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

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

This can be thought of as a hybrid between an "objective" component, $p_M(x \mid \theta_M)$, 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 $\alpha(M \mid \boldsymbol{x})$ for model M is then given by Bayes's formula:

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

where the omitted multiplicative constant is adjusted to ensure $\sum_{M \in \mathcal{M}} \alpha(M \mid \boldsymbol{x}) = 1$. 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 $1 \leq \epsilon \leq 2$ (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 $\log x \le x - 1$, we find

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

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 $\sigma_{Q,i}^2$ for the variance, under Q, of X_i, given $(X_1, \ldots, X_{i-1})$),

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

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

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

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

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

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 $\operatorname{rank}(X) < p$—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 $\operatorname{rank}(X) < p$ 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:

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

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 $i \geq p$:

$$\begin{aligned} A_i &:= \left\{(X^i)^{\mathrm{T}}(X^i)\right\}^{-1}, && (35) \\ \widehat{\boldsymbol{\theta}}_i &:= A_i (X^i)^{\mathrm{T}} \boldsymbol{Y}^i && (36) \end{aligned}$$

and, for $i > p$:

$$\begin{aligned} \eta_i &:= \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\theta}}_{i-1}, && (37) \\ k_i^2 &:= 1 + \boldsymbol{x}_i^{\mathrm{T}} A_{i-1} \boldsymbol{x}_i = (1 - \boldsymbol{x}_i^{\mathrm{T}} A_i \boldsymbol{x}_i)^{-1}, && (38) \\ Z_i &:= k_i^{-1}(Y_i - \eta_i) && (39) \end{aligned}$$

(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

$$\begin{aligned} 1 - 1/k_i^2 &\approx p_1/i, & (49) \\ 1 - 1/h_i^2 &\approx p_2/i. & (50) \end{aligned}$$

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

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

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

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

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

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 $\operatorname{rank}(X) \geq p+1$.

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, $n\overline{Y}^2/\sigma^2 \sim \chi_1^2$, 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

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

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

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

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 $\text{var}\{(\overline{U}_i)^2\} = 8/\nu_i^2 + 48/\nu_i^3 = \lambda_i$, say; and, for $i < j$,

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

Hence

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

(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 $\text{var}\{\sum_{i=p+1}^{n}(\overline{U}_i)^2\}$ 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

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

and, for $i < j$,

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

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.

Bayesian Analysis (2015) **10**, Number 2, pp. 501–504

Comment on Article by Dawid and Musio[*]

Matthias Katzfuss[†] and Anirban Bhattacharya[‡]

The authors consider the interesting and important issue of Bayesian inference based on objective functions other than the likelihood. They focus on model selection in the low-dimensional setting using prequential local proper scoring rules.

1 General non-likelihood-based inference

There is a large and disparate literature on inference based on objective functions other than the likelihood. We will briefly mention some examples here, but we believe that a more thorough review and comparison would be a worthy endeavor.

Numerous objective functions have been proposed to replace the (log-)likelihood in pursuit of various inference goals. Proper scoring rules are a natural choice for serving as such objective functions, due to their property of being minimized (in expectation) under the true model. Depending on the goal of the analysis, certain well-known proper scoring rules can achieve robustness (e.g., continuous ranked probability score, or CRPS), have simple closed-form expressions (e.g., Dawid–Sebastiani score), or do not require densities (e.g., CRPS) or normalizing constants (e.g., Hyvärinen score, as in the present paper). See Gneiting and Katzfuss (2014) for a recent review of these and other scoring rules.

In a frequentist context, examples of approaches falling into this category of scoring-rule-based inference are minimum contrast estimation (e.g., Pfanzagl, 1969; Birgé and Massart, 1993), composite likelihood (e.g., Lindsay, 1988), and M-estimation (e.g., Huber and Ronchetti, 2009). Some further review is given in Dawid et al. (2014).

There have also been related approaches in the Bayesian framework. Shaby (2014) provides a nice review of Bayesian inference using general objective functions and, based on results of Chernozhukov and Hong (2003), he proposes an "open-faced sandwich adjustment" to obtain pseudo-posteriors with properly calibrated frequentist properties. Further, the "Gibbs posterior" (Jiang and Tanner, 2008; Li et al., 2013) has received considerable interest, where the negative log-likelihood is replaced by some "empirical risk" R_n (usually targeting the specific parameter to be estimated) to construct a pseudo-posterior of the form

$$Q(\theta) \propto \exp\{-\lambda R_n(\theta)\}\pi(\theta), \tag{1}$$

where λ is a positive scaling constant (often called "temperature"). Sampling from the pseudo-posterior Q can be performed via standard MCMC algorithms.

[*] Main article DOI: 10.1214/15-BA942.

[†] Department of Statistics, Texas A&M University, katzfuss@tamu.edu

[‡] Department of Statistics, Texas A&M University, anirbanb@stat.tamu.edu

DOI: 10.1214/15-BA942A

2 Objective Bayesian model selection

In objective Bayesian model selection, a discrete prior is assumed on a (finite) class of models, and given a particular model, objective improper priors are placed on the model parameters. While improper priors are commonly used for analysis of a single model, one faces difficulties in comparing models via Bayes factors, since the marginal likelihoods of the competing models are only specified up to arbitrary constants. A number of remedies have been proposed in the literature to deal with this issue, such as fractional Bayes factors (O'Hagan, 1995) and intrinsic Bayes factors (Berger and Pericchi, 1996).

In the present paper, the authors take a different approach, which relies on replacing the (log-)marginal likelihood by a local proper scoring rule. The Hyvärinen score is recommended as a default. From the expression of the Hyvärinen score in the authors' equation (16), it can be seen that the arbitrary constant disappears. The authors look at examples where the Hyvärinen scores are analytically tractable and provide asymptotic orders for the difference in Hyvärinen scores assuming the respective models to be true.

