Context-dependent score based Bayesian information criteria

N.T. Underhill * and J.Q. Smith †

Abstract. In a number of applications, we argue that standard Bayes factor model comparison and selection may be inappropriate for decision making under specific, utility-based, criteria. It has been suggested that the use of scoring rules in this context allows greater flexibility: scores can be customised to a client's utility and model selection can proceed on the basis of the highest scoring model. We argue here that the approach of comparing the cumulative scores of competing models is not ideal because it tends to ignore a model's ability to 'catch up' through parameter learning. An alternative approach of selecting a model on its maximum posterior score based on a plug in or posterior expected value is problematic in that it uses the data twice in estimation and evaluation. We therefore introduce a new Bayesian posterior score information criterion (BPSIC), which is a generalisation of the Bayesian predictive information criterion proposed by Ando (2007). This allows the analyst both to tailor an appropriate scoring function to the needs of the ultimate decision maker and to correct appropriately for bias in using the data on a posterior basis to revise parameter estimates. We show that this criterion can provide a convenient method of initial model comparison when the number of models under consideration is large or when computational burdens are high. We illustrate the new methods with simulated examples and real data from the UK electricity imbalance market.

Keywords: Scoring rules, Bayesian model selection, Information criteria, Utility based model selection

1 Introduction

In practice, model selection is often undertaken in order to make sound predictions of future values of one or more specific quantities of interest. An analyst may be concerned with predicting certain marginals, conditional relationships or other quantities (for example, quantiles), but have relatively little interest in the overall performance of models across the full joint distribution

Consider a high dimensional Gaussian network, which may have been built to exploit a number of promising covariates or to allow the researcher to build up a plausible set of connections between components hierarchically. For example, suppose the analyst is ultimately interested in the dependence of demand, D on price, P. Rather than modelling this directly (which might require the elicitation of complex priors and likelihood specifications beyond the experience of the analyst or subject matter experts) she con-

*University of Warwick n.t.underhill@warwick.ac.uk
†University of Warwick j.q.smith@warwick.ac.uk

siders that is more appropriate to build up a series of intermediate dependencies which are more readily modelled, understood and elicited. The hope is that the relationship of interest can be established more accurately than through a more direct approach.

However, many standard model selection approaches, for example, Bayes factors, are concerned with the performance of the model across the full joint distribution rather than the subset of interest. In our example, the analyst is interested in choosing between models based on their performance in predicting demand based on a change in price, rather than including performance of the forecasts of intermediate variables introduced solely for modelling convenience.

It is well known that the fact that the models being evaluated typically capture the data generating process only in an approximate way presents challenges to the procedure for model selection. Many standard model selection approaches, for example, those using Bayes factors, assume an underlying *M-closed* context (Bernardo and Smith (1994)), in which one of the models under consideration is asserted to be true. In all but the simplest cases, this assumption is dubious. More usually, we operate in an *M-open* context, in which the class of fitted models is simply a convenient proxy for an unknown true model.

Once we accept that the model class under consideration cannot usually be guaranteed to correspond in its entirety to a comprehensive and exact representation of the modeller's belief about what might unfold, the focus immediately shifts to identifying *which aspects* of the model performance are most important to the end user. Rather than seeking to provide a complete and faithful portrayal or explanation of the underlying physical, causative processes, the analyst aims at releasing to her clients a model which is 'the best available and good enough' ('requisite' to use the terminology of Phillips (1982)) to enable predictions to be made within acceptable bounds, where acceptability is defined with reference to the end user's utility function.

With growing access to larger datasets, and the associated interest in 'big data' and 'big models', we argue that such approximations will become more common and necessary. One reason for this is the appetite to exploit large datasets in the absence of a clearly defined structure *a priori*. For example, unsupervised learning (Gharamani (2004)) on a large but incomplete data set provides opportunities to discover new predictive covariates, but also presents challenges in that the search for the best model is likely to be pragmatically, rather than theoretically motivated. Even in those cases where the modeller has access to greater insight into the underlying data structure and relationships, the larger the size and nature of the data set, the less feasible it is to build faithful probability structures over every aspect of the joint distribution.

The model selection decision can be formulated as a decision problem: in the Bayesian predictive context which we consider, Bernardo and Smith (1994) develop this approach. The generic formulation of the problem, as given in Vehtari and Ojanen (2012), is to choose the model M_k which maximises the expected utility:

$$\bar{U}(M_k, \hat{a}_k) = \int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y} \mid D, M_\star) d\tilde{y},$$

where $\hat{a}_k$ is the prediction which would maximise the expected utility if model M_k were to generate future data, and where we denote the true, but unknown, data generating mechanism for the future observations y by $M_\star$. In the example where our principal interest is in the conditional relationship between demand and price, an appropriate utility might be the logarithmic conditional density $f(D \mid P)$.

The major difficulty is in approximating this expectation, as the true data generating process is typically unknown. We could 'score' the model on the basis of the realised utility of its historical one-step ahead forecast – a 'prequential' approach (Dawid (1984)). One disadvantage of this approach is that early forecasting performance may not be a particularly relevant indicator of future performance, particularly where the model has improved through parameter learning (the 'catch up' effect (van Erven et al. (2012)).

An alternative approach is cross-validation, in which a validation set acts as a proxy for future observations, and the expected utility is estimated based on assessing models built from the remaining data. However, this lacks a formal Bayesian foundation and can be computationally intensive and problematic for certain types of model (for example, time series).

This paper introduces a new approach in the form of a context dependent Bayesian information criterion – the *Bayesian Posterior Score Information Criterion (BPSIC)* which is based on a bias corrected posterior predictive expected score which can be tailored to the utility of a model user.

In Section 2 we provide background to common approaches to utility based model selection; Section 3 and the appendix contain the motivation for and derivation of the proposed criterion. Section 4 illustrates the application of the criterion to some stylised examples, together with analysis of situations in which the model is correctly and incorrectly specified under a variety of scoring rules. We also illustrate the advantages of the posterior estimated score as opposed to an assessment of cumulative historical performance when models improve performance through parameter updating in the light of incoming data. In Section 5 we apply the selection criterion to a problem involving the quantification of risk characteristics of imbalance prices in the UK electricity market.

2 Utility based model selection

Vehtari and Ojanen (2012) give a comprehensive survey of established and recent approaches in Bayesian model selection. This is framed in decision theoretic terms as choosing the model M_k which maximises the expected utility:

$$\bar{U}(M_k, \hat{a}_k) = \int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y} \mid D, M_\star) d\tilde{y}, \tag{1}$$

where $\hat{a}_k$ is the prediction which would maximise the expected utility if model M_k were to generate future data, and where we denote the true, but unknown, data generating mechanism for the future observations y by $M_\star$.

A number of appropriate utilities have been proposed in the form of *proper scoring*

rules - functions of the elicited probability density and the values which materialise (Winkler et al. (1996), Gneiting and Raftery (2007)). In the case of point predictive forecasting, Gneiting (2011) provides a comprehensive review of loss functions which target particular aspects of the forecast. For example, the requirement to forecast prudential reserves for regulatory and risk management applications require ensuring a good fit to particular distribution quantiles, for which specific quantile and expectile scoring rules have been developed (see Gneiting (2011) and below for specific examples of these).

More generally, Dawid (2007) formulates the role of proper scoring rules and their related divergences in the context of a general decision problem where we have an outcome space X, action space A, and loss function L. If P, Q are distributions over X, Dawid (2007) defines:

- Bayes act $a_P := \arg\inf_{a \in A} L(P, a)$
- Proper scoring rule $S(x, Q) := L(x, a_Q)$
- Entropy function $H(P) := S(P, P)$
- Divergence function $d(P, Q) := S(P, Q) - H(P)$,

where we denote $L(P, a) = E^P[L(X, a)]$, $S(P, Q) = E^P[L(X, Q)]$, where $X \sim P$.

For example, if we take the logarithmic density loss function $L(x, Q) = -\log Q(x)$, the Bayes act a_P is equal to P, the associated proper scoring rule is the logarithmic scoring rule $S(x, Q) = -\log Q(x)$ and the divergence corresponds to Kullback Leibler divergence. Alternatively, in the case of quadratic loss, the loss function $L(x, a) = (a - x)^2$, the Bayes act a_P corresponds to choosing μ_P, where μ_P is defined to be the mean of the distribution defined by P, and the associated proper scoring rule is $S(x, Q) = (\mu_Q - x)^2$.

Note that in the second of these examples, this scoring rule is not *strictly* proper (see, e.g. Gneiting and Raftery (2007)) in the sense that the expected score will be maximised by any distribution sharing the same mean as the the true distribution, P, and neither is it *local* (see, e.g. Bernardo and Smith (1994)) in that it depends on values of the density function of Q at points other than the observed value x. Bernardo (1979) shows that the only strictly proper local scoring rule is the logarithmic score. For the purposes here, we argue that locality is not a crucial property (although we accept that, from a more theoretical standpoint, departure from locality may be at odds with the likelihood principle - see Bernardo and Smith (1994) for a discussion of this point) – our central concern is in choosing a loss function which targets a particular utility. It is interesting to note that other authors have discarded the need for local scoring rules. For example, Musio and Dawid (2013) advocate non-local scoring rules, albeit in their case to avoid the need to calculate normalising constants.

In the M-closed case in which one of the models under consideration is, in fact, $M_\star$, then under a zero-one utility, Bernardo and Smith (1994) show that the Bayes factor

(Kass and Raftery (1995)) leads to optimal selection. But despite its popularity, in the $M-open$ context which we consider here, automatic use of the Bayes factor to perform model selection gives cause for concern. As Kadane and Dickey (1980) establish, the Bayes factor will be only optimal in those situations in which we have a true model and our utility takes the $0-1$ form.

