Bayesian Model Selection Based on Proper Scoring Rules

A. P. Dawid M. Musio

September 19, 2014

Abstract
Bayesian model selection with improper priors is not well-defined because of the dependence of the marginal likelihood on the arbitrary scaling constants of the within-model prior densities. We show how this problem can be evaded by replacing marginal log likelihood by a homogeneous proper scoring rule, which is insensitive to the scaling constants. Suitably applied, this will typically enable consistent selection of the true model.

Keywords: consistent model selection; homogeneous score; Hyvärinen score; prequential

1 Introduction
The desire for an “objective Bayesian” approach to model selection has produced a wide variety of suggested methods, none entirely satisfactory from a principled perspective. Here we develop an approach based on the general theory of proper scoring rules, and show that, suitably deployed, it can evade problems associated with arbitrary scaling constants, and deliver consistent model selection.

2 Bayesian Model Selection
Let $\mathcal{M}$ be a finite or countable class of statistical models for the same observable $X \in \mathcal{X} \subseteq \mathbb{R}^k$. Each $M \in \mathcal{M}$ is a parametric family, with parameter $\theta_M \in \mathcal{T}_M$, a $d_M$-dimensional Euclidean space; when $M$ obtains, with parameter value $\theta_M$, then $X$ has distribution $P_{\theta_M}$, with Lebesgue density $p_M(x \mid \theta_M)$. Having observed data $X = x$, we wish to make inference about which model $M \in \mathcal{M}$ (and possibly which parameter-value $\theta_M$) actually generated these data.

A subjective Bayesian would begin by assigning a discrete prior distribution over $\mathcal{M}$, with $\alpha(M)$, say, the assessed probability that the true model is $M \in \mathcal{M}$;
and, within each model $M$, a prior distribution $\Pi_M$ for its parameter $\theta_M$ (to be interpreted as describing conditional uncertainty about $\theta_M$, given the validity of model $M$). For simplicity we suppose that $\Pi_M$ has a density function, $\pi_M(\theta_M)$, with respect to Lebesgue measure $d\theta_M$ over $T_M$.

The predictive density function of $X$, given only the validity of model $M$, is

$$p_M(x) = \int_{T_M} p_M(x \mid \theta_M) \pi_M(\theta_M) d\theta_M. \quad (1)$$

This can be thought of as a hybrid between an “objective” component, $p_M(x \mid \theta_M)$, and a “subjective” component, $\pi(\theta_M)$.

Considered as a function of $M \in \mathcal{M}$, for given data $x$, $p_M(x)$ given by (1) — or any function on $\mathcal{M}$ proportional to this—supplies the marginal likelihood function, $L(M)$, over $M \in \mathcal{M}$, based on data $x$:

$$L(M) \propto p_M(x). \quad (2)$$

The posterior probability $\alpha(M \mid x)$ for model $M$ is then given by Bayes’s formula:

$$\alpha(M \mid x) \propto \alpha(M) \times L(M) \quad (3)$$

where the omitted multiplicative constant is adjusted to ensure $\sum_{M \in \mathcal{M}} \alpha(M \mid x) = 1$. In particular, the odds, $\alpha(M_1)/\alpha(M_2)$, in favour of one model $M_1$ versus another model $M_2$, are multiplied, on observing $X = x$, by the Bayes factor $L(M_1)/L(M_2)$.

However, although the Bayes factor is “objective” to the extent that it does not involve the initial discrete prior distribution $\alpha$ over the model space $\mathcal{M}$, it does still depend on the prior densities $\pi_{M_1}$, $\pi_{M_2}$, within the models being compared. As shown in [Dawid (2011)] if the data are independently generated from a distribution $Q$, the log-Bayes factor, $\log L(M_1)/L(M_2)$, behaves asymptotically as $n\{K(Q, M_2) - K(Q, M_1)\} + O_p(n^{\frac{1}{2}})$ when $K(Q, M_2) > K(Q, M_1)$, where $K(Q, M)$ denotes the minimum Kullback-Leibler divergence between $Q$ and a distribution in $M$; while, if $Q$ lies both in $M_1$ and in $M_2$ (so that $K(Q, M_2) = K(Q, M_1) = 0$), with $q(x) \equiv p(x \mid M_1, \theta_1^*) \equiv p(x \mid M_2, \theta_2^*)$ say, we have log Bayes factor

$$\log \frac{L(M_1)}{L(M_2)} = \frac{1}{2} (d_{M_2} - d_{M_1}) \log \frac{n}{2\pi e} + \log \frac{\rho(\theta_1^* \mid M_1)}{\rho(\theta_2^* \mid M_2)} + V, \quad (4)$$

where $\rho(\theta \mid M) = \pi_M(\theta)/\{\det I_M(\theta)\}^{\frac{1}{2}}$ is the “invariantised” prior density with respect to the Jeffreys measure on $M$; $V = O_p(1)$, with asymptotic expectation 0; and the dependence of $V$ on the prior specification is $O_p(n^{\frac{-1}{2}})$.

We thus see that, at any rate for comparing models of different dimension, the dependence of the Bayes factor on the within-model prior specifications is typically negligible compared with the leading term in the asymptotic expansion. Nevertheless many Bayesians have agonised greatly about that dependence, and have attempted to determine an “objective” version of the Bayes factor. The
most obvious approach, of using improper within-model priors, is plagued with
difficulties: the term \( \rho(\theta^* | M) \) is perfectly well-defined when we have a fully
specified prior density, integrating to 1; but when the prior density is non-
integrable this function is specified only up to an arbitrary scale factor—and
(4) will depend on the chosen value of this factor. A variety of ad hoc methods
have been suggested to evade this problem (see for example O’Hagan (1995),
Berger and Pericchi (1996)). These methods are necessarily somewhat subtle—
one might even say contorted—and often do not even respect the leading term
asymptotics of (4).

In Dawid (2011) it was argued that the problem of model selection with
improper priors can largely be overcome by focusing directly on the posterior
odds, rather than the Bayes factor, between models. An alternative approach,
that we develop here, is to replace the Bayes factor by something different (but
related), that is insensitive to the scaling of the prior. For preliminary accounts
of this idea, see Musio and Dawid (2013); Dawid and Musio (2014).