Some clarification regarding practical implementation of the model selection procedure presented here would be helpful. When can we be sure that one model is truly better than another — or in other words, can anything be said about posterior model probabilities (also see Section 3 below)? Can the the necessary quantities be computed for models beyond the simple Gaussian examples considered in the paper?

3 Scaling issues

As indicated in (1) above, the literature on Gibbs posteriors typically includes a multiplicative scaling constant λ on the objective function. The choice of λ is considered a critical issue, as it has a direct effect on the (pseudo-)posterior uncertainty. Shaby (2014) does not consider a multiplicative scaling of the objective function, but his open-face-sandwich correction automatically adjusts for such scaling, and his approach is thus invariant to scaling. Without such a correction, the scaling issue also arises when the objective function is specified to be a proper scoring rule, including the Hyvärinen score. As implicitly acknowledged by the authors in their Footnote 2, the scaling of a proper scoring rule is arbitrary, in that any proper scoring rule is still proper when multiplied by a constant.

In the context of model selection between models M_1 and M_2 with scores S_{M_1} and S_{M_2}, respectively, the scaling constant can arbitrarily inflate or deflate the pseudo Bayes factor,

$$\mathrm{PBF} = \frac{\exp(\lambda S_{M_1})}{\exp(\lambda S_{M_2})}$$

and thus the amount of evidence in favor of M_1 over M_2 (cf. Kass and Raftery, 1995). This also makes it challenging to compute pseudo posterior model probabilities, such as

$$\widetilde{P}(M_1|\mathbf{x}) = \frac{\exp(\lambda S_{M_1})}{\exp(\lambda S_{M_1}) + \exp(\lambda S_{M_2})}. \tag{2}$$

If S_{M_1} is larger than S_{M_2}, (2) can be arbitrarily close to 0.5 or 1 by choosing λ to be very small or very large, respectively.

In light of these scaling issues, how should model selection be calibrated and interpreted? Moreover, is it possible to handle more than two competing models or even high-dimensional settings, where the number of competing models may grow exponentially with the sample size? In the high-dimensional linear regression context, Johnson and Rossell (2012) showed that a number of commonly used procedures (including fractional and intrinsic Bayes factors) assign vanishingly small posterior probabilities to the true model with increasing sample size. The scaling issue may assume an even more important role in such cases.

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: 109–122. MR1394065. doi: http://dx.doi.org/10.2307/2291387. 502

Birgé, L. and Massart, P. (1993). "Rates of convergence for minimum contrast estimators." *Probability Theory and Related Fields*, 97: 113–150. MR1240719. doi: http://dx.doi.org/10.1007/BF01199316. 501

Chernozhukov, V. and Hong, H. (2003). "An MCMC approach to classical estimation". *Journal of Econometrics*, 115: 293–346. MR1984779. doi: http://dx.doi.org/10.1016/S0304-4076(03)00100-3. 501

Dawid, P., Musio, M., and Ventura, L. (2014). "Minimum scoring rule inference." arXiv:1403.3920. 501

Gneiting, T. and Katzfuss, M. (2014). "Probabilistic forecasting." *Annual Review of Statistics and Its Application*, 1(1): 125–151. doi: http://dx.doi.org/10.1146/annurev-statistics-062713-085831. 501

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. Hoboken, NJ: Wiley, 2nd edition. MR2488795. doi: http://dx.doi.org/10.1002/9780470434697. 501

Jiang, W. and Tanner, M. A. (2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining." *The Annals of Statistics*, 36(5): 2207–2231. MR2458185. doi: http://dx.doi.org/10.1214/07-AOS547. 501

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107(498): 649–660. MR2980074. doi: http://dx.doi.org/10.1080/01621459.2012.682536. 503

Kass, R. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430): 773–795. doi: http://dx.doi.org/10.1080/01621459.1995.10476572. 502

Li, C., Jiang, W., and Tanner, M. A. (2013). "General oracle inequalities for Gibbs posterior with application to ranking." *Journal of Machine Learning Research: Workshop and Conference Proceedings 30*, 512–521. 501

Lindsay, B. (1988). "Composite likelihood methods." *Contemporary Mathematics*, 80: 221–239. MR0999014. doi: http://dx.doi.org/10.1090/conm/080/999014. 501

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57: 99–138. MR1325379. 502

Pfanzagl, J. (1969). "On the measurability and consistency of minimum contrast estimates." *Metrika*, 14(1): 249–272. doi: http://dx.doi.org/10.1007/BF02613654. 501

Shaby, B. A. (2014). "The open-faced sandwich adjustment for MCMC using estimating functions." *Journal of Computational and Graphical Statistics*, 23(3): 853–876. MR3224659. doi: http://dx.doi.org/10.1080/10618600.2013.842174. 501, 502

Bayesian Analysis (2015) **10**, Number 2, pp. 505–509

Comment on Article by Dawid and Musio[*],[†]

Christopher M. Hans[‡] and Mario Peruggia[§]

Dawid and Musio present interesting results on how to affect model comparison using proper scoring rules, focusing chiefly on Bayesian model comparison. Among the reasons stated to justify the proposed approach we note:

1. The insensitivity of the procedure to a renormalization of the prior distribution,
2. The flexibility and/or robustness of the method when implemented using a prequential score.

The focus of the article is on the derivation of consistency results for the proper scoring rule methods based both on their implementation through a multivariate score and a prequential score. There are very many such results in the article, but the gist of the argument is that some form of proper scoring rule method can produce a consistent procedure even in cases when the standard Bayesian approach fails to do so or when it fails altogether, as is the case when improper priors are used and Bayes factors cannot be calculated.

Consistent model selection is unquestionably a desirable property as is the formulation of a coherent, universal framework for statistical inference. The Bayesian approach using *proper* priors accomplishes the latter. The proposed proper scoring rule methods mend the complications that arise when the Bayesian approach is used with improper priors. However, the beauty of the coherent Bayesian inferential framework is lost when model comparison is no longer based on the likelihood score. As in all compromises, something is gained at the expense of losing something else, or, as some would say, there is no free lunch!