The Bayes factor criterion that we should select model M_i in preference to M_j if $p(x \mid M_i) > p(x \mid M_j)$ can equivalently be expressed in the form

$$-\log p(x \mid M_i) < -\log p(x \mid M_j) \tag{2}$$

or, if observed data x comprises a series of observations which we can sequence $x_1, x_2, \ldots x_n$, then Dawid (1984) establishes a *prequential* formulation of the form

$$\sum_{k=1}^{n} -\log p(x_k \mid x_1, \ldots, x_{k-1}, M_i) < \sum_{k=1}^{n} -\log p(x_k \mid x_1, \ldots, x_{k-1}, M_j). \tag{3}$$

In other words, we are selecting models based on their cumulative logarithmic score. One problem with this approach is that the logarithmic score will significantly penalise models which differ in the tails of the distribution where the log predictive density is a large negative number. If a user is more interested in performance in the body of the distribution – in our example, suppose we are interested in system planning to meet typical demand levels over a range of price scenarios – then this may place too much emphasis on models which fit the tails accurately.

Suppose, for instance, that we have chosen to use a 'big model' Bayesian network which introduces of a number of 'intermediate' dependencies - through modelled variables $X = (X_1, X_2, \ldots X_n)$, where we denote the parents of X_i by $Pa(X_i)$, the demand D is represented by X_1, price P is represented by X_n, and n is large.

If our interest lies either (a) in forecasting the marginal distribution of the variable X_n, or (b) in the conditional distribution $X_n \mid X_1$, then we are interested in the predictive utility only for these elements. If we denote the true data density by P, and the modelled density by Q, this might, for example, correspond to our interest in minimising (a) the **marginal** Kullback Leibler divergence:

$$K_{X_n}(P, Q) := KL(P(X_n), Q(X_n)).$$

or (b) the **conditional** Kullback Leibler divergence

$$K_{X_n|X_1}(P, Q) := E^P[K_{X_n}(P(X_n \mid x_1), Q(X_n \mid x_1))].$$

However, Bayes factor comparison will automatically include the model's performance on all other relationships, regardless of whether they are pertinent to the decision

maker's utility. For example, in a simple case where the variables are independent, we have

$$KL(P,Q) = \sum_{i=1}^{n} K_{X_i}(P,Q),$$

or more generally we will have:

$$KL(P,Q) = \sum_{i=1}^{n} K_{X_i|Pa(X_i)}(P,Q),$$

so that the good performance of a model on the marginal or conditional dependency of interest will be disguised by poor performance on the other marginals or conditionals which are introduced by the model. If, for example, variable X_i for some $1 < i < n$ has little or no influence over the marginal of interest, but has been poorly modelled, then its introduction will have little impact on the performance of the model on the marginal of interest, but it will have a larger negative impact on the overall model score.

The exact impact of this will depend on the extent to which data contains 'unusual' observations, but for high dimensional settings we can reasonably expect outliers on *some* dimension. In big data contexts, therefore, outlying observations on aspects of the model in which we are not interested can have a distorting impact on decisions taken if we use the logarithmic score across the full joint distribution.

One solution is to choose a more appropriate scoring rule: in this case using the relevant marginal and conditional logarithmic scores might be preferable. Even if, as we will now assume for the remainder of this paper, the analyst has selected a scoring rule appropriate to her application, there remains the problem of how to compute the expected future utility in Equation 1 for each model under consideration. It is this problem to which we now turn our attention.

By analogy with the prequential representation of the Bayes factor as the cumulative log score of a series of one step ahead forecasts, we could compute a cumulative realised score as an estimate for the expected future utility. There is a question around the relevance of historical performance, as assessed via the *cumulative* score, to future performance. van Erven et al. (2012) call this the 'catch up' effect: a model can initially perform poorly - for example, a vague prior has been used to initial data to have a greater influence on parameter updating - but after a period of 'training', it may start to out-perform alternative models. Where a cumulative score is used, the score may fail to identify this change point at an early stage. We illustrate this in Section 4.

An alternative, at the other extreme, is to assess models based on their posterior predictive performance, using the data observed as a proxy for new observations. This is problematic (see the discussion in Aitkin (1991)) in introducing a bias and overfitting: by using the data twice to derive the posterior predictive distribution and

assess this distribution the model can appear to perform better than it would on a truly independent sample.

Two main approaches have been taken which attempt to compensate for these effects. Cross validation (Stone (1974), Arlot and Celisse (2010)) assesses performance by averaging predictive performance based on models built on one subset of the data, with predictive performance judged on the data 'left out' of the the model construction. Bernardo and Smith (1994), Key et al. (1999) and Vehtari (2001) are early examples of the application of cross validation to the estimation of general expected utilities, and more recently Vehtari and Ojanen (2012) have also advocated this approach, although it lacks a formal Bayesian foundation.

One practical drawback of cross-validation is that it can be computationally complex to rebuild models for multiple training samples, particularly where MCMC methods are used to obtain parameter estimates. This may rule it out for initial exploratory model comparison, in situations where a large number of candidate models are being evaluated. Although computationally less expensive variants are available, for example, where random sub-samples are used, there is a danger that they may omit important observations. This could be a particular danger when our aim is to assess the model on its ability to forecast tail quantiles of the distribution. Where smaller samples are used to form the training data, this may not fully reflect the model's ability to 'catch up' with a larger training sample.

The second approach, which we develop in this paper, is to construct an 'information criterion' which seeks to adjust directly for the bias inherent in using the same data twice. We provide a brief explanation of this method in the section below.

3 Score based information criteria

Information criteria are designed to enable model performance to be assessed retrospectively and therefore need to correct for bias in using the same data for parameter estimation and assessment. Akaike (1973) is a landmark: models are assessed on their fit in terms of Kullback Leibler divergence to a true model, while from a Bayesian standpoint, Schwarz (1978) provides the first motivation of the use of an information criterion. The criterion:

$$BIC = k \log n - 2 \log(L) \tag{4}$$

is expressed in terms of the maximised log likelihood, $\log(L)$ of the data together with an adjustment, where k is the number of parameters in the model and n the number of observations. Stone (1977) shows that AIC and leave-one-out cross-validation are asymptotically equivalent methods.

Extending this measure to encompass a wider family of utilities requires criteria which select the model with the smallest expected discrepancy (where the discrepancy is defined by the analyst). Working in a frequentist context, Linhart and Zucchini (1986) provide a substantial development. In brief, a discrepancy function $\Delta(\theta)$ is chosen

based on a consideration of the aspects of importance – popular divergences used for this purpose are Kullback Leibler divergence, Kolmogorov discrepancy and the Pearson chi-squared discrepancy – and the expected discrepancy is estimated as:

$$E^F\left[\Delta(\hat{\theta}\right] \approx E^F\left[\Delta_n(\hat{\theta}) + \mathrm{Tr}(\Omega_n^{-1}\Sigma_n)/n\right], \tag{5}$$

where $\hat{\theta} = \arg\min(\Delta_n(\theta))$, $\Delta_n(\theta)$ is the empirical discrepancy of the observed data, and Ω_n, Σ_n are estimators of the matrix $(\partial^2\Delta(\theta_0)/\partial\theta_i\partial\theta_j)$ and the covariance matrix $(\sqrt{n}\partial\Delta_n(\theta_0)/\partial\theta_i)$.

Claeskens and Hjort (2003) also explore this in a frequentist setting where the emphasis is on establishing one or more parameters or functions of parameters of a distribution through the development of the *focussed information criterion*.

From a Bayesian perspective, use of point value 'plug-in' estimators and subsequent bias correction is problematic: it fails to account for the full uncertainty expressed by the analyst's posterior distribution (see for example, the discussion in Celeux et al. (2006)). The Deviance Information Criterion (Spiegelhalter et al. (2002)) sought to address this by considering the posterior distribution of the data log likelihood. DIC is defined as

$$DIC = \bar{D} + p_D = D(\bar{\theta}) + 2p_D. \tag{6}$$

The first term, denoted by $\bar{D}$, represents the fit as defined by the posterior expectation of the deviance:

$$\bar{D}(\theta) = E_{\theta|y}\left[D(\theta)\right] = E_{\theta|y}\left[-2\log p(y \mid \theta) + 2\log f(y)\right],$$

where $f(y)$ is an arbitrary standardising term which does not impact the model comparison, and the second measures the 'complexity' of the model, defined as

$$p_D = E_{\theta|y}\left[D\right] - D(E_{\theta|y}\left[\theta\right]).$$

Vehtari (2001) suggest an extension of the DIC to cope with arbitary utilities of the form

$$\bar{U}_{DIC} = \bar{u}(E_\theta\left[\theta\right]) + 2(E_\theta\left[\bar{u}(\theta)\right] - \bar{u}(E_\theta\left[\theta\right])) \tag{7}$$

However, Ando (2007) observed that, if the function of the complexity term in DIC is to compensate for bias in the posterior estimation of the model fit, then it is incorrectly calculated. Ando (2007) introduces a Bayesian predictive information criterion (BPIC) defined as

$$BPIC = -2E_{\theta|y}\left[\log L(y \mid \theta)\right] + 2n\hat{b}_\theta, \tag{8}$$

with

$$n\hat{b}_\theta = E_{\theta|y}\left[\log(L(y \mid \theta)\pi(\theta))\right] - \log(L(y \mid \hat{\theta}_n)\pi(\hat{\theta}_n)) + \mathrm{Tr}(J_n^{-1}(\hat{\theta}_n)I_n(\hat{\theta}_n)) + p/2, \tag{9}$$

p representing the dimension of the parameter vector θ, $\hat{\theta} = \arg\max_\theta \pi(\theta \mid y)$ and I_n, J_n as defined in equation 10 below. See also Zhou (2011) for further variants based on alternative estimators.