3 Proper Scoring Rules

The log-Bayes factor for comparing models \( M_1 \) and \( M_2 \) is

\[
\log p_{M_1}(x) - \log p_{M_2}(x).
\]  

(5)

One way of interpreting (5) is as a comparison of the log scores (Good 1952)
of the two predictive density functions, \( p_{M_1}(\cdot) \) and \( p_{M_2}(\cdot) \), for \( X \), in the light of
the observed data \( x \). That is, defining \( S_L(x, Q) = -\log q(x) \), for any proposed
distribution \( Q \) with density function \( q(\cdot) \) over \( X \), and \( x \in X \), we can interpret the log score \( S_L(x, Q) \) as a measure of how badly \( Q \) did at forecasting the
outcome \( x \); then the log-Bayes factor measures by how much the log score for
\( M_1 \) (using the associated predictive density) was better (smaller) than that for
\( M_2 \).

Now the above definition of the log score, \( S_L(x, Q) \), is just one of many
functions \( S(x, Q) \) having the property of being a proper scoring rule (see e.g. 
Dawid (1986)): this is the case if, defining \( S(P, Q) \) as the expected score,
\( E_{X \sim P} S(X, Q) \), when \( X \) has distribution \( P \), \( S(P, Q) \) is minimised, for any given
\( P \), by the “honest” choice \( Q = P \). Associated with any proper scoring rule is a
generalized entropy function:

\[
H(P) := S(P, P),
\]

and a non-negative discrepancy function:

\[
D(P, Q) := S(P, Q) - H(P).
\]

These reduce to the familiar Shannon entropy and Kullback-Leibler discrepancy
when \( S \) is the log score.

In this paper we explore the implications and ramifications, for Bayesian
model selection, of replacing the log score by some other proper scoring rule as
a yardstick for measuring and comparing the quality of statistical models. In particular we shall see that, for a certain class of such proper scoring rules, the problems with improper priors simply do not arise.

4 Prequential application

Let \( X = (X_1, X_2, \ldots), X^n = (X_1, X_2, \ldots, X_n) \). Let \( Q \) be a distribution for \( X \), with induced joint distribution \( Q^n \), having density \( q^n(\cdot) \), for \( X^n \). Then \( q^n \) can be decomposed into its sequence of recursive conditionals:

\[
q^n(x^n) = q(x_1) \times q_2(x_2) \times \ldots \times q(x_n) \tag{6}
\]

where \( q_i(\cdot) \) is the density function of the distribution \( Q_i \) of \( X_i \), given \( X_{i-1} = x_{i-1} \); note that this depends on \( x_{i-1} \), even though the notation omits this.

We can apply a proper scoring rule \( S_i \) (the form of which could in principle even depend on \( x_{i-1} \)) to the \( i \)th term in (6), and cumulate the scores to obtain

\[
S^n(x^n, Q) := \sum_{i=1}^{n} S_i(x_i, Q_i),
\]

where \( Q_i \) is a function of \( x_{i-1} \). It is readily seen that this yields a proper scoring rule for \( X^n \) (strictly proper if every \( S_i \) is).

Define

\[
\Delta^n(x^n; P, Q) := S^n(x^n, Q) - S^n(x^n, P), \tag{7}
\]

and

\[
D^n(x^n; P, Q) := \sum_{i=1}^{n} D_i(P_i, Q_i), \tag{8}
\]

where \( D_i \) is the discrepancy function associated with the component scoring rule \( S_i \). Then \( D^n \) is in fact a function of \( x^{n-1} \).

Now \( D^n \geq 0 \) is non-decreasing, and under suitable conditions we will have \( D^n \to \infty \) a.s. \([P]\). One useful condition for this is the following:

**Lemma 4.1** Suppose that \( P \) and \( Q \) are mutually singular (as distributions for the infinite sequence \( X \)), and for all \( i \) and some \( k > 0 \), \( D_i(P_i, Q_i) \geq kH^2(P_i, Q_i) \), where \( H \) denotes Hellinger distance. Then \( D^n \to \infty \) a.s. \([P]\).

**Proof.** Singularity implies \( \sum_{i=1}^{n} H^2(P_i, Q_i) \to \infty \) a.s. \([P]\). [Kabanov et al. 1977]. \[\square\]

**Remark 4.1** We can replace \( H^2 \) in Lemma 4.1 by any other discrepancy measures dominating (a multiple of) \( H^2 \), including Kullback-Leibler divergence, and \( d_\epsilon \) given by \( d_\epsilon(P, Q) = \int |1 - q(x)/p(x)|^\epsilon p(x) \, dx \) for \( 1 \leq \epsilon \leq 2 \). [Skouras 1998]. This latter is the \( L_1 \) distance for \( \epsilon = 1 \) and the squared \( \chi^2 \)-distance for \( \epsilon = 2 \).
Also, 

\[ U^n := \Delta^n(X^n; P, Q) - D^n(X^n; P, Q) \]  

(9) is a 0-mean martingale under \( P \): indeed it is the difference of the two 0-mean martingales 

\[ S^n(X^n, Q) - S^n(P^n, Q^n) \]  

(10) and 

\[ S^n(X^n, P) - H^n(P^n). \]  

(11) Under suitable and reasonable conditions on the behaviour of the increments 

\[ S_i(x_i, Q_i) - S_i(x_i, P_i) \]  

of \( \Delta_n(P, Q) \), \( |U_n| \) will remain small in comparison with \( D_n \). For example, if the increments are all of similar size, a martingale law of the iterated logarithm (see, e.g., Stout (1970)) would restrict \( \sup_n |U_n| \) to have order \( (n \ln \ln n)^{1/2} \), while \( D_n \) would be of order \( n \). It would then follow that, with \( P \)-probability 1, \( \Delta_n \to \infty \). In such a case, if \( P \) is the true distribution generating the data, then eventually we will have, with probability 1, \( S^n(X^n, P) > S^n(X^n, Q) \). Then choosing the model with the lowest prequential score \( S^n \) will yield a consistent criterion for selecting among a finite collection of distributions for \( X \).

