Bayesian Analysis (2015) **10**, Number 2, pp. 479–499

Bayesian Model Selection Based on Proper Scoring Rules*

A. Philip Dawid† and Monica Musio‡

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 $\boldsymbol{X} \in \mathcal{X} \subseteq \mathcal{R}^k$. 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 θ_M, then $\boldsymbol{X}$ has distribution P_{θ_M}, with Lebesgue density $p_M(\boldsymbol{x} \,|\, \theta_M)$. Having observed data $\boldsymbol{X} = \boldsymbol{x}$, we wish to make inference about which model $M \in \mathcal{M}$ (and possibly which parameter-value θ_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 Π_M for its parameter θ_M (to be interpreted as describing conditional uncertainty about θ_M, given the validity of model M). For simplicity we suppose that Π_M has a density function, $\pi_M(\theta_M)$, with respect to Lebesgue measure $d\theta_M$ over $\mathcal{T}_M$.

*Related articles: DOI: 10.1214/15-BA942A, DOI: 10.1214/15-BA942B, DOI: 10.1214/15-BA942C; rejoinder at DOI: 10.1214/15-BA942REJ.

†University of Cambridge, apd@statslab.cam.ac.uk

‡University of Cagliari, mmusio@unica.it

 DOI: 10.1214/15-BA942

The *predictive density function* of $\boldsymbol{X}$, given only the validity of model M, is

$$p_M(\boldsymbol{x}) = \int_{\mathcal{T}_M} p_M(\boldsymbol{x} \,|\, \theta_M)\, \pi_M(\theta_M)\, d\theta_M. \tag{1}$$

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

Considered as a function of $M \in \mathcal{M}$, for given data $\boldsymbol{x}$, $p_M(\boldsymbol{x})$ 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 $\boldsymbol{x}$:

$$L(M) \propto p_M(\boldsymbol{x}). \tag{2}$$

The posterior probability $\alpha(M \,|\, \boldsymbol{x})$ for model M is then given by Bayes's formula:

$$\alpha(M \,|\, \boldsymbol{x}) \propto \alpha(M) \times L(M) \tag{3}$$

where the omitted multiplicative constant is adjusted to ensure $\sum_{M \in \mathcal{M}} \alpha(M \,|\, \boldsymbol{x}) = 1$. 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 $\boldsymbol{X} = \boldsymbol{x}$, 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 α over the model space $\mathcal{M}$, it does still depend on the prior densities π_{M_1}, π_{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

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

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 $\rho(\theta^* \mid M)$ 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

$$\log p_{M_1}(\boldsymbol{x}) - \log p_{M_2}(\boldsymbol{x}). \tag{5}$$

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 $\boldsymbol{X}$, in the light of the observed data $\boldsymbol{x}$. That is, defining $S_L(\boldsymbol{x}, Q) = -\log q(\boldsymbol{x})$, for any proposed distribution Q with density function $q(\cdot)$ over $\mathcal{X}$, and $\boldsymbol{x} \in \mathcal{X}$, we can interpret the *log-score* $S_L(\boldsymbol{x}, Q)$ as a measure of how badly Q did at forecasting the outcome $\boldsymbol{x}$; 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, $S_L(\boldsymbol{x}, Q)$, is just one of many functions $S(\boldsymbol{x}, Q)$ 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, $\mathrm{E}_{\boldsymbol{X}\sim P} S(\boldsymbol{X}, Q)$, when $\boldsymbol{X}$ 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 *generalised 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.

Standard statistical theory is largely based on the log-score (corresponding to log-likelihood), the Shannon entropy, and the Kullback–Leibler discrepancy. However, a very large part of that theory generalises straightforwardly when these are replaced by some other proper scoring rule, and its associated entropy and discrepancy: see Dawid et al. (2015) for applications of proper scoring rules to general estimation theory. Use of a proper scoring rule other than the log-score typically sacrifices some efficiency for gains in computational efficiency and/or robustness. Because there is a wide variety of

proper scoring rules, this offers greatly increased flexibility. The choice of which specific rule to use may be based on external considerations—for example, derived from the loss function of a real decision problem (Grünwald and Dawid, 2004); or chosen for convenience—for example, for reasons of tractability or robustness (Dawid and Musio, 2014).

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 $\boldsymbol{X} = (X_1, X_2, \ldots)$, $\boldsymbol{X}^n = (X_1, X_2, \ldots, X_n)$. Let Q be a distribution for $\boldsymbol{X}$, with induced joint distribution Q^n, having density $q^n(\cdot)$, for $\boldsymbol{X}^n$. Using a prequential (sequential predictive) approach (Dawid, 1984), decompose q^n into its sequence of recursive conditionals:

$$q^n(\boldsymbol{x}^n) = q_1(x_1) \times q_2(x_2) \times \cdots \times q_n(x_n) \tag{6}$$

where $q_i(\cdot)$ is the density function of the distribution Q_i of X_i, given $\boldsymbol{X}^{i-1} = \boldsymbol{x}^{i-1}$; note that this depends on $\boldsymbol{x}^{i-1}$, even though the notation omits this. We now apply a proper scoring rule S_i (the form of which could in principle even depend on $\boldsymbol{x}^{i-1}$) to the ith term in (6), and cumulate the scores to obtain the *prequential score*