We have previously argued that in many applications, we are interested in estimating the expected divergence of a model, where the divergence is based on the particular scoring rule which reflects the end user's utility. In other words, we seek an analogue of the BPIC which allows us to assess, in a Bayesian fashion, the posterior expected quantity $E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right]$, as a measure of discrepancy from the true data generating process for z, for our chosen scoring rule S.

In particular, suppose that we observe n observations $y = (y_1, y_2, \ldots y_n)$ and we are considering a candidate model M with probability density $f_\theta := f(y \mid \theta)$, prior density $\pi(\theta)$ and posterior density $\pi(\theta \mid y)$, where we are interested in its ability to minimise the expected divergence induced by the scoring function $S(f, z)$ for future observations z from the true data generating process. We define the cumulative score $C_S(y \mid \theta) := \sum_{k=1}^n S(f_\theta, y_k)$.

We denote by $\hat{\theta}_n$ the parameter value which maximises $n^{-1}\log(L(y \mid \theta)\pi(\theta))$, and assume that it is unique. We denote the posterior mean by $\bar{\theta}_n$, and define the matrices:

$$
\begin{aligned}
I_n(\theta) &= \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial(\log f_\theta(y_k) + \log\pi(\theta)/n)}{\partial\theta}\frac{\partial(\log f_\theta(y_k) + \log\pi(\theta)/n)}{\partial\theta^T}\right), \quad (10)\\
J_n(\theta) &= -\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial^2(\log f_\theta(y_k) + \log\pi(\theta)/n)}{\partial\theta\partial\theta^T}\right),\\
J_n^S(\theta) &= -\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial^2(S(f_\theta, y_k) + \log\pi(\theta)/n)}{\partial\theta\partial\theta^T}\right),\\
U_n^S(\theta) &= \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial(S(f_\theta, y_k) + \log\pi(\theta)/n)}{\partial\theta}\right).
\end{aligned}
$$

In Appendix 1 we generalise the proof in Ando (2007) to establish the following result:

Theorem 1 *Under the assumptions that we have a scoring rule S which is such that*

- $n\,\mathrm{cov}(U_n^S(\hat{\theta}_n), \hat{\theta}_n) \approx 0$, $n\,\mathrm{cov}(U_n^S(\hat{\theta}_n), \bar{\theta}_n) \approx 0$,
- $U_n^S(\hat{\theta}_n)$,$(\bar{\theta}_n - \hat{\theta}_n)$ *are uncorrelated,*

if we consider the bias from estimating the posterior expected score

$$b_S := E_y\left[\frac{1}{n}E_{\theta|y}\left[C_S(y \mid \theta)\right] - E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right]\right]$$

then, asymptotically, assuming that the matrix J is non-singular, we have:

$$n\hat{b}_S \approx E_{\theta|y}\left[C_S(y \mid \theta) + \log \pi(\theta)\right] - \left(C_S(y \mid \hat{\theta}_n) + \log \pi(\hat{\theta}_n)\right) \tag{11}$$

$$+\frac{1}{2}\operatorname{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)) + \operatorname{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)I_n(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)) - nU_n^S(\hat{\theta}_n)(\bar{\theta}_n - \hat{\theta}_n)$$

We therefore propose a **Bayesian Posterior Score Information Criterion** defined as:

$$BPSIC = 2E_{\theta|y}\left[C_S(y \mid \theta)\right] - 2n\hat{b}_S, \tag{12}$$

with models with higher values of the BPSIC being preferred.

Note that:

1. The additional requirements on the covariances are necessary to ensure suitable approximations to the derivative terms which are necessary as a result of asymptotic expansion around the posterior mode, which may not equal the parameter value which maximises the scoring function. In practice, many common scoring rules will satisfy this condition. For example, piecewise linear scoring rules have values of the derivative $U_n^S(\hat{\theta})$ dominated by a constant term which depends solely on the number of observations exceeding a quantile estimate, and therefore have an extremely low covariance with the level of the parameter estimate $\hat{\theta}$.

2. All the relevant quantities can be readily computed, for example, from a MCMC posterior sample, enabling its calculation to be incorporated as a standard routine in the initial evaluation of multiple models, without the need to perform the multiple estimation runs required for cross-validation.

3. Computation of $J^S(\hat{\theta}_n)$ requires that the score function should have a finite second derivative. This will not be the case at all points for some scoring functions, for example, absolute loss, or piecewise linear functions. Although, in practice, we have found that this tends not to be problematic in that the points at which the derivatives of the scoring rule do not exist will tend not to concentrate around the posterior mode, we indicate in the examples in the next section how routine modifications can prevent problems occuring.

Use of the BPSIC should readily facilitate an initial comparison of future expected utilities of models. In addition, if, for example, MCMC output is stored, then a future user of a model should be able to re-assess its performance based on an alternative scoring rule with a fairly straightforward recalculation.

In the next section, we illustrate the performance of the BPSIC with three examples based on stylised simulated data. In Section 5 we illustrate the application to the problem of predicting quantiles of UK electricity imbalance prices.

4 Simulation examples

4.1 Performance on different score functions

In this example, we compare estimates obtained using BPSIC to the actual bias. We consider different score functions and situations in which the model is correctly and incorrectly specified. This gives an insight into the performance of the BPSIC approximation in a variety of applications.

In the correctly specified scenario, we consider a normal model M_1 with unknown mean μ, known variance $\sigma^2 = 0.5^2$, and a conjugate prior $\mu \sim N(0, 0.1^2)$ We assume that the true data generating process is normally distributed with mean 0 and variance 0.5^2. In the incorrectly specified scenario, the normal model M_1 has unknown mean μ, known variance $\sigma^2 = 0.5^2$, and a conjugate prior $\mu \sim N(0, 5^2)$ and we assume that the true data generating process is normally distributed with mean 1 and variance 2^2.

We consider four score functions (where we define these as the negative of the corresponding loss function), where the score $S(f_\theta, y_k)$ results under the model expressed by the density function f at the parameter value θ, if the value y_k is observed.

- Logarithmic predictive density. We define $S(f_\theta, y_k) = \log f_\theta(y_k)$. As we have seen, maximising this score corresponds to minimising the Kullback Leibler divergence.
- Quadratic score. We define $S(f_\theta, y_k) = -(\mu(f_\theta) - y_k)^2$, where $\mu(f_\theta)$ denotes the predictive mean of the distribution with density f. Although for the purposes of the normal model example here, this reduces to a scaled version of the log density when a vague prior is chosen, when we include the more informative prior specification we have adopted under the correctly specified model scenario, it also results in a different weighting between prior and score function.
- Absolute loss. We define $S(f_\theta, y_k) = -\left|\mu(f_\theta) - y_k\right|$. We remark that there are undefined second derivatives at $\mu(f_\theta) = y_k$. Any problems encountered can be addressed by approximating this by the Huber loss function, defined as

$$S(f_\theta, y_k) = \begin{cases} -(\mu(f_\theta) - y_k)^2/2, & \text{if } |\mu(f_\theta) - y_k| \\ k(|\mu(f_\theta) - y_k| - k/2) & \text{otherwise.} \end{cases}$$

- Quantile loss This time, our focus of interest is in being able to forecast a speific quantile – perhaps for a risk management application. We select a quantile scoring rule, reflecting a focus on our ability to forecast the 0.95 quantile. A number of quantile scoring rules have been established (see Gneiting and Raftery (2007)); here we make use of the *asymmetric piecewise linear scoring function* (see Gneiting (2011) defined by

$$S(f_\theta, y_k) = (y_k - \tau(f_\theta))(\mathbb{1}(\tau(f_\theta) > y_k) - \tau)$$

where $0 < \tau < 1$ is the quantile of interest, and $\tau(f_\theta)$ denotes the value predicted by the density f_θ.

As with absolute loss, one common feature of these scoring rules is that they are piecewise linear, therefore having undefined second derivatives for some values. For the computation of the BPSIC it is possible to make an adjustment by making use of the *quantile Huber loss* proposed by Aravkin et al. (2014), which takes the form:

$$\rho_\tau(f_\theta, y_k) = \begin{cases} \tau\left|y_k - \tau(f_\theta)\right| - \dfrac{\kappa\tau^2}{2}, & \text{if } y_k - \tau(f_\theta) < -\tau\kappa \\ (1-\tau)\left|y_k - \tau(f_\theta)\right| - \dfrac{\kappa(1-\tau)^2}{2}, & \text{if } y_k - \tau(f_\theta) > (1-\tau)\kappa \\ \dfrac{1}{2\kappa}(y_k - \tau(f_\theta))^2, & \text{otherwise,} \end{cases}$$

where the value of κ is selected by the user as the threshold within which a quadratic approximation replaces the corresponding piecewise linear scoring rule.

The graphs below show the result of comparing the average BPSIC bias with the average actual bias (based on simulating future observations from the true distribution). Figure 1 shows the results in the correctly specified model case; Figure 2 illustrates the incorrectly specified case.

We observe that the bias in the first case where the prior is more informative is lower, reflecting the greater weighting given to the prior compared to the new data. The scale of the bias is dominated by the natural scale of the scores themselves. We speculate that there is an additional effect in that greater bias is likely to be seen when we use scores which are 'closer' to the logarithmic score: asymptotically, this score will be maximised under Bayesian updating. We comment on this in our conclusion.

4.2 Comparison with cross-validation

The previous example showed a reasonable fit between the estimates provided by BPSIC and the average bias across a number of loss functions. However, the practical application of the criterion will depend also on the amount of additional variance introduced through the bias correction.