### 4.1 Application to model selection

The above theory can be applied to the case that \( P, Q \) are the predictive distributions associated with different Bayesian models, \( M \) and \( N \). In particular, suppose we have statistical models 

\[ P = \{ P_\theta : \theta \in \mathcal{T} \} \]  

(12) with prior \( \pi \) over \( \mathcal{T} \); and 

\[ Q = \{ Q_\phi : \phi \in \mathcal{F} \} \]  

(13) with prior \( K \) over \( \mathcal{F} \); and corresponding predictive distributions 

\[ P = \int_\mathcal{T} P_\theta d\Pi(\theta) \]  

(14) 

\[ Q = \int_\mathcal{F} Q_\phi dK(\phi). \]  

(15) Under conditions that allow application of the above results, we will have \( P(A) = 1 \), where \( A \) is the event \( S^n(X^n, Q) - S^n(X^n, P) \to \infty \). Since \( P(A) = \int_\mathcal{T} P_\theta(A) d\Pi(\theta) \), we must have \( P_\theta(A) = 1 \) for \( \theta \in S \), where \( \Pi(S) = 1 \). In particular, if \( \Pi \) has Lebesgue density \( \pi \) that is everywhere positive, then \( P_\theta(A) = 1 \) for almost all \( \theta \in \mathcal{T} \). So the criterion \( S^n \) will choose the correct model with probability 1 under (almost) any distribution in that model. This result generalises the consistency property of log-marginal likelihood (Dawid 1992) to other proper scoring rules.
5 Local scoring rules

We call a scoring rule $S(x, Q)$ local (of order $m$) if it can be expressed as a function of $x$, and of the density function $q(\cdot)$ of $Q$ and its derivatives up to the $m$th order, all evaluated at $x$. Thus the log-score is local of order 0. For the case that the sample space $\mathcal{X}$ is an interval on the real line, Parry et al. (2012) have characterised all proper local scoring rules. It was shown that these can all be expressed as a linear combination of the log score and a “key local” scoring rule, which is a proper local scoring rule that is homogeneous in the sense that its value is unchanged if $q$ and (thus) all of its derivatives are multiplied by some constant $c > 0$. This suggests a way of evading the problematic normalising constant of the compleat Bayesian analysis: if we replace the log-score in (5) by some key local scoring rule, the dependence on the normalising constant will disappear. Indeed, there is no problem in computing such a score even for an “improper” density $q(\cdot)$, having infinite integral over $\mathcal{X}$.

For any $k \geq 1$, the simplest key local scoring rule is the order-2 rule of Hyvärinen (2005):

$$S_H(x, Q) := 2\Delta \log q(x) + \| \nabla \log q(x) \|^2,$$

(16)

where $\nabla$ denotes gradient, and $\Delta$ is the Laplacian operator $\sum_{i=1}^{k} \partial^2/(\partial x_i)^2$.

The associated discrepancy function is

$$D_H(p, q) = \int \| \nabla \log p(x) - \nabla \log q(x)\|^2 p(x) \, dx.$$

(17)

Variations on (16) and (17) can be obtained, on first performing a non-linear transformation of the space $\mathcal{X}$, or equipping $\mathcal{X}$ with the structure of a Riemannian space and reinterpreting $\nabla$, $\Delta$ accordingly (Dawid and Lauritzen 2005). Other key local scoring rules for the multivariate case are considered by Parry (2013). Though such variations can be useful, here we largely confine ourselves to the basic Hyvärinen score $S_H$ of (16). However, there remains some freedom as to how this is applied: for example, we could apply the multivariate score directly to the data, or to a sufficient statistic, or cumulate the 1-dimensional scores associated with each term in the decomposition (6) (Mameli et al. 2014). While such manipulations have no effect on comparisons based on the log score $S_L$, they do typically affect those based on the Hyvärinen score $S_H$. There is thus greater flexibility to apply this in useful ways, e.g. for ease of computation, or to improve robustness to model misspecification.

6 Multivariate Normal Distribution

Consider in particular the case that the distribution $Q$ of $X$ is multivariate normal:

$$X \sim \mathcal{N}_k(\mu, \Sigma),$$

(18)

1Some conditions on the behaviour of densities at the boundary of $\mathcal{X}$ are required in order for (16) to be a proper scoring rule.

2For convenience we have introduced an extra factor of 2.
with density
\[ q(x) \propto \exp\left(-\frac{1}{2}(x - \mu)^T \Phi (x - \mu) \right) \] (19)

where \( \Phi := \Sigma^{-1} \) is the precision matrix, and (in contrast to the usual convention for likelihood functions) the “constants” implicit in the proportionality sign are allowed to depend on the parameters, \( \mu \) and \( \Phi \), but not on \( x \).

We have
\[
\nabla \log q = -\Phi (x - \mu) \\
\Delta \log q = -\text{tr} \Phi
\]
so that, applying (16),
\[
S_H(x, Q) = \|\Phi (x - \mu)\|^2 - 2 \text{tr} \Phi.
\]
(22)

The associated discrepancy between \( P = \mathcal{N}(\mu_P, \Phi_P^{-1}) \) and \( Q = \mathcal{N}(\mu_Q, \Phi_Q^{-1}) \) is
\[
D_H(P, Q) = \text{tr} (\Phi_P - 2\Phi + \Phi_P^{-1} \Phi_Q^2) + \|\Phi_Q (\mu_P - \mu_Q)\|^2.
\]
(23)

The score \( S_H \) may be relatively easy to compute if the model is defined in terms of its precision matrix \( \Phi \), as for a graphical model. Note also that, whereas the log-score \( S_L \) in this case would involve computing the determinant of \( \Phi \), this is not required for \( S_H \).

We can now compare different hypothesised multivariate normal distributions \( Q \) for the observed data \( x \) by means of their associated \( S_H \) scores given by (22).

### 6.1 Univariate case

For the univariate case \( Q = \mathcal{N}(\mu, \sigma^2) \) we get
\[
S_H(x, Q) = \frac{1}{\sigma_Q} \left\{(x - \mu)^2 - 2\sigma^2\right\} \\
D_H(P, Q) = \frac{1}{\sigma_Q^2} \left\{(\sigma_P^2 - \sigma_Q^2)^2 + \mu_P - \mu_Q\right\}.
\]
(24) \hspace{1cm} (25)

In this case the Kullback-Leibler discrepancy is given by
\[
2\text{KL}(P, Q) = \frac{\sigma_P^2}{\sigma_Q^2} + \ln \frac{\sigma_Q^2}{\sigma_P^2} + \frac{(\mu_P - \mu_Q)^2}{\sigma_Q^2} - 1.
\]
(26)

Using \( \ln x \leq x - 1 \), we find
\[
D_H(P, Q) \geq \frac{2}{\sigma_Q^2} \text{KL}(P, Q).
\]
(27)
In the context of §4, where $P$ and $Q$ are both Gaussian processes for $(X_1, X_2, \ldots)$, we can apply Remark 4.1 to deduce that prequential model comparison between $P$ and $Q$ based on the Hyvärinen score will be consistent whenever $P$ and $Q$ are mutually singular, and (writing $\sigma_{Q,i}^2$ for the variance, under $Q$, of $X_i$, given $(X_1, \ldots, X_{i-1})$),

$$\lim \inf_{i \to \infty} \sigma_{Q,i}^2 > 0 \text{ a.s. } [P],$$

and likewise with $P$ and $Q$ interchanged.