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

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

Define

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

and

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

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

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 $\boldsymbol{X}$), 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). □

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

Also,

$$U^n := \Delta^n(\boldsymbol{X}^n; P, Q) - D^n(\boldsymbol{X}^n; P, Q) \tag{9}$$

is a 0-mean martingale under P: indeed, it is the difference of the two 0-mean martingales

$$S^n(\boldsymbol{X}^n, Q) - S^n(P^n, Q^n) \tag{10}$$

and

$$S^n(\boldsymbol{X}^n, P) - H^n(P^n). \tag{11}$$

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 $(n \log\log n)^{\frac{1}{2}}$, 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, $S^n(\boldsymbol{X}^n, P) < S^n(\boldsymbol{X}^n, Q)$. 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 $\boldsymbol{X}$.

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

$$\mathcal{P} = \{P_\theta : \theta \in \mathcal{T}\} \tag{12}$$

with prior Π over $\mathcal{T}$; and

$$\mathcal{Q} = \{Q_\phi : \phi \in \mathcal{F}\} \tag{13}$$

with prior K over $\mathcal{F}$; and corresponding predictive distributions

$$P = \int_{\mathcal{T}} P_\theta \, d\Pi(\theta), \tag{14}$$

$$Q = \int_{\mathcal{F}} Q_\phi \, dK(\phi). \tag{15}$$

Under conditions that allow application of the above results, we will have $P(A) = 1$, where A is the event $S^n(\boldsymbol{X}^n, Q) - S^n(\boldsymbol{X}^n, P) \to \infty$. 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 Π has Lebesgue density π 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 $S(\boldsymbol{x}, Q)$ *local (of order m)* if it can be expressed as a function of $\boldsymbol{x}$, and of the density function $q(\cdot)$ of Q and its derivatives up to the mth order, all evaluated at $\boldsymbol{x}$. 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 property of a key local scoring rule has been found useful in estimation theory. In standard likelihood inference, we need to compute, and differentiate with the respect to the parameter, the log-normalising constant of the statistical model distributions; and this can be computationally prohibitive. But if, instead of log-score, we use a key local scoring rule, the normalising constant simply does not figure in the score, so simplifying computation: for some examples, see Dawid and Musio (2013, 2014). Applied to model selection, 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[1] scoring rule is the order-2 rule of Hyvärinen (2005):[2]

$$S_H(\boldsymbol{x}, Q) := 2\Delta \log q(\boldsymbol{x}) + \|\boldsymbol{\nabla} \log q(\boldsymbol{x})\|^2, \tag{16}$$

where $\boldsymbol{\nabla}$ denotes gradient, and Δ is the Laplacian operator $\sum_{i=1}^{k} \partial^2/(\partial x_i)^2$. The associated discrepancy function is

$$D_H(p, q) = \int \|\boldsymbol{\nabla} \log p(\boldsymbol{x}) - \boldsymbol{\nabla} \log q(\boldsymbol{x})\|^2 \, p(\boldsymbol{x}) \, d\boldsymbol{x}. \tag{17}$$

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 $\boldsymbol{\nabla}$, Δ 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. to ease computation, to improve robustness to model misspecification, or (as in Section 4) to ensure other desirable properties such as consistency.

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

[2] For convenience we have introduced an extra factor of 2.

6 Multivariate Normal Distribution

Consider in particular the case that the distribution Q of $\boldsymbol{X}$ is multivariate normal:

$$\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma), \tag{18}$$

with density

$$q(\boldsymbol{x}) \propto \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Phi(\boldsymbol{x}-\boldsymbol{\mu})\} \tag{19}$$

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, $\boldsymbol{\mu}$ and Φ, but not on $\boldsymbol{x}$.

We have

$$\begin{aligned} \boldsymbol{\nabla} \log q &= -\Phi(\boldsymbol{x}-\boldsymbol{\mu}), & (20)\\ \Delta \log q &= -\operatorname{tr}\Phi & (21) \end{aligned}$$

so that, applying (16),

$$S_H(\boldsymbol{x}, Q) = \|\Phi(\boldsymbol{x}-\boldsymbol{\mu})\|^2 - 2\operatorname{tr}\Phi. \tag{22}$$

The associated discrepancy between $P = \mathcal{N}_k(\boldsymbol{\mu}_P, \Phi_P^{-1})$ and $Q = \mathcal{N}_k(\boldsymbol{\mu}_Q, \Phi_Q^{-1})$ is

$$D_H(P,Q) = \operatorname{tr}\left(\Phi_P - 2\Phi_Q + \Phi_P^{-1}\Phi_Q^2\right) + \left\|\Phi_Q\left(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\right)\right\|^2. \tag{23}$$

The score (22) may be relatively easy to compute if the model is defined in terms of its precision matrix Φ, as for a graphical model. Note also that, whereas the log-score S_L in this case would involve computing the determinant of Φ, this is not required for S_H.