In our next example, we compare the performance of the BPSIC and leave one out cross validation (LOO-CV). We use the same mis-specified example as previously, with the model M_1 having unknown mean μ, known variance $\sigma^2 = 0.5^2$, and a conjugate prior $\mu \sim N(0, 5^2)$ and the true data generating process being normally distributed with mean 1 and variance 2^2.

This time, we select the quantile scoring rule, reflecting a focus on model performance on forecasting the 0.95 quantile. We simulate 2000 scenarios in which a sample of 100

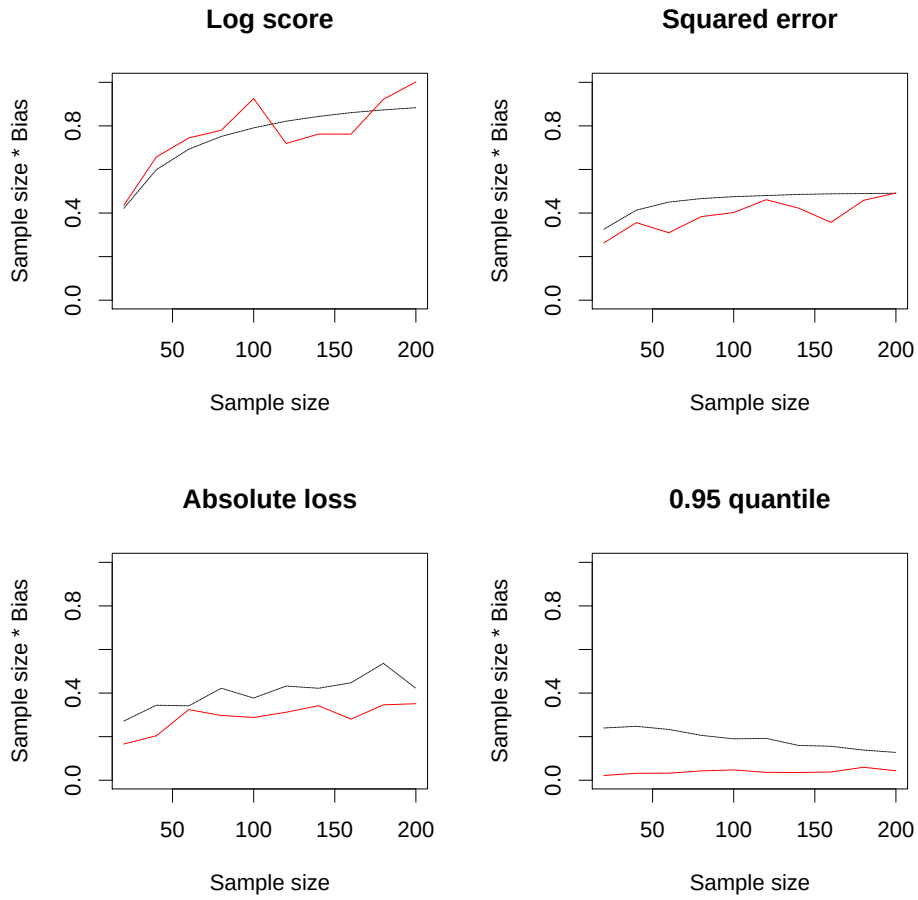

Figure 1: Performance of actual bias compared to asymptotic (BPSIC) bias. The true data generating process is given by a $N(0, 0.5^2)$ distribution. The model being assessed, M_1, is a normal distribution with unknown mean and known variance equal to the true variance. The mean has a conjugate prior $\mu \sim N(0, 0.1^2)$. The figure shows the simulated average actual bias (7,000 simulations, with computation of the relevant expectations for each simulation computed by 1,000 posterior parameter simulations), shown by the solid red line and the average asymptotic bias (dotted black line) under four different loss functions.

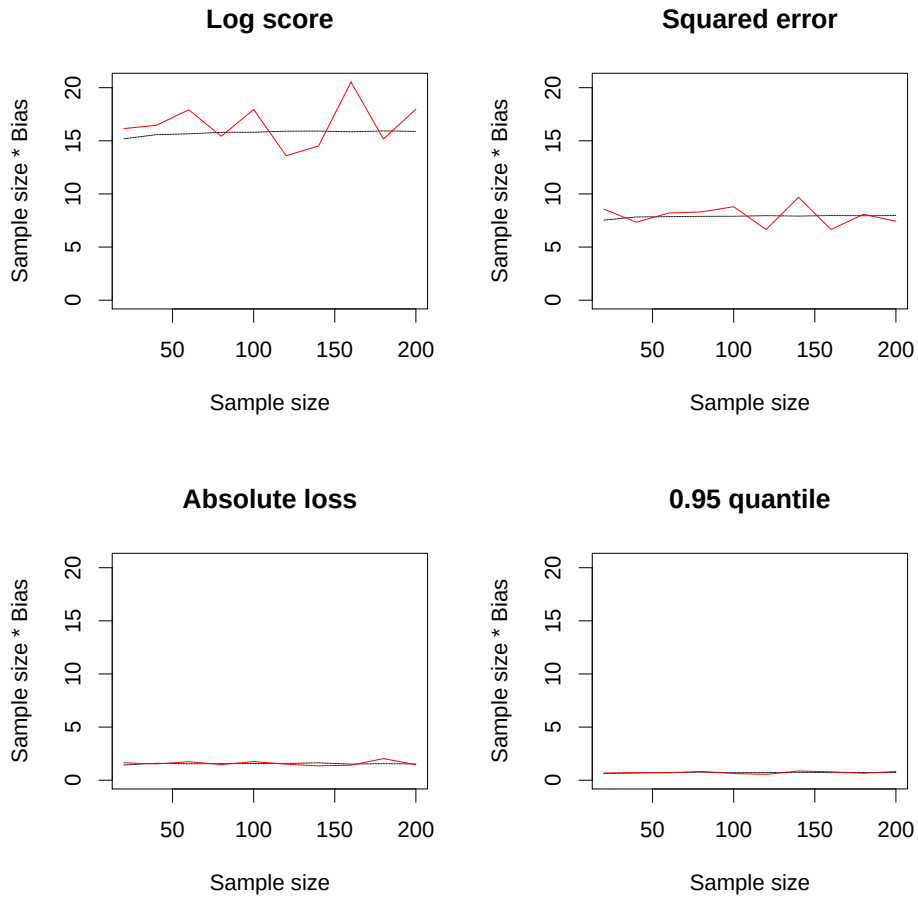

Figure 2: Performance of actual bias compared to asymptotic (BPSIC) bias - misspecified model. The true data generating process is given by a $N(1, 2^2)$ distribution. The model being assessed, M_1, is a normal distribution with unknown mean and fixed variance of 0.5^2. The mean has a conjugate prior $\mu \sim N(0, 5^2)$. The figure shows the simulated average actual bias (7,000 simulations, with computation of the relevant expectations for each simulation computed by 1,000 posterior parameter simulations), shown by the solid red line and the average asymptotic bias (dotted black line) under four different loss functions.

observations is used to generate a LOO-CV score, an (unadjusted) posterior score, the BPSIC, and simulated 'true score'. The results are shown in Figure 3.

The LOO-CV and BPSIC estimation errors are extremely close, with the unadjusted posterior score positively biased. The standard deviation of errors is almost identical between both methods. The amount of the bias correction is shown in the second graph, and we also show the standard deviations of the individual 'one left out' predictive scores which are averaged to form the LOO-CV estimate.

In the situation in which LOO-CV estimates are expensive to obtain, we might be tempted to undertake a randomised selection of a subset of samples. However, in this example, the additional variability introduced by using a smaller LOO sub-sample is significantly in excess of that introduced by the bias adjustment by the BPSIC.

4.3 Posterior averaged performance in terms of 'catching up'

We commented previously that the practice of comparing models on the basis of their cumulative scores may be less than optimal. Informally, if we are interested in making use of the models to make future predictions, we may be less concerned about their early performance than in their more recent 'track record'. This is likely to be particularly pertinent in a high dimensional setting, where we require increasingly large 'training sets' to calibrate model parameters. van Erven et al. (2012) study this 'catch up' effect, and propose a solution in which a prior is placed over a switching distribution governing which model should be used in making predictions at a given point in time. An alternative approach of 'calibrating' the models to a similar level of information on an initial training sample, and then comparing models on their subsequent performance has been proposed in Xu et al. (2014). Both approaches retain the Bayes factor (cumulative log score) as the selection metric but, instead, make adjustments to compensate for the catch up effect itself.

We suggest that an alternative approach is to discard the Bayes factor altogether as being inappropriate for this type of problem - instead we should be estimating expected **future** utility. Here we examine the ability of the BPSIC to assess the model's performance based on its **current state** (that is, taking into account parameter learning) as an alternative to using modified Bayes factor selection. We use log predictive utility here, but other utilities would be similarly applicable.

We suppose that the true model is normally distributed $N(0.2, 1^2)$. We wish to compare two models M_1: a fixed model normally distributed $N(0, 1^2)$ and model M_2 with known variance 1^2 but unknown mean $\mu \sim N(1, 4^2)$. The relatively vague prior on μ in model M_2 means that, assessed on a sequential basis as data is received, M_1 will initially perform better, but after sufficient observations, model M_2 will become the preferred model.