7 Bayesian Model

For the Bayesian the parameter is a random variable, $\Theta$ say. Let the statistical model have density $p(x \mid \theta)$ at $X = x$, when $\Theta = \theta$. If the prior density is $\pi(\theta)$, the marginal density of $x$ is

$$q(x) = \int p(x \mid \theta) \pi(\theta) \, d\theta.$$ 

Then we find

$$\frac{\partial \log q(x)}{\partial x_i} = E \left\{ \frac{\partial \log p(x \mid \Theta)}{\partial x_i} \bigg| X = x \right\}$$

$$\frac{\partial^2 \log q(x)}{\partial x_i^2} = E \left\{ \frac{\partial^2 \log p(x \mid \Theta)}{\partial x_i^2} \bigg| X = x \right\} + \text{var} \left\{ \frac{\partial \log p(x \mid \Theta)}{\partial x_i} \bigg| X = x \right\}$$

where the expectations and variances are taken under the posterior distribution of $\Theta$ given $X = x$, having density $\pi(\theta \mid x) = \frac{p(x \mid \theta) \pi(\theta)}{q(x)}$. This yields

$$S_H(x, Q) = \sum_i \left( E \left\{ \left( \frac{\partial \log p(x \mid \Theta)}{\partial x_i^2} \right)^2 \bigg| X = x \right\} + 2 \left\{ \frac{\partial \log p(x \mid \Theta)}{\partial x_i} \bigg| X = x \right\} \right)$$

$$- \left[ E \left\{ \left( \frac{\partial \log p(x \mid \Theta)}{\partial x_i} \bigg| X = x \right\} \right]^2 \right)$$

$$= E \{ S_H(x, P_\Theta) \mid X = x \}$$

$$+ \sum_i \text{var} \left\{ \frac{\partial \log p(x \mid \Theta)}{\partial x_i} \bigg| X = x \right\}.$$ 

(28)

7.1 Exponential Family

Suppose further that the model is an exponential family with natural statistic $T = t(x)$:

$$\log p(x \mid \theta) = a(x) + b(\theta) + \sum_{j=1}^k \theta_j t_j(x).$$ 

(30)
Define $\mu \equiv \mu(x)$, $\Sigma \equiv \Sigma(x)$ to be the posterior mean-vector and dispersion matrix of $\Theta$, given $X = x$. Then we obtain

$$S_H(x, Q) = 2\Delta a + 2d^T \mu + \|\nabla a + J\mu\|^2 + 2 \text{tr} J\Sigma J^T$$

with $d \equiv d(x) := (\Delta t_j)$, $J \equiv J(x) := (\partial t_j(x)/\partial x_i)$.

For the special case $T = X$, this becomes

$$S_H(x, Q) = 2\Delta a + \|\nabla a + \mu\|^2 + 2 \text{tr} \Sigma.$$

8 Linear model: variance known

Consider the following normal linear model for a data-vector $Y = (Y_1, \ldots, Y_n)^T$:

$$Y \sim N(X\theta, \sigma^2 I), \quad (31)$$

where $X$ ($n \times p$) is a known design matrix of rank $p$, and $\theta \in \mathbb{R}^p$ is an unknown parameter vector. In this section we take $\sigma^2$ as known.

8.1 Multivariate score

Consider giving $\theta$ a normal prior distribution:

$$\theta \sim N(m, V). \quad (32)$$

The marginal distribution $Q$ of $Y$ is then

$$Y \sim N(Xm, XVX^T + \sigma^2 I) \quad (33)$$

with precision matrix

$$\Phi = (XVX^T + \sigma^2 I)^{-1} = \sigma^{-2} \left( I - X \left(X^TX + \sigma^2 V^{-1}\right)^{-1} X^T \right)$$

on applying the Woodbury matrix inversion lemma (equation (10) of [Lindley and Smith (1972)]).

An “improper” prior can now be generated by allowing $V^{-1} \to 0$, yielding

$$\Phi = \sigma^{-2} \Pi$$

where

$$\Pi := I - XAX^T,$$

with $A := (X^TX)^{-1}$, is the projection matrix onto the space of residuals.

Although this $\Phi$ is singular, and thus can not arise from any genuine dispersion matrix, there is no problem in using it in (22). We obtain:

$$S_H(y, Q) = \frac{1}{\sigma^4} \left( R - 2\nu\sigma^2 \right) \quad (34)$$
where \( R \) is the usual residual sum-of-squares for model (31), on \( \nu := n - p \) degrees of freedom. Note that, unlike marginal log-likelihood, this is well-defined, in spite of the fact that we have not specified a “normalising constant” for the improper prior density. This is of course a consequence of the homogeneity of the Hyvärinen score \( S_H \).

The above analysis is not however applicable if \( \text{rank}(X) < p \)—in particular, whenever \( n < p \). Taking \( V^{-1} \to 0 \) is equivalent to using an improper prior density \( \pi(\theta) \equiv c \), with \( 0 < c < \infty \). When \( X \) is of rank \( p \), the integral formally defining the marginal density of \( Y \) is finite for each \( y \) (even though the resulting density is itself improper). However, when \( \text{rank}(X) < p \) this integral is infinite at each \( y \), so that no marginal joint density—even improper—can be defined.

Using the criterion (34) for comparing different normal linear models, all with the same known residual variance \( \sigma^2 \), is equivalent to comparing them in terms of their penalised scaled residual sum-of-squares, \((R/\sigma^2) + 2p\)—which is just Akaike’s AIC for this known-variance case. (However, when \( \sigma^2 \) varies across models, the criterion (34) is no longer equivalent to AIC.)

Now it is well known that AIC is not a consistent model selection criterion. As an example, consider the two models:

\[
M_1 : Y_i \sim N(0, 1) \\
M_2 : Y_i \sim N(\theta, 1).
\]

Then, with \( \bar{Y} \) denoting the sample mean \( \sum_i Y_i / n \), we have \( \text{AIC}_1 = \sum_i Y_i^2 \), \( \text{AIC}_2 = \sum_i (Y_i - \bar{Y})^2 + 2 \), so that \( \text{AIC}_1 - \text{AIC}_2 = n\bar{Y}^2 - 2 \). When \( M_1 \) holds, this is distributed, for any \( n \), as \( \chi^2_1 - 2 \), which has a non-zero probability of being positive, and thus favouring the incorrect model \( M_2 \).

Hence the above approach does not seem an entirely satisfactory solution to the model-selection problem.

### 8.2 Prequential score

In an attempt to restore consistent model selection, we turn to the prequential approach.

In (31), let \( x_i \) be the \( i \)th row of \( X \), and \( X^i \) the matrix containing the first \( i \) rows of \( X \). Assuming \( X \) is non-singular; then \( X^i \) is non-singular if and only if \( i \geq p \).