Then, for those situations in which the Bayesian approach is not broken, two questions arise naturally:

1. When does a proper scoring rule model comparison produce a different answer than a log-score model comparison?
2. For those situations in which the answers are different, can an argument be made for preferring the proper scoring rule method?

This suggests juxtaposing the proposed method to model comparison methods that compare directly the (log-) likelihoods for the various models.

[*] Main article DOI: 10.1214/15-BA942.

[†] This work is supported by the National Science Foundation under award numbers DMS-1310294, SES-1024709, and SES-1424481.

[‡] Department of Statistics, The Ohio State University, Columbus, Ohio, U.S.A., hans@stat.osu.edu

[§] Department of Statistics, The Ohio State University, Columbus, Ohio, U.S.A., peruggia@stat.osu.edu

DOI: 10.1214/15-BA942B

Focusing on the technically simpler situations, such as that of the univariate Gaussian process of Section 6.1, may be helpful to develop some deeper intuition. The addenda in the cumulative prequential delta log-scores of (7) in the article are given by

$$S_{L,i}(x_i, Q_i) - S_{L,i}(x_i, P_i) = \frac{1}{2}\left[\log \sigma^2_{Q_i} - \log \sigma^2_{P_i} + (x_i - \mu_{Q_i})^2/\sigma^2_{Q_i} - (x_i - \mu_{P_i})^2/\sigma^2_{P_i}\right],$$

and the addenda in the cumulative prequential delta Hyvärinen scores are given by

$$S_{H,i}(x_i, Q_i) - S_{H,i}(x_i, P_i) = 2/\sigma^2_{P_i} - 2/\sigma^2_{Q_i} + (x_i - \mu_{Q_i})^2/\sigma^4_{Q_i} - (x_i - \mu_{P_i})^2/\sigma^4_{P_i},$$

where $(\mu_{P_i}, \sigma^2_{P_i})$ and $(\mu_{Q_i}, \sigma^2_{Q_i})$ are the conditional means and variances of x_i given $\mathbf{x}^{i-1}$ (all the observations preceding x_i), under models P and Q, respectively.

Note that, for the case of a covariance stationary Gaussian process, $\sigma^2_{P_i}$ and $\sigma^2_{Q_i}$ are constant in i. As a consequence, the cumulative prequential delta scores based on the Hyvärinen rule and the log-score are perfectly linearly related whenever $\sigma^2_{P_i} = \sigma^2_{Q_i} = \tau^2$. As an example, this is the case for two iid sequences with possibly different means and equal variances and for two zero-mean, AR(1) sequences with possibly different autoregressive parameters and equal innovation variances. The delta scores are also perfectly linearly related if the two covariance stationary Gaussian process have equal conditional means $\mu_{P_i} = \mu_{Q_i}$ and possibly different conditional variances $\sigma^2_{P_i} = \tau^2_P$ and $\sigma^2_{Q_i} = \tau^2_Q$. As an example, this is the case for two iid sequences with equal means and possibly different variances and for two zero-mean, AR(1) sequences with equal autoregressive parameters and possibly different innovation variances.

For Gaussian processes with non-stationary covariance structure, the prequential delta scores based on the Hyvärinen rule and on the log-score may not be perfectly linearly related. Is it then possible to characterize with necessary and sufficient conditions the Gaussian processes for which the two delta scores are perfectly linearly related? When the delta scores are not perfectly linearly related, how do they differ both in a finite-sample and an asymptotic sense?

Regardless of whether the delta scores are or are not perfectly linearly related, there remains the question of how model comparison decisions based on the two scores differ. To address this issue, we look at comparisons between two models and conform to the recommendation made by the authors in Section 4, which is to select the model with the lower prequential score. When using the log-score, this corresponds to using the Bayes decision rule under 0–1 loss and assuming equal prior probabilities for the two models. When using the Hyvärinen score, there does not appear to be any principled way to justify the use of the zero cut-off for the difference in prequential scores, although, if the delta log-score and the delta Hyvärinen scores are perfectly linearly related, such a cut-off for the delta Hyvärinen score can be readily made to correspond to infinitely many Bayes rules under generalized 0–1 loss for an appropriate choice of prior model probabilities.

An inspection of the expressions for $S_{L,i}(x_i, Q_i) - S_{L,i}(x_i, P_i)$ and $S_{H,i}(x_i, Q_i) - S_{H,i}(x_i, P_i)$ reveals that the squared departures of the observations from their conditional means are normalized by the conditional variance in the log-score and by the

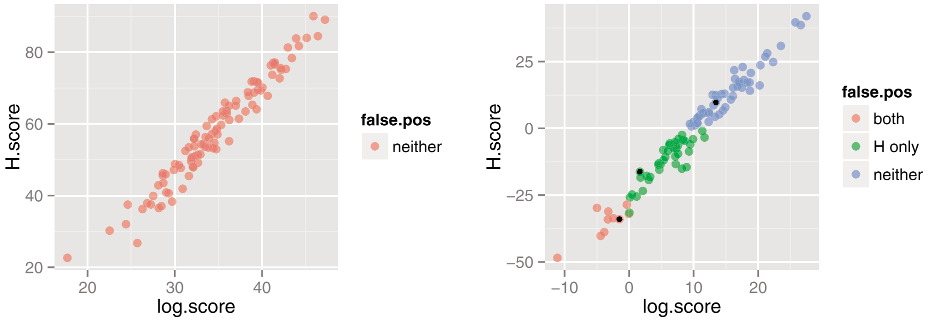

Figure 1: Cumulative prequential delta Hyvärinen scores vs. delta log-scores for 100 simulated data sets. False positive identifications of Q as the data generating model are highlighted by color. The three points in the right panel plotted with a black center correspond to the three time series in Figure 2.

squared conditional variance in the Hyvärinen score. Beside the unnatural fact that the normalized terms are no longer unitless, this suggests that the delta Hyvärinen score may be more sensitive than the delta log-score to the presence of outlying observations when the alternative model has larger variance than the model generating the data.