Figure 4 shows the results based on 1,000 simulations. If the Bayes factor (equivalently cumulative log score) is used then on average, model M_2 will only be chosen when the cumulative log score difference is lower than 0, that is after

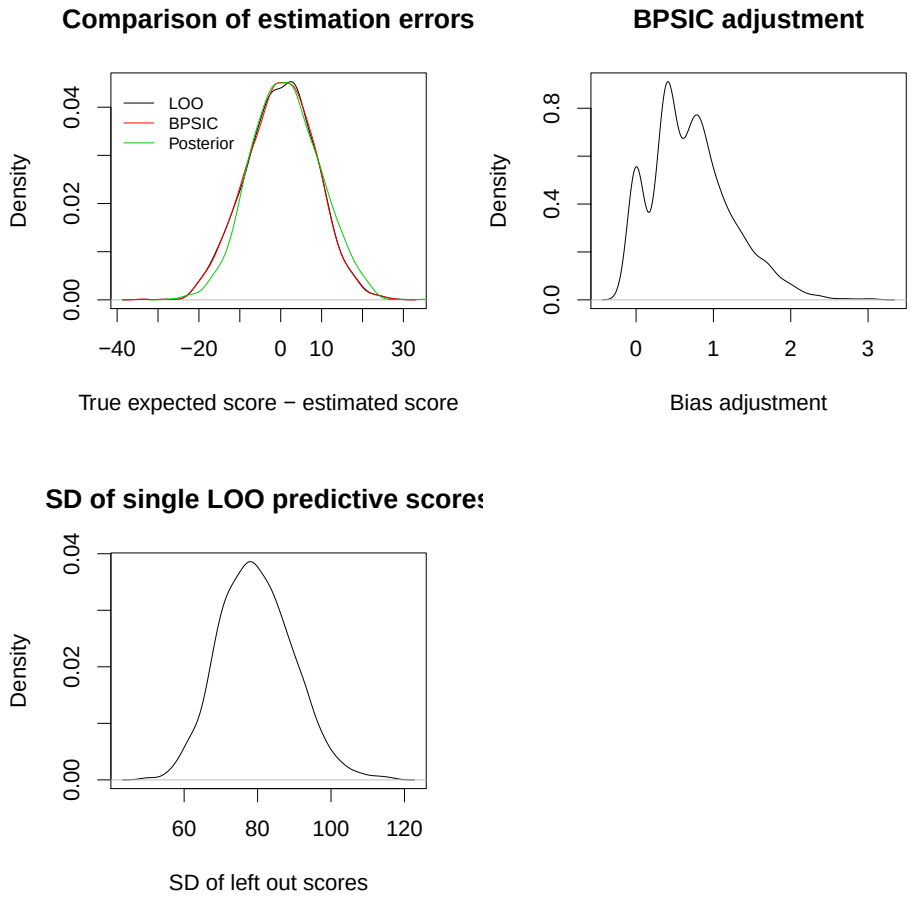

Figure 3: Comparison of the BPSIC with leave one out cross validation. The true data generating process is given by a $N(1, 2^2)$ distribution. The model being assessed, M_1, is a normal distribution with unknown mean and fixed variance of 0.5^2. The mean has a conjugate prior $\mu \sim N(0, 5^2)$. The scoring function chosen is the 0.95 asymmetric piecewise linear quantile loss function. The figure shows the results of 2,000 simulations of a sample size of 100. In each simulation the BPSIC is calculated using a sample of 1,000 from the posterior distribution, and the 'true' scores are calculated on a new sample of 1,000 observations generated from the true distribution. The first graph shows the estimation error resulting from the leave one out, unadjusted posterior and adjusted (BPSIC) score. The second graph shows the variation in the bias adjustment for the BPSIC, and the final graph shows the standard deviation the individual 'left out' scores which are averaged to form the LOO-CV estimate.

approximately 170 observations. Naive assessment based on the uncorrected posterior log score (that is, the posterior Bayes factor of Aitkin (1991)) will result in selecting model M_2 immediately. If the true posterior log score is used (with knowledge of the true data generating process) then model M_2 should be preferred on average much earlier – approximately after 50 observations. We gain a very similar result using the corrected estimate from the BPSIC (bottom left hand graph).

This suggests that if our interest is in future model performance, then unless we have strong reasons to believe that one of the models under considerations is, in fact true, we may be better served by using the BPSIC as a metric. Of course, we would obtain very similar results using cross validation, but, in general, this is much more computationally intensive.

We should remark that one advantage of the cross-validated score over the BPSIC is that it would enable us to better assess performance of the predictive density of the **posterior predictive score**. For many prediction problems, we are more concerned with how the posterior predictive score will perform than we are with the average of the scores across the posterior distribution. Note that, as the BPSIC is defined as an average divergence (and therefore an estimate of $E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right]$), this means that for concave score functions such as the log score, we will have

$$E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right] \leq E_z\left[S(E_{\theta|y}\left[f_\theta\right], z)\right], \tag{13}$$

where $E_{\theta|y}\left[f_\theta\right]$ is the posterior predictive density. In this example, it will mean that if our consideration is when we should employ the posterior predictive distribution, this will typically be earlier than that suggested by the BPSIC (in this example around a sample size of 40).

5 Quantile prediction - UK electricity market imbalance

Where the end goal is to select a model which is used to provide an estimate of 'tail risk', the use of a suitably bias corrected scoring rule can be particularly appropriate. In this section, we apply the BPSIC to the problem of risk management of imbalance exposures to UK electrity market participants. Within the UK, an electricity balancing mechanism is managed by the System Operator (National Grid) to ensure security of supply. Market participants are required to inform the System Operator of their forecast output (in the case of generators) and demand (in the case of suppliers) approximately one hour in advance of each half hour's electricity production.

Typically, it will be necessary for the System Operator to intervene to ensure the actual electricity generated in a given period meets actual demand (which will be different from that implied by the aggregate of forecasts received due to forecast error). In the situation we consider here, the overall system is *short*, in other words it is necessary for the System Operator to seek additional sources of generation (for example, requesting additional short term generation be activated or for certain high users to reduce demand). The cost of these activities is reflected in the *System Buy Price* (SBP) charged to those who have underforecast demand or overforecast supply,

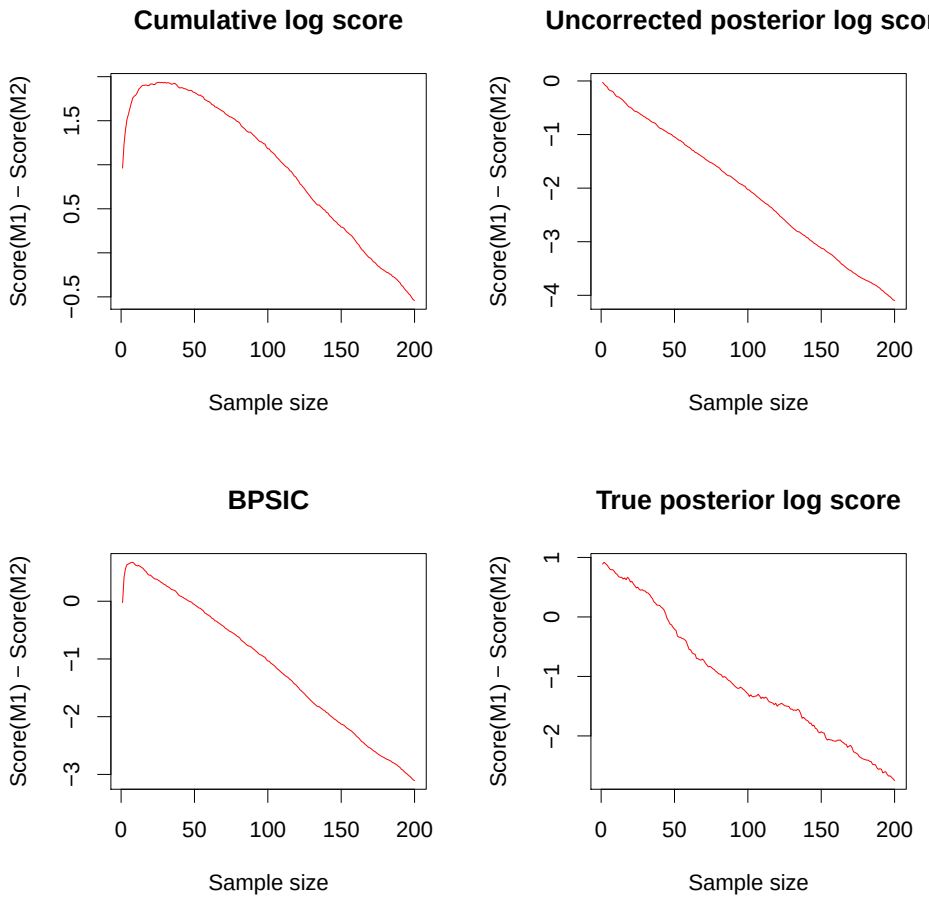

Figure 4: Comparison of the performance of the cumulative log score and posterior average (estimated and true) log scores. The true data generating process is given by a $N(0.2, 1^2)$ distribution. We compare two models: M_1 is a fixed model consisting of a normal distribution $N(0, 1^2)$. Model M_2 is a model with known (true) variance and unknown mean μ, $N(\mu, 1^2)$, where the mean has a conjugate prior $\mu \sim N(1, 4^2)$. The figure shows the comparison scores averaged over 1,000 simulations.

and this will typically be significantly higher than the prevailing market price (in this situation, the System Sell Price (SSP) will reflect a prevailing market price).

Accurate quantification of the amount of risk exposure to imbalance volumes is an important consideration for all market participants. Typically value at risk and associated risk premia charged to contracts which contribute to system imbalance might be based on an assessment of the 95th percentile, in accordance with market risk practice within the financial services industry. Models which forecast this level of risk are to be preferred to those which provide a better general forecast coverage.

