.

As a final comment, I would like to point out that the prequential approach with whatever probability assessment is selected cannot be reduced to the posterior maximization principle. By contrast, the prequential probability is equivalent with a predictively calculated code length, and if the probability assessment is done by the MDL principle, the resulting model and the probability forecasting system have the basic asymptotic optimality properties required from a 'completely successful' explanation of the empirical data sequence, as so eloquently described by Professor Dawid. Moreover, experience shows that the results are generally very good, even non-asymptotically, although attention must be paid to an inherent start-up problem, which arises from poor initial predictions. Unless checked, these can grossly penalize a complex model, which at the later stages may turn out to provide superior predictions.

M. GOLDSTEIN (University of Durham, UK)

I am having a little difficulty understanding the prequential principle. Clearly, the comparison between what theory predicts and what actually happens should be an important component of any methodology. But, during the conference presentation of this paper, I got the impression (perhaps, wrongly) that the prequential principle claimed that this was the only comparison of any importance. Asymptotically, this may be true, but for actual real world observations surely it is false, because it omits our heuristic understanding as to why our various models are performing well or badly.

Perhaps the principle might be clarified by applying it to the following problem. We wish to forecast hours of sunshine in Spain. Our sampling method is to observe the weather each time we come to a Valencia meeting on Bayesian statistics. We have three forecasting methods. Method one always forecasts "very sunny". Method two follows the local weather forecast. Method three is more mysterious but so far has predicted sunshine with amazing accuracy.

On the prequential principle, method three is ahead. But suppose we happen to notice that the conditional forecast of method three for sunshine at Valencia 5 (say) gives sunshine with probability 2. Should not this forecast, which has not yet occurred, and may never be observed, influence our assessment of this method?.

Go back to the comparison of methods one (always forecast sunny) and method two (use weather forecast). Overall, it may be that, on the current data, method one outperforms method two by a wide margin on all reasonable scoring rules. However, our sample of observations on Valencia 4 was taken under somewhat different circumstances to the previous meetings (i.e., earlier in the year). Might this imply that good performance on rule two at Valencia 4 should carry extra weight as compared to the preceding observations? If so, this extra weight derives from our heuristic understanding as to why rule two should outperform rule one under such circumstances. It is this general level of insight into the conceptual strengths and weaknesses of our forecasting methods which does not appear to be captured in prequential type principles.

D.V. LINDLEY (Somerset, UK)

This paper represents a substantial contribution to that method within the Bayesian paradigm that recognizes only observables and dispenses with parameters. It admirably succeeds
in bringing together several strands of theory, yet always remains within the study of observables. I have two related questions. The coding theorem is rarely applicable because of noise in the system. The logarithmic scoring-rule has the defect of ignoring the topology of the space (it is local) and consequently does not allow for one’s being nearly right. Can the ideas be generalized to incorporate these two features of noise and proximity, which themselves may be related?

TRAN VAN HOA (Wollongong University, Australia)

As an econometrician who has done substantial analysis in economic modelling, I find the paper by Professor Dawid interesting and stimulating. It is interesting because it provides me (or other users of the statistical tools) with a good review of the foundation for the empirical assessment of alternative econometric models. It is stimulating because it gives impetus to statistical or econometric modellers to explore further methodologies or modifications to improve their model selection procedure in a context peculiar to their field of interest.

Professor Dawid’s interpretation of the Popperian philosophical attitude associated with prequential analysis is akin to the idea of empiricism popularized by Nobel laureate Milton Friedman (formerly of Chicago University) in relation to economic hypothesis testing: any attempt at describing reality must be measured against empirical evidence. My concern is not with this line of argument but with the fact that, in practically all important economic modelling applications, empirical evidence is not what it should be. I can give a simple example of this. Suppose we are interested in measuring the standard of living of university academics in the United Kingdom. But the concept of “standard of living” is elusive or cannot be measured exactly. As a result, some approximation or even substitute is used. Since measurement errors are usually not known, how much is the result from prequential analysis different from what it should be when this phenomenon occurs?

When we deal with models fitted to data with unknown measurement errors, successful explanations or model failure may not tell the whole story. The problem is compounded when the models themselves are misspecified. In our area of economic research when only so much can be obtained from observed data or specified models, we echo Professor DeGroot’s motto of “practise, practise and practise” with “theorize, theorize and theorize more”. The immediate relevance of this principle in economics is in the determination of the exchange rates in international finance and in the explanation of the business cycles.

REPLY TO THE DISCUSSION

I am very grateful to the discussants for raising a number of interesting points.