This point is illustrated in Figure 1, which is based on 100 simulated data sets of size 101 from a zero-mean Gaussian AR(1) process P with autoregressive parameter ϕ equal to 0.5 and innovation variance equal to 1. The alternative model, Q, is taken to be a zero-mean Gaussian AR(1) process with autoregressive parameter ϕ equal to 0.1 and innovation variance equal to 4. The prequential delta log-scores and delta Hyvärinen scores are built based on the conditional distributions of observations 2 through 101. These distributions are Gaussian with mean equal to ϕ times the preceding observation and variance equal to the innovation variance.

For each simulated data set, we calculate the cumulative prequential delta scores

$$\Delta_L^{101}(\mathbf{x}^{101}; P, Q) = \sum_{i=2}^{101}(S_{L,i}(x_i, Q_i) - S_{L,i}(x_i, P_i))$$

and

$$\Delta_H^{101}(\mathbf{x}^{101}; P, Q) = \sum_{i=2}^{101}(S_{H,i}(x_i, Q_i) - S_{H,i}(x_i, P_i)).$$

Correct identification of the data generating model under score $*$ corresponds to

$$\Delta_*^{101}(\mathbf{x}^{101}; P, Q) > 0.$$

The left panel of Figure 1 shows that both the delta log-score and the delta Hyvärinen score correctly identify model P as the data generating model in all 100 simulations. The right panel corresponds to the same simulated data sets with the exception that,

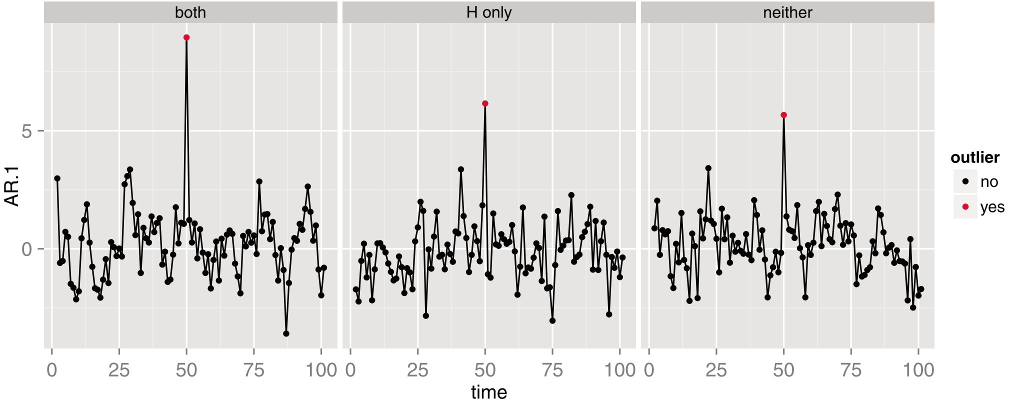

Figure 2: Three sample contaminated times series (with the outlier depicted in red) and their misclassification status according to the cumulative prequential delta Hyvärinen scores and delta log-scores. These three series correspond to the points plotted with a black center in the right panel of Figure 1.

in each data set, the 50th observation in the sequence of 101 is contaminated by adding 7 to it, making the observation an additive outlier. The figure shows that the delta Hyvärinen score is much more sensitive to the presence of the additive outlier. In 10 out of 100 cases both methods incorrectly select model Q, in 50 cases they both correctly select model P, but there are 40 cases in which only the method based on the delta Hyvärinen score incorrectly selects model Q.

Figure 2 displays three sample contaminated times series (with the outlier depicted in red) that were analyzed in the simulations. The first series is misclassified by both delta scores, the second is misclassified by the delta Hyvärinen score only, and the third is correctly classified by both delta scores. These three series correspond to the points plotted with a black center in the right panel of Figure 1.

The normalization by the square of the variance (or an estimate of the variance) appears throughout the article (cf. (34), (43), and (70)) leading one to suspect that in all these situations the Hyvärinen score may similarly be impacted by the presence of outliers. While the dependencies in the data may have played some role in our simulation, we are convinced that the variance normalization is the main issue. In fact, we were able to simulate examples with similar features after setting ϕ equal to zero in both processes, thus eschewing the effect of serial correlations. Such a choice, however, causes the delta scores to be perfectly linearly related and makes the figures harder to decipher due to overplotting, which is why we presented the simulation based on correlated data instead.

Our simulation, following the set-up of Section 6.1, is based solely on a comparison of the likelihoods for the two models. However, it is reasonable to conjecture that the sensitivity of the Hyvärinen score to the presence of outliers will be injected, via the likelihood, also when Bayesian model comparisons are carried out, irrespective of the type of prior distribution specified for the model parameters (proper, improper, subjective, or objective, as the case might be). Related questions are as follows. Is it possible to modify the prequential Hyvärinen score so as to alleviate its sensitivity to the presence of outliers? How does the method behave in the face of other model violations? Are other model comparison methods based on different proper scoring rules not as sensitive to the presence of outliers?

In summary, the authors have proposed an interesting method for performing Bayesian model selection when improper priors are used for within-model parameters by replacing the log marginal likelihood with a proper scoring rule. The method avoids the machinations associated with several of the alternative approaches that the authors mention toward the end of Section 2 at the expense of moving even farther away from the formal Bayesian paradigm. The authors justify their approach in part by proving consistency for model selection in certain settings.

While the paper provides a framework for approaching the problem, important choices still must be made in order to implement the strategy, both with proper and improper priors. We have seen that these choices can have a substantial impact on the finite-sample properties of the methods. Our investigation was limited to the Hyvärinen score, as this is the score most thoroughly discussed in the paper. The authors note that they "confined attention to the most basic homogeneous rule, the Hyvärinen score" for simplicity, but that "there are no clear theoretical grounds for preferring one [homogeneous scoring rule] over another." In light of our investigation above, we wonder whether some theoretical progress might be made by identifying a limited set of properties that might be of interest (e.g., scale invariance, robustness to model violations, etc.) and identifying classes of scoring rules and variants of the prequential score that perform appropriately with respect to one or all of these considerations. We believe that further research in this direction would give the framework a stronger theoretical footing and provide guidance to practitioners who wish to use the methods.