Imbalance data for each day in the period 17th October 2011 to 28th May 2014 was obtained from the Elexon data portal (https://www.elexonportal.co.uk). The dataset used for modelling comprised the Net Imbalance Volume (NIV) (the total amount of electricity (in MWh) by which the system was short or long compared to forecast demand), SBP and SSP (both denominated in GBP/MWh). Prices and balancing behaviour vary throughout the day depending on the degree of demand across the day, and for this exercise, we used data only for Settlement Period 16 (this corresponds to a particular half hourly generation period on each day between 7.30 am and 8.00 am in the winter and between 6.30 am and 7.00 am in the summer). This resulted in a sample size of 438.

We would expect the SBP to become more stressed in periods where the NIV is higher. This might be the case, for example, where a power station suffers an unforeseen outage, as in these situations the System Operator will be required to procure a substantial amount of energy at very short notice, often being forced to make use of extremely high cost sources of generation and/or high bids from commercial generators. The degree of stress can be measured by the ratio of the SBP to a measure of typical prices which are prevailing in the market - for this purpose we use the ratio of SBP to SSP.

Panels a) and b) in Figure 5 shows a plot of the SBP/SSP compared to the NIV. As can be seen, in addition to a positive relationship between the NIV and the SBP/SSP ratio. there is also a significant skew in the residuals, as we would expect from an increasingly expensive 'supply stack' of generation.

The skew-normal distribution (Azzalini (1986)) has been used successfully to reflect skewness without the need for ancillary data transformation and was used to model this aspect of the data. In particular, we selected a linear regression model with skew-normal residuals of the form

$$f_{SN}(y_i;\beta_1,\beta_2,\omega^2,\alpha) = \frac{2}{\omega}\phi\left(\frac{y_i-(\beta_1+\beta_2 x_i)}{\omega}\right)\Phi(\alpha\omega^{-1}(y_i-(\beta_1+\beta_2 x_i))) \qquad (14)$$

MCMC was used to obtain posterior estimates (see Fruhwirth-Schnatter and Pyne (2010)). First, we allowed both slope and intercept terms to be fitted, using a vague normal gamma prior (a diagonal matrix with entries of 0.01 for the precision matrix, a mean of zero and $Gamma(0.01, 0.01)$ distribution). The posterior distributions from

12,000 simulations with a burn in of 4.000 are shown in Figure 5.

As alternative models, we constrained the intercept term, β_1 at different values between 0.2 and 1.2 (each time by setting the mean for β_1 to the desired value, and the corresponding parameter of the prior precision matrix at 1,000,000). BPSIC values were computed for the standard logarithmic score and also the 0.95 quantile scoring rule, reflecting the possible focus of interest of a risk management decision using this model to price an appropriate risk premium. In Figure 6, we show the corresponding BPSIC at various values. In particular, the model with the value of β_1 close to that fitted freely gives the highest BPSIC logarithmic score. However, if our interest is in fitting accurately to the prediction of the 0.95 quantile, the graphs show that models with a lower intercept value provide a greater expected future utility.

Table 1 summarises the information criteria output for the fitted model, together with the highest scoring models under logarithmic and quantile scoring criteria. Although not considered here, it is easy to see how such a table could be extended to include other scoring rules reflecting the diverse utilities of the possible future user base for a model. Such an extension could enable a more informed selection of the most appropriate model implementation and parameterisation for a particular need.

Table 1: Comparison of BPSIC for fitted and fixed intercepts - skew-normal model

	Fitted	Intercept = 0.6	Intercept = 1.0
BPSIC(log score)	-124.1	-340.8	-123.9
Asymptotic bias(log score)	3.7	4.4	4.5
BPSIC(percentile score)	-46.2	-41.5	-46.8
Asymptotic bias(percentile score)	2.5	3.4	3.4

6 Discussion

This paper introduces a Bayesian score based information criterion which is an estimate of the posterior weighted average of out of sample performance for a user defined scoring function relevant to the problem at hand. The effects of the bias inherent in using the observed data for the dual purposes of posterior updating and model selection are controlled through the incorporation of an asymptotic bias correction which is derived by applying the results in Ando (2007).

We believe the measure may enable analysts to incorporate in their model assessments more realistic utility functions which are tailored to the aims of their end users. We also suggest that this technique has a complementary role to play alongside more established procedures, for example, cross validated estimates. However, in order to increase our understanding of the situations in which this approach is likely to be optimally employed, we would be interested in exploring the following areas:

Particularly for small data sets, it seems intuitively plausible that cross-validation will

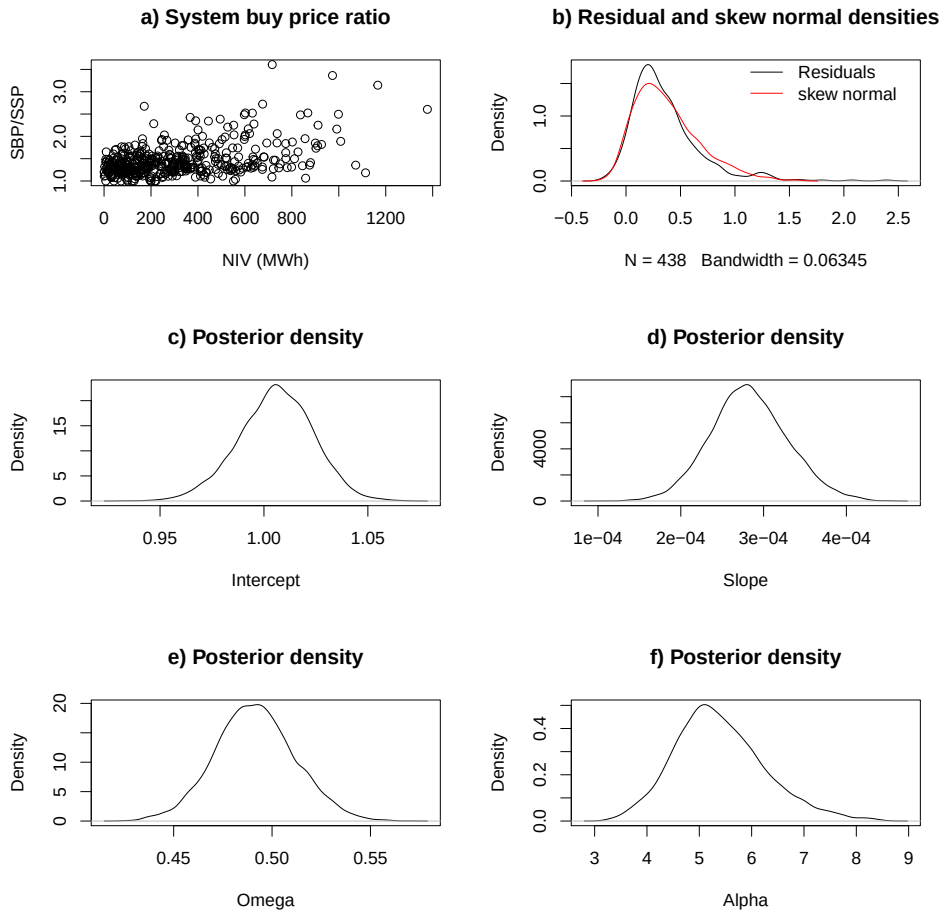

Figure 5: Estimation of the relationship between Net Imbalance Volume and System Buy Price. Panel a) shows data between October 2011 and May 2014 relating to settlement period 16 for those occasions when the system was short. Panel b) compare the residuals from the estimates using the posterior mode linear regression parameters to the skew normal distribution with posterior mode parameters. Panels c) to f) show the posterior parameter estimates obtained from fitting the linear regression model with skew normal residuals estimated using a 12,000 simulation MCMC sample with a burn-in of 4,000 simulations.

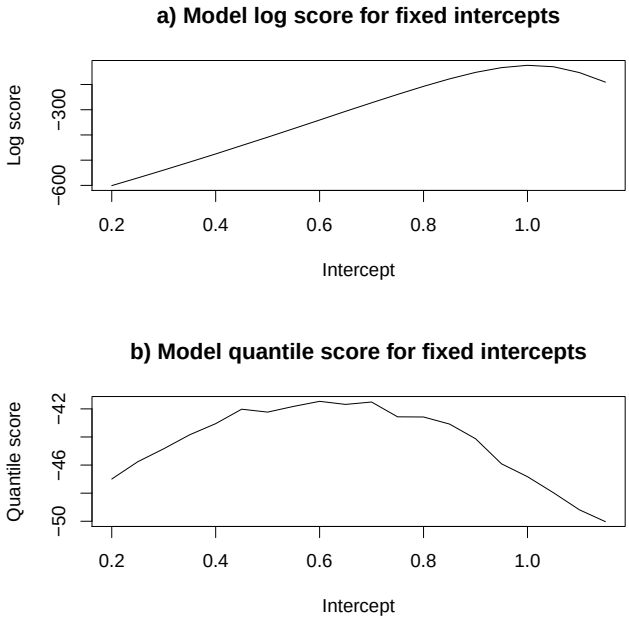

Figure 6: Comparison of the BPSIC obtained for models with fixed intercept values in the linear regression. Note that the maximum BPSIC log score is obtained with an intercept value of approximately 1.0, consistent with the posterior estimates in Figure 5 when a vague prior is placed on the intercept. However, models fitted with lower values of the intercept have higher BPSIC quantile score, reflecting the increased skewness fitted by these models

suffer when influential elements are omitted, and that the bias corrected approach may have some advantages here. It would be interesting to understand better the trade-offs in these situations. Stone (1977) establishes asymptotic equivalence of leave-one-out cross validation and AIC. A further area of research would be to investigate whether similar results can be obtained between cross-validated scoring rules and the analogous BPSIC measure.