We can now compare different hypothesised multivariate normal distributions Q for the observed data $\boldsymbol{x}$ 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

$$\begin{aligned} S_H(x, Q) &= \frac{1}{\sigma^4}\left\{(x-\mu)^2 - 2\sigma^2\right\}, & (24)\\ D_H(P,Q) &= \frac{1}{\sigma_Q^4}\left\{\frac{\left(\sigma_P^2 - \sigma_Q^2\right)^2}{\sigma_P^2} + (\mu_P - \mu_Q)^2\right\}. & (25) \end{aligned}$$

In this case the Kullback–Leibler discrepancy is given by

$$2\mathrm{KL}(P,Q) = \frac{\sigma_P^2}{\sigma_Q^2} + \log\frac{\sigma_Q^2}{\sigma_P^2} + \frac{(\mu_P - \mu_Q)^2}{\sigma_Q^2} - 1. \tag{26}$$

Using $\log x \leq x - 1$, we find

$$D_H(P, Q) \geq \frac{2}{\sigma_Q^2} \mathrm{KL}(P, Q). \tag{27}$$

In the context of Section 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 $\sigma^2_{Q,i}$ for the variance, under Q, of X_i, given $(X_1, \ldots, X_{i-1})$),

$$\liminf_{i \to \infty} \sigma^2_{Q,i} > 0 \quad \text{a. s. } [P],$$

and likewise with P and Q interchanged.

7 Bayesian Model

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

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

Then we find

$$\begin{aligned}
\frac{\partial \log q(\boldsymbol{x})}{\partial x_i} &= \mathrm{E}\left\{ \frac{\partial \log p(\boldsymbol{x} \mid \Theta)}{\partial x_i} \middle| \boldsymbol{X} = \boldsymbol{x} \right\}, \\
\frac{\partial^2 \log q(\boldsymbol{x})}{\partial x_i^2} &= \mathrm{E}\left\{ \frac{\partial^2 \log p(\boldsymbol{x} \mid \Theta)}{\partial x_i^2} \middle| \boldsymbol{X} = \boldsymbol{x} \right\} + \mathrm{var}\left\{ \frac{\partial \log p(\boldsymbol{x} \mid \Theta)}{\partial x_i} \middle| \boldsymbol{X} = \boldsymbol{x} \right\}
\end{aligned}$$

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

$$\begin{aligned}
S_H(\boldsymbol{x}, Q) &= \sum_i \left(\mathrm{E}\left[2\frac{\partial^2 \log p(\boldsymbol{x} \mid \Theta)}{\partial x_i^2} + 2\left\{ \frac{\partial \log p(\boldsymbol{x} \mid \Theta)}{\partial x_i} \right\}^2 \middle| \boldsymbol{X} = \boldsymbol{x} \right] \right. \\
&\qquad \left. - \left[\mathrm{E}\left\{ \frac{\partial \log p(\boldsymbol{x} \mid \Theta)}{\partial x_i} \middle| \boldsymbol{X} = \boldsymbol{x} \right\} \right]^2 \right) \qquad (28) \\
&= \mathrm{E}\left\{ S_H(\boldsymbol{x}, P_\Theta) \middle| \boldsymbol{X} = \boldsymbol{x} \right\} \\
&\qquad + \sum_i \mathrm{var}\left\{ \frac{\partial \log p(\boldsymbol{x} \mid \Theta)}{\partial x_i} \middle| \boldsymbol{X} = \boldsymbol{x} \right\}. \qquad (29)
\end{aligned}$$

7.1 Exponential Family

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

$$\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) = a(\boldsymbol{x}) + b(\boldsymbol{\theta}) + \sum_{j=1}^{k} \theta_j t_j(\boldsymbol{x}). \tag{30}$$

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

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

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

For the special case $\boldsymbol{T} = \boldsymbol{X}$, this becomes

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

8 Linear Model: Variance Known

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

$$\boldsymbol{Y} \sim \mathcal{N}(X\boldsymbol{\theta}, \sigma^2 I), \tag{31}$$

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

8.1 Multivariate Score

Consider giving $\boldsymbol{\theta}$ a normal prior distribution:

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{m}, V). \tag{32}$$

The marginal distribution Q of $\boldsymbol{Y}$ is then

$$\boldsymbol{Y} \sim \mathcal{N}(X\boldsymbol{m}, XVX^{\mathrm{T}} + \sigma^2 I) \tag{33}$$

with precision matrix

$$\begin{aligned} \Phi &= (XVX^{\mathrm{T}} + \sigma^2 I)^{-1} \\ &= \sigma^{-2}\left\{I - X\left(X^{\mathrm{T}}X + \sigma^2 V^{-1}\right)^{-1} X^{\mathrm{T}}\right\} \end{aligned}$$

on applying the Woodbury matrix inversion lemma ((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

$$\Pi := I - XAX^{\mathrm{T}},$$

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

Although this Φ is singular, and thus cannot arise from any genuine dispersion matrix, there is no problem in using it in (22). We obtain

$$S_H(\boldsymbol{y}, Q) = \frac{1}{\sigma^4}\left(R - 2\nu\sigma^2\right) \tag{34}$$