Define, for \( i \geq p \):

\[
A_i := \{ (X^i)^T(X^i) \}^{-1} \\
\hat{\theta}_i := A_i(X^i)^TY^i
\]

and, for \( i > p \):

\[
\eta_i := x_i^T\hat{\theta}_{i-1} \\
k_i^2 := 1 + x_i^TA_{i-1}x_i = (1 - x_i^TA_i x_i)^{-1} \\
Z_i := k_i^{-1}(Y_i - \eta_i)
\]
(where the identity in (39) follows from the Woodbury lemma).

Then for the improper prior \( V^{-1} \to 0 \) with, the predictive distribution of \( Y_i \), given \( Y^{1-1} \), is

\[
Y_i \sim \mathcal{N}(\eta_i, k_i^2 \sigma^2) \quad (i > p).
\]

That is, in the predictive distribution the \( (Z_i : i = p+1, \ldots, n) \) are independent and identically distributed \( \mathcal{N}(0, \sigma^2) \) variables (which property also holds in the sampling distribution, conditionally on \( \theta \)); moreover, \( R = \sum_{i=p+1}^{n} Z_i^2 \).

Note that, under the model (31), \( \eta_i \) has expectation \( x_i^T \theta \) and variance \( k_i^2 - 1 \). So the predictive distribution (41) and the true distribution will be asymptotically indistinguishable (the property of “prequentially consistent” estimation—see Dawid (1984)) if and only if

\[
k_i^2 \to 1 \quad \text{as} \quad i \to \infty.
\]

This we henceforth assume, for any model under consideration.

For \( i > p \), the incremental score (24) associated with (41) is

\[
S_i = \frac{T_i}{k_i^2 \sigma^2}
\]

where

\[
T_i := \frac{Z_i^2}{\sigma^2} - 2.
\]

Under any distribution in the model, the \( T_i \) are independent, with

\[
E(T_i) = -1 \quad \text{(45)}
\]

\[
\text{var}(T_i) = 2. \quad \text{(46)}
\]

As discussed in §4, minimising the cumulative prequential score

\[
S^* := \sum_i S_i \quad \text{(47)}
\]

should typically yield consistent model choice. We investigate this in more detail in §8.4 below.

Expression (43) is only defined for an index \( i \) exceeding the dimensionality of the model. When comparing models of differing dimensionalities, we should ensure the identical criterion is used for each. We could just cumulate the \( S_i \) over indices \( i \) exceeding the greatest model dimension, \( p_{\max} \) say, but this risks losing relevant information. To restore this, we might add to that sum the multivariate score (34) computed, for each model, for the first \( p_{\max} \) observations.

8.3 Multivariate or prequential?

The multivariate score (34) can be expressed as the sum of rescaled incremental scores:

\[
S_H(y, Q) = \frac{1}{\sigma^2} \sum_{i=1}^{n} T_i = \sum_{i=p+1}^{n} k_i^2 S_i, \quad \text{(48)}
\]
and the scaling factor $k^2_i$ has been assumed to satisfy (42). It would thus seem that (48) is asymptotically equivalent to (47), and thus that model selection by minimisation of the multivariate score (34) should be consistent for model choice. However, we have seen that this is not the case.

Further analysis dispels this paradox. The difference between the prequential and the multivariate score, up to time $n$, is

$$S^* - S_H = \frac{1}{\sigma^2} \sum_{i=p}^{n} \left( \frac{1}{k^2_i} - 1 \right) T_i.$$  

(49)

Under any distribution in the model, this has expectation

$$\mathbb{E} \left( \frac{1}{\sigma^2} \sum_{i} \left( 1 - \frac{1}{k^2_i} \right) \right) = \frac{1}{\sigma^2} \sum_{i} x^T_i A_i x_i,$$

and variance

$$\frac{2}{\sigma^4} \sum_{i=1}^{n} \left( x^T_i A_i x_i \right)^2.$$

Suppose the $(x_i)$ look like a random sample from a $p$-variate distribution, with $\mathbb{E} x_i x^T_i = C$. Then, for large $i$,

$$\mathbb{E} (i x^T_i A_i x_i) = \mathbb{E} \left( \sum_{j=1}^{i} x_j x^T_j / i \right)^{-1} x_i x^T_i \approx \text{tr} C^{-1} C = p.$$

So $1-1/k^2_i \approx p/i$; in particular (42) holds. Then $\mathbb{E}(S^* - S_H) \approx (p/\sigma^2) \sum_{i=p}^{n} i^{-1} \approx p (\log n) / \sigma^2$. A similar analysis shows $\text{var}(S^* - S_H) < \infty$. Thus, under the model, $S^* - S_H = O_p\{p(\log n) / \sigma^2\}$. So, contrary to first impressions, the difference between the cumulative prequential score $S^*$ and the multivariate score $S_H$ diverges to infinity (at a logarithmic rate) under any true model.

### 8.4 Prequentially consistency model selection

We now consider the asymptotic behaviour of the cumulative prequential score $S^*$, given by (47), when used to select between two models, $M_1$ and $M_2$, both of the general form (31), when $M_1$ is true. Let these models have respective dimensions $p_1$ and $p_2$, and variances $\sigma^2_1$ and $\sigma^2_2$. Let $Z_i$, $h^2_i$, as defined above, refer to $M_1$, and denote the corresponding quantities for $M_2$ by, respectively, $W_i$, $k^2_i$. Let $S^*_1$, $S^*_2$ denote the cumulative prequential scores for $M_1$, $M_2$ respectively. We assume conditions on the regressors, as discussed above, under which

$$1 - 1/k^2_i \approx p_1/i$$

(50)

$$1 - 1/h^2_i \approx p_2/i.$$  

(51)

Since the $(Y_i)$ are independent normal variables with variance $\sigma^2_i$, and the $(Z_i)$ and $(W_i)$ are, in each case, constructed from the $(Y_i)$ by an orthogonal
linear transformation, we will have

\[ Z_i \sim \mathcal{N}(0, \sigma_1^2) \text{ independently} \quad (52) \]

\[ W_i \sim \mathcal{N}(\nu_i, \sigma_1^2) \text{ independently} \quad (53) \]

where the \((Z_i)\) have mean 0 since \(M_1\) is true, whereas the \((\nu_i)\) may be non-zero.