Rissanen effectively dismisses all scoring rules other than code length (= negative logarithmic score). I find this a most unimaginative position. As Lindley points out, the logarithmic criterion ignores the topology of the space—indeed it can be characterized as the unique proper scoring rule which does so. This locality property will often be undesirable, however. Thus suppose we are attempting to assess the prognosis of a sick person, and consider two possible forecast distributions:

<table>
<thead>
<tr>
<th></th>
<th>Full recovery</th>
<th>Partial recovery</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>.7</td>
<td>.2</td>
<td>.1</td>
</tr>
<tr>
<td>(ii)</td>
<td>.1</td>
<td>.2</td>
<td>.7</td>
</tr>
</tbody>
</table>

If we observe the outcome “partial recovery”, we might well regard distribution (i) as more successful than (ii), in spite of the fact that they both earn the same logarithmic score. This attitude could stem from the intuitive feeling that “full recovery” is closer to “partial recovery” than is “death”, or could be formalized by developing a decision structure, with
suitable loss function, for which the act optimal under distribution (i) would have smaller loss (when in fact the outcome is "partial recovery") than that optimal under (ii). The derivation of a proper scoring rule from such a decision structure is described by Dawid (1986a): it will typically be non-local. If we are concerned with using the data for decision-making, it would thus seem reasonable to use something other than the logarithmic score, or code-length, to assess performance. The prequential approach, unlike that based on coding, can handle this situation with ease (see for example Section 8 of Dawid (1991a)).

This said, there does appear to be something special about the logarithmic score in the particular case that we are convinced of the truth of a regular parametric model. For then choosing the parameter-value to maximize this score will deliver the maximum likelihood estimate, which, being fully efficient, will generally be superior (at least asymptotically) to that based on minimizing some other, loss-based, discrepancy—even when the purpose at hand is to plug in the chosen value and minimize expected loss. But this seemingly universal optimality is not robust to model failure, whereas the direct loss-based assessment typically will be.

Rissanen complains that the "Bayesian principle" (which I suspect few Bayesians would recognize as such) of maximizing posterior probability is inapplicable to the choice of a good prior distribution. But one of the purposes of my paper is to point out that—at any rate within the "small world" of a regular parametric model—any prior with full support is asymptotically as good as any other. The only possible criteria for choice of prior must therefore be subjective. And indeed Proposition 1 provides an argument for using, in the analysis, one's true subjective probabilities. When we move outside the small world, there arises the problem of choosing an optimal formal prior for use with a false model: this raises some new and important issues which deserve deeper attention, but the need to take prior opinion into account is in no way diminished.

The seemingly more objective approach of MDL relies on having a way to describe (encode) a non-random object such as a model-parameter or model-class descriptor. Although one can come up with clever and intuitively sensible coding systems, there remains something ultimately ad hoc about these, quite as arbitrary as the choice of a prior. What the coding approach does have to offer the Bayesian is a new approach to choosing a prior, based on constructing a code rather than direct subjective assessment. This allows application to complex and non-traditional problems, but its philosophical implications deserve further study.

I agree with Rissanen that the prequential approach does not reduce to posterior maximization, notwithstanding any impression to the contrary which may be left by Section 6.4 of the paper. Indeed the "compleat prequentialist" solution to the model selection problem does not attempt to maximize, but simply uses the full mixture distribution $P_0$ as a probability forecasting system. Since, for any $j$, $P_j$ is absolutely continuous with respect to $P_0$, it follows as in Section 3.3 that the probability forecasts made by $P_0$ will be asymptotically indistinguishable from those made by $P_j$, with $P_j$-probability 1, and hence with $P_j, \theta_j$-probability 1 for almost all $\theta_j$. This simple argument does not even require the various \{P_j\} to be mutually singular: if they are not, it is possible that no definitive decision as to the correct model can be made, but this will not matter since, by Jeffrey's Law, all remaining candidates will be making essentially identical forecasts.

Rissanen's last point concerns finite data-sequences and "start-up problems". It seems to me perfectly reasonable that a complex model should be heavily penalized in the initial stages, since the slow rate at which its parameters can be learned means that it may for a long time predict more poorly than a simpler "incorrect" model. In this case I would rather
use the simple model until the data are sufficiently extensive as to demand more detailed description. In general, the complexity of the model used should increase with the amount of data available: frequentist strategies for achieving this are considered by Dawid (1991b).

Lindley asks about the consequences of noise in the transmission system. My initial reaction to this point was that it was attempting to over-stretch the coding metaphor, but Tran Van Hoa’s discussion suggests that there may be a real problem here, related to the possibility that the the receiver is not observing the variables he thinks he is. At one level, one might attempt to model all aspects of the system, including measurement errors and the relationships between variables and their proxies, so as to arrive at a forecast distribution for whatever it is one is in fact observing: then the theories discussed in the paper become directly applicable. However, if the predictions are poor, it will be impossible to identify whether this is due to the poverty of the fundamental theory (about unobservables) or to that of the ancillary assumptions relating fundamental quantities with observables. It is just possible that the theory of communication across a noisy channel may be able to offer some insight into a better statistical treatment of this difficulty, but my feeling is that it is basic and irremovable.

Goldstein raises some sensible objections to the Frequentist Principle—these echo some of the comments of Schervish (1985). But I think these objections are largely answered if one considers the Principle as applying at the level of “likelihood” rather than “posterior”: that is, in so far as the data distinguish between different models, they will not separate those which have yielded identical forecasts. This is not, of course, to say that we cannot have any other reasons for choosing between them, whether based on their internal structure or on our prior beliefs. Asymptotically such additional considerations will be swamped by the data, and thus for extensive data the Frequentist Principle can indeed be regarded as applying to our “posterior” assessments of models; but Goldstein’s comments do point up the need for non-asymptotic circumspection.