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 $\mathrm{rank}(X) < p$—in particular, whenever $n < p$. Taking $V^{-1} \to 0$ is equivalent to using an improper prior density $\pi(\boldsymbol{\theta}) \equiv c$, with $0 < c < \infty$. When X is of rank p, the integral formally defining the marginal density of $\boldsymbol{Y}$ is finite for each $\boldsymbol{y}$ (even though the resulting density is itself improper). However, when $\mathrm{rank}(X) < p$ this integral is infinite at each $\boldsymbol{y}$, 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 σ^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 σ^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:

$$\begin{aligned} M_1 &: Y_i \sim \mathcal{N}(0,1), \\ M_2 &: Y_i \sim \mathcal{N}(\theta,1). \end{aligned}$$

Then, with $\overline{Y}$ denoting the sample mean $\sum_i Y_i/n$, we have $\mathrm{AIC}_1 = \sum_i Y_i^2$, $\mathrm{AIC}_2 = \sum_i (Y_i - \overline{Y})^2 + 2$, so that $\mathrm{AIC}_1 - \mathrm{AIC}_2 = n\overline{Y}^2 - 2$. 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 $\boldsymbol{x}_i$ be the ith row of X, and X^i the matrix containing the first i rows of X. Assuming X is of full rank, then X^i is of full rank if and only if $i \geq p$.

Define, for $i \geq p$:

$$A_i := \left\{ (X^i)^{\mathrm{T}} (X^i) \right\}^{-1}, \tag{35}$$

$$\widehat{\boldsymbol{\theta}}_i := A_i (X^i)^{\mathrm{T}} \boldsymbol{Y}^i \tag{36}$$

and, for $i > p$:

$$\eta_i := \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\theta}}_{i-1}, \tag{37}$$

$$k_i^2 := 1 + \boldsymbol{x}_i^{\mathrm{T}} A_{i-1} \boldsymbol{x}_i = (1 - \boldsymbol{x}_i^{\mathrm{T}} A_i \boldsymbol{x}_i)^{-1}, \tag{38}$$

$$Z_i := k_i^{-1} (Y_i - \eta_i) \tag{39}$$

(where the identity in (38) follows from the Woodbury lemma).

Then for the improper prior (32) with $V^{-1} \to 0$, the predictive distribution of Y_i, given $\boldsymbol{Y}^{i-1}$, is

$$Y_i \sim \mathcal{N}(\eta_i, k_i^2\sigma^2) \quad (i > p). \tag{40}$$

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 $\boldsymbol{\theta}$); moreover, $R = \sum_{i=p+1}^{n} Z_i^2$.

Note that, under the model (31), η_i has expectation $\boldsymbol{x}_i^{\mathrm{T}}\theta$ and variance $k_i^2 - 1$. So the predictive distribution (40) and the true distribution will be asymptotically indistinguishable (the property of "prequentially consistent" estimation—see Dawid (1984)) if and only if

$$k_i^2 \to 1 \text{ as } i \to \infty. \tag{41}$$

This we henceforth assume, for any model under consideration.

For $i > p$, the incremental score (24) associated with (40) is

$$S_i = \frac{T_i}{k_i^2\sigma^2} \tag{42}$$

where

$$T_i := \frac{Z_i^2}{\sigma^2} - 2. \tag{43}$$

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

$$\mathrm{E}(T_i) = -1, \tag{44}$$

$$\mathrm{var}(T_i) = 2. \tag{45}$$

As discussed in Section 4, minimising the cumulative prequential score

$$S^* := \sum_i S_i \tag{46}$$

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

Expression (42) 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:

$$S_H(\boldsymbol{y}, Q) = \frac{1}{\sigma^2} \sum_{i=p+1}^{n} T_i = \sum_{i=p+1}^{n} k_i^2 S_i, \tag{47}$$

and the scaling factor k_i^2 has been assumed to satisfy (41). It would thus seem that (47) is asymptotically equivalent to (46), 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

$$S^* - S_H = \frac{1}{\sigma^2} \sum_{i=p}^{n} \left(\frac{1}{k_i^2} - 1 \right) T_i. \tag{48}$$

Under any distribution in the model, this has expectation

$$\frac{1}{\sigma^2} \sum_i \left(1 - \frac{1}{k_i^2} \right) = \frac{1}{\sigma^2} \sum_i \boldsymbol{x}_i^{\mathrm{T}} A_i \boldsymbol{x}_i,$$

and variance

$$\frac{2}{\sigma^4} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^{\mathrm{T}} A_i \boldsymbol{x}_i \right)^2.$$

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

$$\mathrm{E}\left(i\boldsymbol{x}_i^{\mathrm{T}} A_i \boldsymbol{x}_i \right) = \mathrm{E}\,\mathrm{tr} \left\{ \left(\sum_{j=1}^{i} \boldsymbol{x}_j \boldsymbol{x}_j^{\mathrm{T}} / i \right)^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right\} \approx \mathrm{tr}\, C^{-1} C = p.$$