Bayesian Analysis (2015) **10**, Number 2, pp. 511–515

Comment on Article by Dawid and Musio*

C. Grazian[†], I. Masiani[‡], and C. P. Robert[§]

Abstract. This note is a discussion of the article "Bayesian model selection based on proper scoring rules" by A. P. Dawid and M. Musio, to appear in *Bayesian Analysis*. While appreciating the concepts behind the use of proper scoring rules, we point out here some possible practical difficulties with the advocated approach.

Keywords: Bayesian model choice, proper scoring rules, Bayes factor.

The[1] frustrating issue of Bayesian model selection preventing improper priors (DeGroot, 1982) and hence most objective Bayes approaches has been a major impediment to the development of Bayesian statistics in practice (Marin and Robert, 2007), as the failure to provide a "reference" answer is an easy entry for critics who point out the strong dependence of posterior probabilities on prior assumptions. This was presumably not forecasted by the originator of the Bayes factor, Harold Jeffreys, who customarily and informally used improper priors on nuisance parameters in his construction of Bayes factors (Robert et al., 2009). (The expansion (4) in the paper, while worth recalling, is unlikely to convince such critics.) It is therefore a very welcome item of news that a truly Bayesian approach can allow for improper priors.

As also pointed out in the paper, there exist a wide range of "objective Bayes" solutions in the literature (Robert, 2001), all provided with validating arguments of sorts, but this range by itself implies that such solutions are doomed in that they cannot agree for a given dataset and a given prior.

Finding a criterion that does not depend on the normalising constant of the predictive possibly is the unravelling key to handle improper priors, and we congratulate the authors for this finding of the Hyvärinen score and related proper scoring rules. Some difficulties deriving from the use of improper prior distributions in model choice may be solved by applying the approach proposed in the paper. There are nonetheless some issues with this solution:

- (Calibration difficulties) Once the score value is computed, the calibration of its strength very loosely relates to a loss function, hence makes decision in favour of a model difficult;
- (A clear dependence on parameterisation) Changing x into the transform $\mathfrak{h}(x)$ produces a different score;

[*] Main article DOI: 10.1214/15-BA942.

[†] Università La Sapienza–Roma, Université Paris-Dauphine

[‡] Università La Sapienza–Roma, Université Paris-Dauphine

[§] Université Paris-Dauphine, University of Warwick, xian@ceremade.dauphine.fr

[1] This work is based on a Master research thesis written by the second author under the supervision of the first and third authors, at Université Paris-Dauphine. The first author is a PhD candidate at Università La Sapienza–Roma and Université Paris-Dauphine.

DOI: 10.1214/15-BA942C

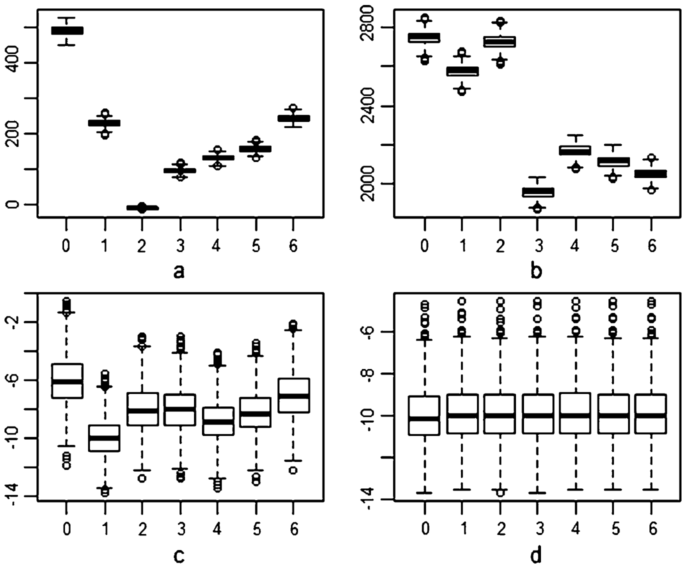

Figure 1: Boxplots over 1,000 simulations of the sample distributions of the scores of seven models under analysis, depending on the true model. Model selection is performed in the case of nested linear normal models. The data was simulated from one of seven nested linear models with up to six covariates. The design matrix is denoted by $\mathbf{X}$. While M_0 is the model that uses zero covariate, M_1 to M_6 use the first, the first two, up to all of the covariates. The values of the covariates were simulated from normal proposals, except for the first column, made of 1's. The data $\mathbf{y} = (y_1, \dots, y_n)$ have distribution $\mathbf{y}|\boldsymbol{\theta} \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma^2)$, with $n = 100$ and $\sigma^2 = 10$. In (a) the true model used for the generations is M_2 (which considers a single regressor), in (b) it is M_3 which considers the first two regressors, in (c) it is M_1 which considers only the constant, while in (d) the correct model is M_0 which has no parameter.

- (A dependence on the dominating measure) As exhibited in the case of exponential families and (30), changing the dominating measure modifies the score function;
- The arbitrariness of the Hyvärinen score, which is indeed independent of the normalising constant, but offers limited arguments in favour of this particular combination of derivatives. Since there exists an immense range of possible score functions, a stronger connection with inferential properties is a clear requirement;
- As noted above, consistency is not a highly compelling argument for the layperson, as it does not help in the calibration and selection of the score. Having an inconsistent multivariate score, while the prequential score remains consistent, is highlighting this difficulty.