Let \( p = \max\{p_1, p_2\} \). Apart from a finite contribution from some initial terms, the difference in prequential scores, up to time \( n \), is

\[ S^*_2 - S^*_1 = \frac{1}{\sigma_2^2} \sum \frac{1}{h_i^2} \left( \frac{W_i^2}{\sigma_2^2} - 2 \right) - \frac{1}{\sigma_1^2} \sum \frac{1}{k_i^2} \left( \frac{Z_i^2}{\sigma_1^2} - 2 \right) \quad (54) \]

where \( \sum \) denotes \( \sum_{i=p+1}^{n} \).

On account of (52) and (53), this has expectation

\[ \mathbb{E}(S^*_2 - S^*_1) = \frac{1}{\sigma_2^2} \sum \frac{\nu_i^2}{h_i^2} + \frac{(\sigma_1^2 - \sigma_2^2)^2}{\sigma_1^4 \sigma_2^4} \sum \frac{1}{h_i^2} + \frac{1}{\sigma_1^2} \sum \left( \frac{1}{k_i^2} - \frac{1}{h_i^2} \right). \quad (55) \]

We now consider various cases for \( M_2 \).

### 8.4.1 \( M_2 \) true

If the true distribution also belongs to \( M_2 \) (as well as to \( M_1 \)), then we must have \( \sigma_2^2 = \sigma_1^2 = \sigma^2 \) say, and \( \nu_i \equiv 0 \). Then (55) reduces to

\[ \mathbb{E}(S^*_2 - S^*_1) = \frac{1}{\sigma^2} \sum \left( \frac{1}{k_i^2} - \frac{1}{h_i^2} \right). \quad (56) \]

On account of (50) and (51), this behaves asymptotically as \( (p_2 - p_1)(\log n)/\sigma^2 + O(\log n) \). Also, an analysis similar to that in §8.3 shows that \( \text{var}(S^*_2 - S^*_1) \) is bounded, so that

\[ S^*_2 - S^*_1 = \frac{(p_2 - p_1) \log n}{\sigma^2} + O_P(\log n). \quad (57) \]

(Compare this with the behaviour of the log Bayes factor in this case, which, in line with (4), is asymptotic to \( \frac{1}{2}(p_2 - p_1) \log n \).)

In particular, when comparing finitely many true models of different dimensions, minimising the cumulative prequential score will consistently favour the simplest true model, at rate \( \propto \log n \).

We now consider cases where \( M_2 \) is false. For simplicity we confine attention to the expected score.

### 8.4.2 Wrong variance

Suppose first that \( M_2 \) has the wrong variance \( \sigma_2^2 \neq \sigma_1^2 \). In this case the first term in (55) is non-negative, the second is positive of order \( n \), and the third term is again of order \( \log n \). The true model \( M_1 \) is thus favoured, at rate \( \propto n \).
8.4.3 Right variance, wrong mean

Suppose now $\sigma_2^2 = \sigma_1^2 = \sigma^2$, but the data-generating distribution does not have the mean-structure of model $M_2$. We note that the log Bayes factor $\ln$ will tend to infinity (almost surely), so selecting the true model $M_1$, if and only if $\sum \nu_i^2 = \infty$.

In this case we have

$$E(S_2^* - S_1^*) = \frac{1}{\sigma^2} \sum \nu_i^2 + \frac{1}{\sigma^2} \sum \left( \frac{1}{h_i^2} - \frac{1}{h_i^2} \right),$$

where $\nu_i \neq 0$ and $h_i^2 \neq h_i^2$.

The first term in (58) is non-negative, while the second term behaves asymptotically as $(p_2 - p_1)(\log n)/\sigma^2$. In particular, if $p_2 > p_1$, then (58) increases at rate at least $(p_2 - p_1)(\log n)/\sigma^2$, so favouring the true model.

However, things are more delicate if $p_2 < p_1$. In this case, if $\sum (\nu_i/h_i)^2$ increases sufficiently slowly — specifically, at rate less than $(p_1 - p_2)\sigma^2(\log n)$ — then the increased simplicity of model $M_2$ more than compensates for the slight inaccuracy in its mean-structure, leading to selection of the slightly incorrect model $M_2$.

The case $p_2 = p_1$ requires a still more delicate analysis, which we shall not pursue here.

**Example** As an example, consider again the comparison of the models $M_1$ and $M_2$ of §8.1.

Under $M_1$, with $Y_i \sim N(0, 1)$, we have $p_1 = 0$, $h_i^2 = 1$, $Z_i = Y_i$. In this special case the cumulative prequential score $S_1^*$ is identical to the multivariate score $S_{H,1}$.

For model $M_2$, with $Y_i \sim N(\theta, 1)$ ($\theta \neq 0$), we have $p_2 = 1$, $h_i^2 = i/(i - 1)$, $W_i = \{(i - 1)/i\})^{1/2}(Y_i - \sum_{j=1}^{i-1}) \sim N(0, 1)$. Although $h_i^2 \rightarrow 1$, $S_2^* - S_{H,2}$ has (under any distribution in $M_2$, and hence also under the simpler model $M_1$) expectation $\sum_{i=1}^{n} i^{-1} \approx \log n$, and bounded variance $2\sum_{i=1}^{n} i^{-2} \approx \pi^2/3$. Since $S_1^* \approx S_{H,1}$, and we have seen that $S_{H,2} - S_{H,1}$ is bounded in probability under $M_1$, $S_2^* - S_1^*$ diverges to infinity (at rate $\log n$) under $M_1$—so consistently selecting the correct model $M_1$.

On the other hand, under $M_2$ we have $S_2^* = \sum_{i=1}^{n}(1 - 1/i)(W_i^2 - 2) = -n + o_p(n)$, while $S_1^* = \sum_{i=1}^{n}(Y_i^2 - 2) = n(\theta^2 - 1) + o_p(n)$, so that $S_2^* - S_1^* = -n\theta^2 + o_p(n)$, which thus diverges to $-\infty$ (this time at rate $n$)—so now consistently selecting the correct model $M_2$.

In summary, although the multivariate score [31] is more straightforward to compute, if consistent model selection is regarded as an important criterion then the prequential score is to be preferred.

9 Linear model: variance unknown

Now suppose we don’t know $\sigma^2$ in [31]. With $\phi = 1/\sigma^2$, we have model density

$$p(y | \theta, \phi) \propto \phi^{\frac{1}{2}} n \exp -\phi \left\{ R + (\theta - \hat{\theta})^\top X^\top X (\theta - \hat{\theta}) \right\}$$

(59)
where $R = y^T \Pi y$, with $\Pi = I - XAX^T$, is the residual sum of squares, on $\nu = n - p$ degrees of freedom.