So $1 - 1/k_i^2 \approx p/i$; in particular (41) holds. Then $\mathrm{E}(S^* - S_H) \approx (p/\sigma^2) \sum_{i=p}^{n} i^{-1} \approx p(\log n)/\sigma^2$. A similar analysis shows $\mathrm{var}(S^* - S_H) < \infty$. Thus, under the model, $S^* - S_H \approx p(\log n)/\sigma^2$. 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 Consistent Model Selection

We now consider the asymptotic behaviour of the cumulative prequential score S^*, given by (46), 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 σ_1^2 and σ_2^2. Let Z_i, k_i^2, as defined above, refer to M_1, and denote the corresponding quantities for M_2 by, respectively, W_i, h_i^2. 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

$$1 - 1/k_i^2 \approx p_1/i, \tag{49}$$

$$1 - 1/h_i^2 \approx p_2/i. \tag{50}$$

Since the (Y_i) are independent normal variables with variance σ_1^2, and the (Z_i) and (W_i) are, in each case, constructed from the (Y_i) by an orthogonal linear transformation, we will have

$$Z_i \sim \mathcal{N}(0, \sigma_1^2) \quad \text{independently,} \tag{51}$$
$$W_i \sim \mathcal{N}(\nu_i, \sigma_1^2) \quad \text{independently,} \tag{52}$$

where the (Z_i) have mean 0 since M_1 is true, whereas the (ν_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

$$S_2^* - S_1^* = \frac{1}{\sigma_2^2} \sum \frac{1}{h_i^2} \left(\frac{W_i^2}{\sigma_2^2} - 2 \right) - \frac{1}{\sigma_1^2} \sum \frac{1}{k_i^2} \left(\frac{Z_i^2}{\sigma_1^2} - 2 \right) \tag{53}$$

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

On account of (51) and (52), this has expectation

$$\mathrm{E}(S_2^* - S_1^*) = \frac{1}{\sigma_2^4} \sum \frac{\nu_i^2}{h_i^2} + \frac{(\sigma_1^2 - \sigma_2^2)^2}{\sigma_1^2 \sigma_2^4} \sum \frac{1}{h_i^2} + \frac{1}{\sigma_1^2} \sum \left(\frac{1}{k_i^2} - \frac{1}{h_i^2} \right). \tag{54}$$

We now consider various cases for M_2.

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 (54) reduces to

$$\mathrm{E}(S_2^* - S_1^*) = \frac{1}{\sigma^2} \sum \left(\frac{1}{k_i^2} - \frac{1}{h_i^2} \right). \tag{55}$$

On account of (49) and (50), this behaves asymptotically as $(p_2 - p_1)(\log n)/\sigma^2 + o(\log n)$. Also, an analysis similar to that in Section 8.3 shows that $\mathrm{var}(S_2^* - S_1^*)$ is bounded, so that

$$S_2^* - S_1^* = \frac{(p_2 - p_1) \log n}{\sigma^2} + o_p(\log n). \tag{56}$$

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

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.

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 (54) 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 $\propto n$—just as for the log-score in the case of proper priors.

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 M_2. We note that the log-Bayes factor (4) 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

$$\mathrm{E}(S_2^* - S_1^*) = \frac{1}{\sigma^4} \sum \frac{\nu_i^2}{h_i^2} + \frac{1}{\sigma^2} \sum \left(\frac{1}{k_i^2} - \frac{1}{h_i^2} \right), \tag{57}$$

where $\nu_i \not\equiv 0$ and $h_i^2 \not\equiv k_i^2$.

The first term in (54) 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 (54) 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 Section 8.1.

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

For model M_2, with $Y_i \sim \mathcal{N}(\theta, 1)$ $(\theta \neq 0)$, we have $p_2 = 1$, $h_i^2 = i/(i-1)$, $W_i = \{(i-1)/i\})^{\frac{1}{2}}(Y_i - \overline{Y}_{i-1}) \sim \mathcal{N}(0, 1)$. Although $h_i^2 \to 1$, $S_2^* - S_{H,2}$ 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 $S_1^* \equiv S_{H,1}$, 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 $S_2^* = \sum_i (1 - 1/i)(W_i^2 - 2) = -n + o_p(n)$, while $S_1^* = \sum_i (Y_i^2 - 2) = n(\theta^2 - 1) + o_p(n)$, 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 (34) 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 σ^2 in (31). With $\phi = 1/\sigma^2$, we have model density

$$p(\boldsymbol{y} \mid \theta, \phi) \propto \phi^{\frac{1}{2}n} \exp -\frac{\phi}{2} \left\{ R + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\mathrm{T}} X^{\mathrm{T}} X (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \right\} \tag{58}$$

where $R = \boldsymbol{y}^{\mathrm{T}} \Pi \boldsymbol{y}$, with $\Pi = I - XAX^{\mathrm{T}}$, is the residual sum of squares, on $\nu = n - p$ degrees of freedom.