Furthermore, the only application of the method presented in the paper is within the setting of the Normal linear model, and we worry that the approach may not be easily

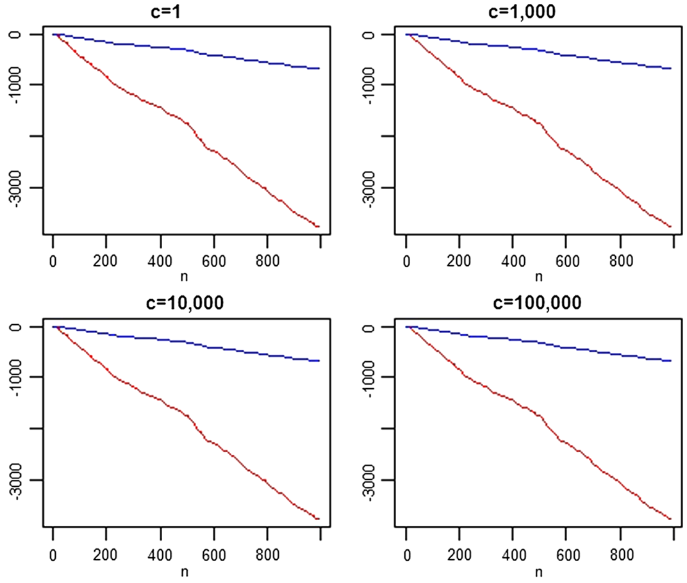

Figure 2: (Linear model) Log-Bayes factor (red) and difference of the scores (blue) as a function of an increasing sample size $n = 1, \ldots, 1000$, and of the prior variance on θ, $V = c\sigma^2$, where $\sigma^2 = 10$ is known. Given simulated data $\mathbf{y} = (y_1, \ldots, y_n)$ with conditional distribution $\mathbf{y}|\theta \sim N(\mathbf{X}\theta, \sigma^2)$, we consider one regressor and two possible models for generating the data: $M_0 : \theta = 0$ and $M_1 : \theta = 1$ when the true model is M_1.

extended to other types of models. In particular, the representation of the precision matrix of the marginal distribution in (33), based on the Woodbury matrix inversion lemma, is essential to easily apply the proper scoring rule approach to model choice with an improper prior, given that an improper prior may then be seen as a limiting version of a conjugate prior and its influence disappears in the following computations. However, the approach overcomes the singularity of the precision matrix of the marginal distribution.

We first performed some simulation studies when applying the proposed method to models that differ from the Normal linear model. When choosing between two different models with no covariates, we observed that the proposed approach can perform well as, for instance, when a Gamma model is opposed to a Normal model (well in the sense of comparing with a standard Bayes factor). However, when a Pareto distribution and a Normal distribution are compared, the approach does not often select the right model when data are generated from the Pareto distribution, while the Bayes factor always yields the right model. In addition, we came to the realisation that the method based on the Hyvärinen scoring rule may not be applied to some models, for example, when data come from a Laplace distribution, which is not differentiable at 0, or for discrete models.

Our simulation studies have also covered linear models, both nested and non-nested. The details of the simulation models are given in the captions of Figures 1–3. The

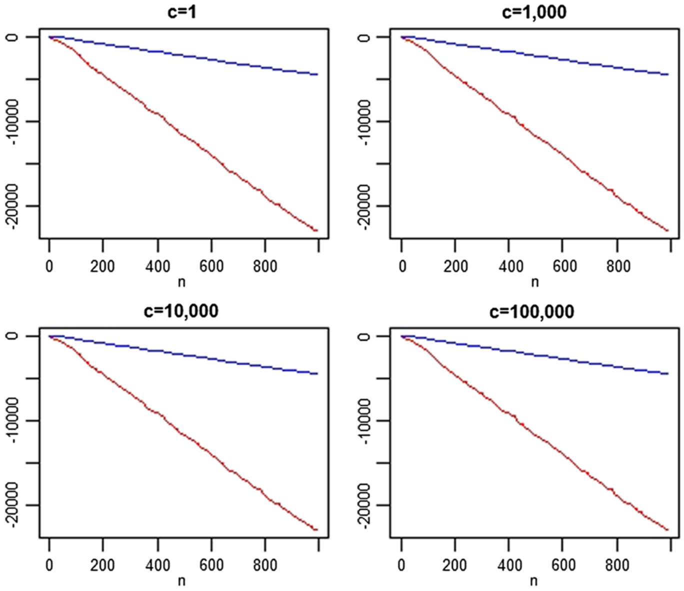

Figure 3: (Nested models) Log-Bayes factor (red) and difference of the scores (blue) as a function of an increasing sample size $n = 1, \dots, 1000$, and of the prior variance on θ, $V = c\sigma^2$, where $\sigma^2 = 10$. The setting is the same as Figure 1, where we consider six possible regressors and we compare model M_3 which considers the first three regressors against model M_6 which considers all the regressors (M_3 is the true model in our simulations).

performance of the multivariate Hyvärinen score when comparing Normal linear models is excellent, as shown in Figure 1, even when using an improper prior, provided the sample size is larger than the number of parameters in the model. Following repeated simulations, we observed that the method is always able to choose the right model. We, however, noticed that, when the true model does not involved covariates, the ability of the method to discriminate between models is reduced. Although this approach shows a consistent behaviour and chooses the right model with higher and higher certainty when the sample size increases, our simulations have also shown that the log-proper scoring rule tends to infinity more slowly than the Bayes factor or than the likelihood ratio. It is approximately four times slower, all priors being equal, as shown in Figures 2 and 3, which represent the comparison between the approach based on the log-Bayes factor and the one based on the difference between the score functions for the case of linear models, both nested (Figure 3) and non-nested (Figure 2).

As a final remark, we would like to point out the alternative and recent proposal of Kamary et al. (2014) for correctly handling partly improper priors in testing settings through the tool of mixture modelling, each model under comparison corresponding to a component of the mixture distribution. Testing is then handled as an estimation problem in an encompassing model. Therein, the authors show consistency in a wide range of situations. We currently appreciate the approach through mixture estimation as the most compelling for the many reasons advanced in Kamary et al. (2014), in particular because the posterior distribution of the weight of a model is easily interpretable and scalable towards selecting this very model or an alternative one. Furthermore, it returns posterior probabilities for the models under comparison without the need to resort to specific prior probability weights.