In a M-open context, there is the vexed question of how, if at all, Bayes updating should proceed. It is not within the scope of this paper to comment further on this *per se*, but rather to remark that it seems to us there is an interesting connection between the amount of bias in updating and assessing a model using the same data, and the similarity between the metric for model assessment and the target divergence which the update seeks to optimise.

For example, standard Bayesian updating results asymptotically in a posterior which minimises the Kullback Leibler divergence to the true model (Berk (1966)). This is the same metric which we use to assess the model under the log score, and therefore the bias here we would expect to be greater, than if we score the model with, for example, a quantile score. Bissiri et al. (2013) proposes application of alternative updating mechanisms to the Bayes rule in M-open contexts. These take the form of $\pi(\theta \mid y) \propto \exp(l(\theta, y))\pi(\theta)$, for a loss function of interest $l(\theta, y)$, acknowledging that in an M-open context it is not necessarily true that one set of parameters will be optimal under all losses. Under such updating procedures, we can derive a similar information criterion to the BPSIC, where the relevant matrices and posterior modes are replaced with their analogues under the alternative loss functions. Here we might expect the bias to be greater if the loss function chosen is comparable to the score function.

We have remarked that our methods may be particularly applicable in big data contexts, where model selection may reflect the decision maker's utility more accurately by using a BPSIC based on, for example, the relevant marginal and conditional logarithmic scores of the variables of interest within a larger model. Another benefit of this relates to the lower bias correction term which is applied in these situations. Typically where cumulative (joint) logarithmic scores are concerned, this will be of the order of the number of parameters in the model, say p. By examination of the BPSIC bias correction term in Equation 11, it will be seen that the corresponding term is $\mathrm{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n))$. So when other, more tailored, scores are used then the bias adjustment will typically be significantly lower, as $J_n^S(\hat{\theta}_n)$ will take zero values on those parameters which do not impact the utility under consideration - for example, in the case of a conditional distribution where parameters separate, the bias correction term will equate to the number of parameters involved in the representation of the relevant conditional distribution.

1 Appendix: Proof of Theorem

The proof uses the method which is introduced in Ando (2007), adjusted to allow for the fact that the estimator $\hat{\theta}_n^\star$ which maximises the scoring function may differ from the posterior mode $\hat{\theta}_n$. In particular, this means that the expansion around the posterior mode requires the relevant first derivatives.

1.1 Notation

We suppose that n observations $y = (y_1, y_2, \dots y_n)$ are generated from the true process with probability density $g(y)$. We assume that a model M with probability density $f_\theta := f(y \mid \theta)$ and prior density $\pi(\theta)$ is being considered as a candidate model for approximating the true data generating process. We denote the likelihood $L(y \mid \theta) = \prod_{k=1}^n f_\theta(y_k)$. We defined $\log \pi_0(\theta) := \lim_{n\to\infty} n^{-1} \log \pi(\theta)$, and assume that this exists. We also assume that $\pi(\theta) = O(1)$. Note that Ando (2007) allows the more general case where $\log \pi(\theta) = O(n)$, which allows the prior distribution to depend on n. We further assume that the parameter vector θ is of dimension p.

Suppose we are interested in the model which minimises the expected divergence induced by the scoring function $S(f, z)$ for observations z from the true data generating process. We define the cumulative score $C_S(y \mid \theta) := \sum_{k=1}^n S(f_\theta, y_k)$.

We denote θ_0, $\hat{\theta}_n$ as the parameter values which maximise $E_z\left[\log(f_\theta(z)\pi_0(\theta))\right]$ and $n^{-1}\log(L(y \mid \theta)\pi(\theta))$ respectively, and assume that these are unique. We define $\bar{\theta}_n$ as

the posterior mean for θ, and define the matrices and estimators:

$$\begin{aligned}
I(\theta) &= E_z\left[\frac{\partial(\log f_\theta(z)+\log\pi_0(\theta))}{\partial\theta}\frac{\partial(\log f_\theta(z)+\log\pi_0(\theta))}{\partial\theta^T}\right], \qquad (15)\\
J(\theta) &= -E_z\left[\frac{\partial^2(\log f_\theta(z)+\log\pi_0(\theta))}{\partial\theta\partial\theta^T}\right],\\
J^S(\theta) &= -E_z\left[\frac{\partial^2(S(f_\theta,z)+\log\pi_0(\theta))}{\partial\theta\partial\theta^T}\right],\\
U^S(\theta) &= E_z\left[\frac{\partial(S(f_\theta,z)+\log\pi_0(\theta))}{\partial\theta}\right],\\
I_n(\theta) &= \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial(\log f_\theta(y_k)+\log\pi(\theta)/n)}{\partial\theta}\frac{\partial(\log f_\theta(y_k)+\log\pi(\theta)/n)}{\partial\theta^T}\right),\\
J_n(\theta) &= -\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial^2(\log f_\theta(y_k)+\log\pi(\theta)/n)}{\partial\theta\partial\theta^T}\right),\\
J_n^S(\theta) &= -\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial^2(S(f_\theta,y_k)+\log\pi(\theta)/n)}{\partial\theta\partial\theta^T}\right),\\
U_n^S(\theta) &= \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial(S(f_\theta,y_k)+\log\pi(\theta)/n)}{\partial\theta}\right).
\end{aligned}$$

Ando (2007) establishes the following results for θ_n and θ_0 assuming appropriate regularity conditions

Lemma 2 *$(\hat{\theta}_n-\theta_0)$ is asymptotically normally distributed as $N(0,n^{-1}J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0))$.*

Proof. See Ando (2007) ■

Lemma 3 *Assuming appropriate additional regularity conditions for the Laplace approximation of the posterior distribution to be valid, we have:*

$$E_y\left[E_{\theta|y}\left[(\theta-\theta_0)(\theta-\theta_0)^T\right]\right] \approx \frac{1}{n}J^{-1}(\theta_0)+\frac{1}{n}J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0) \qquad (16)$$

Proof. See Ando (2007) ■

We now proceed to prove the following theorem:

Theorem 4 *Under the assumptions that we have a scoring rule S which is such that*

- $n\,\mathrm{cov}(U_n^S(\hat{\theta}_n),\hat{\theta}_n)\approx 0$, $n\,\mathrm{cov}(U_n^S(\hat{\theta}_n),\bar{\theta}_n)\approx 0$,

- $U_n^S(\hat{\theta}_n)$,$(\bar{\theta}_n - \hat{\theta}_n)$ *are uncorrelated,*

if we consider the bias from estimating the posterior expected score

$$b_S := E_y\left[\frac{1}{n}E_{\theta|y}\left[C_S(y \mid \theta)\right] - E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right]\right] \tag{17}$$

then, asymptotically,

$$nb_S \approx E_{\theta|y}\left[C_S(y \mid \theta) + \log \pi(\theta)\right] - (C_S(y \mid \hat{\theta}_n) + \log \pi(\hat{\theta}_n))$$

$$+\frac{1}{2}\operatorname{Tr}(J^S(\hat{\theta}_n)J^{-1}(\hat{\theta}_n)) + \operatorname{Tr}(J^S(\hat{\theta}_n)J^{-1}(\hat{\theta}_n)I(\hat{\theta}_n)J^{-1}(\hat{\theta}_n)) - nU_n^S(\hat{\theta}_n)(\bar{\theta}_n - \hat{\theta}_n).$$

Proof. We express the bias as the sum of three expected values:

$$\begin{aligned} E_1 &= E_y\left[\frac{1}{n}E_{\theta|y}\left[C_S(y \mid \theta)\right] - \frac{1}{n}(C_S(y \mid \theta_0) + \log \pi(\theta_0))\right], \\ E_2 &= E_y\left[\frac{1}{n}(C_S(y \mid \theta_0) + \log \pi(\theta_0)) - E_z\left[S(f_{\theta_0}, z) + \log \pi_0(\theta_0)\right]\right], \\ E_3 &= E_y\left[E_z\left[S(f_{\theta_0}, z) + \log \pi_0(\theta_0)\right] - E_z\left[E_{\theta|y}\left[S(f_\theta, z)\right]\right]\right]. \end{aligned} \tag{18}$$

Approximating E_1
To approximate E_1, we perform a Taylor expansion of $C_S(y \mid \theta_0) + \log \pi(\theta_0)$ around the posterior mode $\hat{\theta}_n$, where we obtain:

$$C_S(y \mid \theta_0)+\log \pi(\theta_0) \approx C_S(y \mid \hat{\theta}_n)+\log \pi(\hat{\theta}_n)+n(\theta_0-\hat{\theta}_n)U_n^S(\hat{\theta}_n)-n/2(\theta_0-\hat{\theta}_n)^T J_n^S(\hat{\theta}_n)(\theta_0-\hat{\theta}_n),$$

and so we have

$$E_1 \approx \frac{1}{n}E_y\left[E_{\theta|y}\left[C_S(y \mid \theta)\right] - (C_S(y \mid \hat{\theta}_n) + \log \pi(\hat{\theta}_n))\right]$$

$$-E_y\left[U_n^S(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)\right] + \frac{1}{2}\operatorname{Tr}(E_y\left[J_n^S(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)(\theta_0 - \hat{\theta}_n)^T\right]).$$

Using Lemma 2, and the fact that $J_n^S(\hat{\theta}_n) \to J^S(\theta_0)$ as $n \to \infty$, we have

$$E_1 \approx \frac{1}{n}E_y\left[E_{\theta|y}\left[C_S(y \mid \theta)\right] - (C_S(y \mid \hat{\theta}_n) + \log \pi(\hat{\theta}_n))\right] \tag{19}$$

$$+\frac{1}{2n}\operatorname{Tr}(J^S(\theta_0)J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)) - E_y\left[U_n^S(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)\right].$$