The standard improper prior for this model is $\pi(\boldsymbol{\theta}, \phi) \propto \phi^{-1}$. Multiplying (58) by this and integrating over $(\boldsymbol{\theta}, \phi)$ yields the (improper) joint predictive density[3]

$$p(\boldsymbol{y}) \propto R^{-\frac{1}{2}\nu}, \tag{59}$$

with logarithm (up to a constant)

$$l = -\frac{1}{2} \nu \log R. \tag{60}$$

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

$$\frac{\partial l}{\partial y_i} = -\frac{\nu r_i}{R}, \tag{61}$$

$$\frac{\partial^2 l}{\partial y_i^2} = \nu \left(\frac{2 r_i^2}{R^2} - \frac{\pi_{ii}}{R} \right), \tag{62}$$

and so (noting $\sum_i \pi_{ii} = \nu$) the multivariate score (16) is

$$S_H = -\frac{(\nu - 4)}{\widehat{\sigma}^2} \tag{63}$$

where $\widehat{\sigma}^2 = R/\nu$ is the usual unbiased estimator of σ^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 $J := \widehat{\sigma}^2/(\nu - 4)$.

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

$$J_1 = \frac{(n-1)S^2 + n\overline{Y}^2}{n(n-4)}, \tag{64}$$

$$J_2 = \frac{S^2}{(n-5)} \tag{65}$$

[3] For the integral formally defining this density to be finite at each point we require $\mathrm{rank}(X) \geq p+1$.

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

$$\frac{n\overline{Y}^2}{\sigma^2} > \frac{2n-5}{(n-5)}\,\frac{S^2}{\sigma^2} \approx 2 \tag{66}$$

for large n. But, under M_1, $n\overline{Y}^2/\sigma^2 \sim \chi^2_1$, so that there is a positive probability of the inequality (66) holding, so favouring the more complex model M_2.

9.1 Prequential Score

From (59), as a function of y_i the predictive density of Y_i given $\boldsymbol{y}^{i-1}$ (for $i > p$) is

$$p(y_i \,|\, \boldsymbol{y}^{i-1}) \propto R_i^{-\frac{1}{2}\nu_i} = \left(R_{i-1} + z_i^2\right)^{-\frac{1}{2}\nu_i} \tag{67}$$

where R_i is the residual sum-of-squares based on $\boldsymbol{y}^i$, on $\nu_i := i - p$ degrees of freedom, and $z_i = k_i^{-1}(y_i - \eta_i)$, as given by (37)–(39). Applying the univariate case of (16) now yields (for $i > p$) the incremental score:

$$S_i = \frac{\nu_i\left\{(4+\nu_i)\,Z_i^2 - 2R_i\right\}}{k_i^2 R_i^2} \tag{68}$$

$$= \frac{\left(1+\frac{4}{\nu_i}\right)Z_i^2 - 2s_i^2}{k_i^2 s_i^4}, \tag{69}$$

where $s_i^2 := R_i/\nu_i$ is the residual mean square, based on $\boldsymbol{Y}^i$, 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 Section 8.4 that in this case, when the variance σ^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 $\mathcal{N}(0, \sigma^2)$.

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

$$k_i^2\sigma^2 S_i = \frac{\left(1+\frac{4}{\nu_i}\right)(U_i+1) - 2(\overline{U}_i+1)}{(\overline{U}_i+1)^2} \tag{70}$$

[4] Again, an additional contribution of the form of (63), 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.

with $\overline{U}_i := \nu_i^{-1} \sum_{j=p+1}^{i} U_j$ (where $\nu_i = i - p$). Now $\overline{U}_i = O_p(i^{-\frac{1}{2}})$. Expanding (70) as a power series in $\overline{U}_i$ gives

$$k_i^2 \sigma^2 S_i = \sum_{r=0}^{\infty} (-1)^r \overline{U}_i^r \left\{ (r+1) \left(1 + \frac{4}{\nu_i} \right) (U_i + 1) - 2 \right\} \tag{71}$$

so that

$$\begin{aligned} k_i^2 \sigma^2 S_i - (U_i - 1) &= \frac{4}{\nu_i} + \frac{4U_i}{\nu_i} && (72) \\ &\quad - 2\overline{U}_i \left(U_i + \frac{4}{\nu_i} + \frac{4U_i}{\nu_i} \right) && (73) \\ &\quad + \overline{U}_i^2 \left(1 + 3U_i + \frac{12}{\nu_i} + \frac{12U_i}{\nu_i} \right) && (74) \\ &\quad + O_p(i^{-3/2}). && (75) \end{aligned}$$

Noting

$$\begin{aligned} \mathrm{E}(\overline{U}_i^2) &= 2/\nu_i, && (76) \\ \mathrm{E}(\overline{U}_i U_i) &= 2/\nu_i, && (77) \\ \mathrm{E}(\overline{U}_i^2 U_i) &= 8/\nu_i^2, && (78) \end{aligned}$$

we compute

$$\mathrm{E}\left\{ k_i^2 \sigma^2 S_i - (U_i - 1) \right\} = \frac{2}{i} + O(i^{-3/2}), \tag{79}$$

whence, on account of (41),

$$\mathrm{E}\left(S^* - S_0^* \right) = 2(\log n)/\sigma^2 + O(1) \tag{80}$$

where $S_0^* = \sum_{i=p+1}^{n} (U_i - 1)/(k_i^2 \sigma^2)$ is the cumulative prequential score (46) for the submodel in which the correct variance σ^2 is known.