References

DeGroot, M. (1982). "Discussion of Shafer's 'Lindley's paradox'." *Journal of the American Statistical Association*, 378: 337–339. 511

Kamary, K., Mengersen, K., Robert, C., and Rousseau, J. (2014). "Testing hypotheses as a mixture estimation model." arXiv:1214.2044. 514, 515

Marin, J. and Robert, C. (2007). *Bayesian Core*. Springer-Verlag, New York. MR2723361. 511

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, second edition. MR1835885. 511

Robert, C., Chopin, N., and Rousseau, J. (2009). "Theory of Probability revisited (with discussion)." *Statistical Science*, 24(2): 141–172 and 191–194. 511

Bayesian Analysis (2015) **10**, Number 2, pp. 517–521

Rejoinder*

A. Philip Dawid[†] and Monica Musio[‡]

Abstract. We are deeply appreciative of the initiative of the editor, Marina Vanucci, in commissioning a discussion of our paper, and extremely grateful to all the discussants for their insightful and thought-provoking comments. We respond to the discussions in alphabetical order.

Keywords: consistent model selection, homogeneous score, Hyvärinen score, prequential.

Grazian, Masiani and Robert

Clara Grazian, Ilaria Masiani and Christian Robert (henceforth GMR) point to a number of potential difficulties in our approach.

Calibration We are not sure what GMR mean by the expression "very loosely relates to a loss function." A proper scoring rule $S(x, Q)$ is very strictly a loss function, where the state is the value x of X, and the decision is the quoted distribution Q for X. Moreover (see, for example, Dawid (1986)), given an essentially arbitrary decision problem, with state-space $\mathcal{X}$, decision space $\mathcal{A}$, and loss function $L(x, a)$, we can define $S(x, Q) := L(x, a_Q)$, where a_Q denotes a Bayes act with respect to the distribution Q for X; and this is readily seen to be a proper scoring rule. That is, essentially every decision problem is equivalent to one based on a proper scoring rule. If you take some specified decision problem seriously, you should use the associated proper scoring rule. There is then no problem of calibration.

Dependence on parametrisation GMR are correct in noting that, if we apply a scoring rule after first transforming the state space, we will generally get a non-equivalent result (the log-score is essentially the only exception to this.) However, there will be a new scoring rule for the transformed problem that is equivalent to the original rule for the original problem; see Parry et al. (2012, Section 11) for how a homogeneous score such as that of Hyvärinen transforms. We cannot give any definitive guidance on how to choose an appropriate transformation, though Example 11.1 of the above-mentioned paper suggests that some consideration of boundary conditions may be relevant.

Dependence on dominating measure This is not the case: when constructing the Hyvärinen (or other homogeneous) score, the formula is to be applied to the density with respect to Lebesgue measure.

*Main article DOI: 10.1214/15-BA942.

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

[‡] University of Cagliari, mmusio@unica.it

DOI: 10.1214/15-BA942REJ

Arbitrariness There is indeed a very wide variety of homogeneous proper scoring rules, any one of which will achieve our aim of eliminating the problematic normalising constant. At this point we can do little more than reiterate what we said towards the end of Section 3 of our paper.

Consistency Whether or not a person, lay or otherwise, finds consistency a compelling desideratum is probably a very personal matter. We do find it so. In a related point, we do not see why, in their first paragraph, GMR dismiss the implications of our expansion (4) so uncritically. Indeed, the near identity of the red lines in the four subplots of their own Figure 2, which correspond to very different prior variances, lends support to our conclusion, from (4), that "the dependence of the Bayes factor on the within-model prior specifications is typically negligible."

GMR correctly point out that there are continuous distributions, such as the Laplace distribution, to which we cannot apply the Hyvärinen (or other homogeneous) score. This point deserves further attention. But for discrete models there is a different class of homogeneous proper scoring rules that are appropriate and can be used to the same end of eliminating the normalising constant; see Dawid et al. (2012).

GMR's simulation studies are interesting. In contrast to our own analysis, they appear to show consistency of model selection based on the multivariate version of the Hyvärinen score. We should not complain if our method behaves even better than expected, but we confess we find this puzzling. We must also take issue with their assertion that "the log proper scoring rule tends to infinity [approximately four times] more slowly than the Bayes factor or than the likelihood ratio." It is simply not appropriate to compare absolute values across different scoring rules, since each can be rescaled by an arbitrary positive factor without any consequence for model comparison.

GMR point to the alternative approach of Kamary et al. (2014). However, it seems to us that the part of that paper that relates to handling improper priors could just as readily be applied directly to the Bayes factor. For example, if we are comparing two location models, we might use the identical improper prior (with the identical value for its arbitrary scale factor) for the location parameter in both. Then this scale factor will cancel out in the Bayes factor, so leading to an unambiguous answer. But in any case, this approach is not available unless there are parameters in common between all the models being compared. Our own approach has no such constraint.

Hans and Perrugia

Christopher Hans and Mario Perrugia (HP) only consider "models" without any unknown parameters, so do not directly address our main concern, which was to devise methods for comparing parametric models having possibly improper prior distributions.

They focus on two main issues:

1. Comparisons between the Hyvärinen score and the log-score.
2. Robustness to outliers.

With regard to point 1, HP consider in particular cases where the two scores are linearly related. While we fail to see why this property should be of any fundamental importance (and will pass up their invitation to characterise it), it is worthy of some attention. We do note, however, that, in their analysis of a general covariance stationary Gaussian process, HP err when they say "$\sigma^2_{P_i}$ and $\sigma^2_{Q_i}$ are constant in i." Recall that $\sigma^2_{P_i}$ is not the unconditional variance of X_i under P, but its conditional variance, given $(X_1, \ldots, X_{i-1})$. Their asserted constancy property will hold for an AR(p) process only for $i > p$; while for a general process it will fail, although a limiting value will typically exist.