Approximating E_2
We can ignore the term E_2 as approximately zero, as we have:

$$E_2 \approx E_y\left[S(f_{\theta_0}, y) + \log \pi_0(\theta_0)\right] - E_z\left[S(f_{\theta_0}, z) + \log \pi_0(\theta_0)\right] - \log \pi_0(\theta_0) \log \pi_0(\theta).$$

Approximating E_3
For the term E_3, we perform a Taylor expansion around θ_0. Writing

$$E_3 = E_z\left[S(f_{\theta_0}, z) + \log \pi_0(\theta_0)\right] - E_y\left[E_{\theta|y}\left[E_z\left[(S(f_\theta, z) + \log \pi_0(\theta)\right]\right]\right]$$

$$+E_y\left[E_{\theta|y}\left[\log \pi_0(\theta)\right]\right],$$

if we expand around θ_0, then we have

$$E_3 \approx E_z\left[S(f_{\theta_0}, z) + \log \pi_0(\theta_0)\right] - E_z\left[(S(f_{\theta_0}, z) + \log \pi_0(\theta_0)\right]$$

$$-U^S(\theta_0)E_y\left[E_{\theta|y}\left[(\theta - \theta_0)\right]\right] + \frac{1}{2}\operatorname{Tr}(J^S(\theta_0)E_y\left[E_{\theta|y}\left[(\theta - \theta_0)(\theta - \theta_0)^T\right]\right]) + E_y\left[E_{\theta|y}\left[\log \pi_0(\theta)\right]\right].$$

Applying Lemma 3 and approximating $\log \pi_0(\theta) \approx n^{-1} \log \pi(\theta)$, gives

$$E_3 \approx \frac{1}{2n}\operatorname{Tr}(J^S(\theta_0)(J^{-1}(\theta_0) + J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)) + \frac{1}{n}E_y\left[E_{\theta|y}\left[\log \pi(\theta)\right]\right]$$

$$-U^S(\theta_0)E_y\left[E_{\theta|y}\left[(\theta - \theta_0)\right]\right].$$

Approximating total bias
Combining the terms gives the bias

$$nb_S \approx E_y\left[E_{\theta|y}\left[C_S(y \mid \theta) + \log \pi(\theta)\right] - (C_S(y \mid \hat{\theta}_n) + \log \pi(\hat{\theta}_n))\right]$$

$$+\frac{1}{2}\operatorname{Tr}(J^S(\theta_0)J^{-1}(\theta_0)) + \operatorname{Tr}(J^S(\theta_0)J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)).$$

$$-nE_y\left[U_n^S(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)\right] - nU^S(\theta_0)E_y\left[E_{\theta|y}\left[(\theta - \theta_0)\right]\right].$$

We can rearrange the final two terms as

$$-nE_y\left[U_n^S(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + U^S(\theta_0)(\bar{\theta}_n - \theta_0)\right]$$

$$= -nE_y\left[U_n^S(\hat{\theta}_n)(\bar{\theta}_n - \hat{\theta}_n)\right] + nE_y\left[(U_n^S(\hat{\theta}_n) - U^S(\theta_0))(\bar{\theta}_n - \theta_0)\right],$$

We now make use of our assumption that we are working with a score function S which is such that $n\operatorname{cov}(U_n^S(\hat{\theta}_n), \hat{\theta}_n) \approx 0$ and $n\operatorname{cov}(U_n^S(\hat{\theta}_n), \bar{\theta}_n) \approx 0$.

The first term vanishes where posterior modes and means are equal (as would be the case, for example, under conjugate symmetric priors) and where these are not equal,

the expectation can be approximated to first order by $-nU_n^S(\hat{\theta}_n)(\bar{\theta}_n - \hat{\theta}_n)$, using the assumption that $U_n^S(\hat{\theta}_n)$ and $(\bar{\theta}_n - \hat{\theta}_n)$ are uncorrelated.

If we rewrite the second term as

$$E_y\left[\sqrt{n}(U_n^S(\hat{\theta}_n) - U^S(\theta_0))\sqrt{n}(\bar{\theta}_n - \theta_0)\right]$$

$$= E_y\left[\sqrt{n}(U_n^S(\hat{\theta}_n) - U^S(\theta_0))\right] E_y\left[\sqrt{n}(\bar{\theta}_n - \theta_0)\right] + n\,\mathrm{cov}(U_n^S(\hat{\theta}_n), \bar{\theta}_n)$$

$$\approx n\,\mathrm{cov}(U_n^S(\hat{\theta}_n), \bar{\theta}_n),$$

the final approximation following from the fact that the second expectation is asymptotically zero and the first expectation is bounded.

We therefore approximate the quantities with their estimators:

$$nb_S \approx E_{\theta|y}\left[C_S(y \mid \theta) + \log \pi(\theta)\right] - \left(C_S(y \mid \hat{\theta}_n) + \log \pi(\hat{\theta}_n)\right)$$

$$+\frac{1}{2}\,\mathrm{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)) + \mathrm{Tr}(J_n^S(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)I_n(\hat{\theta}_n)J_n^{-1}(\hat{\theta}_n)) - nU_n^S(\hat{\theta}_n)(\bar{\theta}_n - \hat{\theta}_n).$$

■

References

Aitkin, M. (1991). "Posterior Bayes Factors." *Journal of the Royal Statistical Society Series B-Methodological*, 53(1): 111–142.

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle." In Petrov, B. and Csaki, F. (eds.), *Second international symposium on information theory*, 267–281.

Ando, T. (2007). "Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models." *Biometrika*, 94(2): 443–458.

Aravkin, A., Kambadur, A., Lozano, A., and Luss, R. (2014). "Sparse quantile Huber regression for efficient and robust estimation." *arXiv:1402.4624v1*.

Arlot, S. and Celisse, A. (2010). "A survey of cross-validation procedures for model selection." *Statistics Surveys*, 4: 40–79.

Azzalini, A. (1986). "A class of distributions which includes the normal ones." *Scandinavian Journal of Statistics*, 12: 171–178.

Berk, R. (1966). "Limiting behavior of posterior distributions when the model is incorrect." *Annals of Mathematical Statistics*, 37(1): 51–58.

Bernardo, J. (1979). "Expected Information as Expected Utility." *The Annals of Statistics*, 7: 686–690.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. New York: Wiley.

Bissiri, P., Holmes, C., and Walker, S. (2013). "A General Framework for Updating Belief Distributions." *arXiv:1306.6430v1*, 1: 1–50.

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006). "Rejoinder to 'Deviance Information Criteria for Missing Data Models'." *Bayesian Analysis*, 70.

Claeskens, G. and Hjort, N. (2003). "The focused information criterion (with discussion)." *Journal of the American Statistical Association*, 98: 879–899.

Dawid, A. (1984). "Statistical Theory - The Prequential Approach." *Journal of the Royal Statistical Society Series A - Statistics in Society*, 147: 278–292.

— (2007). "The geometry of proper scoring rules." *Annals of the Institute of Statistical Mathematics*, 59: 77–93.

Fruhwirth-Schnatter, S. and Pyne, S. (2010). "Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions." *Biostatistics*, 11(2): 317–336.

Gharamani, Z. (2004). "Advanced Lectures on Machine Learning." chapter Unsupervised Learning, 72–112. Springer.

Gneiting, T. (2011). "Making and evaluating point forecasts." *Journal of the American Statistical Association*, 106(494): 746–762.

Gneiting, T. and Raftery, A. (2007). "Strictly proper scoring rules, prediction and estimation." *Journal of the American Statistical Association*, 102(477): 359–378.

Kadane, J. and Dickey, J. (1980). *Bayesian Decision Theory and the Simplification of Models*, 245–268. New York: Academic Press.

Kass, R. and Raftery, A. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90(430): 773–795.

Key, J., Pericchi, L., and Smith, A. (1999). "Bayesian Model Choice: What and Why?" In Bernardo, J. (ed.), *Bayesian Statistics 6*, 343–370. Oxford University Press.

Linhart, H. and Zucchini, W. (1986). *Model Selection*. John Wiley and Sons.

Musio, M. and Dawid, P. (2013). "Model Selection with Proper Scoring Rules." In *Cambridge Statistics Initiative One-Day Meeting*.

Phillips, L. (1982). "Requisite decision modelling: a case study." *The Journal of the Operational Research Society*, 33: 303–311.

Schwarz, G. (1978). "Estimating the dimension of a model." *Annals of Statistics*, 6(2): 461–464.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002). "Bayesian measures of complexity and fit." *Journal of the Royal Statistical Society Series B*, 64(4): 583–639.

Stone, M. (1974). "Cross-Validatory Choice and Assessment of Statistical Predictions." *Journal of the Royal Statistical Society Series B (Methodological)*, 36(2): 111–147.

— (1977). "An asymptotic equivalance of choice of model by cross-validation and Akaike's criterion." *Journal of the Royal Statistical Society Series B*, 39(1): 44–47.

van Erven, T., Grunwald, P., and de Rooij, S. (2012). "Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the Akaike information criterion - Bayesian information criterion dilemma." *Journal of the Royal Statistical Society Series B*, 74(2): 1–37.

Vehtari, A. (2001). "Bayesian model assessment and selection using expected utilities." Ph.D. thesis, Helsinki University of Technology.

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6: 142–228.

Winkler, R., Munoz, J., Bernardo, J., Blattenberger, G., and Kadane, J. (1996). "Scoring rules and the evaluation of probabilities." *Test*, 5(1): 1–60.

Xu, X., Lu, P., MacEachern, S., and Xu, R. (2014). "Calibrated Bayes Factors for model comparison." *Biometrika*.

Zhou, S. (2011). "Bayesian model selection in terms of Kullback-Leibler discrepancy." Ph.D. thesis, Columbia University.