In the remainder of this section, we argue that $S^* - S_0^*$ differs from its expectation (80) by $O_p\{(\log n)^{\frac{1}{2}}\}$. Computations have been executed and/or checked using the software *Mathematica*.

On cumulating the term $\propto U_i/\nu_i$ in (72) 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 $\propto \overline{U}_i U_i$ in (73). We find $\mathrm{var}(\overline{U}_i U_i) = 4/\nu_i + O(\nu_i^{-2})$, and $\overline{U}_i U_i$ and $\overline{U}_j U_j$ are uncorrelated for $i \neq j$. Hence on cumulating the term $\overline{U}_i U_i$ in (73) 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 $O_p\{(\log n)^{\frac{1}{2}}\}$ to $S^* - S_0^*$.

There is also a term $\propto \overline{U}_i/\nu_i$ in (73). Since $\overline{U}_i/\nu_i = O_p(i^{-3/2})$, its cumulative sum is $O_p(n^{-\frac{1}{2}})$.

Now consider (74). We look first at the term $\overline{U}_i^2$. We compute $\mathrm{var}\{(\overline{U}_i)^2\} = 8/\nu_i^2 + 48/\nu_i^3 = \lambda_i$, say; and, for $i < j$,

$$\mathrm{Cov}\{(\overline{U}_i)^2, (\overline{U}_j)^2\} = \left(\frac{\nu_i}{\nu_j}\right)^2 \lambda_i.$$

Hence

$$\begin{aligned} \mathrm{var}\left\{\sum_{i=p+1}^{n} (\overline{U}_i)^2\right\} &= \sum_{i=p+1}^{n} \lambda_i + 2\sum_{i=p+1}^{n}\sum_{j=i+1}^{n}\left(\frac{\nu_i}{\nu_j}\right)^2 \lambda_i \\ &\leq 56\left\{\sum_{i=1}^{\nu} i^{-2} + 2\sum_{i=1}^{\nu}\sum_{j=i+1}^{\nu} j^{-2}\right\} \end{aligned}$$

(with $\nu = n - p$), since $\lambda_i \leq 56/\nu_i^2$. We have $\sum_{i=1}^{\infty} i^{-2} < \infty$, and, for large i, $\sum_{j=i+1}^{\nu} j^{-2} < \sum_{j=i+1}^{\infty} j^{-2} \approx i^{-1}$. So $\mathrm{var}\{\sum_{i=p+1}^{n}(\overline{U}_i)^2\}$ is of order $\log n$, and cumulating the term $\overline{U}_i^2$ in (74) again makes a contribution $O_p\{(\log n)^{\frac{1}{2}}\}$ over and above its expectation.

Now consider the term $U_i\overline{U}_i^2$ in (74). We have

$$\mathrm{var}(U_i\overline{U}_i^2) = \frac{24}{\nu_i^2} + \frac{1024}{\nu_i^3} + \frac{4928}{\nu_i^4} \tag{81}$$

and, for $i < j$,

$$\mathrm{Cov}(U_i\overline{U}_i^2, U_j\overline{U}_j^2) = \frac{48(\nu_i + 4)}{\nu_i^2\nu_j^2}. \tag{82}$$

By an argument similar to that for $\overline{U}_i^2$, we find that cumulating the term $U_i\overline{U}_i^2$ in (74) again makes a contribution $O_p\{(\log n)^{\frac{1}{2}}\}$ (over and above its expectation).

Putting everything together, we have

$$S^* - S_0^* = 2(\log n)/\sigma^2 + O_p\{(\log n)^{\frac{1}{2}}\}. \tag{83}$$

Now we have shown in Section 8.4 that, for comparing two true models M_1 and M_2 with known variance σ^2 and respective dimensions $p_1 < p_2$, under conditions on the behaviour of the $(\boldsymbol{x}_i)$, the difference in their cumulative prequential scores S_0^* behaves asymptotically as $(p_2 - p_1)(\log n)/\sigma^2$. Since, from (83), 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 predictive density, 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, with their usual improper priors.

While the literature on "objective" Bayesian model selection contains some valuable discussion of general principles—see, for example, Bayarri et al. (2012)—most of it focuses on explorations and recommendations of appropriate priors, or classes of priors, or relationships between priors, for use in specified circumstances or for specified purposes. When those priors are improper, as is commonly the case, further manipulations and distortions of the Bayes factor are required to produce a well-defined procedure. Our approach here makes no specific recommendations, leaving users free to apply their most favoured prior distributions. Instead, we have introduced a very general procedure, based on homogeneous proper scoring rules, that allows the use of improper priors, however selected, without needing to worry about the arbitrariness of their scaling constants.

There remains the issue of the choice of homogeneous proper scoring rule. There are no clear theoretical grounds for preferring one over another. Purely for simplicity, we have confined attention to the most basic homogeneous rule, the Hyvärinen score, but similar results can be expected for other homogeneous scoring rules. Further theoretical and computational exploration and comparison of the properties of the various methods is clearly required. Such exploration might be extended to their performance in other contexts: for example, issues of consistent model selection when the number of parameters increases with the sample size (Moreno et al., 2010; Johnson and Rossell, 2012).