HP's specific applications do have this constancy property (at least for $i > 1$). In the case they consider of different means and equal variances, the Hyvärinen incremental delta score is just a constant multiple of that for the log-score, and this property extends to the cumulative scores. Since an overall positive scale factor is irrelevant, the two scores are essentially equivalent in this case.

For the other case HP consider, of equal means and different variances, even after rescaling the incremental delta scores will differ by an additive constant, c say. The cumulative scores, to time n, will thus differ by nc, which tends to infinity—an effect that might seem to jeopardise the consistency analysis in our paper. However, the following analysis shows that this is not so. Using HP's formulae, and setting $\xi = \tau^2_P/\tau^2_Q$, consider first the log-score. The incremental delta log-scores are, under P, independent and identically distributed, with expectation $\frac{1}{2}(\xi - 1 - \log \xi) > 0$ and finite variance, so that the difference between the cumulative prequential score for Q and that for P tends to infinity almost surely—so favouring the true model P. Likewise Q will be favoured when it is true. Now consider the Hyvärinen score. Again the incremental delta log-scores under P are independent and identically distributed with finite variance, now with expectation $\tau_q^{-2}(\xi + \xi^{-1} - 2) > 0$; so once again, the true model is consistently favoured.

HP ask whether there is any principled reason for applying the cut-off value 0 to the difference in prequential scores. Well, it seems natural to us to choose the model whose predictions have performed best so far, so indicating that this might continue into the future—although, as the advertisers of financial products are obliged to point out, past success cannot be taken as an infallible guide to future performance. We further note the essential equivalence of this recipe to the machine learning technique of "empirical risk minimisation" in Statistical Learning Theory, which has developed an extensive theory, extending well beyond the case of parametric models, characterising when this will be effective; see Rakhlin et al. (2015); Rakhlin and Sridharan (2015) for application to the general case of dependent sequential observations.

In any case, should one wish to use a cut-off different from 0, there is no impediment to doing so—this would not affect the consistency properties we have investigated, which only rely on the difference of cumulative scores tending to infinity. How the choice of cut-off could relate to differential prior probabilities and utilities is a topic that deserves further consideration.

Turning to HP's point 2, their simulations appear to show that the Hyvärinen score is less robust to additive outliers than the log-score (though we note that in their example the outlier only affects 2 of the 100 summands of the overall score.) Issues of the robustness of minimum score inference have been considered by Dawid et al. (2015), where it is shown that (in an estimation context) certain proper scoring rules do enjoy good robustness properties (generally better than straightforward likelihood). However, these do not include the Hyvärinen score or other homogeneous scores. Thus there may indeed be a conflict between the aim of our current paper, which is to overcome problems associated with improper distributions, and the very different aim of protecting against outliers.

Katzfuss and Bhattacharya

Matthias Katzfuss and Anirban Bhattacharya (KB) are particularly concerned with the question of whether our approach can be tweaked to yield a "pseudo-Bayes factor", where a general score takes the place of log-likelihood. While it would be very nice if this were so, we are a little dubious. As KB point out, there are serious problems related to the arbitrary scaling of a general score. These are compounded when, as for the homogenous cases we consider, the score is a dimensioned quantity. Thus if the basic observable X has the dimension of length, L, then the Hyvärinen score has dimension L^{-2}, so any scale factor, such as λ in their (1.1) or (3.1), would have to have dimension L^2. Otherwise put, whether we are measuring X in nanometers or in parsecs will affect the absolute value of the score (though not the comparisons that form the basis of our method).

There is no reason why our method should not be used to compare a finite number of models, rather than just 2. However, when the number is countably infinite, or grows with sample size, even likelihood-based model selection can fail to be consistent. In that case the problem can sometimes be solved by regularisation, essentially equivalent to introducing prior probabilities over models and selecting on the basis of the posterior model probabilities. Perhaps some analogue of this device might work for more general proper scoring rules.

References

Dawid, A. P. (1986). "Probability Forecasting." In: S. Kotz, N. L. Johnson, and C. B. Read (eds.), *Encyclopedia of Statistical Sciences*, volume 7, 210–218. New York: Wiley-Interscience. MR0892738. 517

Dawid, A. P., Lauritzen, S., and Parry, M. (2012). "Proper Local Scoring Rules on Discrete Sample Spaces." *The Annals of Statistics*, 40: 593–608. MR3014318. doi: http://dx.doi.org/10.1214/12-AOS972. 518

Dawid, A. P., Musio, M., and Ventura, L. (2015). "Minimum Scoring Rule Inference." *Scandinavian Journal of Statistics*, submitted for publication. arXiv:1403.3920 520

Kamary, K., Mengersen, K., Robert, C., and Rousseau, J. (2014). "Testing Hypotheses as a Mixture Estimation Model." arXiv:1412.2044 518

Parry, M. F., Dawid, A. P., and Lauritzen, S. L. (2012). "Proper Local Scoring Rules." *The Annals of Statistics*, 40: 561–92. MR3014317. doi: http://dx.doi.org/10.1214/12-AOS971. 517

Rakhlin, A. and Sridharan, K. (2015). "On Martingale Extensions of Vapnik–Chervonenkis Theory with Applications to Online Learning." In: V. Vovk, H. Papadopoulos, and A. Gammerman (eds.) *Measures of Complexity: Festschrift in Honor of Alexey Chervonenkis*, Chapter 15. Heidelberg: Springer-Verlag, in press. http://www-stat.wharton.upenn.edu/~rakhlin/papers/chervonenkis_chapter.pdf 519

Rakhlin, A., Sridharan, K., and Tewari, A. (2015). "Sequential Complexities and Uniform Martingale Laws of Large Numbers." *Probability Theory and Related Fields*, 161: 111–153. MR3304748. doi: http://dx.doi.org/10.1007/s00440-013-0545-5. 519