The standard improper prior for this model is $\pi(\theta, \phi) \propto \phi^{-1}$. Multiplying by this and integrating over $(\theta, \phi)$ yields the (improper) joint predictive density

$$p(y) \propto R^{-\frac{1}{2}\nu},$$

with logarithm (up to a constant)

$$l = -\frac{1}{2} \nu \log R.$$  

Writing $r = \Pi y$ (the residual vector), we find

$$\frac{\partial l}{\partial y_i} = -\frac{\nu r_i}{R}$$

$$\frac{\partial^2 l}{\partial y_i^2} = \nu \left(\frac{2r^2_i}{R^2} - \frac{\pi_{ii}}{R}\right),$$

and so (noting $\sum_i \pi_{ii} = \nu$) the multivariate score [10] is

$$S_H = -\frac{(\nu - 4)}{\bar{\sigma}^2}$$

where $\bar{\sigma}^2 = R/\nu$ is the usual unbiased estimator of $\sigma^2$. So long as at least one model under consideration has $\nu > 4$ (a very reasonable requirement), choosing a model by minimisation of the predictive score is thus equivalent to minimising $J := \bar{\sigma}^2/(\nu - 4)$.

Again, this model selection criterion is typically inconsistent. Thus consider the comparison between models $M_1$ and $M_2$ of §8.1, now extended to have unknown variance $\sigma^2$. We have

$$J_1 = \frac{(n-1)S^2 + n\bar{Y}^2}{n(n-4)}$$

$$J_2 = \frac{S^2}{(n-5)}$$

where $S^2 := \sum_{i=1}^n (Y_i - \bar{Y})^2/(n-1)$ is a consistent estimate of $\sigma^2$ under either model. Then $M_2$ is preferred if $J_2 < J_1$, which holds when

$$\frac{n\bar{Y}^2}{\sigma^2} > \frac{2n-5}{(n-5)} \frac{S^2}{\sigma^2} \approx 2$$

for large $n$. But, under $M_1$, $n\bar{Y}^2/\sigma^2 \sim \chi_1^2$, so that there is a positive probability of the inequality [67] holding, so favouring the more complex model $M_2$.

\footnote{For the integral formally defining this density to be finite at each point we require $\text{rank}(X) \geq p + 1$.}
9.1 Prequential score

From (60), as a function of \( y_i \) the predictive density of \( Y_i \) given \( y^{i-1} \) (for \( i > p \)) is

\[
p(y_i | y^{i-1}) \propto R_i^{-\frac{4}{\nu_i}} = (R_{i-1} + z_i^2)^{-\frac{4}{\nu_i}}
\]

where \( R_i \) is the residual sum-of-squares based on \( y_i \), on \( \nu_i := i - p \) degrees of freedom, and \( z_i = k_i^{-1}(y_i - \eta_i) \), as given by (38)–(40). Applying the univariate case of (16) now yields (for \( i > p \)) the incremental score:

\[
S_i = \nu_i \left\{ \frac{(4 + \nu_i) Z_i^2 - 2 R_i}{k_i^2 R_i^2} \right\}
\]

\[
= \frac{(1 + \frac{4}{\nu_i}) Z_i^2 - 2 s_i^2}{k_i^2 s_i^4}
\]

where \( s_i^2 := R_i/\nu_i \) is the residual mean square, based on \( Y_i \), under the model. The prequential score is now obtained by cumulating \( S_i \) over \( i \). Once again, under reasonable conditions this can be expected to yield consistent model selection.\(^4\)

We investigate this consistency property further, for the special case of comparing two true models of different dimensions \( p_1 < p_2 \). We saw in §8.4 that in this case, when the variance \( \sigma^2 \) is known (and under reasonable assumptions on the models) the prequential Hyvärinen score prefers the simpler model over the more complex model, at rate \((p_2 - p_1)(\log n)/\sigma^2\).

We consider the asymptotic behaviour of \( S^* := \sum_{i=p+1}^n S_i \) under a distribution in the model.\(^5\) In this case the \((Z_i : i > p)\) are independent and identically distributed as \( N(0, \sigma^2) \).

Writing \( U_i := \frac{Z_i^2}{\sigma^2} - 1 \), so that \( E(U_i) = 0, E(U_i^2) = 2 \), we have

\[
k_i^2 \sigma^2 S_i = \frac{(1 + \frac{4}{\nu_i}) (U_i + 1) - 2(\overline{U}_i + 1)}{(\overline{U}_i + 1)^2}
\]

with \( \overline{U}_i := \nu_i^{-1} \sum_{j=p+1}^i U_j \) (where \( \nu_i = i - p \)). Now \( \overline{U}_i = O_p(i^{-1}) \). Expanding \( \overline{U}_i \) as a power series in \( \overline{U}_i \) gives

\[
k_i^2 \sigma^2 S_i = \sum_{r=0}^{\infty} (-1)^r \overline{U}_i \left\{ (r + 1) \left( 1 + \frac{4}{\nu_i} \right) (U_i + 1) - 2 \right\}
\]

so that

\[
k_i^2 \sigma^2 S_i - (U_i - 1) = \frac{4}{\nu_i} + \frac{4U_i}{\nu_i}
\]

\(^4\)Again, an additional contribution of the form of (64), computed for an initial string of observations, could be incorporated to ensure fair comparison between models of different dimension.

\(^5\)Our analysis is indicative, rather than fully rigorous.
\[-2\bar{U}_i \left( U_i + 4 \frac{4U_i}{\nu_i} \right) \]  
\[= \bar{U}_i^2 \left( 1 + 3U_i + \frac{12}{\nu_i} + \frac{12U_i}{\nu_i} \right) \]  
\[+ O_p(i^{-3/2}). \]  

(74) \hspace{2cm} (75) \hspace{2cm} (76)

Noting

\[E(\bar{U}_i^2) = \frac{2}{\nu_i} \]  
\[E(\bar{U}_i U_i) = \frac{2}{\nu_i} \]  
\[E(\bar{U}_i^2 U_i) = \frac{8}{\nu_i^2}; \]  

we compute

\[E \left\{ k_i^2 \sigma^2 S_i - (U_i - 1) \right\} = \frac{2}{i} + O(i^{-3/2}), \]  

whence, on account of (72),

\[E(S^* - S_0^*) = 2(\log n) / \sigma^2 + O(1) \]  

(80) \hspace{2cm} (81)

where \(S_0^* = \sum_{i=p+1}^{n}(U_i - 1)/(k_i^2 \sigma^2)\) is the cumulative prequential score (47) for the submodel in which the correct variance \(\sigma^2\) is known.

In the remainder of this Section we argue that \(S^* - S_0^*\) differs from its expectation (81) by \(O_p(\log n)\).