References

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian Model Choice with Application to Variable Selection." *The Annals of Statistics*, 40: 1550–1577. MR3015035. doi: http://dx.doi.org/10.1214/12-AOS1013. 497

Berger, J. O. and Pericchi, L. R. (1996). "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association*, 91: 109–122. MR1394065. doi: http://dx.doi.org/10.2307/2291387. 481

Dawid, A. P. (1984). "Statistical Theory—The Prequential Approach (with Discussion)." *Journal of the Royal Statistical Society, Series A*, 147: 278–292. MR0763811. doi: http://dx.doi.org/10.2307/2981683. 482, 489

— (1986). "Probability Forecasting." In: Kotz, S., Johnson, N. L., and Read, C. B. (eds.), *Encyclopedia of Statistical Sciences*, volume 7, 210–218. New York: Wiley-Interscience. MR0892738. 481

— (1992). "Prequential Analysis, Stochastic Complexity and Bayesian Inference (with Discussion)." In: Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, 109–125. Oxford: Oxford University Press. MR1380273. 483

— (2011). "Posterior Model Probabilities." In: Bandyopadhyay, P. S. and Forster, M. (eds.), *Philosophy of Statistics*, 607–630. New York: Elsevier. 480, 481

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*, 22–28. University of Tokyo. 12–16 December 2005. 484

Dawid, A. P. and Musio, M. (2013). "Estimation of Spatial Processes Using Local Scoring Rules." *AStA Advances in Statistical Analysis*, 97: 173–179. MR3045766. doi: http://dx.doi.org/10.1007/s10182-012-0191-8. 484

— (2014). "Theory and Applications of Proper Scoring Rules." *Metron*, 72: 169–183. MR3233147. doi: http://dx.doi.org/10.1007/s40300-014-0039-y. 481, 482, 484

Dawid, A. P., Musio, M., and Ventura, L. (2015). "Minimum Scoring Rule Inference." *Scandinavian Journal of Statistics*, submitted for publication. arXiv:1403.3920 481

Good, I. J. (1952). "Rational Decisions." *Journal of the Royal Statistical Society, Series B*, 14: 107–114. MR0077033. 481

Grünwald, P. D. and Dawid, A. P. (2004). "Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory." *The Annals of Statistics*, 32: 1367–1433. MR2089128. doi: http://dx.doi.org/10.1214/009053604000000553. 482

Hyvärinen, A. (2005). "Estimation of Non-Normalized Statistical Models by Score Matching." *Journal of Machine Learning Research*, 6: 695–709. MR2249836. 484

Johnson, V. E. and Rossell, D. (2012). "Bayesian Model Selection in High-Dimensional Settings." *Journal of the American Statistical Association*, 107: 649–660. MR2980074. doi: http://dx.doi.org/10.1080/01621459.2012.682536. 497

Kabanov, Y. M., Liptser, R. S., and Shiryayev, A. N. (1977). "On the Question of Absolute Continuity and Singularity of Probability Measures." *Mathematics of the USSR. Sbornik*, 33: 203–221. 482

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. MR0415861. 487

Mameli, V., Musio, M., and Dawid, A. P. (2014). "Comparisons of Hyvärinen and Pairwise Estimators in Two Simple Linear Time Series Models." arXiv:1409.3690 MR3233147. doi: http://dx.doi.org/10.1007/s40300-014-0039-y. 484

Moreno, E., Girón, F. J., and Casella, G. (2010). "Consistency of Objective Bayes Factors as the Model Dimension Grows." *The Annals of Statistics*, 38: 1937–1952. MR2676879. doi: http://dx.doi.org/10.1214/09-AOS754. 497

Musio, M. and Dawid, A. P. (2013). "Local Scoring rules: A Versatile Tool for Inference." Paper presented at 59th World Statistics Congress, Hong Kong. http://www.statistics.gov.hk/wsc/STS019-P3-S.pdf 481

O'Hagan, A. (1995). "Fractional Bayes Factors for Model Comparison." *Journal of the Royal Statistical Society, Series B*, 57: 99–138. MR1325379. 481

Parry, M. F. (2013). "Multidimensional Local Scoring Rules." Paper presented at 59th World Statistics Congress, Hong Kong. http://www.statistics.gov.hk/wsc/STS019-P2-S.pdf 484

Parry, M. F., Dawid, A. P., and Lauritzen, S. L. (2012). "Proper Local Scoring Rules." *The Annals of Statistics*, 40: 561–592. MR3014317. doi: http://dx.doi.org/10.1214/12-AOS971. 484

Skouras, K. (1998). "Absolute Continuity of Markov Chains." *Journal of Statistical Planning and Inference*, 75: 1–8. MR1671674. doi: http://dx.doi.org/10.1016/S0378-3758(98)00117-7. 483

Stout, W. F. (1970). "A Martingale Analogue of Kolmogorov's Law of the Iterated Logarithm." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 15: 279–290. MR0293701. 483

Acknowledgments

We thank the Editor, Associate Editor and referees for their helpful feedback on a previous version of this article.