On cumulating the term \(\propto U_i / \nu_i\) in (74) we obtain variance \(\propto \sum_{i=p+1}^{\infty} \nu_i^{-2}\), which is finite. So this yields a contribution that is \(O_p(1)\).

Consider now the term \(\propto \bar{U}_i U_i\) in (74). We find \(\var(\bar{U}_i U_i) = 4 / \nu_i + O(\nu_i^{-2})\), and \(\bar{U}_i U_i\) and \(\bar{U}_j U_j\) are uncorrelated for \(i \neq j\). Hence on cumulating the term \(\bar{U}_i U_i\) from \(i = p + 1\) to \(n\) we get variance \(\approx \sum_{i=p+1}^{n} 4 / \nu_i \approx 4 \log n\). Thus the random variation in this term contributes \(O_p(\log n)\) to \(S^* - S_0^*\).

There is also a term \(\propto \bar{U}_i / \nu_i\) in (74). Since \(\bar{U}_i / \nu_i = O_p(i^{-3/2})\), its cumulative sum is \(O_p(n^{-2})\).

Now consider (75). We look first at the term \(\bar{U}_i^2\). We compute \(\var \{ (\bar{U}_i)^2 \} = 8 / \nu_i^2 + 48 / \nu_i^3 = \lambda_i\), say; and, for \(i < j\),

\[\Cov \{ (\bar{U}_i)^2, (\bar{U}_j)^2 \} = \left( \frac{\nu_i}{\nu_j} \right)^2 \lambda_i. \]

Hence

\[\var \left\{ \sum_{i=p+1}^{n} (\bar{U}_i)^2 \right\} = \sum_{i=p+1}^{n} \lambda_i + 2 \sum_{i=p+1}^{n} \sum_{j=i+1}^{n} \left( \frac{\nu_i}{\nu_j} \right)^2 \lambda_i \leq 56 \left\{ \sum_{i=1}^{\nu} i^{-2} + 2 \sum_{i=1}^{\nu} \sum_{j=i+1}^{\nu} j^{-2} \right\}. \]

17
(with $\nu = n - p$), since $\lambda_i \leq \frac{56}{\nu^2}$. We have $\sum_{i=1}^{n} i^{-2} < \infty$, and, for large $i$, $\sum_{j=i+1}^{n} j^{-2} < \sum_{j=i+1}^{n} j^{-2} \approx i^{-1}$. So $\sum_{i=p+1}^{n} (U_i)^2$ is of order $\log n$, and cumulating the term $U_i$ in (75) again makes a contribution $O_p\{(\log n)^{\frac{1}{2}}\}$ over and above its expectation.

Now consider the term $U_iU_i^2$ in (75). We have

$$\text{var}(U_iU_i^2) = \frac{24}{\nu_i^2} + \frac{1024}{\nu_i^3} + \frac{4928}{\nu_i^4}$$

and, for $i < j$,

$$\text{Cov}(U_iU_i^2, U_jU_j^2) = \frac{48(\nu_i + 4)}{\nu_i^2\nu_j^2}.$$  

By an argument similar to that for $U_i^2$, we find that cumulating the term $U_iU_i^2$ in (75) again makes a contribution $O_p\{(\log n)^{\frac{1}{2}}\}$ (over and above its expectation).

Putting everything together, we have

$$S^* - S_0^* = 2(\log n)/\sigma^2 + O_p(\log n)^{\frac{1}{2}}.$$  

Now we have shown in §8.4 that, for comparing two true models $M_1$ and $M_2$ with known variance $\sigma^2$ and respective dimensions $p_1 < p_2$, under conditions on the behaviour of the $(x_i)$, the difference in their cumulative prequential scores $S_0^*$ behaves asymptotically as $(p_2 - p_1)(\log n)/\sigma^2$. Since, from (84), the difference between the scores for the unknown and known variance cases is $2(\log n)/\sigma^2 + O_p(\log n)$ for any model, the identical behaviour applies in the case that the variance is unknown.

10 Discussion

Replacement of the traditional log-score by a proper scoring rule, applied to the marginal likelihood, supplies a general method for avoiding some of the difficulties associated with the use of improper prior distributions for conducting Bayesian model comparison and selection. In particular, use of a homogeneous scoring rule, such as the Hyvärinen rule, supplies a method for taming the otherwise wild behaviour associated with the arbitrariness of the normalising constant of such a prior distribution. Moreover, when applied prequentially, scoring rule based model selection will typically lead to consistent selection of the true model: we have argued for this property both in general terms and in the context of normal linear models with known or unknown variance.

There remain issues of the choice of proper scoring rule. Further theoretical and computational exploration of the properties of these methods is clearly required.
References

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association, 91*, 109–22.

Dawid, A. P. (1984). Statistical theory—The prequential approach (with discussion). *Journal of the Royal Statistical Society, Series A, 147*, 278–92.

Dawid, A. P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences*, (ed. S. Kotz, N. L. Johnson, and C. B. Read), pp. 210–8. Wiley-Interscience, New York.

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference (with Discussion). In *Bayesian Statistics 4*, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), pp. 109–25. Oxford University Press, Oxford.

Dawid, A. P. (2011). Posterior model probabilities. In *Philosophy of Statistics*, (ed. P. S. Bandyopadhyay and M. Forster), pp. 607–30. Elsevier, New York.

Dawid, A. P. and Lauritzen, S. L. (2005). The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, pp. 22–8. University of Tokyo. 12–16 December 2005.

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron, 72*, 169–83.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B, 14*, 107–14.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research, 6*, 695–709.

Kabanov, Y. M., Lipster, R. S., and Shiryaev, A. N. (1977). On the question of absolute continuity and singularity of probability measures. *Math. USSR-Sb, 33*, 203–21.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with Discussion). *Journal of the Royal Statistical Society, Series B, 34*, 1–41.

Mameli, V., Musio, M., and Dawid, A. P. (2014). Comparisons of Hyvärinen and pairwise estimators in two simple linear time series models.

Musio, M. and Dawid, A. P. (2013). Local scoring rules: A versatile tool for inference. Paper presented at 59th World Statistics Congress, Hong Kong.

O’Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B, 57*, 99–138.

Parry, M. F. (2013). Multidimensional local scoring rules. Paper presented at 59th World Statistics Congress, Hong Kong.

Parry, M. F., Dawid, A. P., and Lauritzen, S. L. (2012). Proper local scoring rules. *Annals of Statistics, 40*, 561–92.
Skouras, K. (1998). Absolute continuity of Markov chains. *Journal of Statistical Planning and Inference*, **75**, 1–8.

Stout, W. F. (1970). A martingale analogue of Kolmogorov’s law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **15**, 279–90.